Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat - Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Perbandingan Algoritma NBC dan SVM dalam Analisis Sentimen Terhadap Dampak Kesehatan Rokok Elektrik

DOI: http://dx.doi.org/10.35889/progresif.v21i2.2903

Hani Rahmawati[1], Isa Faqihuddin Hanif[2*]
[1]Teknik Informatika, Universitas Muhammadiyah Prof. DR. HAMKA, Jakarta, Indonesia
[2]Sistem dan Teknologi Informasi, Universitas Muhammadiyah Prof. DR. HAMKA, Jakarta, Indonesia
[*]e-mail Corresponding Author: isa@uhamka.ac.id

Abstrak

There is no optimal method for accurately classifying public opinion, so an analytical approach is needed that is able to capture the nuances of public sentiment regarding the health impacts of e-cigarettes. This study examines public perception of the health impacts of electronic cigarettes using two classification algorithms: NBC and SVM. Data sourced from social media X (formerly Twitter) underwent stages of data cleaning, sentiment labeling, TF-IDF weighting, and data balancing through the SMOTE technique. Performance evaluation was conducted using four key metrics: accuracy, precision, recall, and f1-score. NBC achieved 80.5% accuracy with high recall despite low precision. In contrast, SVM recorded superior performance with 95.2% accuracy and more consistent balance between precision and recall. Therefore, the Support Vector Machine (SVM) algorithm is recommended as a more effective method for analyzing public sentiment regarding electronic cigarettes.

Keywords: *Electronic Cigarette; Entiment Analysis; Naïve Bayes Classifier; Support Vector Machine.*

Abstrak

Belum adanya metode yang optimal untuk mengklasifikasikan opini publik secara akurat, sehingga diperlukan pendekatan analitik yang mampu menangkap nuansa sentimen masyarakat terhadap dampak kesehatan rokok elektrik. Studi ini mengkaji persepsi publik terhadap dampak kesehatan rokok elektrik dengan menerapkan dua algoritma klasifikasi: NBC dan SVM. Data yang bersumber dari media sosial X (eks Twitter) diproses melalui tahapan pembersihan data, pelabelan sentimen, pembobotan menggunakan TF-IDF, serta penyeimbangan data menggunakan teknik SMOTE. Evaluasi performa dilakukan menggunakan empat metrik utama: *accuracy, precision, recall*, dan *f1-score*. NBC memperoleh akurasi sebesar 80,5% dengan recall tinggi meskipun *precision*-nya rendah. Sebaliknya, SVM mencatat performa superior dengan akurasi 95,2% serta keseimbangan *precision* dan *recall* yang lebih konsisten. Oleh karena itu, algoritma *Support Vector Machine* (SVM) direkomendasikan sebagai metode yang lebih efektif dalam menganalisis sentimen publik terhadap rokok elektrik.

Kata kunci: *Analisis Sentimen; Rokok Elektrik; Naïve Bayes Classifier; Support Vector Machine.*

1. Pendahuluan

Perkembangan teknologi digital dan pola hidup modern mendorong munculnya berbagai inovasi konsumsi tembakau, salah satunya adalah rokok elektrik atau vape. Produk ini kerap dipandang sebagai alternatif yang dianggap lebih aman dibanding rokok konvensional karena tidak melalui proses pembakaran. Pandangan yang berkembang di masyarakat saat ini masih belum mendapat dukungan kuat dari penelitian ilmiah yang komprehensif, terutama mengenai

dampak jangka panjang rokok elektrik terhadap kesehatan. Oleh karena itu, studi lanjutan yang menyoroti respons publik terhadap produk ini sangat diperlukan dari perspektif kesehatan.

Dalam kurun waktu satu dekade, jumlah pengguna rokok elektrik di Indonesia menunjukkan peningkatan yang mencolok. Laporan Global Adult Tobacco Survey (GATS) tahun 2021 mencatat bahwa prevalensi penggunaan naik dari 0,3% pada 2011 menjadi 3% pada 2021 [1]. Bersamaan dengan tren tersebut, media sosial menjadi ruang terbuka bagi masyarakat untuk menyuarakan pandangannya, baik yang mendukung maupun yang menentang. Namun, belum banyak kajian yang secara sistematis mengukur sentimen masyarakat terhadap isu ini. Masalah yang muncul adalah belum adanya metode yang optimal untuk mengklasifikasikan opini publik secara akurat, sehingga diperlukan pendekatan analitik yang mampu menangkap nuansa sentimen masyarakat terhadap dampak kesehatan rokok elektrik.

Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan solusi berupa analisis sentimen berbasis algoritma klasifikasi teks [2]. Dua algoritma yang banyak digunakan dalam proses klasifikasi adalah *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM), keduanya memiliki popularitas tinggi di kalangan peneliti. NBC efektif dalam mengolah teks berkompleksitas rendah dengan jumlah data besar [3], sedangkan SVM dikenal mampu bekerja optimal pada data berdimensi tinggi dengan hasil akurasi yang unggul [4]. Dengan membandingkan kinerja kedua algoritma ini, diharapkan diperoleh pemahaman yang lebih baik mengenai efektivitas masing-masing model dalam mengklasifikasikan sentimen terkait isu kesehatan rokok elektrik. Pendekatan ini didukung oleh berbagai studi terdahulu [3][4], namun masih jarang yang mengintegrasikan keduanya dalam satu kajian spesifik tentang kesehatan.

Penelitian ini secara khusus ditujukan untuk mengevaluasi serta membandingkan performa NBC dan SVM dalam mengklasifikasikan sentimen masyarakat terhadap rokok elektrik, dengan fokus pada aspek kesehatan. Hasil penelitian ini diharapkan dapat memperkaya pengembangan sistem analisis opini publik berbasis AI dan memberikan perspektif yang bermanfaat bagi para pengambil kebijakan maupun tenaga kesehatan dalam memahami pandangan masyarakat.

2. Tinjauan Pustaka

Pada tahun 2022, Dio Rizki Aditya bersama Endang Supriyati dan Tri Listyorini melakukan penelitian dengan mengangkat judul Analisis Sentimen Pengguna Twitter Terhadap Rokok Elektrik (Vape) di Indonesia, penelitian ini menelaah tanggapan masyarakat Indonesia di media sosial terhadap penggunaan rokok elektrik di mana mereka menggunakan metode *Naïve Bayes Classifier* (NBC) sebagai teknik klasifikasinya. Tujuan penelitian ini adalah mengetahui persepsi masyarakat terhadap rokok elektrik melalui data Twitter. Sentimen dibagi menjadi tiga kategori: positif, negatif, dan netral, dengan hasil akurasi mencapai 77,5% dan distribusi sentimen netral mendominasi (77,3%), disusul negatif (11,7%) dan positif (11%). Hasil ini menegaskan potensi media sosial sebagai sumber informasi dalam mengkaji isu kesehatan.

Alman Muhammadin dan Irwan Agus Sobari (2021), melalui penelitian berjudul *Analisis Sentimen pada Ulasan Aplikasi Kredivo dengan Algoritma SVM dan NBC*, membandingkan efektivitas algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier* (NBC) dalam mengklasifikasikan sentimen pengguna terhadap aplikasi Kredivo. Hasil studi menunjukkan bahwa SVM menghasilkan hasil yang lebih optimal dengan akurasi sebesar 83,3%, sementara NBC mencatatkan akurasi 80,8%, yang mengindikasikan bahwa pemilihan algoritma berperan penting dalam menentukan kualitas hasil klasifikasi.

Penelitian oleh Nia Ramadhani Siregar, Prilly Rismawany, Shafiah Azzahra, dan Yuliana Sari (2024) dalam artikel berjudul *Kajian Bahan Kimia Berbahaya pada Rokok Elektrik serta Dampaknya pada Kesehatan* mengulas secara mendalam kandungan zat berbahaya dalam rokok elektrik dan implikasinya terhadap kesehatan pengguna. Melalui pendekatan deskriptif kuantitatif dan studi literatur, ditemukan bahwa rokok elektrik mengandung bahan kimia berbahaya seperti nikotin, formaldehida, acrolein, logam berat (arsenik, kadmium, timbal), serta senyawa karsinogenik lainnya. Dampaknya mencakup peningkatan detak jantung, iritasi saluran pernapasan, kerusakan DNA, penurunan fungsi paru, hingga risiko kanker. Hasil observasi juga menunjukkan bahwa mayoritas pengguna rokok elektrik adalah remaja laki-laki berusia 18–24 tahun yang memiliki kesadaran akan risiko kesehatan, namun tetap memerlukan edukasi dan regulasi yang lebih ketat. Penelitian ini memperkuat pentingnya informasi transparan terkait kandungan dan dampak rokok elektrik sebagai upaya perlindungan kesehatan masyarakat.

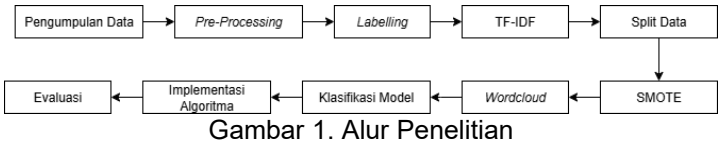
Penelitian oleh Agus Susanto, Muladi Putra Mahardika, dan Heni Purwantiningrum (2023) berjudul *Pemberdayaan Kesehatan Remaja: Edukasi Bahaya Rokok Elektrik bagi Siswa SMA Negeri 2 Tegal* menyoroti pentingnya edukasi tentang bahaya rokok elektrik di kalangan remaja. Melalui kegiatan penyuluhan interaktif kepada 100 siswa kelas 10 dan 11, penelitian ini menunjukkan peningkatan pengetahuan peserta mengenai dampak negatif rokok elektrik terhadap kesehatan. Nilai rata-rata yang diperoleh dari hasil post-test adalah sebesar 13,54, meningkat dari nilai pre-test sebesar 8,86. Materi edukasi meliputi risiko kesehatan yang ditimbulkan oleh rokok elektrik terhadap organ vital seperti sistem pernapasan, otak, jantung, dan paru-paru, serta kandungan kimia berbahaya dalam cairan vape. Penelitian ini menegaskan bahwa edukasi langsung di lingkungan sekolah mampu meningkatkan kesadaran siswa terhadap risiko kesehatan dari penggunaan rokok elektrik.

Studi yang dilakukan oleh Kurnia Ardiansyah Lubis dan tim pada tahun 2024 analisis terhadap persepsi publik mengenai pemindahan ibu kota Indonesia dilakukan melalui penerapan algoritma Naïve Bayes, dengan memanfaatkan data yang telah dikumpulkan sebelumnya dari Twitter. Hasilnya, 74% opini publik tergolong positif, dan akurasi model mencapai 76,30%, memperlihatkan bahwa Naïve Bayes masih menjadi metode yang relevan dalam menganalisis sentimen terhadap isu nasional.

Dari berbagai penelitian tersebut, terlihat bahwa algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) merupakan dua metode yang banyak dimanfaatkan dalam analisis sentimen berbasis media sosial, baik dalam isu layanan, produk digital, hingga kebijakan publik. Di sisi lain, penelitian-penelitian yang berfokus pada rokok elektrik sebagian besar menyoroti aspek kesehatan, seperti kandungan bahan kimia berbahaya dan dampaknya terhadap sistem pernapasan, jantung, serta potensi kanker. Namun, masih sedikit penelitian yang secara khusus mengintegrasikan analisis sentimen dengan topik kesehatan rokok elektrik, apalagi dengan membandingkan performa NBC dan SVM dalam konteks tersebut. Sebagai hasilnya, penelitian ini menyumbangkan perspektif baru melalui menggabungkan dua aspek penting: perbandingan algoritma klasifikasi dalam analisis sentimen, serta isu kesehatan publik terkait rokok elektrik.

3. Metodologi

Dari media sosial X, diikuti oleh tahapan *pre-processing* yang mencakup pembersihan teks, tokenisasi, konversi huruf ke format seragam, penghilangan stopwords, serta stemming untuk menyederhanakan kata. Selanjutnya dilakukan *labeling* sentimen (positif atau negatif), kemudian data diolah menggunakan teknik TF-IDF sebagai pembobotan kata untuk menghasilkan representasi numerik. Proses ini diikuti oleh klasifikasi model dengan NBC dan SVM. Selain klasifikasi, data divisualisasikan dalam bentuk *Wordcloud*. Pada tahap akhir, algoritma diujicobakan dan kinerjanya diukur menggunakan parameter evaluasi berupa akurasi, presisi, *recall*, dan *F1-score*. Gambar 1 menyajikan representasi visual dari alur penelitian.



Gambar 1. Alur Penelitian

3.1 Pengumpulan Data

Pengumpulan data dilakukan untuk menghimpun informasi yang relevan, yang berperan penting dalam menunjang jalannya penelitian. Data dihimpun dari media sosial X melalui metode crawling, dengan periode pengambilan dimulai sejak Oktober 2024 dan masih berlangsung. Total data yang berhasil dikumpulkan berkisar antara 2.000 hingga 2.500 entri, yang mencerminkan opini dan sentimen publik terhadap rokok elektrik.

3.2 *Pre-Processing*

Tahap *pre-processing* berfungsi sebagai tahap pertama dalam upaya mempersiapkan data teks dengan membersihkannya dari komponen-komponen yang tidak memiliki kontribusi signifikan seperti simbol, URL, dan kata-kata yang tidak bermakna. Prosedur ini melibatkan cleansing, tokenisasi, normalisasi huruf, penghapusan stopword, penyaringan token berdasarkan panjang, dan stemming agar data lebih siap untuk dianalisis secara sistematis [5].

a. *Cleansing*
   Pembersihan dilakukan dengan mengeliminasi karakter-karakter yang tidak diperlukan seperti simbol, angka, URL, serta tag HTML yang dapat mengganggu integritas analisis data.
b. *Tokenize*
   Memisahkan teks menjadi elemen-elemen kecil, seperti kata maupun frasa, agar lebih mudah diproses dan dianalisis pada tingkat granular.
c. *Transform Cases*
   Mengubah seluruh tulisan ke format huruf kecil guna memastikan keseragaman penulisan serta meminimalkan potensi ketidaksesuaian akibat penggunaan huruf besar.
d. *Filter Stopwords*
   Menghapus unsur kata yang muncul berulang tetapi tidak relevan untuk penilaian makna misalnya ‘dan,’ ‘atau,’ dan ‘yang,’ untuk menitikberatkan analisis pada kata-kata yang lebih informatif.
e. *Filter Token by Length*
   Memfilter kata-kata yang terlalu pendek (misalnya satu huruf) atau terlalu panjang, sehingga hanya kata-kata yang relevan yang dipertahankan. Serangkaian langkah ini dilakukan untuk menjamin kebersihan data dan memiliki kualitas tinggi untuk mendukung tahap selanjutnya.
f. *Stemming*
   Mengubah kata ke bentuk dasarnya (*root word*), misalnya "berlari" menjadi "lari," guna mengurangi beragam kata dengan arti yang serupa.

3.3. *Labeling*

Labeling merupakan tahap penandaan label sentimen contohnya berupa label positif atau negatif pada data teks, yang dilakukan menggunakan metode berbasis leksikon. Proses ini memanfaatkan kamus kata-kata dengan skor sentimen tertentu untuk menilai emosi dalam teks. Data yang telah dilabeli menjadi terstruktur dan siap digunakan untuk pelatihan model, di mana keakuratan *labeling* memainkan peran penting dalam menjamin ketepatan hasil analisis sentiment.

3.4. TF-IDF

TF-IDF digunakan sebagai teknik penilaian untuk menentukan seberapa penting suatu kata dalam dokumen tertentu dengan mempertimbangkan kemunculannya di seluruh dokumen di seluruh dokumen dalam korpus [6]. TF-IDF membantu menyoroti kata-kata kunci dan sering digunakan sebagai input ke algoritma seperti NBC dan SVM dalam analisis sentimen [7].

3.5. Split Data

Split Data merupakan langkah guna membagi datase ke dalam data yang digunakan untuk proses pelatihan dan evaluasi untuk melatih dan menguji performa model secara adil [8]. Rasio umum pembagiannya adalah 70:30 atau 80:20 supaya model dapat divalidasi dengan data yang benar-benar independen dari data pelatihan [9].

3.6. SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah teknik yang digunakan untuk menyeimbangkan jumlah data antar kelas, khususnya saat data minoritas jauh lebih sedikit dari data mayoritas dengan menciptakan data tiruan pada kategori minoritas [10]. Teknik ini berperan penting dalam menjaga objektivitas model, terutama saat menangani klasifikasi sentimen yang tidak seimbang antara kelas positif dan negatif [11].

3.7. *Wordcloud*

Wordcloud adalah visualisasi kata yang mengacu pada seberapa sering kata muncul dalam teks yang menampilkan kata-kata dengan ukuran sesuai frekuensinya [12]. Ini memudahkan eksplorasi awal data untuk mengenali topik dominan dalam analisis sentiment [13].

3.8. Implementasi Algoritma

Dalam studi ini, diterapkan dua pendekatan klasifikasi algoritmik, yakni *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). NBC dipilih karena kesederhanaannya serta kemampuannya yang baik dalam mengolah data teks berukuran besar meskipun dengan

data pelatihan terbatas, meskipun memiliki kelemahan pada asumsi independensi antar fitur. Sementara itu, SVM mampu bekerja dengan baik pada data yang memiliki fitur dalam jumlah besar dan sifat non-linear melalui penggunaan hyperplane dan fungsi kernel, sehingga mampu menghasilkan akurasi tinggi dalam klasifikasi sentimen.

a. *Naïve Bayes Classifier* (NBC)

NBC merupakan metode pengklasifikasian berbasis probabilistik yang sederhana dan cepat, bekerja efektif pada data teks berukuran besar serta tetap andal meskipun data pelatihan terbatas [14]. Asumsi independensi antar fitur menjadi titik lemah dari pendekatan ini, karena sering kali tidak mencerminkan kondisi data yang sebenarnya [15].

$$P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)} \quad (1)$$

b. *Support Vector Machine* (SVM)

SVM dikenal sebagai algoritma yang tangguh dalam tugas klasifikasi. Ia bekerja dengan membangun hyperplane sebagai pemisah antar kelas, dan dengan fungsi kernel, SVM dapat mengatasi data dengan pola kompleks yang tidak dapat dipisahkan secara linier. [16]. SVM sangat optimal untuk data dalam ruang fitur berdimensi besar dan menghasilkan akurasi tinggi dalam analisis sentiment [17].

$$(x) = sgn\ (w \cdot x + b) \quad (2)$$

3.9. *Confusion Matrix*

Sebagai metode evaluasi klasifikasi, *Confusion Matrix* mengukur kinerja model dengan menghitung jumlah prediksi yang sesuai atau tidak sesuai dengan label aktual [18]. Empat komponen utamanya, TP, TN, FP, dan FN digunakan untuk mengurangi efektivitas model berdasarkan indikator performa seperti akurasi, presisi, recall, dan F1-score [19].

a. *Accuracy*

Accuracy mengacu pada persentase jumlah prediksi yang benar dibandingkan dengan seluruh prediksi yang dilakukan selama proses klasifikasi. Metode ini sangat cocok digunakan pada dataset yang seimbang. Rumusnya:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

b. *Recall*

Recall menghitung kemampuan model dalam menemukan seluruh data yang positif. Metode ini berfokus pada pengurangan kesalahan negatif palsu (false negatives). Rumusnya:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

c. *Precision*

Precision mengukur tingkat keandalan dari prediksi positif. Metode ini penting untuk meminimalkan jumlah kesalahan positif palsu (false positives). Rumusnya:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

d. *F1-Score*

F1-Score menghitung keseimbangan antara precision dan recall dalam satu metrik. Metode ini ideal untuk digunakan pada dataset yang tidak seimbang. Rumusnya:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (6)$$

4. Hasil dan Pembahasan

Pre-processing Pada bab ini disampaikan hasil yang didapat melalui setiap tahapan yang telah dilakukan, mulai dari pengumpulan data, proses pra-pemrosesan, penerapan metode klasifikasi dengan metode NBC dan SVM, hingga tahap evaluasi performa model.

4.1. Pengumpulan Data

Data dikumpulkan secara otomatis dari media sosial X (Twitter) dengan memanfaatkan kata kunci spesifik sebagai dasar pencarian, yaitu “rokok elektrik”, yang dilakukan melalui library *tweet- harvest* di platform Google Colab tanpa memerlukan akses API. Proses ini memudahkan peneliti dalam memperoleh data secara efisien dan cepat. Gambar 2 menunjukkan alur dari proses pengumpulan data.

Gambar 2. Proses Pengumpulan Data

Dari 2.500 tweet yang diminta, berhasil dikumpulkan sebanyak 2.176 tweet. Tweet yang terkumpul masih mengandung elemen-elemen seperti mention (@), hashtag (#), tautan, dan simbol lainnya yang belum dibersihkan. Data mentah tersebut kemudian digunakan sebagai bahan utama untuk tahapan pra- pemrosesan dan analisis sentimen selanjutnya. Tabel 1 menyajikan hasil dari proses pengumpulan data.

Table 1. Hasil Pengumpulan Data

Teks
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c

4.2. *Pre-processing*

Pre-processing adalah tahap awal penting dalam text processing guna membersihkan serta menyiapkan data mentah agar layak digunakan dalam analisis. Proses ini sangat krusial, terutama untuk data media sosial seperti X, karena sering mengandung simbol, tautan, atau kata tidak baku yang dapat mengganggu akurasi analisis.

a. *Cleansing*

Hasil *cleansing* menunjukkan bahwa elemen seperti mention (@), tagar (#), dan tautan telah berhasil dihapus, menjadikan teks lebih bersih dan fokus pada opini pengguna, seperti pandangan terhadap bahaya rokok elektrik dan memberishkan atau menghapus kalimat yang tidak relevan. Proses ini memperbaiki ketepatan analisis sentimen melalui penghapusan komponen yang tidak relevan. Hasil *Cleansing* dapat dilihat pada Tabel 2.

Table 2. Hasil *Cleansing*

Sebelum *Cleansing*	Sesudah *Cleansing*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur Hanya krn ga bau bukan berati ga bahaya
Sebuah penelitian yang diterbitkan dalam HarmReduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c	Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa

b. *Tokenize*

Hasil *tokenize* menunjukkan bahwa kalimat dipecah menjadi kata-kata terpisah, seperti “Rokok elektrik ini lebih bahaya…” menjadi \['Rokok', 'elektrik', 'ini', 'lebih', 'bahaya'], sehingga setiap kata dapat dianalisis secara individual untuk mendukung akurasi dalam analisis sentimen. Hasil *Tokenize* dapat dilihat pada Tabel 3.

Table 3. Hasil *Tokenize*

Sebelum *Tokenize*	Sesudah *Tokenize*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	'Rokok', 'elektrik', 'ini', 'lebih', 'bahaya', 'karena', 'ada', 'wangi', 'yang', 'bikin', 'awareness', 'orang', 'berkurang', 'beda', 'pas', 'kena', 'asap', 'rokok', 'pasti', 'langsung', 'dikibas', 'Temen', 'ku', 'ada', 'yg', 'asma', 'nya', 'lumayan', 'parah', 'kalo', 'ada', 'yg', 'ngvape', 'deket', 'dia', 'pasti', 'itu', 'org', 'itu', 'ku', 'tegur', 'Hanya', 'krn', 'ga', 'bau', 'bukan', 'berati', 'ga', 'bahaya'
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape#Snus#HTP #Kantongnikotin https://t.co/z9B6oWXE9c	'Sebuah', 'penelitian', 'yang', 'diterbitkan', 'dalam', 'Harm', 'Reduction', 'Journal', 'menyebutkan', 'snus', 'dapat', 'menjadi', 'solusi', 'pengurangan', 'dampak', 'merokok', 'untuk', 'kesehatan', 'di', 'Eropa’

c. *Transform Casses*

Hasil *transform cases* menunjukkan bahwa kata-kata dengan huruf kapital seperti "Rokok" dan "Elektrik" telah diubah menjadi huruf kecil agar konsisten, seperti "rokok" dan "elektrik". Transformasi ini mempermudah analisis dan pemetaan kata dalam perhitungan frekuensi atau pembobotan pada analisis sentimen. Tabel 4 menyajikan hasil dari proses *Transform Cases*.

Table 4. Hasil *Transform Casses*

Sebelum *Transform Cases*	Sesudah *Transform Cases*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur hanya krn ga bau bukan berati ga bahaya
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c	sebuah penelitian yang diterbitkan dalam harm reduction journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di eropa

d. Filter *Stopwords*

Hasil filter *stopwords* menunjukkan bahwa kata-kata umum seperti kata hubung dan bantu berhasil dihapus, menyisakan kata-kata penting seperti "rokok", "elektrik", "bahaya", dan "jurnal". Langkah ini bertujuan untuk meningkatkan fokus analisis terhadap inti opini atau informasi dalam tweet. Tabel 5 memperlihatkan data setelah proses filter *stopwords* dilakukan.

Table 5. Hasil Filter *Stopwords*

Sebelum Filter *Stopwords*	Sesudah Filter *Stopwords*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	rokok elektrik bahaya wangi bikin awareness orang berkurang beda pas kena asap rokok langsung dikibas temen ku yg asma nya lumayan parah kalo yg ngvape deket org ku tegur krn ga bau berati ga bahaya
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c	penelitian diterbitkan harm reduction journal snus solusi pengurangan dampak merokok kesehatan eropa

e. Filter *Token By Length*

Hasil filter *token by length* menunjukkan bahwa kata-kata pendek seperti “yg”, “km”, “ga”, “di”, dan “itu” telah dihapus dari teks. Tujuannya adalah agar hanya kata-kata yang lebih bermakna seperti “rokok”, “elektrik”, “bahaya”, dan “penelitian” yang dipertahankan, sehingga analisis sentimen menjadi lebih fokus dan relevan terhadap topik yang dikaji. Hasil Filter *Token by Length* dapat dilihat pada Tabel 6.

Table 6. Hasil Filter *token By Length*

Sebelum Filter *Token by Length*	Sesudah Filter *Token by Length*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	rokok elektrik bahaya wangi bikin awareness orang berkurang beda kena asap rokok langsung dikibas temen asma lumayan parah kalo ngvape deket tegur berati bahaya
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c	penelitian diterbitkan harm reduction journal snus solusi pengurangan dampak merokok kesehatan eropa

f. Stemming

Hasil *stemming* menunjukkan bahwa kata berimbuhan telah berhasil dikembalikan ke bentuk dasarnya, seperti “merasakan” menjadi “rasa” dan “terbitkan” menjadi “terbit”. Proses ini mencegah sistem menganggap kata serupa sebagai entitas berbeda, sehingga meningkatkan konsistensi dan akurasi dalam klasifikasi sentimen. Tabel 7 menyajikan hasil dari proses *stemming*.

Table 7. Hasil *Stemming*

Sebelum *Stemming*	Sesudah *Stemming*
@windaul Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas. Temen ku ada yg asma nya lumayan parah kalo ada yg ngvapendeket dia pasti itu org itu ku tegur. Hanya krn ga bau bukan berati ga bahaya	rokok elektrik bahaya wangi bikin awareness orang kurang beda kena asap rokok langsung kibas temen asma lumayan parah kalo ngvape deket tegur berat bahaya
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa. @Kemenperin_RI #needforalternative #InovasiTembakau #Vape #Snus #HTP #Kantongnikotin https://t.co/z9B6oWXE9c	teliti terbit harm reduction journal snus solusi kurang dampak rokok sehat eropa

Dari 2.176 data tweet hasil crawling, dilakukan *preprocessing* berupa *cleansing*, *tokenizing*, *Transform Casses*, Filter *Stopwords*, Filter *Token By Length* dan *stemming*. Hasilnya, diperoleh 1.464 data yang layak untuk dianalisis lebih lanjut.

4.3. *Labeling*

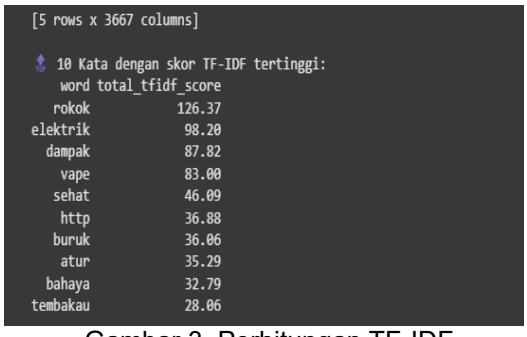
Hasil *labeling* menunjukkan bahwa tweet bernuansa keluhan atau risiko kesehatan diberi label “Negatif”, sedangkan yang mendukung atau solutif diberi label “Positif”. Dari total data, 1.285 tweet berlabel negatif dan 179 positif, mengindikasikan dominasi persepsi negatif terhadap rokok elektrik. Hasil *Labeling* dapat dilihat pada Tabel 8.

Table 8. Hasil *Labeling*

Teks	Sentimen
Rokok elektrik ini lebih bahaya karena ada wangi yang bikin awareness orang berkurang beda pas kena asap rokok pasti langsung dikibas Temen ku ada yg asma nya lumayan parah kalo ada yg ngvape deket dia pasti itu org itu ku tegur Hanya krn ga bau bukan berati ga bahaya	Negatif
Sebuah penelitian yang diterbitkan dalam Harm Reduction Journal menyebutkan snus dapat menjadi solusi pengurangan dampak merokok untuk kesehatan di Eropa	Positif

4.4. TF-IDF

Metode TF-IDF bekerja dengan memberikan bobot pada setiap kata berdasarkan frekuensi relatifnya dalam satu dokumen serta kelangkaannya di seluruh dokumen lainnya, guna menyoroti kata yang paling bermakna. Proses konversi teks ke bentuk numerik dilakukan dengan bantuan *TfidfVectorizer*() dari scikit-learn. Nilai TF-IDF dihitung dan diurutkan untuk mengidentifikasi kata-kata paling signifikan, seperti “rokok”, “elektrik”, “dampak”, dan “vape”. Perhitungan TF-IDF dapat dilihat pada Gambar 3.



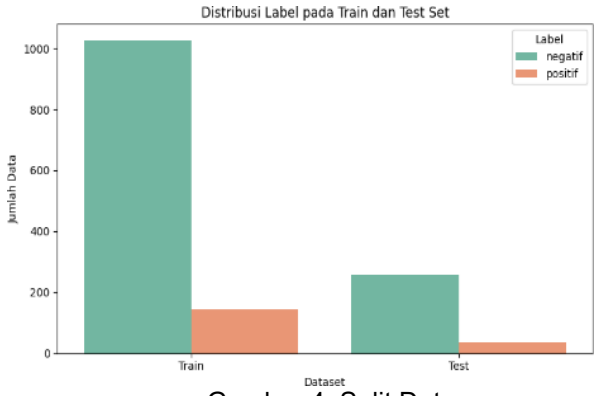
```
[5 rows x 3667 columns]

🔝 10 Kata dengan skor TF-IDF tertinggi:
      word total_tfidf_score
     rokok            126.37
elektrik               98.20
   dampak              87.82
     vape              83.00
    sehat              46.09
     http              36.88
    buruk              36.06
     atur              35.29
   bahaya              32.79
tembakau               28.06
```

Gambar 3. Perhitungan TF-IDF

4.5. Split Data

Split data berperan penting dalam proses pembelajaran mesin dengan memisahkan dataset menjadi dua bagian utama: data latih untuk membentuk model dan data uji untuk mengevaluasi performanya pada data baru. Penelitian ini menggunakan *train_test_split* dari *Scikit-learn* dengan rasio 80% untuk pelatihan dan 20% untuk pengujian, serta menggunakan parameter *stratify* demi menjaga keseimbangan distribusi label. Hasilnya, diperoleh 1.171 data pelatihan dan 293 data pengujian, dengan distribusi label yang tidak seimbang: mayoritas data berlabel negatif. Gambar 4 memperlihatkan tahapan split data.

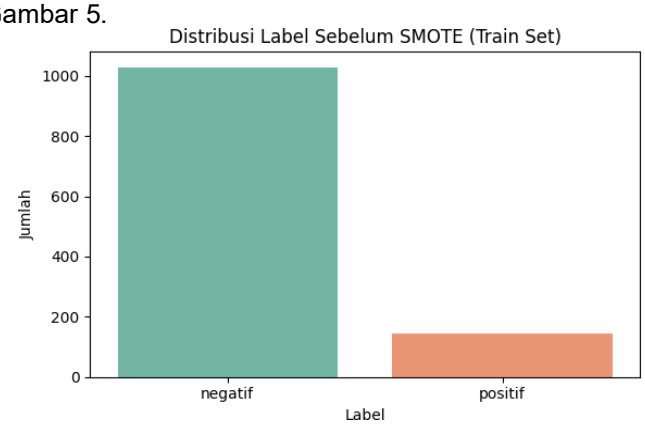


Gambar 4. Split Data

Data pelatihan terdiri dari 1.028 label negatif dan 143 label positif, sedangkan data pengujian terdiri dari 257 label negatif dan 36 label positif. Ini menunjukkan distribusi label yang tidak seimbang.

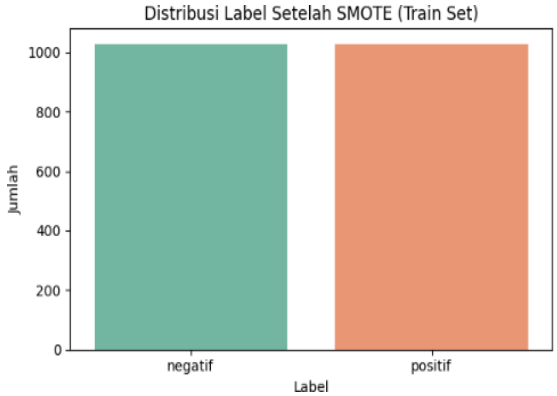
4.6 SMOTE

SMOTE ialah pendekatan *oversampling* guna meratakan distribusi data dengan menciptakan data sintetis pada kelompok minoritas. Dalam kasus ini, data awal menunjukkan ketimpangan antara 1.028 data negatif dan 143 data positif. Distribusi Label Sebelum SMOTE dapat dilihat pada Gambar 5.



Gambar 5. Distribusi Label Sebelum SMOTE

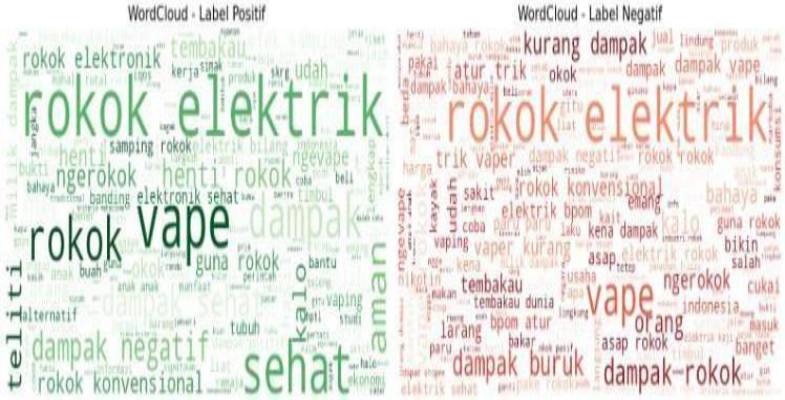
Setelah diterapkan SMOTE, jumlah data positif ditingkatkan menjadi seimbang, yaitu1.028 data, sehingga proses pelatihan model menjadi lebih optimal dan akurasi terhadap kelas minoritas meningkat. Teknik ini diimplementasikan menggunakan library *imbalanced-learn*. Distribusi Label Setelah SMOTE dapat dilihat pada Gambar 6.



Gambar 6. *Distribusi Label Setelah SMOTE*

4.7. *Wordcloud*

Wordcloud berfungsi sebagai visualisasi teks yang menunjukkan seberapa sering kata muncul berdasarkan perbedaan ukuran huruf. Pada label positif, Wordcloud menampilkan kata-kata yang sering muncul, seperti solusi, sehat, serta kurang dampak, yang mencerminkan pandangan positif terhadap rokok elektrik. Sebaliknya, pada label negatif, kata- kata seperti bahaya, buruk, dampak, dan larang mendominasi, mengindikasikan adanya kekhawatiran atau sentimen negatif dari pengguna terhadap penggunaan rokok elektrik. *Wordcloud* ditampilkan pada Gambar 7.



Gambar 7. *Wordcloud* Positif Negatif

4.8. Implementasi Algoritma

Tweet yang telah melalui tahapan preprocessing dan pelabelan kemudian diklasifikasikan menggunakan dua algoritma, yakni *Naïve Bayes Classifier* dan *Support Vector Machine*. Kedua algoritma Pelatihan dilakukan menggunakan data training yang telah diolah dengan teknik oversampling SMOTE, lalu diuji dengan data pengujian yang telah dipisahkan sebelumnya.

a. *Naïve Bayes Classifier* (NBC)

NBC ialah algoritma klasifikasi bersifat statistik yang mengandalkan Teorema Bayes digunakan dengan anggapan bahwa fitur-fitur tidak saling bergantung. Dalam implementasinya, digunakan kelas *MultinomialNB*() dari *library scikit-learn*, yang efektif untuk klasifikasi data teks. Model dilatih menggunakan data hasil *oversampling X_train_smote* dan *y_train_smote*, lalu melakukan prediksi terhadap data uji X_test dan menyimpan hasilnya dalam variabel nb_pred. Gambar 8 memperlihatkan proses implementasi dari algoritma NBC.

```
# === 2 Training Naive Bayes ===
nb = MultinomialNB()
nb.fit(X_train_smote, y_train_smote)
nb_pred = nb.predict(X_test)
```

Gambar 8. Implementasi Algoritma *Naïve Bayes Classifier* (NBC)

b. *Support Vector Machine* (SVM)

Algoritma SVM menyelesaikan tugas klasifikasi dengan mengidentifikasi hyperplane yang dapat memaksimalkan margin antar kelas, guna meningkatkan efisiensi pemisahan data. SVM efektif dalam mengolah data teks berdimensi besar, terutama dengan bantuan kernel seperti 'linear'. Dalam proses pelatihan, digunakan SVC dengan kernel linear pada data yang telah diolah menggunakan SMOTE (*X_train_smote dan y_train_smote*), lalu model digunakan untuk memprediksi data uji (X_test) dan hasilnya disimpan dalam variabel svm_pred. Implementasi Algoritma *Support Vector Machine* (SVM) dapat dilihat pada Gambar 9.

```
# === 3 Training SVM ===
svm = SVC(kernel='linear')
svm.fit(X_train_smote, y_train_smote)
svm_pred = svm.predict(X_test)
```

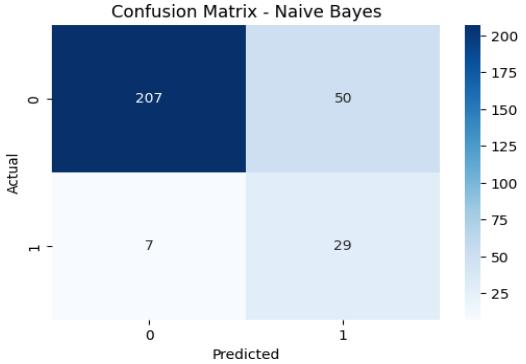
Gambar 9. Implementasi Algoritma *Support Vector Machine* (SVM)

4.9. *Confusion Matrix*

Confusion matrix adalah bentuk representasi dalam tabel yang digunakan untuk mengevaluasi performa model klasifikasi, dengan menunjukkan jumlah prediksi yang tepat dan keliru berdasarkan label sebenarnya. Komponen utama dalam matriks ini mencakup *True Positive*, *True Negative*, *False Positive*, dan *False Negative*, yang menjadi acuan untuk menghitung metrik seperti akurasi, presisi, *recall*, dan *F1-score*.

a. Naïve Bayes Classifier (NBC)

Naïve Bayes Classifier (NBC) adalah algoritma klasifikasi yang bekerja berdasarkan prinsip probabilistik dengan asumsi independensi antar fitur. Evaluasi kinerjanya dilakukan menggunakan *classification_report*() dan *accuracy_score*(), yang menghasilkan akurasi 80,5%, *precision* 36,7%, *recall* 80,6%, dan *f1-score* 50,4%. Berdasarkan *confusion matrix* dengan TP = 29, FP = 50, FN = 7, dan TN = 207, model dinilai efektif dalam mengenali data negatif, namun cenderung tinggi dalam memproduksi prediksi positif yang salah. Visualisasi confusion matrix disajikan pada Gambar 10.



Gambar 10. *Confusion Matrix Naïve Bayes Classifier* (NBC)

Berdasarkan nilai Confusion matrix pada model *Naïve Bayes Classifier (NBC)*, diperoleh metrik evaluasi sebagai berikut:

a. *Accuracy*

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} = \frac{29 + 207}{29 + 50 + 7 + 207} = \frac{236}{293} = 0.805 \; atau \; 80.5\%$$

b. *Precision*

$$Precision = \frac{TP}{TP + FP} = \frac{29}{29 + 50} = \frac{29}{79} = 0.367 \; atau \; 36.7\%$$

c. *Recall*

$$Recall = \frac{TP}{TP + FN} = \frac{29}{29 + 7} = \frac{29}{36} = 0.806 \; atau \; 80.6\%$$

d. *F1-Score*

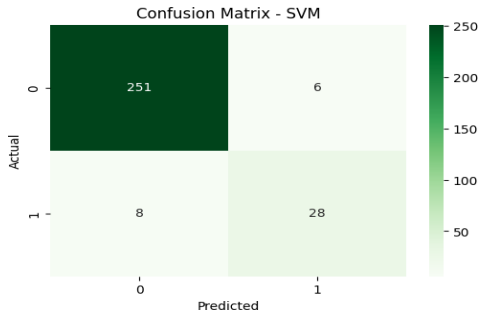
$$F1 - Score = 2 . \frac{Precision \ . \ Recall}{Precision + Recall} = 2 . \frac{0.367 \ . \ 0.806}{0.367 \ + \ 0.806} = 2 . \frac{0.2958}{1.173}$$
$$= 0{,}504 \ atau \ 50.4\%$$

Tingginya nilai *recall* pada model NBC menunjukkan kemampuannya dalam menangkap data positif secara menyeluruh. Namun, hal ini tidak diimbangi dengan precision yang baik, sehingga model cenderung menghasilkan prediksi positif yang berlebihan dan keliru, namun precision-nya rendah karena masih banyak data negatif yang salah diklasifikasikan sebagai positif. Hal ini menandakan perlunya perbaikan agar prediksi model lebih seimbang antara recall dan precision.

b. Support Vector Mechine (SVM)

SVM ialah metode klasifikasi yang memanfaatkan hyperplane sebagai batas pemisah optimal untuk memisahkan kelas-kelas data dalam ruang fitur. Evaluasi terhadap model dilakukan menggunakan fungsi *classification_report*() untuk memperoleh metrik seperti *precision*, *recall*, *f1-score*, serta *accuracy_score*() untuk menghitung akurasi. Berdasarkan *confusion matrix*, model menunjukkan performa klasifikasi yang seimbang, dengan *True Positive* (TP) = 28, *False Positive* (FP) = 6, *False Negative* (FN) = 8, dan *True Negative* (TN) = 251. Hasil ini menunjukkan bahwa SVM efektif dalam mengenali data positif maupun negatif dengan tingkat kesalahan prediksi yang rendah.

Dengan rincian nilai tersebut, diperoleh akurasi sebesar 95,2%, *precision* 82,4%, *recall* 77,8%, dan *f1-score* 79,9%. Temuan ini mengindikasikan bahwa SVM memiliki performa klasifikasi yang kuat dan stabil, dengan tingkat *False Positive* yang rendah serta proporsi *recall* dan *precision* yang seimbang. *Confusion matrix* untuk hasil klasifikasi SVM disajikan pada Gambar 11.



Gambar 11. *Confusion Matrix Support Vector Machine* (SVM)

Berdasarkan nilai Confusion matrix pada model *Support Vector Machine* (SVM), diperoleh metrik evaluasi sebagai berikut:

a. *Accuracy*

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} = \frac{28 + 251}{28 + 251 + 6 + 8} = \frac{279}{293} = 0.952 \ atau \ 95.2\%$$

b. *Precision*

$$Precision = \frac{TP}{TP + FP} = \frac{28}{28 + 6} = \frac{28}{34} = 0.824 \ atau \ 82.4\%$$

c. *Recall*

$$Recall = \frac{TP}{TP + FN} = \frac{28}{28 + 8} = \frac{28}{36} = 0.778 \ atau \ 77.8\%$$

d. *F1-Score*

$$F1 - Score = 2 . \frac{Precision \ . \ Recall}{Precision + Recall} = 2 . \frac{0.824 \ . \ 0.778}{0.824 \ + \ 0.778} = 2 . \frac{0.0640}{1.602}$$
$$= 0.799 \ atau \ 79.9\%$$

Dibandingkan dengan NBC, model SVM menunjukkan performa yang lebih unggul secara keseluruhan dalam mengklasifikasikan data pada dataset yang digunakan.

4.10 Hasil Pengujian dan Pembahasan Kontribusi Penelitian

Hasil pengujian menunjukkan bahwa algoritma *Support Vector Machine* (SVM) memberikan akurasi sebesar 95,2%, sementara *Naïve Bayes Classifier* (NBC) hanya mencapai akurasi 80,5%. Selain itu, nilai *precision, recall*, dan *f1-score* dari SVM juga lebih seimbang dan stabil dibandingkan NBC, yang cenderung memiliki recall tinggi namun precision rendah. Hal ini menandakan bahwa SVM lebih mampu mengklasifikasikan sentimen publik secara akurat terhadap isu kesehatan rokok elektrik.

Temuan ini sejalan dengan penelitian yang menggunakan SVM untuk mengklasifikasikan ulasan produk skincare dan mencatat akurasi sebesar 87% serta f1-score 87,37%. Meskipun objek penelitiannya berbeda, performa SVM yang konsisten tinggi menguatkan hasil penelitian ini, bahwa SVM efektif dalam mengelola data opini berbasis teks, termasuk yang berkaitan dengan isu kesehatan.

Sementara itu, penelitian oleh [4] yang hanya menggunakan NBC untuk mengklasifikasikan sentimen pengguna Twitter terhadap rokok elektrik menunjukkan akurasi sebesar 77,5%, dengan dominasi sentimen netral. Penelitian ini menunjukkan bahwa NBC mampu melakukan klasifikasi, namun belum optimal dalam menangkap kompleksitas opini publik. Penelitian ini memperkuat temuan bahwa NBC cocok digunakan untuk dataset sederhana, namun kurang tangguh jika dibandingkan dengan algoritma seperti SVM dalam menangani data yang lebih kompleks.

Lebih lanjut, konteks kesehatan sebagai latar isu sentimen juga diperkuat dari penelitian oleh [20] yang menekankan pentingnya edukasi terhadap bahaya rokok elektrik di kalangan remaja. Penelitian tersebut menunjukkan bahwa penyuluhan langsung dapat meningkatkan pengetahuan siswa terhadap risiko kesehatan seperti gangguan paru-paru, jantung, otak, dan sistem pernapasan akibat penggunaan rokok elektrik. Hal ini menegaskan bahwa rokok elektrik memang menjadi isu kesehatan yang penting dan layak untuk dianalisis dari sudut pandang opini publik.

Dengan demikian, kontribusi utama dari penelitian ini adalah memberikan perbandingan langsung dan terukur antara NBC dan SVM dalam konteks analisis sentimen terhadap isu kesehatan, khususnya pada rokok elektrik. Tidak hanya menguatkan hasil dari penelitian sebelumnya, penelitian ini juga menambahkan konteks baru dengan menggunakan data aktual dari media sosial X (Twitter) dan menerapkan pendekatan pembobotan TF-IDF serta *balanceing* data menggunakan SMOTE, yang belum banyak dilakukan secara terintegrasi di penelitian terdahulu.

Hasil ini diharapkan dapat memperkuat literatur yang ada dalam bidang analisis sentimen terhadap isu kesehatan, serta memberikan rujukan bagi penelitian lanjutan dalam memilih algoritma klasifikasi yang tepat sesuai konteks dan karakteristik data.

5. Simpulan

Berdasarkan hasil evaluasi model, NBC menunjukkan akurasi sebesar 80,5% dengan keunggulan dalam mendeteksi data positif (*recall* tinggi), namun lemah dalam akurasi prediksi data negatif (*precision* rendah). Sementara itu, algoritma SVM berhasil mencapai akurasi sebesar 95,2% serta menunjukkan keseimbangan yang optimal antara *precision* dan *recall*. Berdasarkan hasil tersebut, dapat disimpulkan bahwa SVM memiliki keunggulan performa dibandingkan NBC dalam konteks analisis sentimen terhadap kesehatan rokok elektrik pada masyrakat. Maka dari itu, SVM direkomendasikan untuk digunakan dalam penelitian sejenis di masa depan.

Daftar Referensi

[1] “Kementerian Kesehatan dan WHO Menerbitkan Laporan Global Adult Tobacco Survey Indonesia 2021,” *WHO INDONESIA*. https://www.who.int/indonesia/id/news/detail/22-08-2024-ministry-of-health-and-who-release-global-adult-tobacco-survey-indonesia-report-2021

[2] E. Apriani, F. Oktavianalisti, L. D. H. Monasari, I. Winarni, and I. F. Hanif, “Analisis Sentimen Penggunaan TikTok Sebagai Media Pembelajaran Menggunakan Algoritma Naïve Bayes Classifier,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1160–1168,

2024, doi: 10.57152/malcom.v4i3.1482.
[3] A. Muhammadin and I. A. Sobari, "Analisis Sentimen Pada Ulasan Aplikasi Kredivo Dengan Algoritma Svm Dan Nbc," *Reputasi J. Rekayasa Perangkat Lunak*, vol. 2, no. 2, pp. 85–91, 2021, doi: 10.31294/reputasi.v2i2.785.
[4] D. R. Aditya, E. Supriyati, and T. Listyorini, "Analisis Sentimen Pengguna Twitter Terhadap Rokok Elektrik (Vape) Di Indonesia Menggunakan Metode Naïve Bayes," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 1, pp. 43–50, 2022, doi: 10.29100/jipi.v7i1.2145.
[5] M. Khadapi and V. Maruli Pakpahan, "Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024," *JUKI J. Komput. dan Inform.*, vol. 6, no. 2, pp. 130–137, 2024, [Online]. Available: https://www.ioinformatic.org/index.php/JUKI/article/view/681
[6] H. D. Ahmad, E. Y. Puspaningrum, and R. Mumpuni, "Studi Performa TF-IDF dan Word2Vec Pada Analisis Sentimen Cyberbullying," *Router J. Tek. Inform. dan Terap.*, no. 2, pp. 94–106, 2024, [Online]. Available: https://doi.org/10.62951/router.v2i2.76
[7] R. Kosasih and A. Alberto, "Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor," *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 6, no. 1, pp. 134–139, 2021, [Online]. Available: https://doi.org/10.30743/infotekjar.v6i1.3893
[8] A. Azrul, A. Irma Purnamasari, and I. Ali, "Analisis Sentimen Pengguna Twitter Terhadap Perkembangan Artificial Intelligence Dengan Penerapan Algoritma Long Short-Term Memory (Lstm)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 413–421, 2024, doi: 10.36040/jati.v8i1.8416.
[9] C. Very *et al.*, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Optimasi Klasifikasi Sentimen Menggunakan Random Forest dengan Preprocessing K-Means Clustering dan SMOTE," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 10, no. 3, pp. 389–400, 2024.
[10] D. Andriyani, Ahmad Faqih, and Sandy Eka Permana, "The Effect of SMOTE Application on Support Vector Machine Performance in Sentiment Classification on Imbalanced Datasets," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 2, pp. 752–757, 2025, doi: 10.59934/jaiea.v4i2.742.
[11] M. R. F. Rahmatullah, P. N. Andono, Affandy, and M. A. Soeleman, "Improving Random Forest Performance for Sentiment Analysis on Unbalanced Data Using SMOTE and BoW Integration: PLN Mobile Application Case Study," *Sci. J. Informatics*, vol. 12, no. 1, pp. 1–10, 2025, doi: 10.15294/sji.v12i1.19295.
[12] J. P. Setiadi and S. Sugiyamta, "Analisis dan Visualisasi Berbasis Web Sentimen Pengguna Jenius Menggunakan Naïve Bayes Classifier," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 1, pp. 245–254, 2024, doi: 10.32493/jtsi.v7i1.37981.
[13] R. Damanhuri and V. A. Husein, "Analisis Sentimen pada Ulasan Aplikasi Access by KAI Berbahasa Indonesia Menggunakan Word-Embedding dan Classical Machine Learning," *J. Masy. Inform.*, vol. 15, no. 2, pp. 97–106, 2024, doi: 10.14710/jmasif.15.2.62383.
[14] M. CAHYO, "Analisis Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes Classifier (Studi Kasus: Program Studi Diploma III Teknologi Bank Darah di," 2024, [Online]. Available: https://eprints.utdi.ac.id/10436/
[15] R. Hidayat, M. Fikry, Y. Yusra, F. Yanto, and E. P. Cynthia, "Penerapan Naïve Bayes Classifier dalam Klasifikasi Sentimen Publik di Twitter terhadap Puan Maharani," *JUKI J. Komput. dan Inform.*, vol. 6, no. 1, pp. 100–108, 2024, doi: 10.53842/juki.v6i1.479.
[16] A. Mudya Yolanda and R. Tri Mulya, "Implementasi Metode Support Vector Machine untuk Analisis Sentimen pada Ulasan Aplikasi Sayurbox di Google Play Store," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 6, no. 2, pp. 76–83, 2024, doi: 10.35580/variansiunm258.
[17] G. R. Ditami, E. F. Ripanti, and H. Sujaini, "Implementasi Support Vector Machine untuk Analisis Sentimen Terhadap Pengaruh Program Promosi Event Belanja pada Marketplace," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 508, 2022, doi: 10.26418/jp.v8i3.56478.
[18] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.
[19] R. Chandra and E. M. Sipayung, "Analisis Sentimen Ulasan Aplikasi Samsat Digital Nasional Menggunakan Algoritma Naive Bayes Classifier," *J. Nas. Teknol. dan Sist. Inf.*, vol. 10, no. 3, pp. 156–164, 2025, doi: 10.25077/teknosi.v10i3.2024.156-164.

[20] A. Susanto, M. P. Mahardika, and H. Purwantiningrum, “Pemberdayaan Kesehatan Remaja : Edukasi Bahaya Rokok Elektrik bagi Siswa SMA Negeri 2 Tegal,” *J. Pengabdi. UNDIKMA*, vol. 4, no. 3, p. 634, 2023, doi: 10.33394/jpu.v4i3.8178.