CESS (Journal of Computing Engineering, System and Science) e-ISSN: 2502-714x
Vol. 9, No. 1, Januari 2024, pp.299-313 p-ISSN: 2502-7131

Contents list available at www.jurnal.unimed.ac.id

CESS
(Journal of Computing Engineering, System and Science)

journal homepage: https://jurnal.unimed.ac.id/2012/index.php/cess

Evaluasi Performa Naive Bayes dan SVM dalam Analisis Sentimen Kendaraan Listrik di Media Sosial Twitter

Evaluating the Performance of Naive Bayes and SVM in Sentiment Analysis of Electric Vehicles on Twitter Social Media

Gigih Nur Hendrawan[1], Harni Kusniyati[2*]

[1,2] Program Studi Teknik Informatika, Universitas Mercu Buana
Jl. Raya Meruya Selatan, Kembangan, Jakarta, DKI Jakarta 11650
email: [1]41518120005@student.mercubuana.ac.id, [2]harni.kusniyati@mercubuana.ac.id

A B S T R A K

Perkembangan teknologi dalam industri otomotif telah mengalami kemajuan yang signifikan, Salah satu pendorong utama perubahan ini adalah kebutuhan untuk mengatasi masalah lingkungan terutama pencemaran udara yang dihasilkan kendaraan bermotor yang berkontribusi terhadap perubahan iklim global. Kendaraan listrik dinilai sebagai salah satu solusi yang lebih ramah lingkungan dan berkelanjutan. Kendaraan listrik adalah jenis kendaraan yang menggunakan listrik sebagai sumber daya utama untuk menggerakkan mesin atau motor yang menggerakkan kendaraan tersebut. Jenis penelitian yang dipergunakan adalah penelitian kuantitatif yang mengacu pada pendekatan penelitian dengan cara mengumpulkan data yang dapat diukur secara numerik atau menggunakan metode statistik untuk menganalisis data tersebut. Tujuan utama penelitian ini adalah untuk memperbandingkan kinerja antara Algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam mengklasifikasikan sentimen masyarakat terhadap kendaraan listrik di media sosial Twitter, dengan fokus pada mengukur tingkat akurasi, recall, dan presisi dari kedua algoritma tersebut. Evaluasi komparatif antara Support Vector Machine (SVM) dan Naive Bayes dalam klasifikasi sentimen data Twitter menunjukkan bahwa SVM secara signifikan lebih unggul dengan akurasi 95.79% dibandingkan Naive Bayes yang memiliki akurasi 87.39%. SVM menonjol dalam mengklasifikasikan sentimen ''negatif'' dan ''positif'' dengan lebih akurat, sementara Naive Bayes cenderung melakukan lebih banyak kesalahan klasifikasi, walaupun SVM menunjukkan hasil yang menjanjikan, terdapat kekhawatiran mengenai overfitting.

Kata Kunci: *Analisi Sentimen, Suppor vector Machine, Naïve Bayes, kendaraan listrik.*

**Penulis Korespondensi:*
email: harni.kusniyati@mercubuana.ac.id

A B S T R A C T
Technological developments in the automotive industry have made significant progress, One of the main drivers of this change is the need to address environmental problems especially air pollution generated by motor vehicles that contribute to global climate change. Electric vehicles are considered as one of the more environmentally friendly and sustainable solutions. An electric vehicle is a type of vehicle that uses electricity as the main power source to drive the engine or motor that drives the vehicle. The type of research used is quantitative research which refers to a research approach by collecting data that can be measured numerically or using statistical methods to analyze the data. The main purpose of this study was to compare the performance between the Naïve Bayes Algorithm and the Support Vector Machine (SVM) in classifying public sentiment towards electric vehicles on social media Twitter, focusing on measuring the level of accuracy, recall, and precision of the two algorithms. A comparative evaluation between Support Vector Machine (SVM) and Naive Bayes in Twitter's sentiment data classification shows that SVM is significantly superior with 95.79% accuracy to Naive Bayes which has 87.39% accuracy. SVM stands out in classifying "negative" and "positive" sentiments more accurately, while Naive Bayes tends to make more misclassifications, although SVM shows promising results, there are concerns about overfitting.

Keywords: *Sentiment Analysis, Support Vector Machine, Naïve Bayes, electric vehicles*

1. PENDAHULUAN

Perkembangan teknologi dalam industri otomotif telah mengalami kemajuan yang signifikan selama beberapa tahun terakhir. Salah satu pendorong utama perubahan ini adalah kebutuhan untuk mengatasi masalah lingkungan terutama pencemaran udara yang dihasilkan kendaraan bermotor yang berkontribusi terhadap perubahan iklim global. Sehingga muncul salah satu kendaraan baru yaitu kendaraan listrik sebagai salah satu solusi yang lebih ramah lingkungan dan berkelanjutan. Kendaraan listrik adalah jenis kendaraan yang menggunakan listrik sebagai sumber daya utama untuk menggerakkan mesin atau motor yang menggerakkan kendaraan tersebut.

Namun seiring dengan kemajuan teknologi ini muncul banyak pro dan kontra dari masyarakat di Indonesia dan seringkali dinyatakan melalui media sosial misalnya seperti *Twitter.* Adapun *Twitter* dianggap sebagai *platform* yang memungkinkan pengguna untuk mengekspresikan pemikiran dan opini mereka dengan lebih bebas, mudah karena aksesibilitas nya, jumlah pengikut yang tidak terbatas dan batas karakter hanya 280 karakter [1]. Hal ini memungkinkan pengguna untuk menyampaikan pesan mereka dengan jelas, singkat dan efektif [2]. Melalui media sosial ini masyarakat berbagi berbagai tanggapan seperti kritik, saran, pengalaman dan informasi terkait kendaraan listrik yang nantinya data tanggapan tersebut dapat dijadikan sebagai sumber informasi yang berharga dalam memahami opini dan reaksi masyarakat terhadap peristiwa tersebut.

Data tanggapan atau opini yang ditulis oleh masyarakat pada media sosial *Twitter* dapat diklasifikasikan menggunakan analisis sentimen [3]. Sentiment analysis adalah suatu teknik yang dipergunakan untuk menganalisis sudut pandang, emosi, dan sikap yang diungkapkan oleh masyarakat terhadap suatu topik [4]. Namun banyaknya informasi dan pendapat di platform ini sering kali sulit diolah dan dianalisis secara manual untuk menentukan sentimen masyarakat terkait fenomena teknologi kendaraan listrik. Maka dari hal tersebut diperlukan

sebuah mesin yang dapat melakukan analisis sentimen secara otomatis dan mengklasifikasikan sentimen dari pendapat masyarakat tersebut menjadi negatif dan positif yaitu dengan melakukan klasifikasi teks.

Salah satu algoritma klasifikasi yang sering digunakan adalah algoritma Naïve *Bayes*. Algoritma *Naïve Bayes* adalah algoritma untuk mengklasifikasikan data dengan cara yang sangat sederhana dalam mengasumsikan klasifikasi atribut [5]. Algoritma ini banyak dipergunakan dan dikenal memiliki perhitungan sederhana dan tingkat akurasi yang baik [6],[7]. Selain itu terdapat metode lain yang sering digunakan dalam klasifikasi text yaitu *Support Vector Machine* (SVM). SVM adalah model yang berasal dari teori pembelajaran statistika yang digunakan untuk tugas klasifikasi dan regresi. SVM dikenal karena kemampuannya dalam memisahkan dua kelas data dengan batas keputusan yang optimal.

Berdasarkan masalah tersebut maka peneliti akan membandingkan hasil evaluasi kinerja dari Algoritma *Naïve Bayes* dan *Support Vector Machine (SVM)* dalam mengklasifikasikan sentimen masyarakat pada media sosial *Twitter* terhadap kendaraan listrik. Dengan penelitian ini diharapkan dapat memperoleh pengetahuan mengenai algoritma mana yang lebih baik dalam mengklasifikasikan sentimen masyarakat terhadap kendaraan listrik juga dapat memberikan pengetahuan mengenai sentimen masyarakat indonesia saat ini terkait teknologi kendaraan listrik.

2. DASAR/TINJAUAN TEORI

2.1. Analisis Sentimen

Menurut Penelitian Al-Ayyoub, dkk yang dikutip pada penelitian [8] bahwasanya Analisis sentimen adalah ranah penelitian yang terus berkembang dan berada di persimpangan berbagai cabang ilmu seperti Penambangan Data, Pemrosesan Bahasa Alami *(Natural Language Processing/NLP),* dan Pembelajaran Mesin (Machine Learning). Fokus utamanya adalah untuk mengidentifikasi dan mengekstraksi sentimen atau perasaan yang terkandung dalam kalimat berdasarkan kontennya. Dalam analisis sentimen, peneliti berusaha untuk memahami apakah teks memiliki sentimen positif, negatif, atau netral, serta tingkat ekspresinya. Kemudian menurut Vinodhini dan Chandrasekaran yang dikutip pada penelitian (Sari & Hayuningtyas, 2019) Analisis sentimen adalah bidang ilmu yang bersifat interdisipliner dimana pendekatannya melibatkan penggunaan perspektif dari berbagai disiplin ilmu yang relevan secara bersamaan dan terintegrasi dalam pemecahan masalah.

2.2. *Text Mining*

Menurut Penelitian Retno, dkk [10] *Text mining* adalah istilah yang digunakan untuk merujuk pada metode yang bertujuan menghasilkan informasi baru yang lebih spesifik atau mudah dimengerti dari sejumlah dokumen. Secara umum, *text mining* melibatkan proses ekstraksi inti atau pokok dari dokumen-dokumen teks yang tidak terstruktur. Proses *text mining* melibatkan berbagai teknik preprocessing teks, seperti pencarian, ekstraksi data, dan kategorisasi.

2.3. TF – IDF

Pembobotan TF-IDF adalah proses yang mengubah teks menjadi data numerik dengan memberikan bobot pada setiap kata. Ini menggunakan metrik statistik untuk menilai pentingnya kata dalam dokumen. Komponen TF mengukur seberapa sering kata muncul dalam dokumen, sementara DF mengukur seberapa umum kata itu dalam dokumen-dokumen lane. IDF adalah kebalikannya dan mengindikasikan seberapa unik kata itu dalam korpus data.

TF-IDF digunakan untuk menilai pentingnya kata dalam sebuah dokumen dengan menggabungkan frekuensi kemunculan kata dalam dokumen (TF) dan inversi frekuensi dokumen (IDF) berdasarkan pencarian atau query yang digunakan. Proses ini melibatkan tokenisasi, penghapusan *stopwords,* dan *stemming* untuk menghasilkan nilai TF-IDF yang mencerminkan signifikansi kata dalam konteks dokumen tersebut [13] [14].

2.4. *Naïve Bayes*

Algoritma *Naïve Bayes* digunakan untuk mengklasifikasikan data dengan cara yang cukup sederhana, yaitu dengan mengasumsikan independensi antara atribut-atribut yang digunakan dalam klasifikasi. Meskipun sederhana Algoritma *Naïve Bayes* sering digunakan dalam berbagai aplikasi *machine learning* karena kemampuannya yang terbukti memberikan tingkat akurasi yang tinggi dengan perhitungan yang relatif mudah.[15]. Berikut adalah rumus *Naïve Bayes* dalam menghitung probabilitas fitur yaitu [16] :

$$P(w|positif/negatif) = \frac{\left(nk\left|\frac{positif}{negatif}\right.\right)+1}{\left(n,\left|\frac{pos}{neg}\right.\right)+|kosakata|}$$

Keterangan:

$P(w\|positif/negatif)$	: Peluang kemunculan kata pada kategori
Wk	: Kata yang muncul pada sebuah kategori
$\left(nk\left\|\frac{positif}{negatif}\right.\right)+1$	: Jumlah frekuensi kemunculan kata pada kategori
Nk	: Kemunculan setiap kata pada kategori
$\|kosakata\|$	: Jumlah semua kata dari semua kategori

Kemudian Untuk rumus yang digunakan dalam menghitung probabilitas kategori pada data training yaitu sebagai berikut:

$$P\left(\frac{pos}{neg}\middle|d\right) = p\left(\frac{pos}{neg}\right) * \prod iP(a1|\frac{pos}{neg})$$

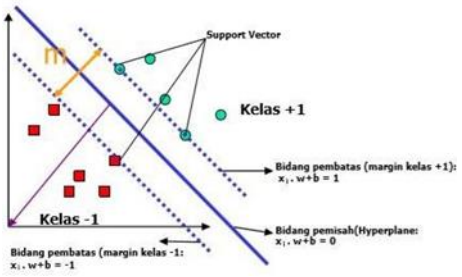
Keterangan:

$p\left(\frac{pos}{neg}\right)$	: Peluang kemunculan kata pada kategori atau kelas dengan data uji
d	: Dokumen data baru
$* \prod iP(a1\|\frac{pos}{neg})$	: Peluang kemunculan ai pada kategori atau kelas
ai	: Kata baru yang akan diuji

2.5. SVM

Support Vector Machine (SVM) termasuk algoritma pembelajaran mesin populer yang digunakan supaya tugas klasifikasi dan regresi. SVM adalah jenis algoritma pembelajaran terawasi yang dapat menangani data linier dan non linier. Ide dasar SVM adalah menemukan hyperplane dalam ruang berdimensi tinggi yang memisahkan kelas yang berbeda dengan jarak sejauh mungkin. Dengan kata lain, SVM mencoba menemukan batas keputusan terbaik yang dapat mengklasifikasikan titik data dengan benar sambil mempertahankan jarak maksimum antara batas keputusan dan titik data terdekat. Untuk mencapai hal ini, SVM menggunakan teknik yang disebut trik kernel, yang memetakan kumpulan data asli ke ruang fitur berdimensi lebih tinggi di mana hyperplane diskriminan lebih mudah ditemukan. Kernel SVM yang paling umum dipakai termasuk kernel linier, kernel polinomial, serta kernel fungsi basis radial (RBF).

Dalam SVM, tujuannya adalah mencari manfaat klasifikasi paling bagus supaya tidak menyamakan anggota antar dua kelas untuk data pelatihan. Ukuran pengertian fungsi klasifikasi yang "optimal" bisa direalisasikan dengan cara geometris. Pada dataset yang dapat dipisahkan dengan linear, manfaat klasifikasi linear sesuai dengan hyperplane f(x) pemisah yang melampaui pusat dua kelas memisahkannya.



Gambar 1. Margin Hyperplane

SVM (*Support Vector Machine*) adalah metode yang digunakan untuk melakukan klasifikasi atau regresi. Terdapat beberapa variasi SVM yang berbeda, tetapi yang paling umum adalah SVM dengan fungsi kernel. Rumus dasar SVM untuk klasifikasi adalah sebagai berikut:

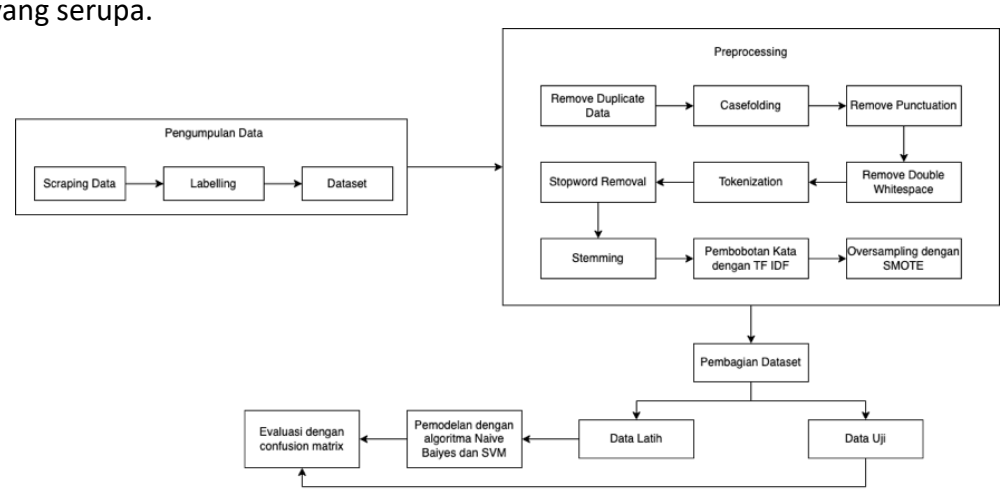
$$K(X_1, X_j = \ tanh\ ax_1 + X_j\ + \beta$$

Keterangan:

K : *Kernel*
X & Y : *Vector input space*
d :*quadratic*
α : *scalar parameter*

3. METODE

Bab ini menjelaskan metodologi yang digunakan dalam penelitian ini, yang bertujuan untuk menyediakan sebuah kerangka kerja yang dapat diikuti untuk mereplikasi studi dalam kondisi yang serupa.



Gambar 2. Alur Tahapan Penelitian

Proses penelitian diawali dengan pengumpulan data, yang melibatkan scraping data dari *twitter* dan kemudian melabeli data tersebut untuk membentuk sebuah dataset yang akan

digunakan dalam penelitian. Selanjutnya, dataset tersebut diproses melalui serangkaian tahap preprocessing, yang mencakup penghapusan data duplikat, *casefolding,* tokenisasi, dan proses pembersihan lainnya seperti penghapusan tanda baca dan spasi berlebih serta *stemming*.

Selanjutnya adalah pemberian bobot pada kata-kata melalui pendekatan TF-IDF (*Term Frequency-Inverse Document Frequency*) dan menangani ketidakseimbangan kelas dengan teknik *oversampling* menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Dataset yang telah diolah dibagi menjadi dua: data latih dan data uji, yang mana data latih digunakan untuk membangun model klasifikasi menggunakan algoritma Naive Bayes dan SVM (Support Vector Machine), sementara data uji digunakan untuk mengevaluasi kinerja model tersebut dengan menggunakan confusion matrix. Alur tahapan penelitian ini dirancang untuk memastikan bahwa model klasifikasi yang dikembangkan dapat secara efektif mengidentifikasi dan mengklasifikasikan data sesuai dengan label yang telah ditentukan.

3.1. Pengumpulan Data

3.1.1. Scraping

Penelitian ini melakukan pengumpulan data dari *Twitter* menggunakan *tweet*-harvest dengan kata kunci yang berkaitan dengan kendaraan listrik. Data yang diambil adalah tweet yang diposting dengan rentang waktu dari 1 Januari 2018 hingga 11 November 2023.

3.1.2. Labelling

Pelabelan dataset adalah tahap penting dalam analisis sentimen yang melibatkan penentuan kelas atau sentimen yang sesuai untuk setiap tanggapan *Twitter* dalam dataset. Dalam penelitian ini terdapat tiga kategori kelas yang digunakan yaitu positif, negatif dan netral. Proses pelabelan dilakukan secara manual dengan melihat kata pada tanggapan yang ada. Selanjutnya data yang telah dilabeli dengan baik akan digunakan untuk melatih dan menguji model analisis sentimen untuk mengklasifikasikan sentimen masyarakat terkait kendaraan listrik.

3.2. Preprocessing

Setelah dilakukan pengumpulan data *tweet* dan labelling maka akan dilakukan tahapan *preprocessing*. *Preprocessing* adalah tahapan awal dalam pengolahan data di mana data mentah atau tidak terstruktur diubah menjadi data yang lebih terstruktur dan siap digunakan untuk analisis lebih lanjut [18]. Pada penelitian ini dilakukan enam tahapan *Preprocessing* meliputi:

3.2.1. *Remove Duplicate Data*

Sebagai langkah awal dalam tahap *preprocessing*, perlu dilakukan penghapusan data duplikat untuk memastikan kebersihan dan konsistensi dataset. Penghapusan ini dilakukan dengan tujuan menghindari adanya pengaruh berlebihan dari informasi yang sama dalam proses analisis sentimen.

3.2.2. *Casefolding*

Dalam tahap ini, dilakukan proses *casefolding* untuk mengubah semua teks ke dalam bentuk huruf kecil (Rusdiaman & Rosiyadi, 2019). Tujuan dari casefolding adalah untuk memastikan konsistensi dalam penanganan teks, sehingga kata yang ditulis dengan huruf besar atau kecil memiliki representasi yang seragam.

3.2.3. *Remove Punctuation*

Di bawah ini merupakan hasil dari *remove punctuation* atau penghapusan tanda baca. Tujuan dari penghapusan tanda baca ini biasanya untuk menyederhanakan teks sehingga

hanya informasi kata yang tersisa, yang memudahkan pemrosesan lebih lanjut (Kurniasari, 2020).

3.2.4. *Remove Double Whitespace*

Tahap selanjutnya yaitu *remove double whitespace* atau penghapusan spasi ganda. Spasi ganda ini bisa mengganggu proses analisis teks atau pemrosesan bahasa alami (NLP) karena dapat menyebabkan inkonsistensi dalam mengidentifikasi batasan kata.

3.2.5. *Tokenization*

Selanjutnya yaitu tahapan tokenisasi yang dimana tahapan ini adalah proses dalam memecah teks menjadi bagian-bagian kecil. Tujuan dari tokenisasi adalah untuk memudahkan analisis lebih lanjut dari teks dengan membuatnya menjadi bagian-bagian yang dapat dikelola dan diproses oleh algoritma (Lestari et al., 2023).

3.2.6. *Stopword Removal*

Selanjutnya adalah tahapan *stopword removal* atau penghapusan kata yang tidak penting. Dalam pembuatan model NLP, menghilangkan *stopwords* sering kali dapat meningkatkan performa model dengan mengurangi dimensi data input dan menghindari pembelajaran dari fitur-fitur yang kurang informatif (Wibawa et al., 2021). Berikut merupakan hasil dari tahapan *stopword removal*.

3.2.7. *Stemming*

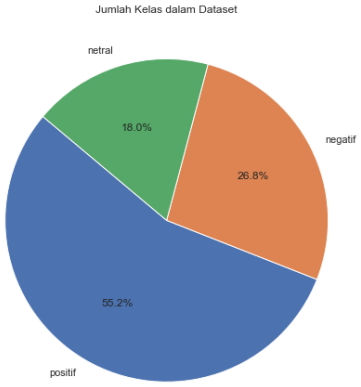
Tahapan selanjutnya dalam preprocessing yaitu *stemming*. *Stemming* adalah proses dalam pemrosesan bahasa alami (*Natural Language Processing* - NLP) untuk mengurangi kata-kata ke bentuk dasar.

3.2.8. Pembobotan TF – IDF

Selanjutnya sebelum proses implementasi akan dilakukan terlebih dahulu pembobotan TF-IDF untuk menentukan bobot kata pada dokumen. Metode ini menggabungkan dua faktor kunci yaitu frekuensi kata dalam dokumen (TF) dan sejauh mana kata tersebut umum atau jarang muncul dalam seluruh kumpulan dokumen (IDF). Dengan melakukan TF-IDF maka akan membantu mengidentifikasi kata-kata yang memiliki kontribusi yang tinggi terhadap makna atau isi dari suatu dokumen tertentu.

3.2.9. *Oversampling* dengan SMOTE

Selanjutnya yaitu tahapan *oversampling*. teknik ini digunakan untuk menangani ketidakseimbangan kelas dalam kumpulan data. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam setiap kelas tidak seimbang; yaitu, satu kelas (kelas minoritas) memiliki jumlah sampel yang jauh lebih sedikit dibandingkan dengan kelas lain (kelas mayoritas) (Rodríguez-Torres et al., 2022).



Gambar 3. *Distribusi Kelas Sebelum Oversampling*

Dapat dilihat dari gambar 3 yang mana distribusi kelas positif jauh lebih banyak dibanding kelas netral dan negative. Oleh sebab itu, peneliti melakukan *oversampling* untuk menangani ketidak seimbangan distribusi kelas. Tujuan dari oversampling adalah untuk meningkatkan jumlah sampel dalam kelas minoritas sehingga jumlah sampel antar kelas menjadi lebih seimbang. Hal ini dilakukan untuk memastikan bahwa model pembelajaran mesin tidak bias terhadap kelas mayoritas dan dapat belajar fitur dari kelas minoritas dengan lebih efektif. Adapun Teknik yang digunakan yaitu SMOTE dimana cara kerjanya yaitu sampel baru dari kelas minoritas dibuat secara sintetis dengan cara interpolasi antara sampel minoritas yang ada. Berikut merupakan hasil dari oversampling dengan SMOTE.

3.3. Pembagian dataset

Sebelum masuk ke tahap *processing* atau pemodelan dengan algoritma SVM dan Naïve baiyes, perlu dilakukan pembagian data terlebih dahulu. Data dibagi menjadi 80% untuk data latih dan 20% data uji. Tujuan dari pembagian ini agar algoritma dapat melakukan pembelajaran menggunakan data latih. Lalu model yang telah dibangun dapat diuji menggunakan data uji. Pembagian data ini menghasilkan 1425 rows untuk data latih dan 357 rows untuk data uji.

3.4. Pemodelan Dengan *Naïve Bayes* dan SVM

Tahapan selanjutnya yaitu pemodelan dengan algoritma naïve baiyes dan Support Vector Machine (SVM). Untuk melakukan pemodelan dengan SVM, peneliti menggunakan metode Grid Search CV untuk mengoptimalkan parameter model. Grid Search CV menyediakan kerangka kerja yang sistematis untuk mengeksplorasi berbagai kombinasi parameter guna menemukan konfigurasi optimal berdasarkan metrik performa yang ditentukan. Adapun parameter yang digunakan untuk melakukan tuning parameter dengan Grid Search CV adalah sebagai berikut.

Tabel 1. Parameter tuning SVM

Parameter	Value
C	0.1, 1, 10
Kernel	Linear, rbf, poly

Sedangkan untuk pemodelan dengan algoritma naïve baiyes, peneliti menggunakan Multinomial Naive Bayes yang merupakan salah satu varian dari algoritma Naive Bayes yang cocok untuk digunakan pada data yang berdistribusi multinomial, yang umumnya adalah data frekuensi kata dalam teks

3.5. Evaluasi

Pada tahapan ini dilakukan pengujian *model performance* dari kedua model yang telah dihasilkan menggunakan confusion matrix. *Confussion matrix* adalah metode penghitungan yang membandingkan hasil klasifikasi dengan data sebenarnya. Matrik ini menunjukkan tingkat akurasi dalam persentase dan berguna sebagai acuan untuk menilai performa algoritma klasifikasi (Hasanah et al., 2019).

Tabel 2. *Confusion Matrix*

Confusion Matrix		Predicted		
		Class 1	Class 2	Class 3
Actual	*Class 1*	*TP*	C_{12}	C_{13}
	Class 2	C_{21}	*TP*	C_{23}
	Class 3	C_{31}	C_{32}	*TP*

Confusion Matrix dalam kinerjanya dapat dinilai dengan empat parameter: TP, FN, FP, dan TN. TP (*True Positive*) adalah sejumlah data yang terbukti benar dan dideteksi oleh sistem sebagai benar [19]. FN (*False Negative)* adalah sejumlah data yang terbukti salah dan dideteksi oleh sistem sebagai salah. FP (*False Positive*) adalah sejumlah data terbukti benar namun dideteksi oleh sistem sebagai salah. TN (*True Negative*) ialah sejumlah data terbukti salah dan dideteksi oleh sistem sebagai benar

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengumpulan data

4.1.1. Hasil Scraping

Pada tahap ini, peneliti menggunakan tweet-harvest untuk melakukan *crawling* data dari Twitter, dengan beberapa kata kunci yaitu kendaraan listrik, mobil listrik, motor listrik, EV, kendaraan elektrik, PLN EV mobil hybrid, baterai elektrik, pengisian daya mobil listrik, teknologi EV, perkembangan kendaraan listrik, infrastruktur pengisian daya, kendaraan ramah lingkungan, emisi nol, mobilitas berkelanjutan, kendaraan listrik terbaru, otomotif elektrik, mobil listrik terjangkau, mobil listrik mewah, konsep mobil listrik, inovasi kendaraan listrik.

Proses pengumpulan data dilakukan dalam rentang waktu mulai dari 1 Januari 2018 hingga 11 November 2023. Hasil dari proses crawling ini berhasil mengumpulkan sebanyak 1133 baris data terkait dengan topik kendaraan listrik.

Data yang diperoleh mencakup beragam informasi seperti teks tweet, tanggal dan waktu posting, serta metadata lainnya. Namun, pada penelitian ini hanya diambil data teks saja. Keberhasilan pengumpulan data menjadi kunci dalam memahami sentimen pengguna Twitter terhadap kendaraan listrik selama periode waktu yang ditentukan. Berikut merupakan lima sampel data yang telah diperoleh dari hasil *crawling*.

Tabel 3. *Sampel Hasil Crawling*

No.	Teks
1.	Debat Tim Kampanye Prabowo, Anies, dan Ganjar Soal Kendaraan Listrik https://t.co/AKAvunM4tn #TempoBisnis
2.	Promo Kepala Semprotan Air Tanpa Listrik Alat Buat Cuci Mobil Motor Kendaraan Semprotan Pistol Air Nozzle Spray Hose Water Gun C02 dengan harga Rp54.900. Klik link dibawah ini kk terimaksh https://t.co/xvJU7DrTYk https://t.co/f0mt6mX56s
3.	@jokowi Mesin dan kendaraan listrik tuk di pelajari.... buat bahan jualan 2024 ya pa? Eh.... Btw Itenas kampus pertama di jabar yang udah ngembangin mobil listrik pa coba liat deh dah ada banyak hasil

4.1.2. Hasil *Labelling*

Dalam tahap labelling data, peneliti melakukan proses *labelling* secara manual untuk memberikan label sentimen pada setiap tweet yang telah dikumpulkan dari Twitter. Metode *labelling* manual dipilih dengan tujuan untuk memastikan akurasi dan keakuratan hasil, mengingat kompleksitas dan variasi dalam ekspresi sentimen manusia yang sulit dipahami oleh algoritma otomatis.

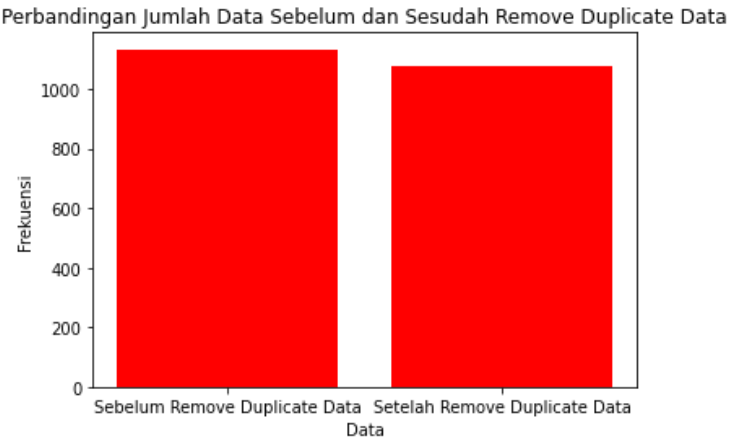
Metode *labelling* ini mengklasifikasikan sentimen ke dalam tiga kelas utama, yaitu positif, negatif, dan netral. Berikut merupakan lima sampel hasil *labelling*.

Tabel 4. *Sample Hasil Labelling*

Teks	Class
Debat Tim Kampanye Prabowo, Anies, dan Ganjar Soal Kendaraan Listrik https://t.co/AKAvunM4tn #TempoBisnis	netral
Promo Kepala Semprotan Air Tanpa Listrik Alat Buat Cuci Mobil Motor Kendaraan Semprotan Pistol Air Nozzle Spray Hose Water Gun C02 dengan harga Rp54.900. Klik link dibawah ini kk terimaksh https://t.co/xvJU7DrTYk https://t.co/f0mt6mX56s	netral
@jokowi Mesin dan kendaraan listrik tuk di pelajari.... buat bahan jualan 2024 ya pa? Eh.... Btw Itenas kampus pertama di jabar yang udah ngembangin mobil listrik pa coba liat deh dah ada banyak hasil	negatif

4.2. Preprocessing

Tahap awal preprocessing adalah penghapusan duplikat. Proses ini penting untuk memastikan bahwa setiap sampel dalam dataset adalah unik, sehingga menghindari bias dalam analisis selanjutnya. Dari visualisasi pada gambar, terlihat bahwa jumlah data berkurang setelah duplikat data dihapus.



Gambar 4. Perbandingan Jumlah Data Sebelum dan Sesudah Penghapusan Duplikat Data

Selanjutnya yaitu tahapan *Casefolding*, tokenisasi, dan proses pembersihan seperti penghapusan tanda baca dan spasi berlebih serta *stemming*. merupakan tahapan esensial dalam preprocessing data teks yang bertujuan untuk menghomogenkan dataset, memudahkan analisis, dan meningkatkan kinerja algoritma pembelajaran mesin. *Casefolding* mengkonversi semua teks ke huruf kecil untuk menghilangkan distingsi antara penggunaan huruf besar dan kecil, sedangkan tokenisasi memecah teks menjadi unit-unit dasar (token) yang memfasilitasi pemrosesan dan analisis lebih lanjut. Proses pembersihan mengeliminasi tanda baca dan spasi yang tidak perlu, yang bisa mengganggu proses analisis teks dan mengurangi efektivitas model prediktif yang dikembangkan. Keseluruhan proses ini memastikan konsistensi dalam data, yang vital untuk pengolahan alami bahasa dan tugas-tugas klasifikasi teks.

	clean_sw	stemmed
0	debat, tim, kampanye, prabowo, anies, dan, gan...	debat tim kampanye prabowo anies dan ganjar so...
1	promo, kepala, semprotan, air, tanpa, listrik,...	promo kepala semprot air tanpa listrik alat bu...
2	jokowi, mesin, dan, kendaraan, listrik, tuk, d...	jokowi mesin dan kendara listrik tuk di ajar b...
3	live, streaming, launching, program, menuju, 1...	live streaming launching program tuju 100 kend...
4	dinas, perhubungan, dishub, kota, surabaya, me...	dinas hubung dishub kota surabaya proyeksi jum...
...	...	...
1072	kolaborasi, gelar, ev, fun, day, di, bandung, ...	kolaborasi gelar ev fun day di bandung pln sia...
1073	kolaborasi, gelar, ev, fun, day, di, bandung, ...	kolaborasi gelar ev fun day di bandung pln sia...
1074	dorong, program, kendaraan, listrik, pln, gela...	dorong program kendara listrik pln gelar ev fu...
1075	nurrohmah, mantaplah, pln, sudah, melakukan, d...	nurrohmah mantap pln sudah laku digitalisasi s...
1076	halo, alumni, itb, dan, indonesia, electric, v...	halo alumni itb dan indonesia electric vehicle...

Gambar 5. Hasil Preprocessing

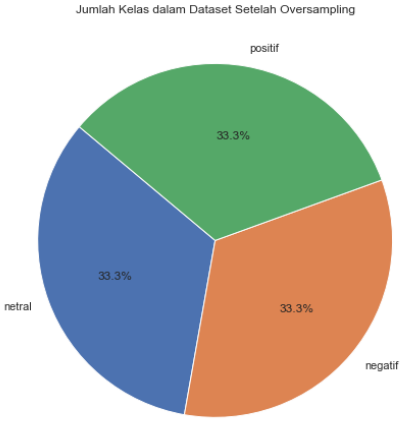
Selanjutnya yaitu pembobotan kata dengan TF iDF. Seperti yang telah dijelaskan pada sub bab sebelumnya, Tahapan ini digunakan dalam pemrosesan bahasa alami dan pencarian informasi untuk menilai seberapa penting suatu kata dalam sebuah dokumen yang merupakan bagian dari kumpulan dokumen (korpus). Berikut merupakan hasil dari pembobotan kata dengan TF iDF.

	00	000	00789mutiara	020	028	09	090	10	100	1001	...	yukpakaimolis	yush	zakat	zaman	zero	zetribandaro	zinchenko	zone	zul	zuli
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.172819	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1072	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1073	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1074	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1075	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1076	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.140519	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

1077 rows × 4514 columns

Gambar 6. *Hasil TF iDF*

Tahapan terakhir dari preprocessing yaitu *oversampling* dimana tahapan ini digunakan untuk menyeimbangkan jumlah kelas. Dapat dilihat pada gambar 7 bahwa jumlah dari ketiga kelas telah seimbang.



Gambar 7. *Distribusi Kelas Setelah Oversampling*

4.3. Pemodelan dengan algoritma SVM dan Naïve Baiyes

Setelah melakukan seluruh tahapan preprocessing, selanjutnya yaitu pemodelan dengan algoritma SVM. Gambar 7 merupakan hasil pemodelan dengan SVM.

```
grid_search.score(X_test,y_test)
0.957983193277311

grid_search.score(X_train, y_train)
1.0
```

Gambar 8. *Skor Latih dan Skor Uji SVM*

Hasil yang didapatkan pada gambar 7 menunjukkan bahwa model yang dilatih mampu mencapai akurasi sempurna (100%) pada data latih dan akurasi yang sangat tinggi pada data uji (sekitar 96%). Akurasi sempurna pada data latih dapat menunjukkan bahwa model telah mempelajari fitur-fitur data dengan sangat baik, meskipun ada potensi terjadinya fenomena overfitting, di mana model terlalu menyesuaikan diri dengan data latih sehingga *performance* nya pada data yang belum pernah dilihat bisa berkurang. Namun, nilai akurasi yang tinggi pada data uji menandakan bahwa model memiliki kemampuan generalisasi yang baik, dengan asumsi distribusi data uji yang representatif terhadap populasi umum.

```
# Tampilkan parameter terbaik yang ditemukan oleh grid search
print("Parameter Terbaik:", grid_search.best_params_)
Parameter Terbaik: {'C': 10, 'kernel': 'rbf'}
```

Gambar 9. *Parameter terbaik yang dihasilkan*

Berdasarkan output dari Grid Search pada gambar 8, diketahui bahwa parameter 'C' dengan nilai 10 dan kernel 'rbf' (Radial Basis Function) merupakan konfigurasi yang paling efektif untuk model ini dalam kumpulan data yang digunakan. Nilai 'C' yang tinggi menunjukkan preferensi model terhadap margin klasifikasi yang lebih lebar dengan toleransi yang lebih rendah terhadap kesalahan klasifikasi pada data latih.

Selain itu, dapat dilihat pada gambar 9, algoritma naïve baiyes pada penelitian ini mendapatkan akurasi sebesar 87% pada data uji dan 94% pada data latih. hasil ini mengalami overfitting seperti processing dengan SVM yang telah dilakukan sebelumnya.

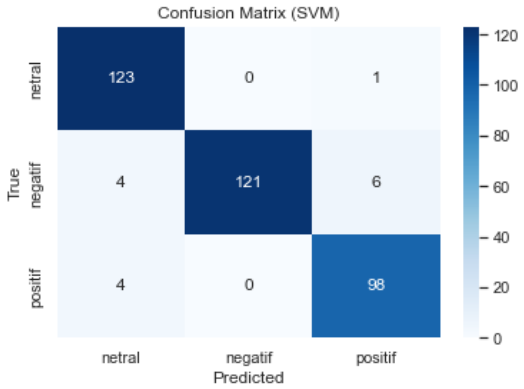
```
model_naive_bayes.score(X_test,y_test)
0.8739495798319328

model_naive_bayes.score(X_train, y_train)
0.9375438596491228
```

Gambar 10. *Skor Uji dan Skor Latih Naïve Baiyes*

4.4. *Evaluasi Model SVM*

Pada penelitian ini, model klasifikasi yang dikembangkan menggunakan metode *Support Vector Machine* (SVM) telah menunjukkan performa yang sangat bagus, dengan akurasi keseluruhan mencapai 95.79%.

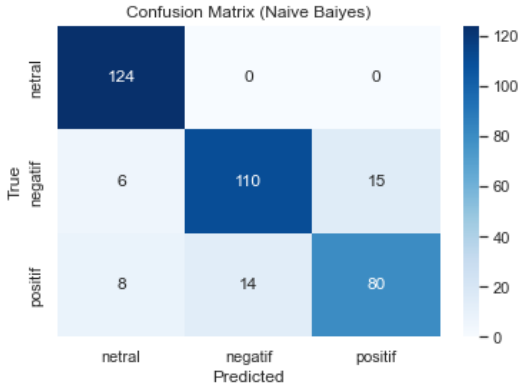


Gambar 11. Confusion matrix *(SVM)*

Confusion matrix yang dihasilkan dari model SVM menunjukkan distribusi prediksi yang komprehensif terhadap kelas sebenarnya. Model berhasil mengklasifikasikan 123 instansi dengan tepat sebagai kelas 'netral', dan 121 instansi sebagai kelas 'negatif'. Sejalan dengan itu, 98 instansi berhasil dikategorikan secara benar sebagai kelas 'positif'. Terdapat indikasi kesalahan klasifikasi yang minimal, dengan hanya 1 instansi kelas 'netral' yang salah diklasifikasikan sebagai 'positif', 4 instansi kelas 'negatif' yang salah diklasifikasikan sebagai 'netral', dan 6 instansi kelas 'positif' yang salah diklasifikasikan sebagai 'negatif'. Selanjutnya, terdapat 4 instansi kelas 'netral' yang salah diklasifikasikan sebagai 'negatif', namun tidak ada kesalahan klasifikasi dari 'positif' ke 'netral'.

4.5. *Evaluasi Model Naïve Baiyes*

Pada penelitian ini, model klasifikasi yang dikembangkan menggunakan metode Naïve Baiyes telah menunjukkan dengan akurasi keseluruhan mencapai 87.39%.



Gambar 12. Confusion Matrix *(Naïve Baiyes)*

Berdasarkan *confusion matrix* yang disertakan untuk model Naive Bayes, dapat dilihat bahwa model tersebut memiliki kemampuan prediksi yang baik untuk kelas 'netral' dengan semua 124 instansi diklasifikasikan dengan benar, tidak ada kesalahan klasifikasi (false positives atau false negatives). Untuk kelas 'negatif', model mengklasifikasikan 110 instansi dengan benar, tetapi terdapat 6 false negatives (instansi 'negatif' yang salah diklasifikasikan sebagai 'netral') dan 15 false positives (instansi dari kelas lain yang salah diklasifikasikan sebagai 'negatif'). Pada kelas 'positif', terdapat 80 prediksi yang benar, dengan 14 false negatives (instansi 'positif' yang salah diklasifikasikan sebagai 'negatif') dan 8 false positives (instansi 'positif' yang salah diklasifikasikan sebagai 'netral').

5. KESIMPULAN

Dalam penelitian ini, evaluasi komparatif antara Support Vector Machine (SVM) dan Naive Bayes dalam klasifikasi sentimen data Twitter menunjukkan bahwa SVM secara signifikan lebih unggul dengan akurasi 95.79% dibandingkan Naive Bayes yang memiliki akurasi 87.39%. SVM menonjol dalam mengklasifikasikan sentimen 'negatif' dan 'positif' dengan lebih akurat, sementara Naive Bayes cenderung melakukan lebih banyak kesalahan klasifikasi, seperti yang terlihat dalam confusion matrix. Walaupun SVM menunjukkan hasil yang menjanjikan, terdapat kekhawatiran mengenai overfitting, yang ditandai oleh perbedaan performa antara data latih dan uji, suatu isu yang juga perlu dipertimbangkan dalam penggunaan Naive Bayes, menunjukkan pentingnya mengevaluasi kemampuan generalisasi kedua model terhadap data yang belum dikenal.

REFERENSI

[1] E. Nofianto, Fitriyah, and Supratiwi, “Media Sosial sebagai Sarana Pendidikan Politik oleh Pejabat Publik (Studi pada Akun Media Sosial Nur Hidayat Sardini) Eri,” *Jurnal Ilmiah*, vol. 10, no. 2, pp. 1–94, 2019, doi: 10.33087/jiubj.v23i1.3060.

[2] S. N. J. Fitriyyah, N. Safriadi, and E. E. Pratama, “Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes,” *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 5, no. 3, p. 279, 2019, doi: 10.26418/jp.v5i3.34368.

[3] Alfandi Safira and F. N. Hasan, “Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier,” *ZONAsi: Jurnal Sistem Informasi*, vol. 5, no. 1, pp. 59–70, 2023, doi: 10.31849/zn.v5i1.12856.

[4] I. R. Afandi, I. F. Hanif, F. N. Hasan, E. Sinduningrum, Z. Halim, and N. Pratiwi, “Analisis Sentimen Opini Masyarakat Terkait Penyelenggaraan Sistem Elektronik Menggunakan Metode Logistic Regression,” *Jurnal Linguistik Komputasional*, vol. 5, no. 2, pp. 77–84, 2022, doi: https://doi.org/10.26418/jlk.v5i2.103.

[5] M. Siddik, Hendri, R. N. Putri, and Y. Desnelita, “Klasifikasi Kepuasan Mahasiswa Terhadap Pelayanan Perguruan Tinggi Menggunakan Algoritma Naïve Bayes Classification,” vol. 3, pp. 1–23, 2020, doi: 10.31539/intecoms.v3i2.1654.

[6] R. Sari and R. Y. Hayuningtyas, “Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pada Wisata TMII Berbasis Website,” *Indonesian Journal on Software Engineering (IJSE)*, vol. 5, no. 2, pp. 51–60, 2019, doi: 10.31294/ijse.v5i2.6957.

[7] E. Indrayuni, “Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes,” *Jurnal Khatulistiwa Informatika*, vol. 7, no. 1, pp. 29–36, 2019, doi: 10.31294/jki.v7i1.5740.

[8] H. Tuhuteru, “Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine,” *Information System Development (ISD)*, vol. 5, no. 2, pp. 7–13, 2020.

[9] R. Sari and R. Y. Hayuningtyas, “Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pada Wisata TMII Berbasis Website,” vol. 5, no. 2, pp. 51–60, 2019.

[10] A. Retno, T. Hayati, P. Y. Saputra, and A. M. Sastri, “Sistem Koreksi Kesalahan Pengetikan Kata Kunci dalam Pencarian Artikel Menggunakan Algoritma Jaro-Winkler,” *2019*, pp. 60–65.

[11] Oto.com, "Kendaraan Listrik Kendaraan Electrik - Masa depan mobilitas," oto.com. [Online]. Available: https://www.oto.com/kendaraan-listrik

[12] N. Hendrastuty *et al.*, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 6, no. 3, pp. 150–155, 2021.

[13] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *Journal of Intelligent System and Computation*, vol. 1, no. 1, pp. 43–49, 2019, doi: 10.52985/insyst.v1i1.36.

[14] M. A. Rofiqi, Abd. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 1, no. 2, pp. 58–64, 2019, doi: 10.28926/ilkomnika.v1i2.18.

[15] R. Y. Hayuningtyas, "Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita," vol. 6, no. 1, pp. 18–22, 2019.

[16] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.

[17] S. Ratna, "Pengolahan Citra Digital Dan Histogram Dengan Phyton Dan Text Editor Phycharm," *Technologia: Jurnal Ilmiah*, vol. 11, no. 3, p. 181, 2020, doi: 10.31602/tji.v11i3.3294.

[18] L. Hermawan and M. Bellaniar Ismiati, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *Jurnal Transformatika*, vol. 17, no. 2, p. 188, 2020, doi: 10.26623/transformatika.v17i2.1705.

[19] Ainurrohmah, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 4, pp. 493–499, 2021.