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

Mochamad Yamin Amzah1, Kusnadi2, Luhur Bayuaji, S.T, M.Eng, Ph.D3

123Universitas Budi Luhur, Jl. Ciledug Raya, Kec. Pesanggrahan, Kota Jakarta Selatan, Indonesia

*2111601080@student.budiluhur.ac.id1, 2111601213@student.budiluhur.ac.id2, luhur.bayuaji@budiluhur.ac.id3*

**Abstrack** - In this digital era, as the internet becomes more widespread, the use of social media has also increased. One of the social media platforms that has gained high popularity is X, which is often used by the public as a place to channel opinions or opinions through a post. These posts can be used as sentiment analysis or opinion mining material that is useful for identifying trends in the market or to find out opinions from the public. Classification algorithms, namely Support Vector Machine (SVM), are often used as classification techniques in sentiment analysis, but have a problem, namely the level of accuracy that is not optimal. This is due to the significant number of attributes in the dataset. Usually, attributes related to sentiment analysis in text have large dimensions so that feature selection is needed to improve the performance of SVM. This research is intended to optimize the accuracy of the SVM algorithm by utilizing the feature selection gain ratio with the object of sentiment analysis research. From the research results obtained an increase in SVM accuracy with gain ratio. With threshold weight gain ratio > 0.0001 (1732 features), a significant increase in accuracy is obtained from 61.63% to 71.51%. As for the threshold weight > 0.002 (518 features), there is a slight increase in accuracy from 61.63% to 62.79%. As for the selection of the best features with the gain ratio, the accuracy is better than the best features obtained from the information gain, namely the information gain gets an accuracy of 56.40 and the gain ratio with a weight > 0.0001 gets a result of 71.51%.

**Keywords** - *Sentiment Analysis, Support Vector Machine*, *Gain Ratio, and Information gain*.

*I***ntisari** - Pada era digital ini seiring penyebaran internet semakin luas, pemanfaatan media sosial juga mengalami peningkatan. Salah satu platform media sosial yang mendapatkan popularitas tinggi adalah X yang kerap digunakan oleh masyarakat sebagai wadah untuk menyalurkan opini atau pendapat melalui sebuah *Posts*. *Posts* tersebut dapat dijadikan bahan analisis sentimen atau *opinion mining* yang berguna untuk identifikasi kecenderungan hal dipasar ataupun untuk mengetahui opini dari masyarakat. Algoritma klasifikasi yaitu *Support Vector Machine* (SVM) kerap digunakan sebagai teknik klasifikasi dalam analisis sentimen, namun memiliki masalah yaitu tingkat akurasi yang belum optimal. Hal ini disebabkan oleh adanya jumlah atribut yang signifikan dalam dataset. Biasanya atribut-atribut yang terkait dengan analisis sentimen pada teks memiliki dimensi yang besar sehingga diperlukan seleksi fitur atau *feature selection* untuk meningkatkan kinerja dari SVM. Penelitian ini dimaksudkan untuk mengoptimasi tingkat akurasi dari algoritma SVM dengan pemanfaatan *feature selection* *gain ratio* dengan obyek penelitian analisis sentimen. Dari hasil penelitian diperoleh peningkatan akurasi SVM dengan *gain ratio*. Dengan bobot *threshold* *gain ratio* > 0,0001 (1732 fitur), diperoleh peningkatan akurasi yang signifikan dari 61,63% menjadi 71,51%. Adapun untuk bobot *threshold* > 0,002 (518 fitur), terdapat sedikit peningkatan akurasi dari 61,63% menjadi 62,79%. Adapun *Feature selection* terbaik dengan *gain ratio* mendapatkan akurasi lebih baik dibandingkan dengan fitur terbaik yang diperoleh dari *information gain*, yaitu *information gain* mendapatkan akurasi 56,40 dan *gain ratio* dengan bobot > 0,0001 memeroleh hasil 71,51%.

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

1. **Pendahuluan**

Perkembangan internet yang semakin pesat pada era digital saat ini membuat media sosial tidak hanya digunakan untuk berbagi informasi atau berita, tetapi juga sebagai platform untuk menyampaikan pendapat dan lainnya. Salah satu media sosial yang populer adalah X, yang sebelumnya dikenal sebagai Twitter, namun berubah nama menjadi X pada Juli 2023. X sering digunakan oleh masyarakat untuk mengekspresikan opini, perasaan, atau emosi. Konten yang diposting oleh pengguna dalam bentuk teks dibatasi hingga 280 karakter. Najib menyatakan bahwa dengan batasan ini, postingan (sebelumnya disebut tweet) pengguna cenderung lebih efektif dan ekspresif, sehingga konten memiliki nilai emosional atau sentimen yang baik untuk analisis sentimen [1].

Berdasarkan data statistik dari databoks.katadata per April 2023, Indonesia menempati peringkat keenam dalam jumlah pengguna X terbanyak di dunia. Media sosial sering dijadikan sarana kampanye politik oleh tokoh publik. Penelitian Hafizullah menunjukkan bahwa strategi kampanye yang digunakan oleh tokoh-tokoh publik, terutama politisi, melibatkan berbagai platform media sosial. Salah satu contohnya adalah penggunaan hashtag di Twitter dan inisiatif petisi di Facebook, di mana komentar yang ditinggalkan dapat memperoleh reaksi "suka" atau "tidak suka" dari para pendukung mereka [2].

Algoritma-algoritma klasifikasi seperti *Naive Bayes*, *Support Vector Machine* (SVM), *Maximum Entropy Classifier*, dan *K-Nearest Neighbor* bisa diterapkan dalam analisis sentimen [3]. Berdasarkan berbagai klasifikasi algoritma, peneliti memilih algoritma SVM. SVM memiliki keunggulan dalam penerapan pada data berdimensi tinggi. Namun, algoritma ini memiliki kelemahan dalam menangani data berjumlah besar. Umumnya, atribut dalam sentimen teks sangat banyak, sehingga jika semua atribut digunakan, kinerja SVM dapat menurun dan menyebabkan akurasi rendah.

*Feature selection*

Upaya untuk mengoptimalkan kinerja SVM dilakukan melalui *Feature selection* yang tepat.. Hafidzillah [2] menyatakan bahwa *Feature selection* bekerja dengan mengurangi atau menghapus fitur yang kurang relevan untuk mempermudah algoritma dalam memproses klasifikasi teks sentimen serta meningkatkan akurasi. Maulana dan Mandiri dalam penelitiannya menggunakan dataset Cornell dan Stanford menunjukkan bahwa *Support Vector Machine* berbasis information gain dapat meningkatkan akurasi sebesar 0,16% terhadap *dataset Stanford*. Dalam penelitian terkait analisis sentimen dengan SVM, algoritma *Naïve Bayes* dan SVM dibandingkan untuk melihat bagaimana keduanya dapat dioptimalkan dengan mengombinasikan bersama algoritma *Particle Swarm Optimization* (PSO). Algoritma *Naïve Bayes* dan PSO memiliki tingkat akurasi sebesar 79,07%, sementara SVM dan PSO memiliki tingkat akurasi sebesar 81,16% [4].

Penelitian oleh Somanti dan Aprilian menunjukkan bahwa tingkat akurasi SVM mengalami peningkatan ketika menggunakan *feature selection Information Gain* dibandingkan dengan *Chi Square*. Akurasi SVM dengan *Information Gain* meningkat dari 69,36% menjadi 72,45%[5]. Penelitian yang dilakukan oleh Sandi dkk mengklasifikasikan bantuan langsung tunai (BLT) menggunakan metode *naïve bayes* dan *gain ratio*. Metode ini merupakan pengembangan dari seleksi *fitur information gain*. Hasil penelitian menunjukkan peningkatan performa dalam proses pengujian menggunakan *confusion matrix* dengan validasi melalui metode *10 fold cross-validation*. Tingkat akurasi yang diperoleh adalah 86,58%, dengan presisi 94,17% dan *recall* 89,68%. [6].

Berdasarkan sejumlah penelitian yang telah disebutkan sebelumnya, terdapat peluang untuk meningkatkan akurasi algoritma SVM dengan menerapkan metode seleksi fitur menggunakan *gain ratio*. *Gain Ratio* merupakan pengembangan dari *Information Gain* yang berfungsi mengoptimalkan nilai yang dinormalisasi untuk sebuah fitur dalam klasifikasi. Diharapkan SVM berbasis *gain ratio* dapat menghasilkan nilai yang berbeda dan memperbaiki tingkat akurasi, sehingga lebih baik daripada penelitian sebelumnya. Dengan merujuk pada tinjauan objek penelitian dan struktur konseptual yang telah diuraikan sebelumnya, hipotesis dari hasil penelitian ini adalah diduga terdapat peningkatan akurasi algoritma SVM setelah menerapkan teknik *feature selection* *gain ratio*.

1. **Signifikansi Studi**
2. 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[5] | *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 *information gain*, akurasi meningkat menjadi 72,45%, yang merupakan peningkatan sebesar 3,08%. *Information gain* juga menunjukkan hasil yang lebih baik dibandingkan dengan metode chi-square. |
| 2 | Reza Maulana [8] | Peningkatan Akurasi Analisis Sentimen Review Film Menggunakan *Support Vector Machine* Berbasis *Information gain*.  Hasil penelitian menunjukkan bahwa SVM berbasis *Information gain* 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[9] | Analisis Sentimen terhadap Ulasan Film Berbahasa Inggris Menggunakan Metode *Support Vector Machine* dengan Feature Selection *Information gain*  Penelitian ini menunjukkan bahwa kombinasi *Stopword* dan *Stemming* meningkatkan akurasi hingga 86,12%. Meskipun *Information gain* (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 [10] | 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%. |

1. Tinjauan Pustaka
2. 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 [11].

1. 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 [12].

1. 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 [13].

1. 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 [14].

1. 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 [15].

1. 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 [16].

1. 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 [17].

1. Information gain

*Information gain* 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 [18].

1. 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 [19].

1. 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 [20].

1. Metode Penelitian

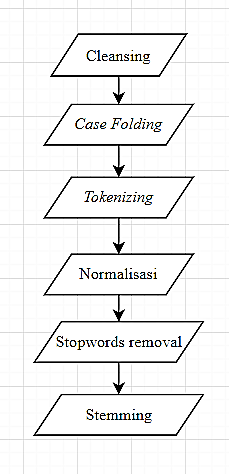
Penelitian ini bertujuan untuk membuat sebuah model analisis sentimen menggunakan metode *Support Vector Machine* (SVM) dengan menggunakan seleksi fitur *gain ratio*. Model ini akan diuji untuk melihat seberapa akuratnya dalam mengklasifikasikan sentimen dari postingan di *platform* X. Data yang digunakan 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.

1. 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.

1. Term weighting

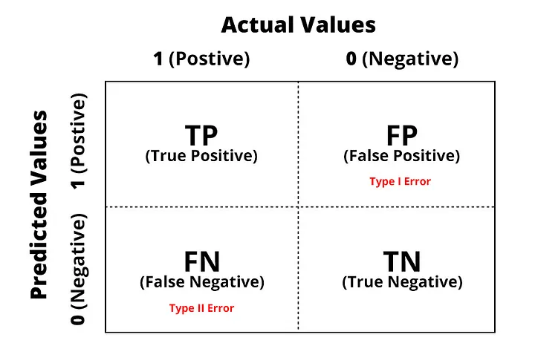
Kata-kata yang sudah dipangkas akarnya dan dihitung kepentingannya melalui proses *Term weighting*. Penimbangan istilah adalah teknik yang digunakan dalam menganalisis teks dan pengolahan bahasa alami untuk memberi skor pada kata-kata.

1. 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.

1. 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.

1. 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.

1. Instrumentasi

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

1. 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.
2. 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.
3. 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.

1. 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.

1. **Hasil dan Pembahasan**
   1. *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 ii  
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.

* 1. *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 IiI  
data hasil *preprocessing*

|  |  |
| --- | --- |
| ***Sentiment*** | ***TextTweet*-*Cleaned*** |
| *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 |

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.

1. *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 Iv  
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.

1. *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 *information gain* dan *gain ratio*. Dengan demikian, akurasi performa nilai SVM akan dibandingkan antara SVM tanpa *feature selection*, SVM dengan *feature selection information gain*, dan SVM dengan *feature selection gain ratio*. Setelah jumlah fitur diperoleh untuk gain ratio yaitu dengan penentuan bobot threshold *information gain* sebesar 0,0001 dan 0,002. Maka lakukan *Feature selection* terbaik menggunakan code dibawah ini.

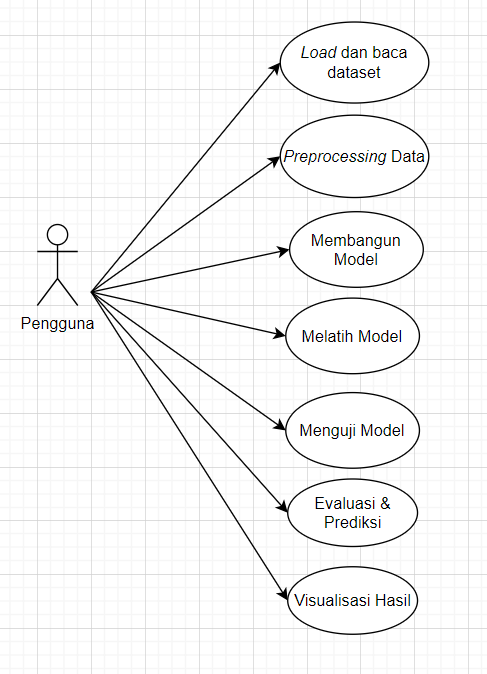
|  |  |
| --- | --- |
| k\_best\_info\_gain = SelectKBest(mutual\_info\_classif, k=num\_top\_features) # Pilih sejumlah fitur teratas  X\_train\_selected\_info\_gain = k\_best\_info\_gain.fit\_transform(X\_train\_tfidf, y\_train)  X\_test\_selected\_info\_gain = k\_best\_info\_gain.transform(X\_test\_tfidf) | (1) |

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.

* 1. *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*.

1. *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* *information gain* yaitu diperlihatkan oleh tabel 5 berikut:

Tabel v  
fitur terbaik *information gain* yang di peroleh

|  |  |  |
| --- | --- | --- |
| No | Feature Name | *Information gain* 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 |
| 11 | Kalah | 0,052415 |

Setelah menyajikan fitur terbaik yang diperoleh pada *feature selection* *information gain* 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')  svm\_classifier.fit(X\_train\_selected\_info\_gain, y\_train)  y\_pred\_svm= svm\_classifier.predict(X\_test\_selected\_info\_gain) | (2) |

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.

* 1. *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.

* 1. *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 information gain*. 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 information gain* 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 %.

1. *Evaluasi Pembadingan*

Berikut ini adalah evaluasi perbandingan model yang dibangun berdasarkan model SVM tanpa feature selection, SVM dengan TF-IDF, SVM dengan *Information gain*, 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 *information gain* 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.

1. *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*.

A screenshot of a computer

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

Gambar 6. Tampilan hasil prediksi sentimen

Pada gambar 6, terlihat bahwa aplikasi menampilkan hasil prediksi mengenai klasifikasi sentimen dengan dua kemungkinan nilai, yaitu Positif atau Negatif.

* 1. *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. Yang kedua adalah menggunakan metoda lain untuk feature selection selain *Information gain* ataupun Gain Ratio sebagai alternatif untuk *Feature selection* yang optimal. Terakhir, melakukan perbandingan algoritma klasifikasi SVM dengan algoritma lain sehingga dapat menganalisis algoritma dan metoda terbaik yang memberikan hasil prediksi klasifikasi sentimen lebih akurat.

1. **Kesimpulan**

Dari uraian permasalahan, tinjauan pustaka, studi sebelumnya, penelitian terkait, serta metode penelitian dalam meningkatkan kinerja *Support Vector Machine* (SVM) dengan pemilihan fitur *gain ratio*, dapat disimpulkan bahwa kinerja SVM dapat ditingkatkan melalui strategi tersebut. Dengan pemilihan bobot yang sesuai pada gain ratio, terbukti mendapatkan akurasi lebih baik dibandingkan dengan fitur yang diperoleh dari *feature selection Information gain*. Adapun proses pembobotan TF-IDF tidak terbukti bisa meningkatkan perfoma dan akurasi dibandingkan dengan SVM tanpa TF-IDF. Peningkatan akurasi signifikan terjadi pada feature selection gain ratio dengan bobot threshold 0,0001 (1732 fitur) yaitu dari 61,63% menjadi 71,51%. Sementara untuk bobot threshold 0,002 (518 fitur) akurasi mengalami sedikit peningkatan yaitu dari 61,63% menjadi 62,79%. Hal ini mengindikasikan bahwa: pertama, model SVM dengan *feature selection gain ratio* lebih baik dibandingkan dengan model SVM tanpa *gain ratio*. Kedua, dengan pemilihan *threshold* bobot *gain ratio* yang sesuai dapat memeroleh tingkat akurasi yang lebih baik dibandingkan dengan fitur terbaik yang diperoleh dari *feature selection* *Information gain.*

Mengacu kepada hasil penelitian optimasi SVM dengan *feature selection gain ratio*, dapat disampaikan beberapa saran untuk penelitian lebih lanjut, seperti melakukan penelitian lebih lanjut pada Validasi dengan beragam *dataset*, untuk memastikan keandalan model, penting untuk menguji model ini pada beragam dataset seperti misalnya data komentar social media, data sebuah artikel, dan lainnya. Adapun untuk kasus sentimen analisis, perlu dicoba dengan menambahkan jumlah dataset yang akan dijadikan sebagai bahan penelitian.

**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 Informatics J.*, 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,” *J. Pengemb. Teknol. …*, 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 Int. Conf. Power, Control Comput. Technol. 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,” *J. Penelit. Ilmu dan Teknol. Komput.*, vol. 12, no. 2, pp. 1–11, 2020.

[5] O. Somantri and D. Apriliani, “Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 537, 2018, doi: 10.25126/jtiik.201855867.

[6] S. A. Sandi, Y. Novianto, and S. A. Sandi, “Klasifikasi Kelayakan Keluarga Penerima Bantuan Langsung Tunai Menggunakan Gain Ratio Dan Naïve Bayes Jurnal Informatika Dan Rekayasa Komputer ( JAKAKOM ),” *J. Inform. Dan Rekayasa Komput.*, vol. 3, no. April, pp. 433–442, 2023.

[7] A. Sharma and S. Dey, “Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis,” *Int. J. Comput. Appl.*, no. June, pp. 15–20, 2012.

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

[9] 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 Eng.*, vol. 8, no. 5, pp. 10138–10152, 2021.

[10] 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 Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 2, pp. 314–324, 2022, doi: 10.29100/jipi.v7i2.2675.

[11] G. A. Buntoro, R. Arifin, G. N. Syaifuddiin, A. Selamat, O. Krejcar, and H. Fujita, “Implementation of a Machine Learning Algorithm for Sentiment Analysis of Indonesia‘s 2019 Presidential Election,” *IIUM Eng. J.*, vol. 22, no. 1, pp. 78–92, 2021, doi: 10.31436/IIUMEJ.V22I1.1532.

[12] 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.

[13] 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),” *J. Inform. dan Teknol. Inf.*, vol. 18, no. 1, pp. 49–60, 2021, doi: 10.31515/telematika.v18i1.4341.

[14] 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,” *J. Varian*, vol. 6, no. 1, pp. 89–96, 2022, doi: 10.30812/varian.v6i1.2205.

[15] S. Sunardi, A. Fadlil, and D. Prayogi, “Face Recognition Using Machine Learning Algorithm Based on Raspberry Pi 4b,” *Int. J. Artif. Intell. Res.*, vol. ISSN, no. 1, pp. 2579–7298, 2022, doi: 10.29099/ijair.v7i1.321.

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

[17] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, “A Review of Feature Selection and Classification Approaches for Heart Disease Prediction,” *IJITEE*, vol. 4, no. 3, 2020.

[18] 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. J. Comput.*, vol. 5, no. 2, pp. 107–118, 2020, doi: 10.21108/indojc.2020.5.2.499.

[19] F. N. Fajriyan, M. Ahsan, and W. Harianto, “Komparasi Tingkat Akurasi Information Gain Dan Gain Ratio Pada Metode K-Nearest Neighbor,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 1, pp. 386–391, 2022, doi: 10.36040/jati.v6i1.4694.

[20] Visitor Analytics, “Term Frequency Inverse Document Frequency (TF-IDF),” *Visit. Anal.*, no. December, 2023, [Online]. Available: https://www.visitor-analytics.io/es/glosario/t/term-frequency-inverse-document-frequency-tf-idf

[21] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayes dan Pembobotan Emoji,” *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.