# PEMODELAN SENTIMEN KOMENTAR YOUTUBE TERKAIT KEBIJAKAN HILIRISASI NIKEL MENGGUNAKAN ALGORITMA SUPERVISED LEARNING

**Ditta Adelia, Uce Indahyanti, Yunianita Rahmawati, Suprianto**
Informatika, Universitas Muhammadiyah Sidoarjo
Jl. Raya Gelam No. 250, Pagerwaja, Gelam, Kec. Candi, Kab. Sidoarjo, Jawa Timur, Indonesia
*uceindahyanti@umsida.ac.id*

**ABSTRAK**

Kebijakan hilirisasi nikel Indonesia sejak 2020 menuai respon beragam, mulai dari dukungan kedaulatan ekonomi hingga kritik dampak lingkungan dan ketimpangan distribusi manfaat. Platform YouTube dengan 139 juta pengguna aktif menjadi ruang diskusi publik yang dinamis terkait kebijakan ini. Namun, analisis sentimen sering menghadapi kendala ketidakseimbangan kelas yang menyebabkan model bias terhadap kelas mayoritas dan gagal mendeteksi kelas minoritas. Penelitian ini bertujuan menganalisis sentimen publik terhadap kebijakan hilirisasi nikel melalui komentar YouTube dan mengevaluasi efektivitas metode *SMOTE-Tomek Links* dalam menangani ketidakseimbangan kelas. Dataset 7.850 komentar YouTube yang telah melalui tahap *text processing* dan pelabelan, menghasilkan distribusi tidak seimbang (81,04% negatif vs 18,96% positif). Tiga algoritma *supervised learning* (*Naive Bayes*, *Random Forest*, *Support Vector Machine*) diuji pada dua skenario yaitu sebelum dan sesudah penerapan *SMOTE-Tomek Links*. Hasil menunjukkan bahwa model terbaik diperoleh menggunakan SVM rasio 80:20 dengan *SMOTE-Tomek Links*, mencapai akurasi 96,37% dan *recall* kelas positif 89%. Metode ini efektif mengurangi bias dengan penurunan selisih *recall* dari 14% menjadi 9%. Analisis sentimen mengungkapkan opini negatif didominasi kekhawatiran kerusakan lingkungan dan ketimpangan ekonomi, sementara opini positif berfokus pada kemajuan industri dan kedaulatan negara.

***Kata kunci :*** *Analisis Sentimen, Hilirisasi Nikel, YouTube, SMOTE-Tomek Links, Support Vector Machine*

## 1. PENDAHULUAN

Indonesia sebagai salah satu negara yang memiliki cadangan nikel terbesar di dunia mencapai 22% dari total cadangan global. Angka ini menempatkan Indonesia sebagai eksponen utama dalam transisi energi hijau, khususnya sebagai penyedia bahan baku utama baterai kendaraan listrik [1]. Meski demikian, selama puluhan tahun Indonesia hanya berperan sebagai pemasok bijih nikel mentah dengan nilai ekonomi yang rendah. Menyadari potensi tersebut, Pemerintah Indonesia menerapkan kebijakan hilirisasi melalui larangan ekspor bijih nikel mentah sejak awal tahun 2020. Kebijakan ini bertujuan mengalihkan rantai nilai dari ekspor bahan mentah menjadi produk olahan bernilai tinggi, guna menciptakan lapangan pekerjaan, menarik investasi asing untuk pembangunan fasilitas pengolahan (*smelter*), serta memperkuat kedaulatan ekonomi nasional [2].

Implementasi kebijakan hilirisasi nikel tidak selalu mendapatkan respon yang sepenuhnya positif dari masyarakat. Sebagian pihak mendukung kebijakan ini sebagai upaya melepaskan diri dari ketergantungan ekspor bahan mentah [3]. Namun, tidak sedikit yang mengkritisi dampak negatif yang ditimbulkan, seperti kerusakan lingkungan, pencemaran air dan udara di wilayah pertambangan, dominasi pemodal asing dalam kepemilikan *smelter*, hingga ketimpangan distribusi manfaat ekonomi yang tidak merata bagi masyarakat lokal [4]. Dinamika pro-kontra ini berpotensi menciptakan ketegangan sosial, jika tidak dikelola dengan baik dapat menghambat keberlanjutan kebijakan tersebut.

Era digital telah mengubah cara masyarakat menyampaikan aspirasi terhadap kebijakan publik. Platform media sosial, khususnya YouTube, menjadi ruang diskusi publik yang dinamis dan masif [5]. Data *We Are Social* 2025 menunjukkan YouTube merupakan platform media sosial terpopuler kedua di Indonesia dengan 139 juta pengguna aktif [6]. Berbeda dengan platform lain, komentar YouTube memungkinkan ekspresi opini yang lebih naratif dan detail, menjadikan sumber data yang kaya untuk memahami sentimen publik [7]. Analisis sentimen, salah satu cabang *Natural Language Processing* (NLP), menawarkan pendekatan sistematis untuk mengekstraksi dan mengklasifikasikan opini dari teks tidak terstruktur ke dalam kategori sentimen guna mengidentifikasi pola persepsi masyarakat [8].

Beberapa penelitian telah mengkaji analisis sentimen menggunakan media sosial dengan berbagai algoritma. Pratama dan Febriawan [9] menelaah sentimen kebijakan hilirisasi industri menggunakan Naive Bayes pada Twitter, namun menemukan model cenderung bias pada kelas mayoritas akibat ketidakseimbangan data. Destiyanti dkk [10] membandingkan SVM dan *Naive Bayes* dengan teknik *SMOTE* pada isu pertambangan lokal, di mana SVM terbukti lebih unggul dalam menangani data terbatas. Studi lain oleh Miftahusalam dkk [11] menerapkan *Random Oversampling* dan *Undersampling* pada tiga algoritma klasifikasi, namun teknik tersebut belum cukup efektif meningkatkan akurasi model secara signifikan. Marzuki dkk [12] menerapkan kombinasi *SMOTE* dan *Tomek Links* dengan *Random Forest* pada

analisis sentimen pariwisata, membuktikan efektivitas metode hibrida dalam menangani data tidak seimbang.

Permasalahan utama yang sering ditemukan dalam analisis sentimen media sosial adalah ketidakseimbangan distribusi kelas (*imbalanced dataset*), yang mana satu kategori sentimen cenderung mendominasi [13]. Kondisi serupa juga ditemukan dalam dataset penelitian ini. Dari 7.850 komentar YouTube yang terkumpul, terdapat ketidakseimbangan signifikan dengan 6.362 komentar termasuk sentimen negatif dan 1.488 komentar termasuk sentimen positif, dengan selisih 4.874 komentar. Kondisi ini menyebabkan kinerja model menjadi bias dan cenderung hanya memprediksi kelas mayoritas dengan baik, sementara performa pada kelas minoritas menurun [10]. Model yang dilatih pada data tidak seimbang cenderung mengabaikan kelas minoritas, padahal dalam konteks analisis kebijakan publik, kemampuan mendeteksi sentimen minoritas sama pentingnya dengan sentimen mayoritas untuk memberikan gambaran persepsi masyarakat.

Berdasarkan tinjauan terhadap penelitian terdahulu, teridentifikasi beberapa gap penelitian. Pertama, belum ada penelitian yang secara spesifik menganalisis sentimen kebijakan hilirisasi nikel di Indonesia menggunakan komentar YouTube. Mayoritas studi sebelumnya menggunakan Twitter dengan dataset terbatas dan berfokus pada isu lokal, bukan kebijakan nasional. Kedua, penanganan ketidakseimbangan kelas (*imbalanced dataset*) seringkali belum optimal dalam mendeteksi kelas minoritas. Ketiga, performa model pada penelitian terdahulu masih perlu ditingkatkan, khususnya dalam mendeteksi kelas minoritas yang ditandai dengan rendahnya nilai *recall* dan akurasi keseluruhan.

Oleh karena itu, penelitian ini berupaya mengisi gap tersebut dengan menggunakan dataset YouTube yang lebih representatif terkait kebijakan hilirisasi nikel Indonesia. Membandingkan performa tiga algoritma *supervised learning* yaitu *Naive Bayes*, *Random Forest*, dan *Support Vector Machine* (SVM). Lebih lanjut, penelitian ini mengevaluasi efektivitas metode *SMOTE-Tomek Links* dalam menangani ketidakseimbangan kelas pada dua skenario pengujian yaitu sebelum dan sesudah penyeimbangan data. Hasil penelitian diharapkan dapat memberikan wawasan bagi pembuat kebijakan dalam memahami persepsi publik, sekaligus menjadi evaluasi agar kebijakan yang diterapkan selaras dengan harapan masyarakat.

## 2. TINJAUAN PUSTAKA

### 2.1. *SMOTE-Tomek Links*

*SMOTE-Tomek Links* merupakan metode hibrida yang menggabungkan teknik *oversampling* dan *undersampling* untuk mengatasi ketidakseimbangan distribusi kelas. Metode ini bekerja dalam dua tahap, yaitu SMOTE (*Synthetic Minority Over-sampling Technique*) dan *Tomek Links* yang bekerja secara berurutan untuk menghasilkan dataset yang lebih seimbang dan bersih dari *noise*. *SMOTE* bekerja dengan menciptakan data sintetis untuk kelas minoritas melalui interpolasi linear antara instance minoritas dengan tetangga terdekatnya dalam ruang fitur [14]. Rumus pembentukan data baru pada *SMOTE* dapat dilihat pada Persamaan (1).

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \quad (1)$$

Keterangan:
$x_i$ : Vektor fitur sampel data kelas minoritas.
$x_{zi}$ : Salah satu $k$-tetangga dari $x_i$.
$\lambda$ : Bilangan acak antara 0 dan 1.

Setelah *oversampling*, tahap kedua adalah penerapan *Tomek Links* untuk membersihkan pasangan data yang ambigu di batas keputusan (*decision boundary*) antar kelas [12]. Jika ditemukan pasangan terdekat dari kelas berbeda, data tersebut dihapus untuk mengurangi *noise*.

### 2.2. *Naive Bayes*

*Naive Bayes* adalah metode klasifikasi yang didasarkan pada teori probabilitas dengan asumsi bahwa setiap fitur atau kata bersifat bebas (*independent*) satu sama lain [5]. Perhitungan probabilitas dilakukan menggunakan Teorema Bayes ditunjukkan pada Persamaan (2).

$$P(C_i \mid x) = \frac{P(x \mid C_i) \cdot P(C_i)}{P(x)} \quad (2)$$

Keterangan:
$P(C_i \mid x)$ : Probabilitas posterior (peluang $C_i$ benar jika diberikan data $x$).
$P(C_i)$ : Probabilitas prior (peluang awal hipotesis $C_i$).
$P(x \mid C_i)$ : *Likelihood* (Peluang muncul data $x$ jika hipotesis $C_i$ benar).
$P(x)$ : Probabilitas data $x$.

### 2.3. *Random Forest*

*Random Forest* merupakan metode klasifikasi *ensemble* yang membangun sejumlah pohon keputusan pada saat pelatihan. Untuk membangun pohon keputusan, algoritma ini mengukur ketidakmurnian data menggunakan indeks Gini (*Gini Impurity*) [15]. Rumus perhitungan *Gini Impurity* ditunjukkan pada Persamaan (3).

$$Gini = 1 - \sum_{i=1}^{C} \pi_i^2 \quad (3)$$

Keterangan:
$C$ : Jumlah kelas dalam suatu *node*.
$\pi_i$ : Probabilitas relatif kelas ke-$i$.

### 2.4. *Support Vector Machine* (SVM)

SVM bekerja dengan mencari *hyperplane* (garis pemisah) terbaik yang memisahkan dua kelas data dalam ruang vektor berdimensi tinggi [11]. Fungsi keputusan klasifikasi linier pada SVM dinyatakan dalam Persamaan (4).

$$f(x) = sign\,(w \cdot x + b) \quad (4)$$

Keterangan:
$w$ : Vektor bobot (*weight vector*) *hyperplane*.
$x$ : Vektor input data (nilai *TF-IDF*).
$b$ : Parameter bias.

### 2.5. Evaluasi Model

Kinerja model klasifikasi dievaluasi menggunakan *confusion matrix*. Metode ini merepresentasikan perbandingan antara hasil prediksi model dengan label data yang sebenarnya dalam bentuk matriks [16]. Tabel *confusion matrix* dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

| Actual | Prediction | |
|---|---|---|
| | **Positive** | **Negative** |
| **Positive** | True Positive (TP) | False Negative (FN) |
| **Negative** | False Positive (FP) | True Negative (TN) |

Nilai-nilai parameter tersebut digunakan untuk mengukur kinerja model melalui metrik akurasi, presisi, *recall*, dan *F1-Score*. Akurasi digunakan untuk mengukur proporsi prediksi benar dari keseluruhan data, dihitung menggunakan Persamaan (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Presisi mengukur ketepatan prediksi untuk suatu kelas, dihitung dengan Persamaan (6).

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

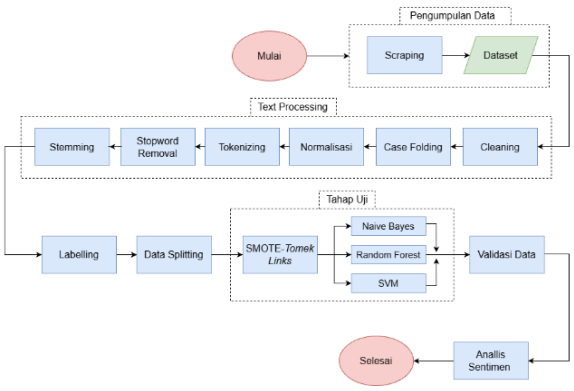
*Recall* mengukur kemampuan model menemukan semua *instance* dari suatu kelas. Rumus *recall* dapat dilihat pada Persamaan (7).

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

*F1-Score* merupakan *harmonic mean* dari *precision* dan *recall*. Rumus *F1-Score* dapat dilihat pada Persamaan (8).

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

## 3. METODE PENELITIAN




Gambar 1. Alur Penelitian

Penelitian ini menerapkan pendekatan kuantitatif berbasis *supervised learning* sebagai metode pemodelan dalam menganalisis komentar publik terkait kebijakan hilirisasi nikel di Indonesia. Adapun tahapan dalam penelitian dapat dilihat pada Gambar 1, alur penelitian dimulai dari pengumpulan data. Kemudian diproses melalui *text processing*, sebelum dilakukan pelabelan. Data yang telah bersih dibagi menjadi data *training* dan *testing*, lalu diseimbangkan menggunakan *SMOTE-Tomek Links*. Selanjutnya, dilakukan pemodelan. Tahap akhir adalah analisis sentimen dan evaluasi performa model.

### 3.1. Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dalam rentang waktu 4 September 2019 hingga 9 Agustus 2025. Sumber data berasal dari platform YouTube yang relevan dengan topik kebijakan hilirisasi nikel di Indonesia. Salah satu contoh referensi video yang digunakan adalah *Jelajah Hilirisasi Nikel, Dari Jantung Morowali Menuju Dunia - YouTube*. Proses pengumpulan data (*scraping*) dilakukan terhadap 51 video menggunakan bahasa pemrograman Python dengan memanfaatkan YouTube Data API v3. Melalui proses tersebut, berhasil dihimpun sebanyak 15.014 data mentah. Untuk keperluan analisis sentimen, penelitian ini membatasi penggunaan data hanya dengan mengekstrak atribut isi komentar. Sampel data hasil *scraping* dapat dilihat pada Tabel 2.

Tabel 2. Contoh Dataset

| No | Text |
|---|---|
| 1 | stop tambang nikel dengan dalih *green energy*, padahal merusak lingkungan yang sudah hijau ☹ |
| 2 | Ayoooo perluas lagi hilirisasi.... |
| 3 | Hilirisasi merugikan bangsa Indonesia, pemerintahnya kurang tepat dikelola |
| 4 | Kebijakan Program hilirisasi memang jempollll.....tetapi pelaksanaannya harus sesuai rencana jangan sampai bocor sampai 5jt ton tuhhh...☹ |
| 5 | Yaudah gausah dihilirisasi timbang merusak alam jual nikel dlm bentuk mentah saja👎👎👎 |

### 3.2. *Text Processing*

Data mentah yang diperoleh bersifat tidak terstruktur dan mengandung banyak *noise* seperti bahasa tidak baku, singkatan, dan karakter tidak relevan [10]. Oleh karena itu dilakukan text processing yang terdiri dari enam tahapan, yaitu:

a. *Cleaning*: Menghapus karakter tidak relevan seperti URL, tag HTML, mention, hashtag, emoji, angka, dan karakter non-alfanumerik termasuk tanda baca.
b. *Case Folding*: Mengonversi seluruh karakter menjadi huruf kecil (*lowercase*).
c. *Normalization*: Mengganti kata tidak baku, singkatan, dan bahasa gaul *(slang)* menjadi bentuk baku sesuai kaidah Bahasa Indonesia.

d. *Tokenizing*: Memisahkan kalimat menjadi token-token kata.
e. *Stopword Removal*: Menghapus kata umum yang sering muncul atau tidak memiliki makna sentimen yang signifikan. Kata negasi seperti 'tidak' dipertahankan untuk menjaga konteks sentimen, namun dikecualikan dari *word cloud* agar visualisasi lebih representatif.
f. *Stemming*: Mengubah kata berimbuhan ke bentuk dasar untuk mengurangi variasi kosakata yang memiliki makna sama.

### 3.3. Labelling

Setelah data bersih, dilakukan proses pelabelan untuk menentukan kelas sentimen dari setiap komentar. Pelabelan dilakukan secara otomatis menggunakan metode berbasis leksikon dengan memanfaatkan kamus *InSet (Indonesia Sentiment Lexicon)*. InSet merupakan kamus sentimen Bahasa Indonesia yang berisi daftar kata berserta bobot polaritasnya [17].

### 3.4. Pembagian Data

Dataset berlabel dibagi menjadi dua subset yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih model dalam mengidentifikasi pola, sedangkan *data testing* berfungsi mengecek seberapa mampu model dalam memprediksi data baru. Penelitian ini menggunakan rasio pembagian 70:30, 80:20, dan 90:10 guna menentukan proporsi paling optimal. Komposisi pembagian data dapat mempengaruhi performa model, di mana penggunaan minimal 70% data latih cenderung memberikan hasil yang lebih stabil dan akurat [18].

### 3.5. Ekstraksi Fitur

Data teks ditransformasi menjadi representasi numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). TF (Term Frequency) mengukur frekuensi kemunculan kata dalam dokumen, sedangkan IDF (*Inverse Document Frequency*) mengukur seberapa unik kata tersebut di seluruh korpus [19].

### 3.6. Skenario Pengujian

Setiap algoritma diuji dalam dua skenario pengujian untuk mengevaluasi dampak penyeimbangan data. Skenario pertama dilakukan menggunakan data asli, sedangkan skenario kedua menerapkan *SMOTE-Tomek Links* untuk menangani masalah ketidakseimbangan kelas. Penerapan *SMOTE-Tomek Links* hanya pada data latih setelah proses pembagian data, sedangkan data uji tetap dibiarkan orisinal guna mencegah terjadinya kebocoran data (*data leakage*). Klasifikasi dilakukan menggunakan tiga algoritma *Supervised Learning*.

### 3.7. Validasi Data

Validasi atau evaluasi model dilakukan untuk mengukur sejauh mana kehandalan model yang telah dibuat. Pengukuran kinerja model dilakukan menggunakan *confusion matrix*. Berdasarkan matriks tersebut, dihitung nilai metrik evaluasi yang meliputi akurasi, presisi, *recall*, dan *F1-Score*.

### 3.8. Analisis Sentimen

Tahap terakhir adalah analisis sentimen, di mana hasil klasifikasi dari model terbaik dianalisis lebih lanjut. Analisis ini bertujuan untuk menginterpretasikan kecenderungan opini publik terhadap kebijakan hilirisasi nikel, apakah cenderung positif (mendukung) atau negatif (menolak), serta memberikan wawasan yang bermanfaat bagi pemangku kepentingan.

## 4. HASIL DAN PEMBAHASAN

### 4.1. Text Processing

Data mentah yang terkumpul selanjutnya melalui tahapan *text processing* untuk membersihkan *noise* dan mengubah data menjadi format yang terstruktur. Tabel 3 menunjukkan hasil transformasi data pada setiap tahapan *text processing* menggunakan salah satu komentar sebagai contoh.

Tabel 3. Contoh *Text Processing*

| Tahapan | Hasil |
|---|---|
| Data Awal | stop tambang nikel dengan dalih green energy, padahal merusak lingkungan yang sudah hijau ☹ |
| *Cleaning* | stop tambang nikel dengan dalih green energy padahal merusak lingkungan yang sudah hijau |
| *Case Folding* | stop tambang nikel dengan dalih green energy padahal merusak lingkungan yang sudah hijau |
| *Normalisasi* | berhenti tambang nikel dengan dalih hijau energi padahal merusak lingkungan yang sudah hijau |
| *Tokenizing* | ['berhenti', 'tambang', 'nikel', 'dengan', 'dalih', 'hijau', 'energi', 'padahal', 'merusak', 'lingkungan', 'yang', 'sudah', 'hijau'] |
| *Stopword Removal* | ['berhenti', 'tambang', 'nikel', 'dalih', 'hijau', 'energi', 'merusak', 'lingkungan', 'hijau'] |
| *Stemming* | ['henti', 'tambang', 'nikel', 'dalih', 'hijau', 'energi', 'rusak', 'lingkungan', 'hijau'] |

Proses *text processing* menghasilkan beberapa perubahan pada dataset. Dari 15.014 komentar awal, terdapat 330 komentar duplikat yang dihapus, dan 6.834 komentar yang tidak mengandung teks relevan setelah proses filtering. Dataset akhir yang siap untuk tahap pelabelan berjumlah 7.850 komentar.

### 4.2. Pelabelan Sentimen

Pelabelan sentimen dilakukan menggunakan pendekatan lexicon-based dengan kamus InSet (Indonesia Sentiment Lexicon) untuk

mengklasifikasikan komentar ke dalam kelas positif atau negatif berdasarkan skor polaritas kata. Proses pelabelan bekerja dengan mencocokkan setiap kata dalam komentar hasil *text processing* dengan kamus InSet, kemudian menjumlahkan bobot polaritas seluruh kata yang cocok. Komentar diklasifikasikan sebagai sentimen positif jika total skor lebih dari nol, dan negatif jika kurang dari nol. Dari total 7.850 komentar, diperoleh 6.362 data diklasifikasikan sebagai sentimen negatif dan 1.488 data sebagai sentimen positif. Sampel hasil pelabelan dapat dilihat pada Tabel 4.

Tabel 4. Hasil Pelabelan Sentimen

| Text Processing | Skor | Sentimen |
|---|---|---|
| henti tambang nikel dalih energi hijau rusak lingkungan hijau | -5 | Negatif |
| ayo luas hilirisasi | 3 | Positif |
| hilirisasi rugi bangsa indonesia pemerintah kurang tepat kelola | -2 | Negatif |

Berdasarkan distribusi data di atas, terlihat adanya *imbalanced dataset* antara kelas mayoritas (negatif) dan kelas minoritas (positif). Dominasi sentimen negatif mengindikasikan bahwa mayoritas respons publik di YouTube cenderung berisi kritik terhadap dampak lingkungan dan implementasi kebijakan hilirisasi nikel di Indonesia.

### 4.3. Pembagian Data

Dataset dibagi menjadi data *training* dan data *testing* menggunakan tiga rasio. Contoh pada rasio 90:10, sebanyak 90% dari total dataset atau 7.065 komentar dialokasikan sebagai data *training*. Sementara itu, 10% sisanya atau sebanyak 785 komentar digunakan sebagai data *testing*. Pendekatan perhitungan yang sama juga diterapkan pada rasio 80:20 dan 70:30. Distribusi data untuk setiap skenario disajikan pada Tabel 5.

Tabel 5. Pembagian Data

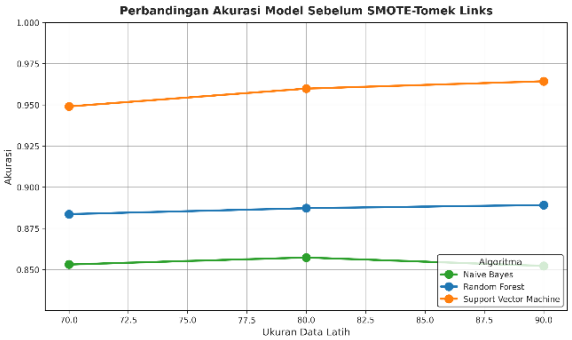
| Rasio | Distribusi Data | | Total |
|---|---|---|---|
| | *Training* | *Testing* | |
| 90 : 10 | 7.065 | 785 | 7.850 |
| 80 : 20 | 6.280 | 1.570 | 7.850 |
| 70 : 30 | 5.495 | 2.355 | 7.850 |

### 4.4. Skenario 1 (Tanpa *SMOTE-Tomek Links*)

Pengujian pertama dilakukan tanpa penerapan teknik penyeimbangan data untuk mengidentifikasi performa baseline dari setiap algoritma. Ketiga algoritma dilatih dan diuji pada tiga rasio pembagian data. Hasil evaluasi performa model disajikan pada Tabel 6.

Tabel 6. Hasil Klasifikasi Skenario 1

| Rasio | Algoritma | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|---|
| 90:10 | Naive Bayes | 85,22% | 87,50% | 85,22% | 81,13% |
| | Random Forest | 88,92% | 89,11% | 88,92% | 87,42% |
| | Support Vector Machine | 96,43% | 96,43% | 96,43% | 96,34% |
| 80:20 | Naive Bayes | 85,73% | 87,87% | 85,73% | 82,01% |
| | Random Forest | 88,73% | 88,81% | 88,73% | 87,22% |
| | Support Vector Machine | 95,99% | 95,96% | 95,99% | 95,88% |
| 70:30 | Naive Bayes | 85,31% | 87,25% | 85,31% | 81,34% |
| | Random Forest | 88,37% | 87,97% | 88,37% | 87,03% |
| | Support Vector Machine | 94,90% | 94,84% | 94,90% | 94,74% |




Gambar 2. Perbandingan Akurasi (Skenario 1)

Berdasarkan Tabel 6, algoritma SVM secara konsisten mengungguli kedua algoritma lainnya pada seluruh rasio pembagian data, dengan akurasi tertinggi 96,43% dicapai pada rasio 90:10. Pada rasio tersebut, SVM juga mencatat nilai *precision* 96,43%, *recall* 96,43%, dan *F1-Score* 96,34%. *Random Forest* menempati posisi kedua dengan akurasi berkisar 88,37% - 88,92%, sementara *Naive Bayes* memiliki akurasi terendah pada rentang 85,22% - 85,73%. Pola ini konsisten di ketiga rasio, menunjukkan bahwa SVM paling efektif dalam menangani ketidakseimbangan kelas tanpa teknik penyeimbangan data.

Ditinjau dari pengaruh rasio pembagian data, akurasi model menunjukkan stabilitas yang baik meskipun terdapat variasi proporsi data training. Rata-rata akurasi seluruh model pada rasio 90:10 mencapai 90,19%, pada rasio 80:20 sebesar 90,15%, dan pada rasio 70:30 sebesar 89,53%. Perbedaan akurasi antar rasio relatif kecil (maksimal 0,66 persentase), mengindikasikan bahwa ketiga model cukup robust terhadap variasi ukuran data training. Namun, akurasi tinggi ini tidak serta-merta menunjukkan performa model yang sempurna, terutama dalam mendeteksi kelas minoritas.

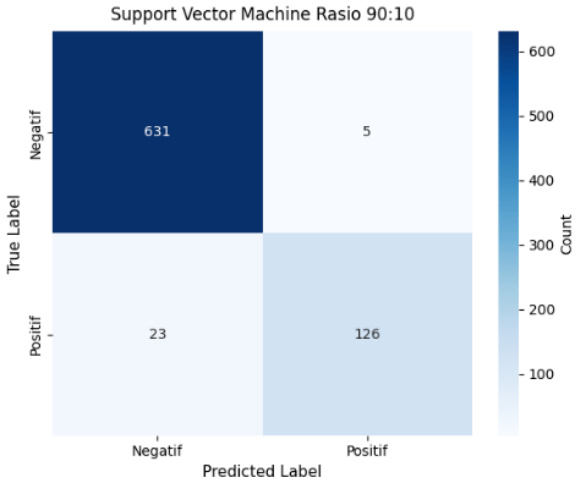
Meskipun akurasi keseluruhan terlihat tinggi, analisis terhadap *classification report* mengungkapkan

masalah terkait bias model terhadap kelas mayoritas. Akurasi yang tinggi dapat menyesatkan karena model bisa mencapai nilai tersebut hanya dengan memprediksi kelas mayoritas (negatif) dengan sangat baik, sambil mengabaikan kelas minoritas (positif). Untuk mengidentifikasi tingkat bias ini, dilakukan analisis *recall* per kelas pada rasio dengan performa terbaik (90:10) sebagaimana disajikan pada Tabel 7.

Tabel 7. *Recall* Skenario 1 (Rasio 90:10)

| Algoritma | *Recall* | | Selisih |
|---|---|---|---|
| | Negatif | Positif | |
| *Naive Bayes* | 100% | 22% | 78% |
| *Random Forest* | 99% | 46% | 53% |
| SVM | 99% | 85% | 14% |

Pada rasio 90:10, terlihat jelas ketidakseimbangan performa antar kelas. SVM memiliki recall kelas negatif 99% namun recall kelas positif hanya 85%, artinya meskipun model berhasil mengidentifikasi 99% dari sentimen negatif dengan benar, masih terdapat 15% sentimen positif yang gagal dideteksi. Pola bias lebih parah terlihat pada *Naive Bayes* dengan *recall* positif hanya 22% (selisih 78%), yang berarti model gagal mendeteksi hampir empat per lima dari sentimen positif. *Random Forest* menunjukkan performa yang sedikit lebih baik dengan *recall* positif 46%, namun masih mengindikasikan bias signifikan (selisih 53%) terhadap kelas mayoritas. *Confusion matrix* untuk model terbaik divisualisasikan pada Gambar 3.




Gambar 3. *Confusion Matrix* Hasil Uji Terbaik

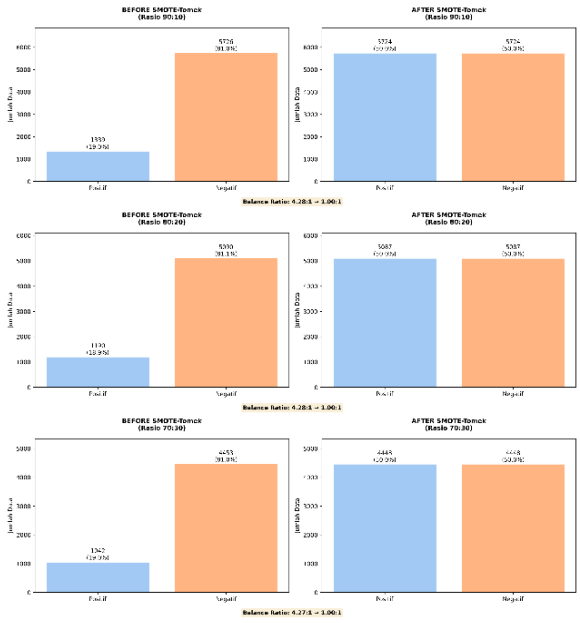
Pada Gambar 3, dapat dilihat rincian kinerja model klasifikasi SVM pada rasio 90:10 terhadap 785 data *testing*. Dari total 636 data bersentimen negatif, model mampu memprediksi dengan sangat akurat sebanyak 631 data (*True Negative*) dan hanya meleset 5 data yang salah diprediksi sebagai positif (*False Positive*). Sebaliknya, pada kelas minoritas yang berjumlah 149 data sentimen positif, model hanya mampu memprediksi dengan benar sebanyak 126 data (*True Positive*), sementara 23 data lainnya salah diklasifikasikan sebagai sentimen negatif (*False Negative*). Tingginya angka *False Negative* dibandingkan *False Positive* ini membuktikan adanya kecenderungan bias model terhadap kelas mayoritas. Hal ini menegaskan bahwa nilai akurasi keseluruhan yang tinggi tidak cukup untuk mengevaluasi model pada *imbalanced dataset*, sehingga diperlukan teknik penyeimbangan data untuk meningkatkan kemampuan model dalam mendeteksi kedua kelas secara proporsional.

### 4.5. Skenario 2 (*SMOTE-Tomek Links*)

Untuk mengatasi masalah ketidakseimbangan kelas yang teridentifikasi pada Skenario 1, diterapkan teknik penyeimbangan data yaitu *SMOTE-Tomek Links* pada data training. *SMOTE* mensintesis data baru untuk kelas minoritas melalui interpolasi k-nearest neighbors (k=5). *Tomek Links* membersihkan *noise* dengan menghapus pasangan data dari kelas berbeda yang terlalu dekat satu sama lain. Tabel 8 menunjukkan perbandingan jumlah sampel sebelum dan sesudah penerapan *SMOTE-Tomek Links*.

Tabel 8. Perbandingan Distribusi Data Training

| Rasio | Kelas | *SMOTE-Tomek Links* | |
|---|---|---|---|
| | | Sebelum | Sesudah |
| 90 : 10 | Negatif | 5.726 | 5.724 |
| | Positif | 1.339 | 5.724 |
| 80 : 20 | Negatif | 5.090 | 5.087 |
| | Positif | 1.190 | 5.087 |
| 70 : 30 | Negatif | 4.453 | 4.448 |
| | Positif | 1.042 | 4.448 |



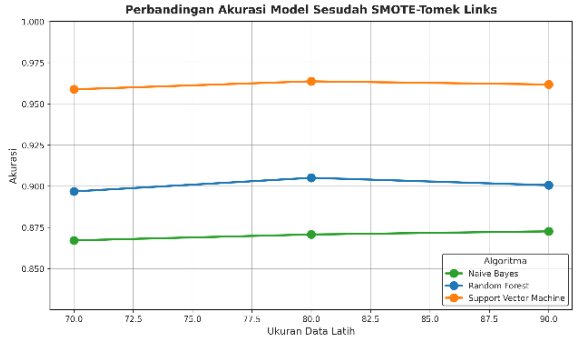
Gambar 4. Perbandingan Rasio Data

Tabel 8 dan Gambar 4 menunjukkan perbandingan distribusi data sebelum dan sesudah penerapan *SMOTE-Tomek Links*. Sebagai contoh pada rasio 90:10, kelas minoritas (positif) yang awalnya berjumlah 1.339 sampel disintesis oleh tahap *SMOTE* sebanyak 4.385 data baru. Bersamaan dengan itu, tahap *Tomek Links* menghapus 2 sampel *noise* dari

kelas mayoritas (negatif) yang awalnya berjumlah 5.726 sampel. Proses ini secara efektif menghasilkan distribusi akhir yang seimbang dengan rasio 1:1, yakni masing-masing 5.724 sampel. Pola penyeimbangan serupa juga terjadi secara konsisten pada rasio 80:20 dan 70:30. Metode ini hanya diterapkan pada data *training* untuk mencegah kebocoran data (*data leakage*). Selanjutnya, ketiga algoritma dilatih ulang pada data *training* yang telah seimbang dan dievaluasi menggunakan data *testing* asli. Hasil pengujian model setelah penerapan SMOTE-Tomek Links disajikan pada Tabel 9.

Tabel 9. Hasil Klasifikasi Skenario 2

| Rasio | Algoritma | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|---|
| 90:10 | Naive Bayes | 87,26% | 87,13% | 87,26% | 87,19% |
| | Random Forest | 90,06% | 89,61% | 90,06% | 89,48% |
| | Support Vector Machine | 96,18% | 96,14% | 96,18% | 96,15% |
| 80:20 | Naive Bayes | 87,07% | 86,89% | 87,07% | 86,98% |
| | Random Forest | 90,51% | 90,09% | 90,51% | 90,06% |
| | Support Vector Machine | 96,37% | 96,33% | 96,37% | 96,34% |
| 70:30 | Naive Bayes | 86,71% | 86,50% | 86,71% | 86,60% |
| | Random Forest | 89,68% | 89,19% | 89,68% | 89,25% |
| | Support Vector Machine | 95,88% | 95,84% | 95,88% | 95,85% |




Gambar 5. Perbandingan Akurasi (Skenario 2)

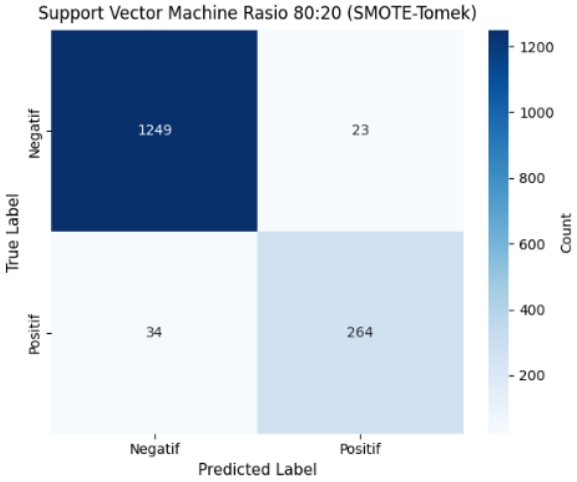
Hasil menunjukkan bahwa SVM tetap mempertahankan posisi terbaik dengan akurasi tertinggi 96,37% pada rasio 80:20. Analisis *recall* per kelas pada rasio dengan performa terbaik (80:20) disajikan pada Tabel 10.

Tabel 10. *Recall* Skenario 2 (Rasio 80:20)

| Algoritma | *Recall* | | Selisih |
|---|---|---|---|
| | Negatif | Positif | |
| *Naive Bayes* | 92% | 64% | 28% |
| *Random Forest* | 97% | 64% | 33% |
| SVM | 98% | 89% | 9% |

Dapat dilihat pada tabel 10 rasio 80:20 (model terbaik), SVM memiliki *recall* kelas negatif 98% dan *recall* kelas positif 89%, dengan selisih hanya 9%. Peningkatan tertinggi terlihat pada *Naive Bayes* dengan *recall* positif meningkat 42% (dari 22% menjadi 64%) dan *Random Forest* meningkat 18% (dari 46% menjadi 64%).

Berdasarkan Gambar 6, kinerja SVM (80:20) terhadap 1.570 data *testing* menunjukkan hasil prediksi yang jauh lebih proporsional. Pada kelas mayoritas (negatif), model memprediksi 1.249 *True Negative* dan 23 *False Positive*. Signifikansi terlihat pada kelas minoritas (positif), di mana model berhasil memprediksi 264 *True Positive* dengan hanya 34 *False Negative*. Tingginya keberhasilan deteksi pada kelas minoritas ini mengonfirmasi penurunan bias model secara drastis setelah dilakukan penyeimbangan data.




Gambar 6. *Confusion Matrix* Hasil Uji Terbaik

Model terbaik di Skenario 1 (SVM 90:10) mencapai akurasi 96,43% dengan *recall* positif 85%, sementara model terbaik di Skenario 2 (SVM 80:20) mencapai akurasi 96,37% dengan *recall* positif 89%. Meskipun terjadi penurunan akurasi sebesar 0,06 poin persentase, peningkatan *recall* positif menunjukkan bahwa model menjadi lebih seimbang dalam mendeteksi kedua kelas. Penyesuaian ini sangat menguntungkan karena menghasilkan model yang jauh lebih seimbang dalam mendeteksi kedua kelas sentimen. Dengan demikian, *SMOTE-Tomek Links* terbukti efektif sebagai solusi untuk menangani ketidakseimbangan data dalam analisis sentimen.

### 4.6. Analisis Kata

Hasil dari proses ini berupa visualisasi untuk mengetahui informasi umum mengenai data dengan menggunakan *word cloud*. Berikut merupakan *word cloud* data komentar positif dan negatif terhadap kebijakan hilirisasi nikel.


Gambar 7. *Word Cloud* Komentar Positif

Berdasarkan Gambar 7, dapat dilihat bahwa semakin besar ukuran kata dalam *word cloud* semakin sering muncul dalam komentar positif. Beberapa kata yang mendominasi yaitu "nikel", "hilirisasi", "negara", "maju", "indonesia", "kerja", "usaha", "rakyat", dan "ekonomi". Kemunculan kata "tidak" dalam konteks positif mengindikasikan penggunaan frasa pembelaan seperti "tidak salah" atau "tidak merugikan". Kata "hilirisasi", "maju", "kerja", dan "usaha" yang dominan menunjukkan bahwa komentar positif mencerminkan dukungan masyarakat terhadap kebijakan hilirisasi nikel sebagai upaya memajukan industri, menciptakan lapangan kerja.


Gambar 8. *Word Cloud* Komentar Negatif

Sedangkan untuk komentar negatif pada Gambar 8 kata-kata yang sangat dominan adalah "rusak", "negara", "asing", "rakyat", "indonesia", "nikel", "lingkungan", dan "pejabat". Kata "rusak" yang paling menonjol mengindikasikan kekhawatiran utama masyarakat terhadap kerusakan lingkungan akibat aktivitas hilirisasi nikel. Kemunculan kata "asing" dan "rakyat" menunjukkan kritik terhadap keterlibatan pihak asing dan ketimpangan distribusi manfaat ekonomi yang tidak dirasakan oleh rakyat lokal. Kata "salah", "korupsi", dan "pejabat" mencerminkan sentimen negatif terhadap tata kelola dan implementasi kebijakan oleh pemerintah.

## 5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa opini publik di YouTube terhadap kebijakan hilirisasi nikel mengalami ketidakseimbangan kelas (*imbalanced dataset*), di mana sentimen negatif mendominasi sebesar 81,04% dibandingkan sentimen positif yang hanya 18,96% (rasio 4,3:1). Secara kualitatif, sentimen negatif menyoroti isu kerusakan lingkungan dan ketimpangan distribusi manfaat ekonomi, sedangkan sentimen positif mencerminkan dukungan terhadap kemajuan industri dan kedaulatan negara. Kondisi data yang memicu bias pada model klasifikasi ini berhasil diatasi menggunakan metode hibrida *SMOTE-Tomek Links*. Dalam komparasi algoritma, Support Vector Machine (SVM) secara konsisten mengungguli Naive Bayes dan Random Forest. Model terbaik diperoleh menggunakan algoritma SVM pada rasio pembagian data 80:20 setelah penerapan *SMOTE-Tomek Links*, yang menghasilkan akurasi tertinggi sebesar 96,37%. Keberhasilan metode ini dibuktikan dengan tingginya nilai *recall* pada kelas minoritas (positif) yang mencapai 89%, serta *selisih* performa antar kelas yang berhasil direduksi menjadi hanya 9%. Hal ini memastikan model menjadi seimbang dan tidak lagi bias terhadap kelas mayoritas. Saran untuk penelitian selanjutnya adalah mencoba metode *Deep Learning* seperti LSTM atau BERT untuk menangkap konteks kalimat yang lebih kompleks, serta memperluas sumber data ke platform media sosial lain untuk mendapatkan perspektif yang lebih beragam.

## DAFTAR PUSTAKA

[1] W. P. Sahputra, B. A. Badia, M. I. Putra, F. C. Putra, dan A. A. Aji, “Rekayasa Proses Ekstraksi dan Pengolahan Bijih Nikel: Teknologi, Tantangan, dan Prospek Masa Depan,” *KAPALAMADA J. Multidisipliner*, vol. 4, no. 02, hal. 243–255, 2025, doi: 10.62668/kapalamada.v4i02.1546.

[2] R. B. Santoso, D. F. Moenardy, R. Muttaqin, dan D. Saputera, “Pilihan Rasional Indonesia dalam Kebijakan Larangan Ekspor Bijih Nikel,” *Indones. Perspect.*, vol. 8, no. 1, hal. 154–179, 2023.

[3] M. Hidayat dan U. Budiyanto, “Sentimen Analisis Tentang Hilirisasi Industri Berdasarkan Opini Masyarakat Di Twitter Menggunakan Metode K-Nearest Neighbor,” *Pros. Semin. Nas. …*, vol. 2, no. September, hal. 826–835, 2023.

[4] Norlaila, W. W. Winarno, dan E. T. Luthfi, “Analisis Sentimen Masyarakat Tentang Tambang Di Indonesia Pada Twitter Menggunakan Data Mining,” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, hal. 1091–1099, 2024, doi: 10.29100/jipi.v9i3.5402.

[5] A. A. Ningtyas, A. Solichin, dan R. Pradana, “Analisis Sentimen Komentar Youtube Tentang Prediksi Resesi Ekonomi Tahun 2023 Menggunakan Algoritme Naïve Bayes,” *Bit (Fakultas Teknol. Inf. Univ. Budi Luhur)*, vol. 20, no. 1, hal. 9, 2023, doi: 10.36080/bit.v20i1.2317.

[6] W. A. Social, “Digital 2025: Indonesia,” We Are Social & Meltwater.

[7] M. B. Al Hakiki dan Y. Darmi, “Penerapan Algoritma Machine Learning SVM dan NBC pada Sentimen Analisis Komentar Youtube Program Pengaduan Masyarakat Lapor Mas

Wapres," *J. Ilmu Multidisiplin*, vol. 4, no. 1, hal. 396–410, 2025, doi: 10.38035/jim.v4i1.884.

[8] M. I. Prayugah, U. Indahyanti, dan N. Ariyanti, "Analisis sentimen publik pada pemerintah dalam serangan ransomware dengan pendekatan smote," vol. 8, no. 2, hal. 333–343, 2024, doi: 10.35145/joisie.v8i2.4764.

[9] A. B. Pratama dan D. Febriawan, "Analisis Sentimen Terkait Hilirisasi Industri Pada Opini Masyarakat X dengan Menggunakan Naive Bayes," *J. Inf. Syst. Res.*, vol. 6, no. 2, hal. 1444–1453, 2025, doi: 10.47065/josh.v6i2.6811.

[10] F. Destiyanti, A. I. Hadiana, dan F. R. Umbara, "Penerapan Metode Support Vector Machine dan SMOTE untuk Klasifikasi Sentimen Publik Terhadap Polisi Republik Indonesia," *Jumanji*, vol. 8, no. 1, hal. 1–15, 2024.

[11] A. Miftahusalam, A. F. Nuraini, A. A. Khoirunisa, dan H. Pratiwi, "Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine Pada Analisis Sentimen Twitter Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, hal. 563–572, 2022, doi: 10.34123/semnasoffstat.v2022i1.1410.

[12] K. Marzuki, L. G. Rady Putra, H. Hairani, L. Z. A. Mardedi, dan J. X. Guterres, "Performance Improvement of The Random Forest Method Based on Smote-Tomek Link on Lombok Tourism Analysis Sentiment," *J. Bumigora Inf. Technol.*, vol. 5, no. 2, hal. 151–158, 2024, doi: 10.30812/bite.v5i2.3166.

[13] A. Nurhopipah dan C. Magnolia, "Perbandingan Metode Resampling pada Imbalanced Dataset untuk Klasifikasi Komentar Program MBKM," *JUPIKOM (Jurnal Publ. Ilmu Komput. dan Multimedia)*, vol. 1, no. 2, hal. 9–22, 2022.

[14] A. A. Nurrahman, M. Mauladi, dan A. Rahman, "Analisis Sentimen Masyarakat terhadap Kenaikan Harga Bahan Bakar Minyak Menggunakan Support Vector Machine dan SMOTE," *J. Tek. Inform.*, vol. 4, no. 2, hal. 396–410, 2025, doi: https://doi.org/10.56211/sudo.v4i2.908.

[15] P. W. S. Aji, Suprianto, dan R. Dijaya, "Prediksi Penyakit Stroke Menggunakan Metode Random Forest." *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)* 4.4 (2023): 916-924.

[16] E. Tohidi, R. P. Herdiansyah, E. Wahyudin, dan Kaslani, "Analisa sentimen komentar video YouTube di channel TVOneNews tentang calon presiden Prabowo Subianto menggunakan Support Vector Machine," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, hal. 660–667, 2024.

[17] Y. Findawati, U. Indahyanti, Y. Rahmawati, dan R. Puspitasari, "Sentiment Analysis of Potential Presidential Candidates 2024: A Twitter-Based Study," *Acad. Open*, vol. 8, no. 1, hal. 1–17, 2023, doi: 10.21070/acopen.8.2023.7138.

[18] H. Bichri, A. Chergui, dan M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *IJACSA*, vol. 15, no. 2, hal. 331–339, 2024.

[19] B. W. Rauf, "Sentimen Analisis Pertambangan Di Konawe Utara Dengan Metode Naïve Bayes," *Pros. Semin. Nas. Pemanfaat. Sains dan Teknol. Inf.*, vol. 1, no. 1, hal. 97–102, 2023.