Optimasi Algoritma *Support Vector Machine* Dengan Menggunakan *Feature Selection Gain Ratio* Untuk Analisis Sentimen

Mochamad Yamin Amzah[1], Kusnadi[2], Luhur Bayuaji[3]
[123] Universitas Budi Luhur, Jl. Ciledug Raya, Kec. Pesanggrahan, Kota Jakarta Selatan, Indonesia
E-mail: 2111601080@student.budiluhur.ac.id[1], 2111601213@student.budiluhur.ac.id[2], luhur.bayuaji@budiluhur.ac.id[3]

Abstract – The ease of internet access has had a positive impact on the increase in the number of social media users in Indonesia. One of the most widely used applications is X or Twitter. Users often upload posts that contain opinions or sentiments, which trigger debates and discussions. This is interesting to analyze as a study of sentiments or opinions that are trending in society. For this analysis, algorithms such as Support Vector Machine (SVM) are required, which are often used for sentiment analysis. However, SVM lacks in accuracy due to the large number of similar words in the dataset. Words related to sentiment analysis usually have large dimensions, so feature selection is needed to improve SVM performance. This research aims to optimize SVM accuracy by using Feature Selection Gain Ratio. The object of research is a dataset related to the 2017 DKI elections from GitHub. The results showed an increase in SVM accuracy with Feature Selection Gain Ratio. With threshold weight gain ratio > 0.0001 (1732 features), accuracy increases from 61.63% to 71.51%. For threshold weights > 0.002 (518 features), the accuracy increased from 61.63% to 62.79%. Feature selection with Feature Selection Gain Ratio gain ratio produces better accuracy than gain ratio, namely 56.40% with gain ratio and 71.51% with gain ratio for weights > 0.0001. The implications of these findings show that the use of Feature Selection Gain Ratio can improve the accuracy of SVM in sentiment analysis. Social media practitioners can utilize this technique to gain more accurate insights from user data. Further research can focus on developing sentiment analysis algorithms with more sophisticated feature selection techniques for various applications on social media platforms.

Keywords - Sentiment Analysis, Support Vector Machine, Feature Selection Gain Ratio, and Gain ratio.

Intisari – Kemudahan akses internet berdampak positif pada peningkatan jumlah pengguna media sosial di Indonesia. Salah satu aplikasi yang paling banyak digunakan adalah X atau Twitter. Pengguna sering mengunggah postingan yang mengandung opini atau sentimen, yang memicu perdebatan dan diskusi. Hal ini menarik untuk dianalisis sebagai bahan kajian sentimen atau opini yang sedang tren di masyarakat. Untuk analisis ini, diperlukan algoritma seperti Support Vector Machine (SVM), yang sering digunakan untuk analisis sentimen. Namun, SVM memiliki kekurangan dalam tingkat akurasi karena banyaknya kata yang mirip dalam dataset. Kata-kata terkait analisis sentimen biasanya memiliki dimensi yang besar, sehingga diperlukan *Feature Selection Gain Ratio* untuk meningkatkan kinerja SVM. Penelitian ini bertujuan mengoptimalkan akurasi SVM dengan menggunakan *Feature Selection Gain Ratio*. Objek penelitian adalah dataset terkait Pilkada DKI 2017 dari GitHub. Hasil penelitian menunjukkan peningkatan akurasi SVM dengan *Feature Selection Gain Ratio*. Dengan bobot *threshold gain ratio* > 0,0001 (1732 fitur), akurasi meningkat dari 61,63% menjadi 71,51%. Untuk bobot *threshold* > 0,002 (518 fitur), akurasi meningkat dari 61,63% menjadi 62,79%. Seleksi fitur dengan *Feature Selection Gain Ratio gain ratio* menghasilkan akurasi lebih baik dibandingkan *gain ratio*, yaitu 56,40% dengan information gain dan 71,51% dengan gain ratio untuk bobot > 0,0001. Implikasi dari temuan ini menunjukkan bahwa penggunaan *Feature Selection Gain Ratio* dapat meningkatkan akurasi SVM dalam analisis sentimen. Praktisi media sosial dapat memanfaatkan teknik ini untuk mendapatkan wawasan lebih akurat dari data pengguna. Penelitian lebih lanjut dapat fokus pada pengembangan algoritma analisis sentimen dengan teknik seleksi fitur yang lebih canggih untuk berbagai aplikasi di *platform* media sosial.

Kata Kunci - Analisis Sentimen, *Support Vector Machine*, *Feature Selection Gain Ratio*, dan *Gain ratio*.

I. PENDAHULUAN

Perkembangan internet dan era digital telah mengubah fungsi media sosial dari sekadar *platform* untuk berbagi informasi menjadi ruang untuk menyampaikan opini dan ekspresi. Salah satu *platform* media sosial yang populer untuk menyampaikan opini adalah X, yang sebelumnya dikenal sebagai Twitter. X menjadi wadah bagi masyarakat untuk mengekspresikan opini, perasaan, atau emosi dalam batasan 280 karakter. Batasan ini diyakini meningkatkan efektivitas dan ekspresivitas postingan, membuat konten memiliki nilai emosional atau sentimen yang baik untuk analisis sentimen [1].

Data statistik dari databoks.katadata per April 2023 menunjukkan bahwa Indonesia menempati peringkat keenam dalam jumlah pengguna X terbanyak di dunia. Media sosial juga sering digunakan sebagai alat kampanye politik oleh tokoh publik. Penelitian sebelumnya menunjukkan bahwa strategi kampanye politik, terutama oleh politisi, melibatkan berbagai platform media sosial seperti X. Di *platform* ini, pengguna sering memberikan opini, menggunakan *hashtag*, dan menandatangani petisi yang berkaitan dengan salah satu partai politik atau kandidat tertentu. Postingan-potingan ini dapat memicu reaksi atau sentimen dari pengguna lainnya, sehingga mewakili beragam pendapat dan perasaan dalam ruang publik digital [2].

Fenomena ini menarik untuk dianalisis sebagai bahan kajian sentimen atau opini yang sedang tren di masyarakat karena mencerminkan dinamika interaksi sosial dalam ruang digital. Dalam analisis sentimen, data yang dihasilkan dari interaksi pengguna media sosial seperti X dapat memberikan wawasan yang berharga tentang pola pikir, preferensi, dan perasaan kolektif yang ada di masyarakat, yang dapat memberikan pandangan yang lebih mendalam tentang tren dan dinamika sosial yang sedang berlangsung.

Analisis sentimen memerlukan penerapan algoritma yang sesuai. Beberapa algoritma yang umum digunakan meliputi *Naive Bayes*, *Support Vector Machine* (SVM), *Maximum Entropy Classifier*, dan *K-Nearest Neighbor* [3]. Dalam konteks ini, peneliti memilih SVM sebagai algoritma untuk analisis sentimen. Pemilihan algoritma ini karena keunggulannya dalam menangani data berdimensi tinggi. Meskipun demikian, SVM tetap mengalami gagguan ketika menangani data berjumlah besar dan mengalami penurunan akurasi karena banyaknya kata yang mirip dalam dataset. Untuk mengatasi kendala ini, diperlukan upaya untuk mengoptimalkan kinerja SVM.

Upaya meningkatkan kinerja SVM, dilakukan dengan memanfaat *Feature Selection Gain Ratio*. Menurut Hafidzillah [2], *Feature Selection Gain Ratio* berfungsi mengurangi atau menghapus kata yang kurang relevan, mempermudah algoritma dalam memproses klasifikasi teks sentimen dan meningkatkan akurasi. Sejalan dengan penelitian sebelumnya, penelitian Maulana dan Mandiri [4] menggunakan *dataset Cornell* dan *Stanford* menemukan bahwa SVM berbasis *Feature Selection Gain Ratio* dapat meningkatkan akurasi sebesar 0,16 % terhadap *dataset Stanford*. Adapun pada penelitian terkait analis sentimen dengan SVM.

Adapun penelitian terkait analis sentimen dengan SVM, dilakukan oleh Ratino [5] membandingkan bagaimana algoritma *Naïve Bayes* dan SVM dapat dioptimasi tingkat akurasinya dengan mengombinasikan bersama algoritma *Particle Swarm Optimization* (PSO). Hasil penelitiannya menunjukkan bahwa Algoritma *Naives Bayes* dan PSO memiliki tingkat akurasi sebesar 79,07 % sementara SVM dan PSO tingkat akurasi sebesar 81,16%. Selanjutnya, penelitian Somanti dan Apriliani [6] menunjukkan bahwa tingkat akurasi SVM mengalami peningkatan ketika menggunakan *Feature Selection Gain Ratio* dibandingkan dengan *Chi Square*. Akurasi SVM dengan *Gain ratio* meningkat dari 69,36% menjadi 72,45%.

Berdasarkan sejumlah penelitian yang telah disebutkan sebelumnya, diketahui bahwa terdapat peluang untuk meningkatkan akurasi algoritma SVM dengan berbasis *Feature*

Selection Gain Ratio. *Feature Selection Gain Ratio* merupakan pengembangan dari *Gain ratio* yang berfungsi mengoptimalkan nilai yang dinormalisasi untuk sebuah fitur dalam klasifikasi. SVM berbasis berbasis *Feature Selection Gain Ratio* dapat menghasilkan nilai yang akurasi yang lebih tinggi dan memberikan hasil yang lebih baik daripada penelitian sebelumnya. Merujuk pada penelitian terdahulu, hipotesis dari penelitian ini adalah terdapat peningkatan akurasi algoritma SVM setelah menerapkan teknik *feature selection gain ratio*.

II. SIGNIFIKANSI STUDI

A. *Penelitian Terdahulu*

Tabel 1 memaparkan hasil penelitian sebelumnya yang dijadikan sebagai acuan dalam penelitian ini.

TABEL I
UKURAN FONT UNTUK MAKALAH

Nomor	Penulis	Penelitian Terdahulu
1	Oman Somantri, Dyah Aprilian[6]	*Support Vector Machine* Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal. Hasil akurasi menggunakan SVM mencapai 69,36%. Setelah menerapkan seleksi fitur dengan *gain ratio*, akurasi meningkat menjadi 72,45%, yang merupakan peningkatan sebesar 3,08%. *Gain ratio* juga menunjukkan hasil yang lebih baik dibandingkan dengan metode chi-square.
2	Reza Maulana [7]	Peningkatan Akurasi Analisis Sentimen Review Film Menggunakan *Support Vector Machine* Berbasis *Gain ratio*. Hasil penelitian menunjukkan bahwa SVM berbasis *Gain ratio* meningkatkan akurasi pada *dataset Cornell* (dari 83,05% menjadi 85,65%) dan pada dataset Stanford (dari 86,46% menjadi 86,62%), membuktikan efektivitasnya dalam analisis sentimen ulasan film.
3	Ration Noor Hafidz Sita Angraeni Windu Gata[4]	Sentimen Analisis Informasi Covid-19 menggunakan *Support Vector Machine* dan Naïve Bayes. Metode SVM dan *Naïve Bayes* dapat digunakan untuk analisis sentimen komentar di Instagram. Namun, penerapan algoritma PSO terbukti meningkatkan akurasi. Hasil akurasi adalah: Naïve Bayes 78,02%, Naïve Bayes dengan PSO 79,07%, SVM 80,23%, dan SVM dengan PSO 81,16%.
4	Nauffan Muti Hibattullah, Said Al Faraby[8]	Analisis Sentimen terhadap Ulasan Film Berbahasa Inggris Menggunakan Metode *Support Vector Machine* dengan Feature Selection *Gain ratio* Penelitian ini menunjukkan bahwa kombinasi *Stopword* dan *Stemming* meningkatkan akurasi hingga 86,12%. Meskipun *Gain ratio* (IG) menurunkan akurasi, IG membantu mengatasi overfitting. *Support Vector Machine* (SVM) dengan kernel Linear efektif untuk klasifikasi analisis sentimen ulasan film..
5	Arief Riski Indra Pratama, Siti Amalia Latipah, dan Betha Nurina Sari [9]	Optimasi Klasifikasi Curah Hujan Menggunakan *Support Vector Machine* (Svm) Dan *Recursive Feature Elimination* (Rfe) Penelitian ini menunjukkan bahwa optimasi SVM dengan RFE meningkatkan akurasi dari 77% menjadi 79%.

B. *Tinjauan Pustaka*

1. *Machine Learning (ML)*

Pembelajaran Mesin (*Machine Learning atau ML*) merupakan cabang dari Kecerdasan Buatan (*Artificial Intelligence atau AI*) yang fokus pada pemahaman struktur data serta mencocokkan data tersebut dengan model yang dapat dipahami dan diaplikasikan oleh manusia[10]

2. *Data Mining*

Data mining adalah proses yang menggunakan teknik-teknik statistik, matematika, dan kecerdasan buatan untuk mengekstraksi serta mengidentifikasi informasi dan pola-pola yang muncul dari kumpulan data yang sangat besar. Pola-pola ini dapat berupa aturan bisnis, kesamaan, korelasi, tren, atau model prediksi [11].

3. *Teks Mining*

Text mining adalah proses yang bertujuan untuk mengekstraksi informasi yang berharga dari teks yang tidak memiliki struktur atau terdiri dari banyak dokumen. Dengan melakukan text mining, kita dapat mengidentifikasi pola, tren, dan hubungan yang tidak terlihat secara langsung dalam teks yang tidak beraturan [12].

4. *Analisis Sentimen*

Analisis sentimen, juga dikenal sebagai penambangan opini, melibatkan pengambilan, penguraian, dan pemrosesan data teks secara otomatis untuk mengumpulkan informasi mengenai pendapat yang termuat dalam kalimat. Pemahaman sentimen digunakan untuk mengetahui apakah seseorang cenderung memiliki pandangan yang negatif atau positif terhadap suatu situasi atau objek tertentu [13].

5. *Konsep Klasifikasi Machine Learning*

Dalam *machine learning*, klasifikasi merupakan metode di mana algoritma atau model diprogram untuk mengenali serta mengelompokkan objek atau data ke dalam kategori-kategori tertentu berdasarkan ciri-ciri yang dimilikinya[14]

6. *Support Vector Machine (SVM)*

Metode pembelajaran berbimbing yang dikenal sebagai *Support Vector Machine* (SVM) merupakan salah satu dari berbagai metode yang dikembangkan oleh Vladimir Vapnik. SVM membangun suatu bidang pemisah, atau serangkaian bidang pemisah, dalam ruang yang memiliki dimensi yang tinggi (bahkan mungkin tak terhingga) yang berguna dalam penyelesaian masalah klasifikasi atau regresi [15].

7. *Feature selection*

Feature selection atau seleksi fitur merupakan proses penting dalam analisis data yang bertujuan untuk mengidentifikasi dan memilih sekumpulan fitur yang paling relevan dan informatif dari data asli. Tujuannya adalah untuk mengurangi kompleksitas data dengan menghilangkan fitur-fitur yang tidak relevan atau berlebihan, serta meningkatkan performa model dengan hanya mempertimbangkan fitur-fitur yang paling signifikan [14]

8. *Gain ratio*

Gain ratio adalah suatu metode dalam memilih fitur. Fungsinya adalah menentukan seberapa pentingnya suatu fitur terhadap variabel target atau kelas yang ingin diprediksi. Biasanya, teknik ini digunakan untuk memilih atribut yang relevan dalam algoritma klasifikasi atau dalam konteks seleksi fitur [16].

9. *Gain Ratio*

Metode *Gain Ratio* merupakan penyempurnaan dari Information Gain. Metode ini mengatasi kesulitan terkait atribut yang memiliki banyak nilai atau kelas. Dengan menggunakan *Gain Ratio*, peringkat pada atribut yang memiliki variasi nilai yang banyak menjadi lebih seimbang [17].

10. *Term Frequency-Inverse Document Frequency (TF-IDF)*

Metode *Term Frequency-Inverse Document Frequency (TF-IDF)* merupakan teknik yang digunakan dalam pengolahan dan klasifikasi teks. Teknik ini menghitung tingkat signifikansi suatu kata dalam satu dokumen dibandingkan dengan semua dokumen dalam koleksi tersebut [18].

C. Metode Penelitian

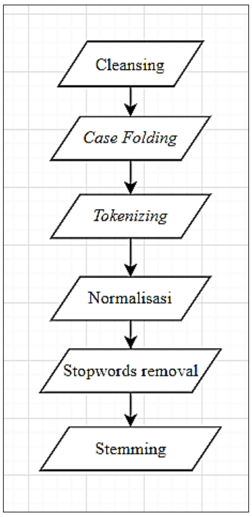
Penelitian ini bertujuan untuk membuat sebuah model analisis sentimen menggunakan metode *Support Vector Machine* (SVM) dengan menggunakan seleksi fitur *gain ratio*. Model ini akan diuji untuk melihat seberapa akuratnya dalam mengklasifikasikan sentimen dari postingan di *platform* X. Data yang digunakan berasal dari dataset publik yang tersedia di GitHub yang berkaitan dengan analisis sentimen terhadap *tweet* mengenai pilkada DKI 2017, yang disebut sebagai postingan. Setelah proses penyaringan, terdapat 879 data dalam dataset, dengan 438 data menunjukkan sentimen negatif dan 441 data menunjukkan sentimen positif. Data ini digunakan sebagai data latih untuk membangun model analisis sentimen, yang akan digunakan untuk mengklasifikasikan sentimen pada postingan baru. Setelah model terbentuk, dilakukan pengujian menggunakan metode *Confusion Matrix*.

1. Pengumpulan Data

Data yang digunakan berasal dari sumber publik yang ditemukan di GitHub. Data ini berfokus pada analisis sentimen terkait Pilkada DKI 2017. Dataset ini tersedia dalam format file CSV.

2. Teks Processing

Teks processing dilakukan untuk memersiapkan data sehingga sudah siap untuk dilakukan proses selanjutnya. Tujuan pokok dari *Teks processing* adalah menjalankan langkah-langkah pembersihan dan transformasi dokumen sehingga mendapatkan representasi yang lebih terstruktur dan relevan. Hal ini bertujuan untuk memfasilitasi kinerja optimal algoritma klasifikasi, menghasilkan hasil yang lebih unggul. Rinciannya mengenai *Teks processing* dokumen disajikan pada gambar 1 dibawah ini.



Gambar 1. Proses *Teks processing* Dokumen

Dari gambar 1, dapat disimpulkan bahwa *terks processing* atau proses pemrosesan teks melibatkan enam langkah. Langkah pertama adalah *Celansing*, yang berfokus pada membersihkan dokumen dari karakter yang tidak diinginkan dan noise. Langkah kedua disebut *Case Folding*, di mana semua huruf dalam teks diubah menjadi huruf kecil atau besar. Langkah ketiga adalah *Tokenizing*, yang bertujuan untuk memecah teks menjadi bagian-bagian kecil seperti kata-kata atau frasa. Langkah keempat disebut Normalisasi, yang bertujuan untuk mengubah variasi kata menjadi bentuk dasar. Langkah kelima adalah *Stowords Remuval*, yang berarti menghapus kata-kata umum yang tidak memberikan makna. Terakhir, langkah keenam adalah *Stemming*, yang bertujuan untuk menghapus imbuhan kata sehingga menyisakan bentuk dasar kata tersebut.

3. Term weighting

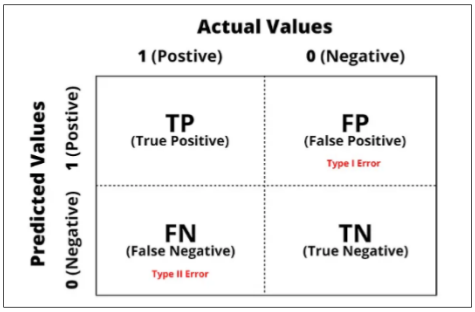
Kata-kata yang sudah dipangkas akarnya dan dihitung kepentingannya melalui proses *Term weighting*. Penimbangan istilah adalah teknik yang digunakan dalam menganalisis teks dan pengolahan bahasa alami untuk memberi skor pada kata-kata.

4. *Klasifikasi dan Validasi*

Proses selanjutnya adalah membuat model dengan membagi data menjadi dua bagian, yaitu bagian untuk latihan (80%) dan bagian untuk pengujian (20%). Data yang digunakan sudah diberi label sentimen negatif dan positif. Data tersebut akan digunakan untuk melatih model analisis sentimen menggunakan algoritma SVM. Model ini akan digunakan untuk mengklasifikasikan sentimen dari data uji, apakah termasuk dalam kategori negatif atau positif.

5. *Evaluasi*

Tujuan evaluasi adalah mengukur kinerja model klasifikasi. Ini dilakukan dengan membandingkan hasil prediksi model dengan nilai sebenarnya dari data yang diuji menggunakan *Confusion Matrix*. *Confusion Matrix* memiliki variabel sebagai berikut:



Gambar 2. *Confusion matrix*

Penjelasan singkat tentang angka dalam matriks pada Gambar 2 di atas adalah sebagai berikut. TP (*True Positive*) adalah jumlah contoh yang sebenarnya positif dan diprediksi sebagai positif oleh model. TN (*True Negative*) adalah jumlah contoh yang sebenarnya negatif dan diprediksi sebagai negatif oleh model. FP (*False Positive*) adalah jumlah contoh yang sebenarnya negatif namun salah diprediksi sebagai positif oleh model. FN (*False Negative*) adalah jumlah contoh yang sebenarnya positif namun salah diprediksi sebagai negatif oleh model.

6. *Metode Pengumpulan Data*

Dalam penelitian ini, metode untuk mengumpulkan data adalah dengan menggunakan dataset publik yang tersedia di GitHub. Dataset tersebut berisi analisis sentimen terkait Pilkada DKI 2017 dan disajikan dalam format file CSV.

7. *Instrumentasi*

Dalam penelitian ini, beberapa *instrument* yang digunakan berupa perangkat keras dan perangkat lunak dibawah ini.

a. Selain digunakan untuk menyusun laporan tesis, perangkat keras pada penelitian ini digunakan juga untuk proses pengambilan data, mengembangkan metoda dan teknik penelitian, membuat model, melakukan pengujian model. Berikut adalah peralatan keras yang terlibat dalam penelitian ini: Intel Core i5 11th Gen 8 CPU, RAM 16 GB DDR4 dan SSD 120 GB, HDD 500GB.
b. Berikut adalah perangkat lunak yang digunakan untuk mendukung penelitian ini: Sistem Operasi Windows 11 64 bit, Tools Jupiter Notebook untuk proses load dataset, preprocessing data, pemodelan, dan pengujian, dan Google chrome web browser.

8. *Teknik Analisis*

Dalam proses analisis, langkah-langkah berikut dijalankan. Pertama, dataset yang digunakan adalah dataset publik yang berisi informasi tentang tokoh-tokoh publik. Kedua, dilakukan proses *pre-processing* terhadap data, termasuk *cleansing*, *case folding*, *tokenizing*, penghapusan *stopword*, dan *stemming*. Ketiga, data dibagi menjadi dua bagian, yakni 80% untuk data latih dan 20% untuk data uji. Keempat, dataset sudah diberi label sentimen positif dan negatif. Kelima, model awal diuji menggunakan data uji, dan hasilnya dibandingkan antara metode seleksi fitur menggunakan metoda information gain dan gain ratio. Terakhir, dilakukan evaluasi yang disajikan dalam bentuk tabel dan grafik, yang kemudian akan digunakan untuk menarik kesimpulan dan memberikan saran dalam proses analisis selanjutnya.

9. *Teknik Perancangan*

Teknik perancangan untuk memperoleh prediksi perasaan terhadap data yang belum memiliki label perasaan adalah: langkah pertama, dataset yang telah diproses dan diberi label perasaan akan digunakan sebagai data latih. Kemudian, data yang belum memiliki label akan digunakan sebagai data uji.

III. HASIL DAN PEMBAHASAN

A. *Analisis Koleksi Dokumen*

Dataset yang digunakan untuk melakukan analisis sentimen dalam penelitian ini berasal dari sumber publik di *GitHub*. Spesifiknya, *dataset* ini terkait dengan *posts* yang berkaitan dengan pilkada DKI 2017. Dari *dataset* tersebut, diperoleh data sentimen untuk kategori positif sebanyak 441 data, dan data dengan sentimen negatif sebanyak 438 data. Sehingga total ada berjumlah 879 dengan dua buah kolom. Pada *dataset* tersebut sudah dilakukan penggantian *emoticon* yang ada dengan tag/penanda special, sehingga dataset hanya berupa *text*. Contoh analisis data set disajikan pada tabel 2 berikut:

TABEL IIi
ANALISIS DOKUMEN BEBERAPA POST DATA DALAM DATA SET

Sentiment	***TextTweet***
Negative	Banyak akun kloning seolah2 pendukung #agussilvy mulai menyerang paslon #aniessandi dengan opini dan argumen pmbenaran..jangan terkecoh
Negative	#agussilvy bicara apa kasihan yaa...lap itu air matanya wkwkwkwk
Negative	sudah boleh Ngakak? survey mu jauh panggang dari api ! #Ahy
Negative	Karna yang belebihan itu tidak baik ya bah #AHY
Negative	Gara2 e-KTP dikorupsi, fotoku yg semula mirip #AHY, pelan tapi pasti skrang malah mirip -Stickmoticons”
Negative	Double LOL. @aniesbaswedan nuduh @AgusYudhoyono miskin ide, sekarang dia copy #AHY. @SBYudhoyono http://m.republika.co.id
Negative	Ingat! Tidak bisa modal OMONGAN saja. Tapi kerja nyata! #agusharimurtiyudhoyono #AHYâ€¦ https://www.instagram.com/p/BROjg0gBpYa/
Positive	Doa ku.. Semoga suaranya mas Agus-Sylvi beneran ke #Badja #PilkadaDKI2017 yakin kan mas Agus.. Aku pada mu..
Positive	Terima kasih mas @AgusYudhoyono, karenamu anak muda telah berani menyuarakan perubahan. Kami tetap bersamamu #YNWA
Positive	Dalam Pilkada DKI, AHY juga mampu menunjukkan kualitasnya sebagai seorang pemimpin muda yang berjiwa ksatria. (AHY mendunia)
Positive	(12) AHY justru menitipkan pentingnya menjaga pluralisme, kebangsaan dan kesatuan dalam persaingan Pilkada. Persatuan lebih penting.
Positive	AHY memang kalah di pilkada tetapi MENANG di hati rakyat.
Positive	Salam kagum buat AHY. Masih muda, berani pensiun dini, berani menantang ahok di pilkada, disaat semua pada takut, termasuk @ridwankamil

Berdasarkan informasi pada tabel tabel 2 diatas, terdapat dua buah kolom, *pertama* kolom sentimen yang memuat informasi mengenai sentimen yang terkandung dalam suatu teks dengan pilihan nilai yang dibatasi hanya pada dua kategori, yaitu "*Negative*" dan "*Positive*". Kolom sentimen ini memiliki peran penting sebagai penanda atau label dalam proses analisis data, memungkinkan pengelompokan atau pemahaman terhadap aspek evaluatif terkait dengan teks yang bersangkutan. Kedua, kolom *TextTweet* yang mencakup *post* atau opini seseorang dalam format teks. Berdasarkan data yang telah dikumpulkan, konten *post* telah mengalami proses konversi, yang melibatkan transformasi *emoticon* ke dalam bentuk teks. Pada bagian kolom TextTweet dalam dataset yang telah disajikan, terdapat informasi tambahan seperti *repost*, *hyperlink*, *hashtag*, dan angka. Keberadaan elemen-elemen tersebut menunjukkan bahwa data belum sepenuhnya bersih dari unsur-unsur yang mungkin memengaruhi proses analisis. Oleh karena itu, diperlukan langkah-langkah *preprocessing data* sebagai tahap awal untuk memastikan kebersihan dan konsistensi data sebelum dilakukan tahap pemodelan. Proses *preprocessing* ini melibatkan penanganan elemen-elemen tambahan tersebut agar data siap digunakan secara optimal dalam rangka pengembangan model atau analisis lebih lanjut.

B. Data Preparation

Sebelum memulai pemodelan, langkah pertama yang perlu dilakukan adalah persiapan data. Proses ini mencakup pengumpulan, penyaringan, dan pengaturan data yang relevan untuk memastikan keakuratan dan keandalan informasi yang akan digunakan dalam analisis. Tahapan ini mencakup membersihkan data dari gangguan, mengubah variabel, dan menggabungkan dataset jika diperlukan. Dengan melakukan persiapan data yang teliti, diharapkan penelitian ini dapat menghasilkan hasil analisis yang akurat dan dapat diandalkan, serta memberikan kontribusi yang penting dalam bidang studi yang sedang dipelajari. Berikut adalah langkah-langkahnya.

1. Text Processing

Pemrosesan data bersumber dari “data-clean.csv” atau data yang telah melalui hasil *preprocessing*. Kemudian data diolah pada tahap berikutnya. Contoh text processing disajikan pada tabe 2 di bawah ini:

TABEL IIIII
DATA HASIL *PREPROCESSIN*

Sentiment	***TextTweet***
Negative	akun kloning dukung serang paslon opini argumen pmbenaran kecoh
Negative	bicara kasihan ya lap air mata
Negative	ngakak survey mu panggang api
Negative	gara e ktp korupsi foto pelan stickmoticons
Negative	double lol nuduh miskin ide copy
Negative	tidak modal omong kerja nyata
Positive	doa ku moga suara mas agus sylvi beneran mas agus mu
Positive	terima kasih mas karena anak muda berani suara ubah sama
Positive	pilkada dki ahy kualitas pimpin muda jiwa ksatria ahy dunia
Positive	ahy titip jaga pluralisme bangsa satu saing pilkada satu
Positive	ahy kalah pilkada menang hati rakyat
Positive	salam kagum ahy muda berani pensiun berani tantang ahok pilkada saat takut
Negative	akun kloning dukung serang paslon opini argumen pmbenaran kecoh

Dari hasil *preprocessing* yang sudah dilakukan pada tabel 3, maka diperoleh jumlah *data* yang akan diproses pada tahapan selanjutnya yaitu total sebanyak 859 *data* dengan dua pembagian yakni: data klasifikasi sentimen *positive* sebanyak 425 data dan data klasifikasi sentimen *negative* sebanyak 434 data.

2. Term Weighting

Selanjutnya dilakukan proses *term weighting* atau pembobotan kata. Digunakan untuk mengevaluasi pentingnya sebuah kata dalam sebuah dokumen relatif terhadap keseluruhan koleksi dokumen. Metode ini memberikan bobot kepada kata-kata berdasarkan frekuensinya dalam dokumen tertentu dan keunikannya di seluruh koleksi dokumen. Dibawah ini hasil term weighting dengan menggunakan beberapa sample data. Untuk kebutuhan perhitungan data pembobotan, maka diperlukan proses terlebih dahulu dari konversi kolom Sentiment, dari semula bertipe *string* menjadi *integer*. Sentimen *Negative* dikonversi menjadi bernilai 0 dan Sentimen *Positive* dikonversi menjadi bernilai 1.

TABEL IVV
KONEVERSI NILAI SENTIMEN TAHAP TERM WEIGHTING

Sentiment	***TextTweet-Cleaned***
0	akun kloning dukung serang paslon opini argumen pmbenaran kecoh
0	bicara kasihan ya lap air mata
0	ngakak survey mu panggang api
0	gara e ktp korupsi foto pelan stickmoticons
0	double lol nuduh miskin ide copy
0	tidak modal omong kerja nyata
1	doa ku moga suara mas agus sylvi beneran mas agus mu
1	terima kasih mas karena anak muda berani suara ubah sama
1	pilkada dki ahy kualitas pimpin muda jiwa ksatria ahy dunia
1	ahy titip jaga pluralisme bangsa satu saing pilkada satu
1	ahy kalah pilkada menang hati rakyat
1	salam kagum ahy muda berani pensiun berani tantang ahok pilkada saat takut

Dari proses konversi nilai sentiment di tabel 4 diatas, selanjutnya Sebelum dilakukan proses split data yaitu dengan skema data train train sebesar 80%, dan testing sebesar 20%. Dari total dataset 859 data, hasil *split data* yang dilakukan adalah Data train sebanyak 687 data dengan pembagian jumlah sentimen positive sebanyak 340 dan sentimen *negative* sebanyak 347 data dan Data testing sebanyak 172 data. Kemudian dilakukan pemecahan kalimat menjadi kata / *term* dan dihitung bobot nilainya menggunakan *TF-IDF*. Dari hasil pemecahan kalimat menjadi kata, diperoleh sebanyak 1732 kata atau *term* dimana akan menjadi sebuah fitur yang selanjutnya harus di seleksi. *Term* inilah yang akan diproses perhitungan bobot / *term weighting* dengan menggunakan TF-IDF.

3. *Feature Selection*

Kemudian setelah diperoleh nilai TF-IDF, dilakukan proses *feature selection* yang bertujuan memilih fitur optimal. Proses ini mempertahankan fitur-fitur paling berpengaruh dan memerbaiki performa model. Dalam penelitian ini menggunakan dua metoda *feature selection* yaitu *gain ratio* dan *gain ratio*. Dengan demikian, akurasi performa nilai SVM akan dibandingkan antara SVM tanpa *feature selection*, SVM dengan *Feature Selection Gain Ratio*, dan SVM dengan *Feature Selection Gain Ratio*. Setelah jumlah fitur diperoleh untuk gain ratio yaitu dengan penentuan bobot threshold *gain ratio* sebesar 0,0001 dan 0,002. Maka lakukan *Feature selection* terbaik menggunakan code dibawah ini.

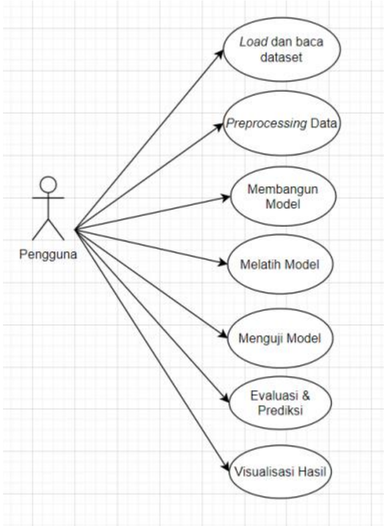
```
k_best_info_gain = SelectKBest (mutual_info_classif,                (1)
k=num_top_features) # Pilih sejumlah fitur teratas
   X_train_selected_info_gain = k_best_info_gain.fit_transform
(X_train_tfidf, y_train)
   X_test_selected_info_gain = k_best_info_gain transform (X_test_tfidf)
```

Adapun nilai dari variable *num_top_features* tersebut adalah menggunakan nilai yang diperoleh pada proses penentuan threshold bobot yang dilakukan sebelumnya yaitu bernilai 1732 dan 518. Setelah itu dilakukan proses pemodelan menggunakan nilai fitur terbaik yang diperoleh, yang akan dijelaskan pada tahapan pemodelan berikut ini.

C. *Pemodelan*

1. *Perancangan Sistem*

Perancangan sistem dapat digambarkan dengan *diagram use case* dibawah ini, yang memberikan gambaran komprehensif mengenai perjalanan dari awal hingga akhir dalam mengembangkan, menerapkan, dan menguji model klasifikasi, dalam hal ini klasifikasi sentimen pada konteks pembelajaran mesin.



Gambar 3. *Use Case Diagram*

Penjelasan dari *diagram use case* gambar 3 adalah *Load* dan baca *dataset* berarti pengguna dapat membaca dataset sentimen analisis. *Preprocessing data* berarti pengguna dapat melakukan pemrosesan data awal seperti memerika data yang bersifat null, *cleansing data*, normalisasi, dan langkah lainnya terkait *preprocessing data*. Membangun model berarti pengguna dapat membangun model menggunakan algoritma SVM dengan *Feature Selection Gain Ratio*. Melatih model berarti pengguna dapat melatih model algoritma SVM dengan *Feature Selection Gain Ratio* dengan *data training*. Menguji Model berarti pengguna dapat melakukan uji terhadap model yang sudah dibangun menggunakan SVM dengan *Feature Selection Gain Ratio*. Evaluasi dan Prediksi berarti pengguna dapat melakukan evaluasi hasil kinerja model SVM dengan *Feature Selection Gain Ratio*. Visualisasi Hasil berarti pengguna dapat memvisualisasikan hasil prediksi dalam sebuah *page user interface*.

2. Penerapan Model SVM

Setelah dilakukan berbagai tahapan dari data preprocessing (seperti *cleansing*, *normalisasi*, *case folding*, dan lainnya), kemudian telah dilakukan juga *term weighting* dan *feature selection*, maka selanjutnya dilakukan pengklasifikasian dengan menggunakan algoritma SVM menggunakan sebelas fitur terbaik yang diperoleh pada *feature selection gain ratio* yaitu diperlihatkan oleh tabel 5 berikut:

TABEL V
FITUR TERBAIK *GAIN RATIO* YANG DI PEROLEH

No	Feature Name	*Gain ratio* Value
1	Tidak	0,100340
2	Menang	0,083212
3	Ahy	0,082693
4	Anies	0,082254
5	Ahok	0,077809
6	Pilkada	0,067778
7	Jakarta	0,066121
8	Dukung	0,061499
9	Dki	0,056644
10	Sandi	0,056104

Setelah menyajikan fitur terbaik yang diperoleh pada *feature selection gain ratio* pada tabel tabel 5 diatas, langkah selanjutnya adalah membangun model SVM dengan jumlah fitur terbaik yang diperoleh. Code yang dibutuhkan untuk proses pembentukan model adalah sebagai berikut.

```
svm_classifier = svm.SVC(kernel='linear')                                   (2)
svmclassifier.fit (X_train_selected_info_gain, y_train)
y_pred_svm=svm_classifier.predict(X_test_selected_info_gain)
```

Proses pembentukan model SVM tersebut menggunakan fungsi fit dan predict adalah dua fungsi utama yang digunakan saat bekerja dengan model. Fungsi "*fit*" digunakan untuk melatih model pada dataset pelatihan, sementara fungsi "*predict*" digunakan untuk membuat prediksi menggunakan model yang telah dilatih.

D. Hasil Pengujian

Pada tahap ini, pengujian dilakukan untuk mengevaluasi kinerja metode yang telah dikembangkan. Evaluasi performa dilakukan secara menyeluruh dengan menggunakan metrik akurasi, presisi, dan recall. Akurasi mengukur proporsi dari total prediksi yang benar terhadap keseluruhan data yang dievaluasi. Presisi memberikan informasi tentang sejauh mana model akurat dalam memprediksi data yang sebenarnya positif, dibandingkan dengan total prediksi positif yang dilakukan. Sementara itu, recall menggambarkan kemampuan model untuk mendeteksi dengan benar data positif dari keseluruhan data positif yang sebenarnya.

1. Confusion matrix

Berdasarkan *Confusion matrix*, dapat dilihat seberapa banyak data yang berhasil diklasifikasikan dengan benar (*true positive dan true negative*), serta jumlah kesalahan yang terjadi dalam klasifikasi (*false positive dan false negative*). Laporan klasifikasi kemudian memberikan informasi detail mengenai metrik evaluasi seperti presisi, *recall*, dan *F1-score* untuk setiap kelas. Ini sangat penting untuk memvalidasi dan meningkatkan kinerja serta keandalan model klasifikasi yang digunakan dalam analisis data.

Hasil *Confusion Matrix* pada 172 sampel data uji, terdapat 63 prediksi positif yang tepat dan 27 prediksi positif yang salah. Sementara itu, terdapat 60 prediksi negatif yang tepat dan 22 prediksi negatif yang salah. Pada kasus lain, terdapat 53 prediksi positif yang tepat dan 32 prediksi positif yang salah, serta 55 prediksi negatif yang tepat dan 32 prediksi negatif yang salah.

E. Perbandingan Model

Dari model yang telah dibuat, maka selanjutnya akan dilakukan pembandingan dengan model lain yang akan digunakan untuk memprediksi klasifikasi sentimen sebuah *posts*.

1. Metode Pembandingan

Pada tahap ini, akan dikembangkan metoda pembanding dari model sebelumnya yaitu SVM dengan *feature selection Gain Ratio*. Beberapa model yang dibangun untuk pembanding adalah SVM tanpa *feature selection* dan TF-IDF, SVM tanpa *feature selection* dengan pembobotan TF-IDF dan SVM dengan *Feature Selection Gain Ratio*. Tujuan dari model pembanding ini adalah untuk membandingkan performa dalam analisis data dan mengevaluasi perbedaan kemampuan dalam melakukan suatu prediksi, dalam hal ini pengklasifikasian sentimen.

Berdasarkan perhitungan, model SVM tanpa *feature selection* dan tanpa TF-IDF menghasilkan tingkat akurasi sebesar 61,63%. Dengan demikian sebesar 61,63% dari keseluruhan data telah diklasifikan dengan benar oleh model. Kemudian *precision* yaitu untuk mengklasifikasikan sentimen positif dari semua yang di prediksi positif, adalah sebesar 57,69%. Dengan kata lain mengindikasikan bahwa sekitar 57,69% dari sentimen yang diprediksi sebagai positif oleh model SVM, merupakan benar sentimen dengan klasifikasi positif. Adapun *recall* merupakan kemampuan model untuk menemukan sebagian besar sentimen positif dari keseluruhan klasifikasi yang sebetulnya tergolong sentimen positif, adalah sebesar 73,31 %.

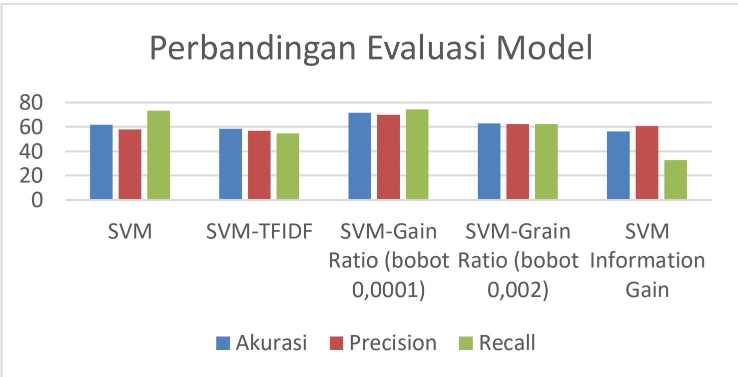
Sehingga tingkat akurasi, precision, dan recall untuk SVM tanpa *feature selection* dapat dilihat pada tabel berikut.

Selanjutnya yang akan dibandingkan yaitu SVM tanpa *feature selection*, namun menggunakan pembobotan TF-IDF. Untuk dilakukan pengujian *model* SVM dengan TF-IDF, maka terlebih dahulu harus dilakukan mekanisme pembobotan TF-IDF yang sudah dibahas diperoleh tingkat akurasi sebesar 58,72% untuk klasifikasi model SVM dengan TF-IDF. Nilai *Precision* sebesar 56,96%, dan *recall* sebesar 54,87%.

Kesimpulan dari pembandingan ini adalah model SVM dengan *Feature Selection Gain Ratio* mengalami penurunan tingkat akurasi dibandingkan dengan SVM tanpa *feature selection*, yaitu menurun dari 61,63% menjadi 56,40%. Artinya sebesar 56,40% dari keseluruhan data telah diklasifikan dengan benar oleh model. Kemudian *precision* yaitu untuk mengklasifikasikan sentimen positif dari semua yang di prediksi positif, adalah sebesar 60,85%. Dengan kata lain mengindikasikan bahwa sekitar 60,85% dari sentimen yang diprediksi sebagai positif oleh model SVM, merupakan benar sentimen positif. Adapun *recall* merupakan kemampuan model untuk menemukan sebagian besar sentimen positif dari keseluruhan klasifikasi yang sebetulnya tergolong sentimen positif, adalah sebesar 32,94 %.

2. Evaluasi Pembadingan

Berikut ini adalah evaluasi perbandingan model yang dibangun berdasarkan model SVM tanpa feature selection, SVM dengan TF-IDF, SVM dengan *Gain ratio*, dan SVM dengan *Gain Ratio*.



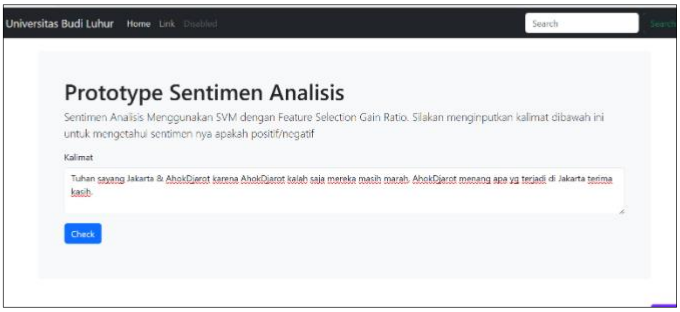
Gambar 4. Perbandingan Evaluasi Model

Dari gambar 4 diatas, diperoleh perbandingan evaluasi model yang cukup menarik. SVM dengan *feature selection gain ratio* mengalami peningkatan akurasi baik itu *gain ratio* yang menggunakan bobot *threshold* 0,02 (518 fitur) dan 0,0001 (1732 fitur). Keduanya mengalami peningkatan akurasi SVM dari semula 61.63% meningkat menjadi 62,79% dan 71,51%. Namun jika dibandingkan SVM dengan feature selection *gain ratio* dan SVM dengan pembobotan TF-IDF megalami penuruan akurasi. Dengan kata lain tidak selalu SVM dan pemanfaatan TF-IDF atau pun penggunaan *feature selection* dapat menghasilkan akurasi yang lebih baik. Dengan demikian, dari temuan ini menunjukkan bahwa pendekatan yang menggabungkan seleksi fitur SVM-*Gain Ratio* memberikan hasil yang lebih baik dalam tugas klasifikasi sentimen.

3. Implementasi Prototype Aplikasi

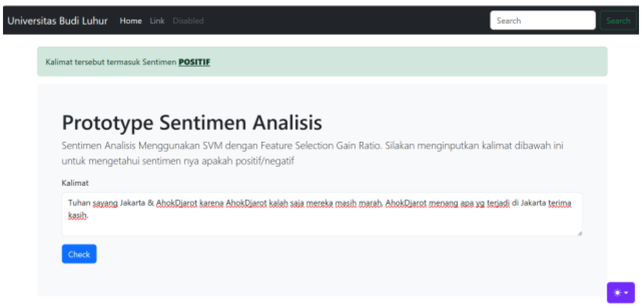
Berdasarkan desain yang telah disusun, maka menghasilkan *Prototype* dengan struktur *client-server*. Perangkat lunak dan perangkat keras sistem dibagi menjadi dua komponen utama: klien (client) dan server. Kedua komponen ini saling berhubungan untuk melaksanakan berbagai fungsi dan memberikan layanan kepada pengguna

Prototype yang dibuat menggunakan *python flask* dan HTML sehingga memungkinkan pengguna untuk mengakses via browser. Adapun untuk desain antarmuka yang *user friendly* dan *responsive*, penulis menggunakan *framework* bootstrap. Aplikasi dirancang untuk memudahkan pengguna memeriksa sentimen dari sebuah kalimat yang diinput oleh pengguna. Berikut tampilan *user interface* sebagai *Prototype* yang telah penulis kembangkan. Dengan tampilan seperti pada gambar dibawah ini, pengguna diharuskan menginput suatu kalimat kemudian berdasarkan model yang sudah di *train*, akan menghasilkan prediksi apakah kalimat yang di *input* oleh pengguna tergolong sentimen *positive* atau *negative*.



Gambar 5. Halaman Utama *Prototype*

Gambar 5 menampilkan halaman utama yang digunakan untuk menguji model dan memprediksi apakah kalimat yang dimasukkan oleh pengguna memiliki sentimen positif atau negatif berdasarkan model yang telah dibuat dan diuji. Langkah-langkah untuk menggunakan halaman tersebut adalah memasukkan kalimat yang ingin diperiksa atau diprediksi sentimennya, kemudian menekan tombol "Check". Setelah itu, hasil prediksinya akan muncul seperti yang terlihat pada Gambar 7 di bawah ini.



Gambar 6. Tampilan hasil prediksi sentiment

Pada Gambar 6, terlihat aplikasi menampilkan hasil prediksi klasifikasi sentimen dengan dua kemungkinan nilai, yaitu Positif atau Negatif. Ini menunjukkan kemampuan aplikasi untuk mengidentifikasi dan membedakan sentimen yang terkandung dalam teks, memberikan pandangan yang jelas tentang apakah konten dipersepsikan secara positif atau negatif oleh algoritma analisis sentimen.

F. Aspek Penelitian Lanjut

Dikarenakan keterbatasan kemampuan penulis dan waktu dalam membuat penelitian ini, penulis berharap apabila ada penelitian lanjutan yang sejenis, dapat mempertimbangkan hal untuk dilakukan pada penelitiannya, yang pertama adalah proses *preprocessing data* lebih selektif dan optimal sehingga dipastikan data yang akan diolah tidak banyak *noise* seperti kata *slang* atau kata non-formal lainnya. Untuk proses *preprocessing* yang lebih optimal, dapat menggunakan kamus KBBI jika tersedia, supaya dipastikan kata yang akan diproses adalah yang kata formal yang memiliki arti. Selanjutnyaa, gunakan metoda lain untuk *feature selection* selain *Gain ratio* sebagai alternatif untuk *SVM* yang optimal. Terakhir, lakukan perbandingan algoritma klasifikasi SVM dengan algoritma lain sehingga dapat menganalisis algoritma dan metoda terbaik yang memberikan hasil prediksi klasifikasi sentimen lebih akurat.

IV. KESIMPULAN

Hasil penelitian menegaskan bahwa hipotesis penelitian terbukti benar, dengan menunjukkan peningkatan signifikan dalam akurasi algoritma SVM setelah penerapan teknik *feature selection gain ratio*. Proses optimasi dengan memilih bobot yang sesuai pada *gain ratio* berhasil meningkatkan kinerja SVM dalam analisis sentimen. Dengan penggunaan *feature selection gain ratio*, terbukti bahwa SVM menghasilkan akurasi yang lebih baik dibandingkan dengan SVM tanpa *feature selection gain ratio*, terutama pada bobot *threshold* 0,0001 (1732 fitur), meningkat dari 61,63% menjadi 71,51%. Kesimpulan ini menegaskan bahwa teknik *feature selection gain ratio* efektif dalam meningkatkan akurasi algoritma SVM dalam analisis sentimen, sesuai dengan hipotesis penelitian.

Pentingnya penelitian ini tidak hanya terletak pada konteks akademis, tetapi juga dalam dampak sosial dan teknologi yang lebih luas. Peningkatan akurasi dalam analisis sentimen dapat memberikan kontribusi yang signifikan dalam pengambilan keputusan di berbagai bidang, baik dalam konteks sosial maupun teknologi. Detail statistik yang disajikan dalam penelitian ini, termasuk peningkatan akurasi dengan bobot threshold yang berbeda, memberikan pemahaman yang lebih dalam tentang kinerja algoritma dalam konteks analisis sentimen. Diskusi tentang mengapa penggunaan TF-IDF tidak meningkatkan performa juga memberikan wawasan yang berharga tentang faktor-faktor yang memengaruhi kinerja algoritma. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi metodologis, tetapi juga wawasan yang berguna bagi pengembangan teknologi analisis sentimen di masa depan.

REFERENSI

[1] A. C. Najib, A. Irsyad, G. A. Qandi, and N. A. Rakhmawati, “Perbandingan Metode Lexicon-based dan SVM untuk Analisis Sentimen Berbasis Ontologi pada Kampanye Pilpres Indonesia Tahun 2019 di Twitter,” *Fountain of Informatics Journal*, vol. 4, no. 2, p. 41, 2019, doi: 10.21111/fij.v4i2.3573.

[2] M. Hafidzullah, S. Sutrisno, and M. Marji, “Seleksi Fitur dengan Information Gain pada Identifikasi Jenis Attention Deficit Hyperactivity Disorder Menggunakan Metode Modified K-Nearest Neighbor,” *Jurnal Pengembangan Teknologi …*, vol. 3, no. 11, pp. 10444–10452, 2019.

[3] S. Pandey, H. Tekchandani, and S. Verma, “A literature review on application of machine learning techniques in pancreas segmentation,” *2020 1st International Conference on Power, Control and Computing Technologies, ICPC2T 2020*, vol. 4, no. 2, pp. 401–405, 2020, doi: 10.1109/ICPC2T48082.2020.9071443.

[4] Ratino, N. Hafidz, S. Anggraeni, and W. Gata, “Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes,” *Jurnal Penelitian Ilmu dan Teknologi Komputer*, vol. 12, no. 2, pp. 1–11, 2020.

[5] Ratino, N. Hafidz, S. Anggraeni, and W. Gata, “Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes,” *Jurnal JUPITER*, vol. 12, no. 2, pp. 1–11, 2020.

[6] O. Somantri and D. Apriliani, “Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, p. 537, 2019, doi: 10.25126/jtiik.201855867.

[7] R. Maulana, “Peningkatan Akurasi Analisis Sentimen Review Film Menggunakan Support Vector Machine Berbasis Information Gain,” Nusa Mandiri, 2019.

[8] N. M. Hibattullah and S. Al Faraby, “Analisis Sentimen terhadap Ulasan Film Berbahasa Inggris Menggunakan Metode Support Vector Machine dengan Feature Selection Information Gain,” *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 10138–10152, 2021.

[9] A. R. I. Pratama, S. A. Latipah, and B. N. Sari, "Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (Svm) Dan Recursive Feature Elimination (Rfe)," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 2, pp. 314–324, 2022, doi: 10.29100/jipi.v7i2.2675.

[10] A. Tedyyana, O. Ghazali, and O. Purbo, "Model Design of Intrusion Detection System on Web Server Using Machine Learning Based," in *Proceedings of the 11th International Applied Business and Engineering Conference, ABEC 2023, September 21st, 2023, Bengkalis, Riau, Indonesia*, EAI, 2024. doi: 10.4108/eai.21-9-2023.2342879.

[11] O. Pahlevi and A. Amrin, "Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease," *SinkrOn*, vol. 5, no. 1, p. 51, 2020, doi: 10.33395/sinkron.v5i1.10589.

[12] A. S. Aribowo and S. Khomsah, "Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19) Implementasi Text Mining Untuk Deteksi Emosi Menggunakan Metode Leksikon (Studi Kasus: Twit Tentang Covid-19)," *Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 1, pp. 49–60, 2021, doi: 10.31515/telematika.v18i1.4341.

[13] S. Siswanto, Z. Mar'ah, A. S. D. Sabir, T. Hidayat, F. A. Adhel, and W. S. Amni, "The Sentiment Analysis Using Naïve Bayes with Lexicon-Based Feature on TikTok Application," *Jurnal Varian*, vol. 6, no. 1, pp. 89–96, 2022, doi: 10.30812/varian.v6i1.2205.

[14] A. Tedyyana, O. Ghazali, and O. W. Purbo, "Machine learning for network defense: automated DDoS detection with telegram notification integration," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 2, p. 1102, May 2024, doi: 10.11591/ijeecs.v34.i2.pp1102-1109.

[15] S. Saikin, S. Fadli, and M. Ashari, "Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results," *JISA(Jurnal Informatika dan Sains)*, vol. 4, no. 1, pp. 22–27, 2021, doi: 10.31326/jisa.v4i1.881.

[16] E. B. Setiawan and I. M. Mubaroq, "The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine," *Ind. Journal on Computing*, vol. 5, no. 2, pp. 107–118, 2020, doi: 10.21108/indojc.2020.5.2.499.

[17] F. N. Fajriyan, Moh. Ahsan, and W. Harianto, "Komparasi Tingkat Akurasi Information Gain Dan Gain Ratio Pada Metode K-Nearest Neighbor," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 1, pp. 386–391, 2022, doi: 10.36040/jati.v6i1.4694.

[18] Visitor Analytics, "Term Frequency Inverse Document Frequency (TF-IDF)," *Visitor Analytics*, no. December, 2023.