**Jurnal Infortech**
Volume 7 No. 2 Desember 2025
E-ISSN: 2715-8160

# Penerapan *Naïve Bayes* Berbasis *SMOTE* Dan *Decision Tree* Untuk Analisis Sentimen Timnas Indonesia

**Ghudzamir 'Ammar Ibrahim Pasaribu [1], Ghofar Taufik[2], Giatika Chrisnawati[3], Indarti[4], Dewi Laraswati[5], Dinda Ayu Muthia[6]**

[1,2,3,4,5,6]Universitas Bina Sarana Informatika
e-mail: [1]ghudzamirammar.as@gmail.com, [2]ghofar.gft@bsi.ac.id, [3]giatika.gcw@bsi.ac.id, [4]indarti.ini@bsi.ac.id, [5]dewi.dwl@bsi.ac.id, [6]dinda.dam@bsi.ac.id

| Diterima | Direvisi | Disetujui |
|---|---|---|
| 14-10-2025 | 17-11-2025 | 22-12-2025 |

**Abstrak** - Penelitian ini bertujuan menganalisis persepsi publik terhadap kinerja Tim Nasional Indonesia pada masa kepelatihan Patrick Kluivert melalui analisis sentimen komentar penggemar di Instagram. Media sosial dipilih karena menjadi wadah utama bagi penggemar untuk menyampaikan opini dan dukungan secara langsung. Sebanyak 300 komentar dikumpulkan dari akun resmi Timnas Indonesia menggunakan teknik data scraping. Data tersebut kemudian melalui proses pembersihan dan pelabelan sebelum dianalisis. Untuk menyeimbangkan distribusi data, digunakan metode *Synthetic Minority Over-sampling Technique (SMOTE)*. Klasifikasi sentimen dilakukan dengan membandingkan dua algoritma machine learning, yaitu *Naïve Bayes* dan *Decision Tree*, guna menilai kinerja masing-masing dalam mengenali sentimen positif maupun negatif. Hasil penelitian menunjukkan bahwa model *Decision Tree* memiliki performa lebih baik, khususnya dengan nilai recall tinggi pada kelas positif, yang menunjukkan kemampuannya dalam mendeteksi sentimen positif secara akurat. Meskipun demikian, model ini masih lemah dalam mengidentifikasi sentimen negatif. Secara keseluruhan, temuan ini menggambarkan antusiasme dan kecintaan besar penggemar terhadap Timnas Indonesia, terlepas dari hasil pertandingan yang diperoleh.

Kata Kunci: Analisis Sentimen, *Naive Bayes, Decision Tree, SMOTE*

***Abstract*** *- This study aims to analyze public perception of the Indonesian National Team's performance during Patrick Kluivert's coaching tenure through sentiment analysis of fan comments on Instagram. Social media was chosen because it is the primary platform for fans to express opinions and support directly. A total of 300 comments were collected from the official Indonesian National Team account using data scraping techniques. The data was then cleaned and labeled before analysis. To balance the data distribution, the Synthetic Minority Oversampling Technique (SMOTE) was used. Sentiment classification was performed by comparing two machine learning algorithms, Naïve Bayes and Decision Tree, to assess their respective performance in recognizing positive and negative sentiment. The results showed that the Decision Tree model performed better, particularly with a high recall value in the positive class, indicating its ability to accurately detect positive sentiment. However, this model was still weak in identifying negative sentiment. Overall, these findings illustrate the great enthusiasm and love fans have for the Indonesian National Team, regardless of match results.*

*Keywords: Sentiment Analysis, Naive Bayes, Decision Tree, SMOTE*

## PENDAHULUAN

Media sosial telah berkembang menjadi salah satu medium komunikasi paling berpengaruh di era digital, menyediakan ruang bagi miliaran pengguna untuk berbagi informasi, berdiskusi, dan mengekspresikan opini terhadap berbagai isu, termasuk dunia olahraga. Keberadaannya melahirkan aliran data teks dalam jumlah besar yang mencerminkan opini dan emosi publik secara *real-time*. Instagram, sebagai salah satu *platform* dengan pengguna aktif terbanyak di Indonesia, menjadi ruang interaksi penting antara penggemar sepak bola dan Tim Nasional (Timnas) Indonesia. Melalui kolom komentar, para penggemar menyalurkan dukungan, kritik, maupun kekecewaan mereka terhadap performa tim. Aktivitas ini menciptakan data berharga yang dapat dianalisis untuk memahami pola pikir dan sentimen masyarakat terhadap tim kesayangan mereka.

Analisis sentimen terhadap komentar penggemar menjadi langkah strategis untuk memahami dinamika opini publik. Dalam konteks sepak bola Indonesia, Timnas memiliki posisi yang istimewa dan sering menjadi simbol kebanggaan nasional. Oleh karena itu, setiap kemenangan, kekalahan, atau pergantian pelatih selalu memicu gelombang emosi yang intens di media sosial.

Komentar yang muncul pun beragam dari apresiasi hingga kritik tajam yang menunjukkan perlunya pendekatan analitik untuk menata dan menafsirkan data tersebut secara sistematis. Analisis semacam ini tidak hanya memberikan gambaran objektif tentang opini publik, tetapi juga dapat menjadi masukan berharga bagi pihak manajemen tim dalam merancang strategi komunikasi dan menjaga hubungan baik dengan pendukung.

Kajian sebelumnya menunjukkan bahwa penerapan algoritma *Machine Learning* telah memberikan hasil yang signifikan dalam analisis sentiment (Hasibuan & Allistair, 2022). Penelitian sebelumnya menunjukkan efektivitas algoritma *Naïve Bayes* dalam mengklasifikasikan sentimen dengan akurasi tinggi (Maulana, Fahmi, Imran, & Hidayati, 2024). Ini adalah algoritma pembelajaran tanpa usaha yang memanfaatkan aturan Bayes bersama dengan premis atau asumsi tinggi yang karakteristiknya bergantung pada kemandirian yang disediakan oleh kelas (Anand et al., 2022).

Namun, performa algoritma berbeda tergantung pada karakteristik data dan konteks media sosial yang digunakan. Karena itu, penelitian ini berupaya memperluas kajian tersebut dengan menerapkan analisis serupa pada platform *Instagram*, yang memiliki pola interaksi dan gaya bahasa berbeda dari *Twitter*.

Salah satu tantangan utama dalam analisis sentimen adalah ketidakseimbangan data *(imbalanced dataset)*, di mana jumlah komentar positif dan negatif tidak seimbang. Bias ini sering terjadi, misalnya ketika kemenangan menghasilkan lebih banyak sentimen positif. Kondisi ini berpotensi membuat model pembelajaran mesin cenderung mengabaikan kelas minoritas (sentimen negatif). Untuk mengatasi hal tersebut, penelitian ini menggunakan metode *SMOTE (Synthetic Minority Over-sampling Technique)*, yaitu teknik yang secara sintetis menambah jumlah data minoritas agar distribusi kelas menjadi seimbang (Cahyaningtyas, Nataliani, & Widiasari, 2021). Teknik ini bekerja dengan meningkatkan jumlah contoh data pada kelas minoritas melalui pembangkitan titik data sintetis dari tetangga terdekatnya menggunakan jarak *Euclidean* (Karamti et al., 2023). Dengan demikian, model dapat belajar lebih efektif dari kedua jenis sentimen.

Penelitian ini juga membandingkan dua algoritma klasifikasi teks yang populer, yaitu *Naïve Bayes* dan *Decision Tree*. *Naïve Bayes* dikenal karena kesederhanaannya dan kemampuannya menangani data teks berdimensi tinggi, dengan prinsip dasar pada teorema *Bayes* yang mengasumsikan independensi antar fitur (Wahyuni, 2022). Walaupun asumsi tersebut jarang sepenuhnya terpenuhi, model ini tetap menunjukkan performa baik pada berbagai studi sebelumnya. Sementara itu, *Decision Tree* bekerja dengan membangun struktur hierarkis berbasis keputusan, membagi data menurut fitur paling informatif, dan menghasilkan model yang mudah diinterpretasi (Fatkhudin, Artanto, Safli, & Wibowo, 2024). Keunggulan *Decision Tree* terletak pada transparansinya, meskipun memiliki risiko *overfitting* jika tidak dikontrol dengan baik (Khasanah, Salim, Afni, Komarudin, & Maulana, 2022).

Dengan membandingkan kedua algoritma tersebut, penelitian ini berupaya tidak hanya menentukan model dengan tingkat akurasi terbaik, tetapi juga memahami pola dan karakteristik data sentimen penggemar Timnas Indonesia di Instagram. Fokus penelitian pada era kepelatihan Patrick Kluivert (sebelum dipecat) memberikan konteks yang menarik, mengingat masa ini menandai fase transisi dan pembentukan gaya permainan baru. Setiap perubahan dalam strategi atau hasil pertandingan di bawah kepemimpinannya memicu reaksi beragam dari penggemar, yang terekam melalui komentar-komentar di media negati. Analisis terhadap periode ini dapat menggambarkan bagaimana persepsi penggemar terbentuk terhadap keputusan pelatih dan hasil pertandingan, sekaligus memberikan masukan empiris bagi federasi dan manajemen Timnas dalam evaluasi kebijakan.

Rumusan masalah utama dari penelitian ini adalah bagaimana analisis sentimen terhadap komentar Instagram dapat digunakan untuk menilai persepsi publik terhadap performa Timnas Indonesia selama masa kepelatihan Patrick Kluivert, dengan menerapkan algoritma *Naïve Bayes* dan *Decision Tree* yang dioptimalkan menggunakan *SMOTE*. Pendekatan ini diharapkan mampu menghasilkan model klasifikasi yang tidak hanya akurat, tetapi juga mampu menangkap pola emosional publik secara representatif.

Nilai kebaruan dari penelitian ini terletak pada kombinasi tiga aspek utama: (1) penggunaan data dari *platform Instagram* yang memiliki dinamika komunikasi visual dan teks berbeda dari media sosial lain; (2) penerapan teknik *SMOTE* untuk mengatasi ketidakseimbangan data sentimen yang sering diabaikan dalam studi sebelumnya; dan (3) perbandingan langsung antara dua algoritma klasik *machine learning* untuk menentukan pendekatan paling efektif dalam konteks analisis opini publik terhadap olahraga nasional.

Secara teoretis, penelitian ini memperkaya literatur mengenai penerapan *machine learning* dalam bidang analisis sentimen olahraga, khususnya di Indonesia. Secara praktis, hasilnya diharapkan memberikan wawasan bagi pengelola Timnas dan pemangku kebijakan dalam memahami persepsi publik serta menyusun strategi komunikasi yang lebih adaptif. Secara akademis, penelitian ini dapat menjadi referensi bagi studi-studi lanjutan yang berfokus pada hubungan antara performa olahraga dan emosi penggemar di media sosial.

Dengan demikian, penelitian ini tidak hanya berfokus pada pemrosesan data tekstual, tetapi juga pada upaya memahami dimensi sosial dan emosional dari dukungan penggemar. Melalui pendekatan analisis sentimen berbasis *machine learning*, penelitian ini berupaya menangkap dinamika hubungan antara Timnas Indonesia dan para pendukungnya, sekaligus memperlihatkan bahwa

kecintaan terhadap tim nasional tidak semata bergantung pada kemenangan, melainkan juga pada rasa memiliki dan semangat kebersamaan yang terbangun di ranah digital.

## METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksploratif dan eksperimen yang negat pada analisis sentimen pengguna *Instagram* terhadap performa Timnas Indonesia pada era kepelatihan Patrick Kluivert. Pendekatan ini dipilih karena mampu mengidentifikasi pola negatif dari data teks dalam jumlah besar serta mengukur tingkat akurasi dari model klasifikasi yang digunakan. Untuk mencapai tujuan tersebut, penelitian ini dilakukan melalui serangkaian proses yang sistematis mulai dari pengumpulan data hingga evaluasi performa model klasifikasi.

Adapun langkah-langkah penelitian yang dilaksanakan adalah sebagai berikut:

1. Identifikasi Masalah dan Studi Literatur
   Tahap awal dimulai dengan mengidentifikasi permasalahan yang relevan, yaitu perlunya pemahaman terhadap persepsi negatif atas performa Timnas Indonesia. Selanjutnya dilakukan kajian pustaka terhadap penelitian terdahulu yang berkaitan dengan analisis sentimen, algoritma *Naïve Bayes* dan *Decision Tree*, serta teknik *SMOTE*, guna memperkuat landasan teori dan metodologi yang akan digunakan (Cahyono, Informatika, & Yogyakarta, 2024).
2. Pengumpulan Data *(Data Acquisition)*
   Data komentar diambil dari platform media negati Instagram yang relevan dengan Timnas Indonesia pada era kepelatihan Patrick Kluivert. Proses pengambilan data dilakukan dengan menggunakan *tools web scraping* seperti *Python* (dengan *library Selenium* atau *BeautifulSoup*) atau layanan seperti *Phantombuster*. Data dikumpulkan berdasarkan *hashtag, mention,* maupun akun resmi yang memuat konten pertandingan, wawancara, atau berita tim nasional. Pengumpulan dilakukan dengan memperhatikan aspek etika digital dan privasi publik.
3. Pra-pemrosesan Data *(Preprocessing)*
   Data mentah yang diperoleh selanjutnya melalui tahapan *preprocessing* untuk meningkatkan kualitas data teks agar layak untuk dianalisis (Ginantra, Yanti, Prasetya, Sarasvananda, & Wiguna, 2022). Proses ini mencakup beberapa langkah:
   a. *Cleaning*: Menghapus karakter non-alfabet, *URL*, *emoticon*, negati, dan angka yang tidak relevan (Kurnia, Purnamasari, & Saputra, 2023).
   b. *Case Folding*: Mengubah semua huruf menjadi huruf kecil untuk standarisasi (Novitasari et al., 2022).
   c. *Tokenizing*: Memecah kalimat menjadi potongan kata (token) (Sriani, Lubis, & Harahap, 2023).
   d. *Stopword Removal*: Menghilangkan kata-kata umum yang tidak memberikan makna penting dalam analisis (seperti “yang”, “di”, “ke”, dll.).
   e. *Stemming*: Mengubah kata ke bentuk dasarnya, seperti “bermain” menjadi “main”, “dilatih” menjadi “latih”.
4) *Labelling* Sentimen
   Setelah *preprocessing*, komentar diklasifikasikan secara manual atau semi-otomatis ke dalam dua kelas sentimen utama, yaitu positif dan negatif. Penentuan ini didasarkan pada kosakata yang mengandung muatan emosi, intensi kalimat, serta konteks penggunaannya. *Labeling* sangat penting sebagai dasar pembelajaran algoritma dalam tahap *supervised learning* (Ningsih, Alfianda, Rahmaddeni, & Wulandari, 2024).
5. Pembagian Data *(Data Splitting)*
   Data yang telah diberi label kemudian dibagi menjadi dua bagian, yaitu *data training* (80%) dan *data testing* (20%). Pembagian ini bertujuan agar model dapat belajar dari data latih dan diuji pada data uji untuk mengukur kemampuan generalisasi model terhadap data baru.
6. Penyeimbangan Data dengan *SMOTE*
   Jika terjadi ketidakseimbangan kelas (jumlah komentar positif jauh lebih banyak atau lebih sedikit dari negatif), maka digunakan negati *SMOTE (Synthetic Minority Over-sampling Technique)*. *SMOTE* bekerja dengan cara menghasilkan data sintetis untuk kelas minoritas berdasarkan kemiripan data eksisting (Iskandar & Nataliani, 2021), sehingga distribusi kelas menjadi lebih seimbang. Hal ini bertujuan untuk menghindari bias model terhadap kelas mayoritas saat proses pelatihan.
7. Klasifikasi Sentimen dengan Algoritma *Machine Learning*
   Pada tahap ini, data yang telah seimbang dan dibagi digunakan untuk pelatihan dua model klasifikasi:
   a. *Naïve Bayes* : Algoritma probabilistik yang mengandalkan teorema *Bayes* untuk memprediksi kemungkinan suatu kelas berdasarkan fitur teks yang diamati (Fatkhudin et al., 2024).
   b. *Decision Tree* : Algoritma klasifikasi berbasis struktur pohon keputusan yang membagi data berdasarkan atribut dengan gain informasi tertinggi (Cahyaningtyas et al., 2021).
   Kedua model diuji untuk mengetahui performa dalam mengklasifikasikan komentar kedalam sentimen positif atau negatif.
8. Evaluasi Model
   Setelah klasifikasi, dilakukan evaluasi performa masing-masing model menggunakan metrik pengujian sebagai berikut:
   a. Akurasi: Proporsi prediksi yang benar terhadap total prediksi.
   b. Presisi: Tingkat ketepatan model dalam memprediksi sentimen tertentu.
   c. *Recall*: Kemampuan model dalam

menemukan semua data yang relevan dari satu kategori.

d. *F1-Score*: Rata-rata harmonis antara presisi dan *recall*.

Evaluasi ini digunakan untuk mengetahui model mana yang lebih unggul dan seberapa baik model dapat mengklasifikasikan sentimen.

9. Analisis Hasil dan Penarikan Kesimpulan
Tahap akhir melibatkan analisis terhadap hasil evaluasi dan membandingkan performa kedua algoritma. Penulis menarik kesimpulan terkait efektivitas model, implikasi hasil analisis terhadap persepsi negatif, serta memberikan saran pemanfaatan hasil penelitian baik secara teoritis maupun praktis.

## HASIL DAN PEMBAHASAN

Hasil penelitian ini mencakup beberapa tahap, mulai dari proses *scrapping data*, pembersihan, pelabelan, hingga klasifikasi menggunakan algoritma *Naïve Bayes* dan *Decision Tree*. Tahap awal adalah mengumpulkan 300 data komentar dari Instagram yang kemudian diolah melalui serangkaian tahapan yang ketat untuk memastikan kualitas data. Proses data *scraping* dilakukan untuk mendapatkan data mentah berupa teks dari komentar-komentar pengguna Instagram terkait performa Timnas Indonesia di era kepelatihan Patrick Kluivert. Data mentah ini kemudian dilanjutkan ke tahap *cleaning* dan *labeling* untuk membersihkan data dari simbol khusus dan memberikan label sentimen positif atau negatif secara manual. Setelah proses ini, data siap untuk diproses lebih lanjut menggunakan *RapidMiner*.

### 1. *WordList to Data*

*WordList to Data* menyajikan kata apa yang banyak muncul dikomentar instagram.

| Row No. | word | in documents | total | in class (po... | in class (ne... |
|---|---|---|---|---|---|
| 1 | pemain | 19 | 26 | 17 | 9 |
| 2 | timnas | 23 | 24 | 20 | 4 |
| 3 | indonesia | 18 | 18 | 14 | 4 |
| 4 | pelatih | 13 | 16 | 14 | 2 |
| 5 | garuda | 11 | 12 | 11 | 1 |
| 6 | kalo | 8 | 11 | 10 | 1 |
| 7 | towel | 10 | 11 | 2 | 9 |
| 8 | kalah | 10 | 10 | 3 | 7 |
| 9 | fifa | 8 | 9 | 5 | 4 |
| 10 | masuk | 8 | 9 | 5 | 4 |
| 11 | rank | 9 | 9 | 3 | 6 |
| 12 | turun | 9 | 9 | 0 | 9 |

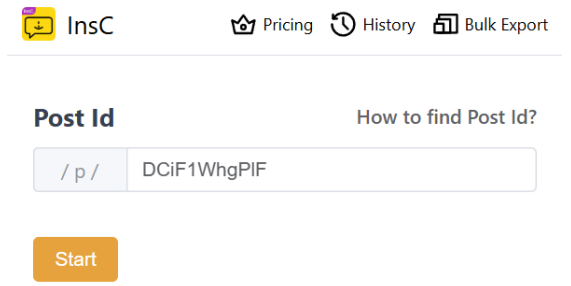
Sumber: Hasil Penelitian Tahun 2025 *(RapidMiner)*
Gambar 1. Hasil *WordList to Data*

Dari hasil *WordList to Data* pada gambar 3 di atas terdapat kata yang paling banyak muncul pada data, seperti “pemain”, “timnas”, “indonesia”, dan lain-lain. Untuk memvisualisasikan kata apa yang paling banyak muncul dari gambar 2 dapat menggunakan *Plot Type Wordcloud* pada menu *Visualizations* di *operators Wordlist to Data.*


Sumber: Hasil Penelitian Tahun 2025 *(RapidMiner)*
Gambar 2. Hasil dari *Wordcloud*

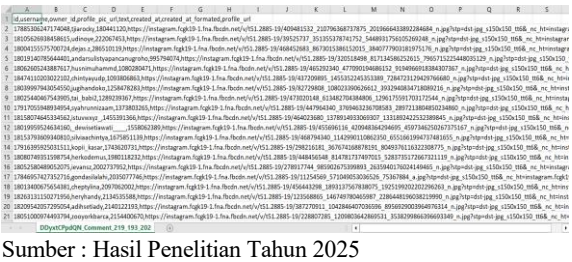
### 2. *Scrapping Data*

*Scrapping data* adalah sebuah negati pengumpulan data dengan menggunakan ekstensi *Google Chrome*, yaitu *InsC – Instagram Comment. InsC – Instagram Comment* merupakan salah satu *Web Scrapper* yang berguna untuk melakukan proses pengambilan data secara efisien. *Web Scrapping* dilakukan dengan cara *copy link* dari postingan yang ingin diambil data komentarnya, dari hasil *scrapping* data didapatkan 230 data. Setelah itu melakukan *export*, format data yang telah diunduh berupa *CSV*.




Sumber: Hasil Penelitian Tahun 2025
Gambar 3. *Google Chrome InsC – Instagram Comment*

Hasil komentar *Instagram* yang sudah melalui *scrapping* dapat diunduh yang hasilnya dapat dilihat pada gambar berikut:



Sumber : Hasil Penelitian Tahun 2025
Gambar 4. File Hasil Unduh *(.csv)* dari Proses *Scrapping*

### 3. *Cleaning* dan *Labeling*

Untuk tahap selanjutnya dilakukan *Cleaning* dan *Labeling*, pada tahap *Cleaning* akan dilakukan penghapusan terhadap simbol, *emoticon* dan *mention*.

Tabel 1. Hasil Proses *Cleaning*

| Sebelum | Sesudah |
|---|---|
| thanks coach Patrickkluivert9 thx pak @erickthohir ♡ 🙏 | thanks coach thx pak |

Sumber : Hasil Penelitian Tahun 2025

Selanjutnya, dilakukan pelabelan secara manual, dimana setiap teks diberikan label sentimen positif dan negatif.

Tabel 2. Hasil Proses *Labeling*

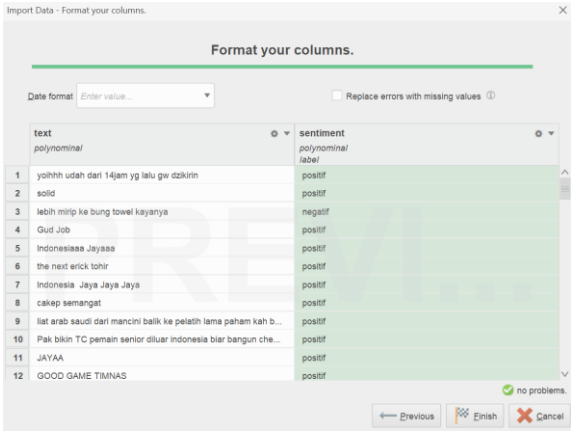
| Teks | Sentimen |
|---|---|
| thanks coach thx pak | Positif |
| Ganti Patrick Udah kebanyakan ngopi dianya | Negatif |

Sumber : Hasil Penelitian Tahun 2025

Hasil dari proses *labelling* didapatkan 136 data positif dan 94 data negatif.

**4. Pra Pemrosesan Data**

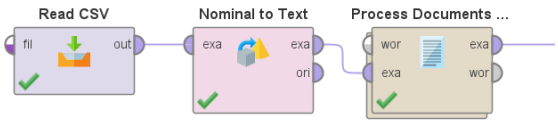
Pra-pemrosesan dilakukan untuk menyiapkan data sebelum analisis sentimen. Pra-pemrosesan dilakukan menggunakan *RapidMiner*. Data yang sudah di *Cleaning* dan *Labeling*, dimasukkan pada *software RapidMiner* dengan *Operators Read CSV*.




Sumber: Hasil Penelitian Tahun 2025

Gambar 5. *Labeling* pada Kolom Sentimen *(RapidMiner)*

Setelah memberikan label pada kolom sentiment, tahap selanjutnya adalah mengubah atribut *polynominal* menjadi atribut *text* menggunakan *Operators Nominal to Text* dan dilakukan pemrosesan dokumen menggunakan *Operators Process Documents* pada *RapidMiner*.

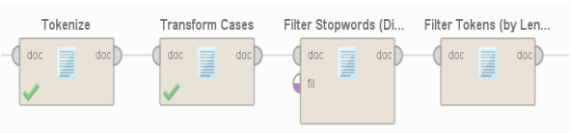



Sumber: Hasil Penelitian Tahun 2025

Gambar 6. Proses Dokumen Pada *RapidMiner*

Didalam *process document* terdiri dari berbagai *subprocess*, yaitu:

1) *Tokenize*
   Proses memecah teks menjadi bagian kecil yang disebut token. Dengan memisahkan teks menjadi elemen-elemen yang lebih *manageable, tokenize* memungkinkan algoritma untuk menganalisis, mengklasifikasikan, dan mengolah informasi dengan lebih efisien, menjadikan pemahaman bahasa menjadi lebih terstruktur dan sistematis. Seperti contoh kalimat: "Ganti Patrick Udah kebanyakan ngopi dianya".
   Menjadi: "Ganti", "Patrick", "Udah", "kebanyakan", "ngopi", "dianya".
2) *Transform Cases*
   Proses mengubah format huruf dalam teks, seperti mengalihkannya menjadi huruf kapital semua, huruf kecil semua, atau hanya huruf pertama yang kapital. Seperti contoh kalimat : "Ganti", "Patrick", "Udah", "kebanyakan", "ngopi", "dianya".
   Menjadi : "ganti Patrick udah kebanyakan ngopi dianya".
3) *Filter Stopwords (Dictionary)*
   Berfungsi untuk menghapus kata-kata yang tidak relevan, proses ini menggunakan *list* bahasa Indonesia yang diambil dari web *Kaggle.com*.
   Seperti contoh kalimat : "yoihhh udah dari 14jam yg lalu gw dzikirin".
   Menjadi : "yoihhh udah jam yg gw dzikirin"
4) *Filter Tokens (by Length)*
   Proses menyaring kata-kata, kita menghapus yang memiliki terlalu sedikit atau terlalu banyak karakter, dengan batasan default minimal 4 karakter dan maksimal 25 karakter.
   Seperti contoh kalimat : "yoihhh udah jam yg gw dzikirin".
   Menjadi : "yoihhh udah dzikirin"




Sumber: Hasil Penelitian Tahun 2025

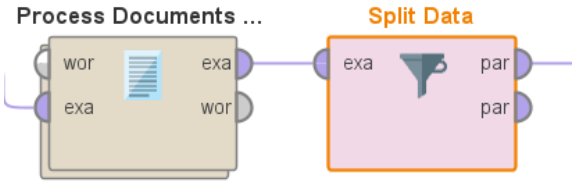
Gambar 7. *Process Documents From Data (RapidMiner)*

Proses didalam *Operators Process Documents* yang berisi *Operators Tokenize, Transform Cases, Filter Stopwords (Dictionary)* dan *Filter Tokens (by Length)* Setelah menyusun model seperti yang dijelaskan di atas dalam *RapidMiner*, menghasilkan kalimat yang lebih sederhana. Setelah pra-pemrosesan data selesai, langkah selanjutnya adalah pembagian dataset.

**5. Pembagian Dataset *(Data Split)***

Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 70:30, 80:20, 90:10. Data

pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk evaluasi model. Pembagian dataset dilakukan dengan menggunakan operator *Split Data*.

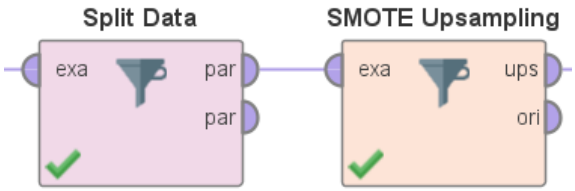



Sumber: Hasil Penelitian Tahun 2025
Gambar 8. *Split Data (RapidMiner)*

*Operators Split Data* untuk membagi data yang sudah melalui Proses *Documents Data*, pembagian data set dengan ratio 0.7 (70%) dan 0.3 (30%).

**6. Penanganan Ketidakseimbangan Kelas**

Untuk mengatasi ketidakseimbangan kelas, metode *SMOTE* diterapkan. Hasil penerapan *SMOTE* menunjukkan peningkatan jumlah data pada kelas minoritas, sehingga proporsi data menjadi lebih seimbang.



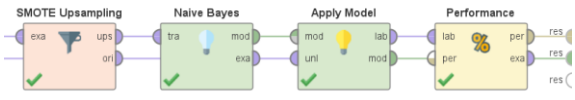

Sumber: Hasil Penelitian Tahun 2025
Gambar 9. *SMOTE Upsampling (RapidMiner)*

Dilakukan penyeimbangan kelas menggunakan *Operators SMOTE Upsampling,* agar terjadi keseimbangan kelas antara label positif dan negatif.

**7. Klasifikasi Naïve Bayes dan Decision Tree**

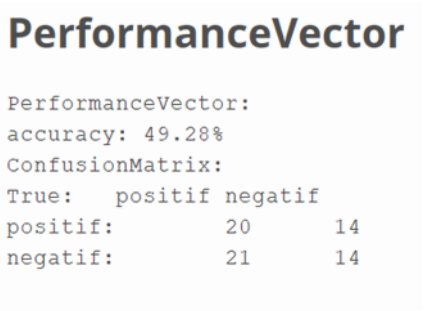
Dua model klasifikasi yang diandalkan adalah *Naïve Bayes* dan *Decision Tree*, masing-masing dengan pendekatan yang khas.

1) *Naïve Bayes*
Gunakan *operators Naïve Bayes, Apply Model* dan *Performance* pada *RapidMiner* untuk mendapatkan hasil klasifikasi *Naïve Bayes*.




Sumber: Hasil Penelitian Tahun 2025
Gambar 10. *Operators Naïve Bayes, Apply Model dan Performance*

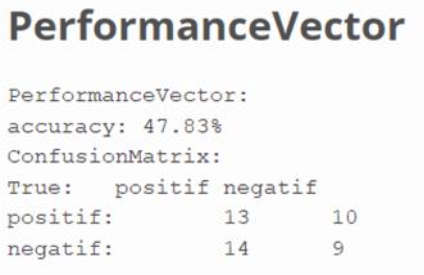
Kemudian dilakukan pengujian akurasi :
a. Pengujian akurasi perbandingan 70:30




Sumber: Hasil Penelitian Tahun 2025
Gambar 11. *Performance Naive Bayes ratio* 70:30

Model memiliki akurasi sebesar 49.28%. Ini menunjukkan bahwa model hanya benar dalam memprediksi sekitar setengah dari keseluruhan data. Terdapat 20 prediksi positif yang benar *(true positives)*. Terdapat 14 prediksi negatif yang benar *(true negatives)*. *Precision* untuk kelas positif adalah 58.82%, yang menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 58.82% adalah benar. *Recall* untuk kelas positif adalah 48.78%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 48.78% di antaranya. Dari 21 prediksi negatif, hanya 14 yang benar. Ini menghasilkan *precision* untuk kelas negatif sebesar 40.00%. Secara keseluruhan, hasil ini menunjukkan bahwa model *Naïve Bayes* memiliki performa yang cukup rendah.

b. Pengujian akurasi perbandingan 80:20




Sumber: Hasil Penelitian Tahun 2025
Gambar 12. *Performance Naive Bayes ratio* 80:20

Model memiliki akurasi sebesar 47.83%. Ini menunjukkan bahwa model hanya benar dalam memprediksi sekitar setengah dari keseluruhan data. Terdapat 13 prediksi positif yang benar *(true positives)*. Terdapat 10 prediksi negatif yang benar *(true negatives)*. *Precision* untuk kelas positif adalah 56.52%, yang menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 56.52% adalah benar. *Recall* untuk kelas positif adalah 48.15%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 48.15% di antaranya. Dari 14 prediksi negatif, hanya 9 yang benar. Ini menghasilkan *precision* untuk kelas negatif sebesar 39.13%. Secara keseluruhan, hasil ini menunjukkan bahwa model

*Naïve Bayes* memiliki performa yang lebih rendah dari rasio sebelumnya.

c. Pengujian akurasi perbandingan 90:10

**PerformanceVector**

PerformanceVector:
accuracy: 39.13%
ConfusionMatrix:

| True: | positif | negatif |
|---|---|---|
| positif: | 5 | 5 |
| negatif: | 9 | 4 |

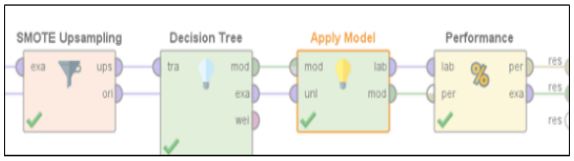
Sumber: Hasil Penelitian Tahun 2025

Gambar 13. *Performance Naive Bayes ratio* 90:10

Model memiliki akurasi sebesar 39.13%. Ini menunjukkan bahwa model hanya benar dalam memprediksi sekitar setengah dari keseluruhan data. Terdapat 5 prediksi positif yang benar *(true positives)*. Terdapat 5 prediksi negatif yang benar (*true negatives*). *Precision* untuk kelas positif adalah 50.00%, yang menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 50.00% adalah benar. *Recall* untuk kelas positif adalah 35.71%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 35.71% di antaranya. Dari 9 prediksi negatif, hanya 4 yang benar. Ini menghasilkan *precision* untuk kelas negatif sebesar 30.77%. Secara keseluruhan, hasil ini menunjukkan bahwa model *Naïve Bayes* memiliki performa yang lebih rendah dari 2 rasio sebelumnya.

2) *Decision Tree*

Begitupun dengan *Decision Tree,* menggunakan *operators Decision Tree, Apply Model* dan *Performance.*




Sumber: Hasil Penelitian Tahun 2025

Gambar 14. *Operators Decision Tree, Apply Model dan Performance*

Kemudian dilakukan pengujian akurasi yang sama dengan pengujian *Naïve Bayes* sebelumnya, yaitu :

a. Pengujian akurasi perbandingan rasio 70:30

**PerformanceVector**

PerformanceVector:
accuracy: 63.77%
ConfusionMatrix:

| True: | positif | negatif |
|---|---|---|
| positif: | 37 | 21 |
| negatif: | 4 | 7 |

Sumber: Hasil Penelitian Tahun 2025

Gambar 15. *Performance Decision Tree ratio* 70:30

*PerformanceVector* dari *Operators Decision Tree* memiliki akurasi sebesar 63.77%, yang menunjukkan bahwa model ini benar dalam memprediksi sekitar 63.77% dari keseluruhan data. Terdapat 37 prediksi positif yang benar *(true positives).* Terdapat 21 prediksi negatif yang benar *(true negatives)*. *Precision* untuk kelas positif adalah 63.79%, menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 63.79% adalah benar. *Recall* untuk kelas positif adalah 90.24%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 90.24% di antaranya. Dari 4 prediksi negatif, terdapat 7 yang benar *(true negatives)*, dengan *precision* untuk kelas negatif sebesar 63.64%.

b. Pengujian akurasi perbandingan rasio 80:20

**PerformanceVector**

PerformanceVector:
accuracy: 60.87%
ConfusionMatrix:

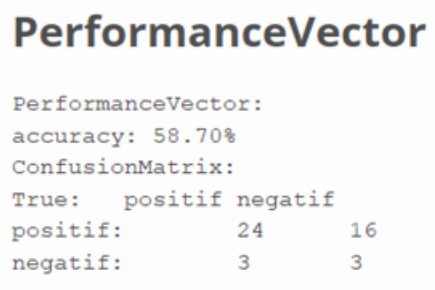
| True: | positif | negatif |
|---|---|---|
| positif: | 12 | 7 |
| negatif: | 2 | 2 |

Sumber: Hasil Penelitian Tahun 2025

Gambar 16. *Performance Decision Tree ratio* 80:20

Model memiliki akurasi sebesar 58.70%, yang menunjukkan bahwa model ini benar dalam memprediksi sekitar 58.70% dari keseluruhan data. Terdapat 24 prediksi positif yang benar *(true positives)*. Terdapat 16 prediksi negatif yang benar (true negatives). *Precision* untuk kelas positif adalah 60.00%, menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 60.00% adalah benar. *Recall* untuk kelas positif adalah 88.89%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 88.89% di antaranya. Dari 3 prediksi negatif, terdapat 3 yang benar (*true negatives*), dengan *precision* untuk kelas negatif sebesar 50.00%.

c. Pengujian akurasi perbandingan rasio 90:10




Sumber: Hasil Penelitian Tahun 2025
Gambar 17. *Performance Decision Tree ratio* 90:10

Memiliki akurasi sebesar 60.87%, yang menunjukkan bahwa model ini benar dalam memprediksi sekitar 60.87% dari keseluruhan data. Terdapat 12 prediksi positif yang benar *(true positives)*. Terdapat 7 prediksi negatif yang benar *(true negatives)*. *Precision* untuk kelas positif adalah 63.16%, menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 63.16% adalah benar. *Recall* untuk kelas positif adalah 85.71%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 85.71% di antaranya. Dari 2 prediksi negatif, terdapat 2 yang benar *(true negatives),* dengan *precision* untuk kelas negatif sebesar 50.00%.

Secara keseluruhan, hasil ini menunjukkan bahwa model *Decision Tree* memiliki performa yang lebih baik dibandingkan model *Naïve Bayes* sebelumnya, terutama dalam hal recall untuk kelas positif yang cukup tinggi. Namun, recall untuk kelas negatif hanya 25%, yang menunjukkan bahwa model ini mungkin kesulitan mengenali beberapa contoh negatif.

**8. Hasil Klasifikasi**

Setelah tahap analisis selesai, diperoleh nilai-nilai yang mencerminkan pola dan karakteristik dari data. Evaluasi yang dilakukan menghasilkan metrik performa seperti tingkat akurasi, presisi, dan *recall* yang membantu dalam memahami cara yang lebih efektif untuk mengklasifikasikan data.

Berikut hasil klasifikasi akurasi sentimen dari model *Naïve Bayes* dapat dilihat berikut ini :

**accuracy: 49.28%**

| | true positif | true negatif | class precision |
|---|---|---|---|
| pred. Positif | 20 | 14 | 58.82% |
| pred. Negatif | 21 | 14 | 40.00% |
| class recall | 48.78% | 50.00% | |

Sumber: Hasil Penelitian Tahun 2025 (RapidMiner)
Gambar 18. *Akurasi Naïve Bayes*

Sedangkan untuk hasil klasifikasi akurasi untuk model *Decision Tree* adalah :

**accuracy: 63.77%**

| | true positif | true negatif | class precision |
|---|---|---|---|
| pred. Positif | 37 | 21 | 63.79% |
| pred. Negatif | 4 | 7 | 63.67% |
| class recall | 90.24% | 25.00% | |

Sumber: Hasil Penelitian Tahun 2025 *(RapidMiner)*
Gambar 18. Akurasi *Decision Tree*

Model *Decision Tree* memberikan hasil yang lebih baik dalam hal akurasi, *precision*, dan *recall* untuk kelas positif dibandingkan dengan *Naïve Bayes.*

## KESIMPULAN

Dari hasil analisa sentimen menggunakan metode *Naïve Bayes* berbabis *SMOTE* dan *Decision Tree* berbasis *SMOTE* serta menggunakan rasio 70:30, 80:20 dan 90:10 diatas, dapat disimpulkan bahwa penulis mengambil rasio 70:30 sebagai hasil penelitian kali ini ini dikarenakan memiliki nilai akurasi sebesar 49.23% untuk *Naive Bayes* dan *Decision Tree* memiliki nilai akurasi sebesar 63.77%.

*Naïve Bayes*, memiliki akurasi sebesar 49.28%, ini menunjukan bahwa model hanya benar dalam memprediksi sekitar setengah dari keseluruhan data. Terdapat 20 prediksi positif yang benar *(true positives).* Terdapat 14 prediksi negatif yang benar *(true negatives). Precision* untuk kelas positif adalah 58.82%, yang menunjukkan bahwa dari semua prediksi dipositif yang dibuat, sekitar 58.82% adalah benar. *Recall* untuk kelas positif adalah 48.78%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 48.78% di antaranya. Dari 21 prediksi negatif, hanya 14 yang benar. Ini menghasilkan *precision* untuk kelas negatif sebesar 40.00%. secara keseluruhan, hasil ini menunjukan bahwa model *Naïve Bayes* memiliki performa yang cukup rendah.

*Decision Tree,* memiliki akurasi sebesar 63.77%, yang menunjukkan bahwa model ini benar dalam memprediksi sekitar 63.77% dari keseluruhan data. Terdapat 37 prediksi positif yang benar *(true positives).* Terdapat 21 prediksi negatif yang benar *(true negatives). Precision* untuk kelas positif adalah 63.79%, menunjukkan bahwa dari semua prediksi positif yang dibuat, sekitar 63.79% adalah benar. *Recall* untuk kelas positif adalah 90.24%, berarti dari semua kasus positif yang sebenarnya, model ini dapat mengenali sekitar 90.24% di antaranya. Dari 4 prediksi negatif, terdapat 7 yang benar *(true negatives),* dengan precision untuk kelas negatif sebesar 63.64%.

Secara keseluruhan, hasil ini menunjukkan bahwa model *Decision Tree* memiliki performa

yang lebih baik dibandingkan model *Naïve Bayes* sebelumnya, terutama dalam hal *recall* untuk kelas positif yang cukup tinggi. Namun, *recall* untuk kelas negatif hanya 25%, yang menunjukkan bahwa model ini mungkin kesulitan mengenali beberapa contoh negatif.

## REFERENSI

Anand, M. V., Kiranbala, B., Srividhya, S. R., C., K., Younus, M., & Rahman, M. H. (2022). Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/2436946

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *Aiti*, *18*(2), 173–184. https://doi.org/10.24246/aiti.v18i2.173-184

Cahyono, N., Informatika, S., & Yogyakarta, U. A. (2024). Analisis Sentimen Komentar Instagram Pada Program Kampus. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 2372–2381.

Fatkhudin, A., Artanto, F. A., Safli, N. A., & Wibowo, D. (2024). Decision Tree Berbasis SMOTE Dalam Analisis Sentimen Penggunaan Artificial Intelligence Untuk Skripsi. *REMIK: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *8*(April), 494–505. Retrieved from https://www.jurnal.polgan.ac.id/index.php/remik/article/view/13531%0Ahttps://www.jurnal.polgan.ac.id/index.php/remik/article/download/13531/2453

Ginantra, N. L. W. S. R., Yanti, C. P., Prasetya, G. D., Sarasvananda, I. B. G., & Wiguna, I. K. A. G. (2022). Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan K-NN. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *11*(3), 205–215. https://doi.org/10.23887/janapati.v11i3.49450

Hasibuan, E., & Allistair, E. (2022). Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play. *JTS (Jurnal Teknik Dan Science*, *1*(3), 13–24.

Iskandar, J. W., & Nataliani, Y. (2021). Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(6), 1120–1126. https://doi.org/10.29207/resti.v5i6.3588

Karamti, H., Alharthi, R., Anizi, A. Al, Alhebshi, R. M., Eshmawi, A. A., Alsubai, S., & Umer, M. (2023). Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach. *Cancers*, *15*(17). https://doi.org/10.3390/cancers15174412

Khasanah, N., Salim, A., Afni, N., Komarudin, R., & Maulana, Y. I. (2022). Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes. *Technologia : Jurnal Ilmiah*, *13*(3), 207. https://doi.org/10.31602/tji.v13i3.7312

Kurnia, Purnamasari, I., & Saputra, D. D. (2023). Analisis Sentimen Dengan Metode Naïve Bayes, SMOTE Dan Adaboost Pada Twitter Bank BTN. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *7*(2), 235–242. https://doi.org/10.35870/jtik.v7i3.707

Maulana, B. A., Fahmi, M. J., Imran, A. M., & Hidayati, N. (2024). Analisis Sentimen Terhadap Aplikasi Pluang Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 375–384. https://doi.org/10.57152/malcom.v4i2.1206

Ningsih, W., Alfianda, B., Rahmaddeni, & Wulandari, D. (2024). Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 556–562.

Novitasari, I., Basuki Kurniawan, T., Arrova Dewi, D., Ilmu Komputer, F., Bina Darma, U., Selatan, S., … Vokasi, F. (2022). Analysis of Public Sentiment Towards Ruang Guru Tweets Using the Naive Bayes Classifier (NBC) Algorithm / Analisis Sentimen Masyarakat Terhadap Tweet Ruang Guru Menggunakan Algoritma Naive Bayes Classifier (NBC). *Jurnal Mantik*, *6*(3), 3308–3318. Retrieved from https://iocscience.org/ejournal/index.php/mantik/article/view/2887

Sriani, Lubis, A. H., & Harahap, Y. F. (2023). Analisis Sentimen Masyarakat Terhadap Resesi Ekonomi Global 2023 Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Ilmiah Elektronika Dan Komputer*, *16*(2), 442–450. Retrieved from http://journal.stekom.ac.id/index.php/elkom□page442

Wahyuni, W. (2022). Analisis Sentimen terhadap Opini Feminisme Menggunakan Metode Naive Bayes. *Jurnal Informatika Ekonomi Bisnis*, *4*, 148–153. https://doi.org/10.37034/infeb.v4i4.162