**JURNAL FASILKOM**
Volume 15 No. 2 | Agustus 2025: 202-212

P-ISSN : 2089-3353
E-ISSN : 2808-9162

# Analisis Sentimen Publik Terhadap Kampanye Pengurangan Sampah Plastik Menggunakan Algoritma Naïve Bayes

Nanda Dwi Husna Sadikin[1], Sari Susanti[2]
[1,2]Program Studi Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya
[1]nandadwihs@gmail.com *, [2]sarisusanti@ars.ac.id

***Abstract***

*Plastic waste is one of the most alarming sources of environmental pollution in Indonesia. Various campaigns to reduce plastic are intensified, especially through social media X (previously Twitter). This research aims to analyze public sentiment towards the campaign using Naïve Bayes algorithm. The data used are 4,351 Indonesian tweets collected during November 2024-April 2025 with the keywords "plastic waste", "reduce plastic", and "plastic pollution". The analysis process follows the CRISP-DM stages. Data was processed through preprocessing stages such as cleaning, tokenization, and stemming, and then automatically labeled using IndoBERT, a BERT-based pre-trained model specifically designed for sentiment classification in Indonesian. Features were extracted with TF-IDF, and the model was trained using three data sharing scenarios (60:40, 70:30, and 80:20). To overcome the class imbalance found, Random Over Sampling (ROS) technique was applied. ROS works by duplicating data from minority classes to balance the training data distribution, which aims to allow the model to learn more fairly. Evaluation results show that without ROS, the model is unable to recognize the positive sentiment class (recall = 0). However, after ROS was applied, the performance of the model improved significantly, especially for minority classes, with recall increasing to more than 59%. This improvement proves that ROS is effective in improving model performance for sentiment analysis on imbalanced data. This research is expected to contribute to the utilization of machine learning to understand public perception and support the formulation of more effective environmentalpolicies.*

*Keywords: sentiment analysis, naïve bayes, plastic waste, text mining, random over sampling*

**Abstrak**

Sampah plastik menjadi salah satu sumber pencemaran lingkungan yang mengkhawatirkan di Indonesia. Berbagai kampanye pengurangan plastik digencarkan, terutama melalui media sosial X (sebelumnya Twitter). Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap kampanye tersebut menggunakan algoritma *Naïve Bayes*. Data yang digunakan berupa 4.351 *tweet* berbahasa Indonesia yang dikumpulkan selama November 2024–April 2025 dengan kata kunci "sampah plastik", "kurangi plastik", dan "polusi plastik". Proses analisis mengikuti tahapan CRISP-DM. Data diproses melalui tahapan *preprocessing* Data diproses melalui tahapan *preprocessing* seperti *cleaning*, tokenisasi, dan *stemming*, kemudian dilabeli secara otomatis menggunakan IndoBERT, sebuah model pra-terlatih berbasis BERT yang dirancang khusus untuk klasifikasi sentimen dalam bahasa Indonesia. Fitur diekstraksi dengan TF-IDF, dan model dilatih menggunakan tiga skenario pembagian data (60:40, 70:30, dan 80:20). Untuk mengatasi ketidakseimbangan kelas yang ditemukan, diterapkan teknik *Random Over Sampling* (ROS). ROS bekerja dengan menduplikasi data dari kelas minoritas untuk menyeimbangkan distribusi data latih, yang bertujuan agar model dapat belajar secara lebih adil. Hasil evaluasi menunjukkan bahwa tanpa ROS, model tidak mampu mengenali kelas sentimen positif (*recall* = 0). Namun setelah ROS diterapkan, performa model meningkat signifikan, terutama pada kelas minoritas, dengan kenaikan recall hingga lebih dari 59%. Peningkatan ini membuktikan bahwa ROS efektif dalam memperbaiki performa model untuk analisis sentimen pada data yang tidak seimbang. Penelitian ini diharapkan dapat memberikan kontribusi dalam pemanfaatan *machine learning* untuk memahami persepsi publik dan mendukung perumusan kebijakan lingkungan yang lebih efektif.

Kata kunci: analisis sentimen, naïve bayes, sampah plastik, text mining, *random over sampling*



## 1. Pendahuluan

Permasalahan lingkungan, khususnya yang berkaitan dengan sampah plastik, semakin menjadi perhatian global karena dampaknya yang merugikan terhadap ekosistem dan kesehatan manusia. Di Indonesia, data Sistem Informasi Pengelolaan Sampah Nasional (SIPSN, 2024) mencatat bahwa berdasarkan rekapitulasi nasional melalui Sistem Informasi Pengelolaan Sampah Nasional (SIPSN) Kementerian Lingkungan Hidup dan Kehutanan (KLHK) tahun 2024, sekitar 6,63 juta ton sampah plastik dihasilkan setiap tahun, yang menyumbang 19,64% dari total timbulan sampah nasional [1]. Menurut Asosiasi Industri Plastik Indonesia (INAPLAS) dan Badan Pusat Statistik (BPS) pada tahun 2019, jumlah tersebut bahkan diperkirakan mencapai 64 juta ton per tahun, dengan sekitar 85.000 ton kantong plastik mencemari lingkungan [2]. Tren ini diprediksi meningkat hingga 16% dari total komposisi sampah nasional pada 2025

P-ISSN : 2089-3353
E-ISSN : 2808-9162

[3]. Akumulasi sampah plastik mencemari tanah dan air, merusak kesuburan lahan, serta mengancam mikroorganisme dan biota laut [4]. Di sisi lain, tingginya ketergantungan terhadap plastik dalam industri makanan, rumah tangga, dan kesehatan membuat pengurangannya menjadi tantangan tersendiri [5]. Untuk meningkatkan kesadaran publik, kampanye pengurangan plastik banyak disuarakan melalui media sosial, khususnya platform X (sebelumnya *Twitter*), menggunakan tagar seperti #KurangiPlastik dan #NoPlastik [6]. Namun, tanggapan publik terhadap kampanye ini sangat beragam, mulai dari dukungan hingga penolakan, yang terekam dalam ekspresi bahasa pengguna media sosial [7].

Studi sebelumnya menunjukkan bahwa media sosial dapat dimanfaatkan untuk menganalisis opini publik secara *real-time*. Misalnya, [8] menganalisis kampanye 3M selama pandemi Covid-19 melalui Twitter menggunakan analisis sentimen dan *Social Network* Analysis. Penelitian tersebut menunjukkan dominasi sentimen kebingungan dan keberadaan aktor utama dalam jaringan komunikasi digital, namun juga menyoroti lemahnya kohesi jaringan dalam kampanye tersebut. Kajian lain dilakukan oleh [9] yang menganalisis sentimen masyarakat terhadap kendaraan listrik di Indonesia menggunakan metode Naïve Bayes. Hasil penelitian tersebut memberikan wawasan penting bagi pemerintah serta industri otomotif dalam merancang kebijakan yang mendukung transisi energi bersih.

Meskipun demikian, kajian ilmiah yang secara spesifik mengukur sentimen publik terhadap kampanye pengurangan sampah plastik di media sosial x (sebelumnya *twitter*), khususnya di Indonesia, masih terbatas. Penelitian ini penting dilakukan untuk mengisi celah tersebut, dengan menganalisis persepsi publik terhadap kampanye pengurangan plastik menggunakan pendekatan analisis sentimen. Teknik ini merupakan bagian dari *Natural Language Processing* (NLP) yang mampu mengklasifikasikan opini publik ke dalam kategori positif, negatif, atau netral [10]. Salah satu algoritma yang banyak digunakan dalam analisis sentimen adalah *Naïve Bayes*, yang dikenal karena efisiensinya dalam memproses data teks pendek dan kemudahan implementasinya [11], [12]

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk: (1) menganalisis persepsi publik terhadap kampanye pengurangan plastik di media sosial X; (2) mengklasifikasikan sentimen publik ke dalam kategori positif, negatif, dan netral; serta (3) mengevaluasi kinerja algoritma *Naïve Bayes* dalam mengklasifikasikan opini tersebut. Kontribusi utama penelitian ini adalah memberikan bukti empiris terkini mengenai respons publik terhadap kampanye lingkungan berbasis media sosial di Indonesia, sekaligus memperkaya literatur analisis sentimen pada isu lingkungan yang masih jarang dibahas. Selain itu, hasil penelitian ini dapat menjadi acuan praktis bagi pengambil kebijakan dan pelaku kampanye dalam merancang strategi komunikasi yang lebih efektif, tepat sasaran, dan berbasis data.

## 2. Metode Penelitian

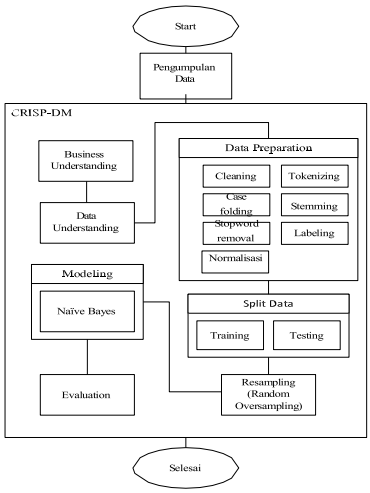
Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap opini masyarakat mengenai isu pengurangan sampah plastik di media sosial X (sebelumnya *Twitter*). Untuk mencapai tujuan tersebut, kami mengadopsi pendekatan CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Kerangka kerja ini dipilih karena menyediakan proses yang sistematis dan terstruktur, sangat relevan untuk mengorganisasi alur kerja dari pengumpulan data teks hingga evaluasi performa model prediktif.

Proses penelitian ini mengikuti lima dari enam tahapan utama CRISP-DM [13]:

1. *Business Understanding*: Tahap ini berfokus pada pemahaman tujuan penelitian, yaitu analisis persepsi publik, dan perumusan masalah data mining.
2. Data *Understanding*: Data relevan dari media sosial X dikumpulkan dan dieksplorasi untuk mengidentifikasi masalah kualitas data dan mendapatkan wawasan awal.
3. Data *Preparation*: Fase ini mencakup aktivitas untuk membangun dataset akhir, yang meliputi pembersihan data, pemilihan atribut, dan transformasi data.
4. *Modeling*: Berbagai teknik pemodelan dipilih dan diterapkan. Dalam penelitian ini, kami menggunakan algoritma Naïve Bayes dan mengkalibrasi parameternya.
5. *Evaluation*: Model yang telah dibangun dievaluasi secara mendalam untuk memastikan bahwa model tersebut memenuhi tujuan dan kriteria keberhasilan yang telah ditetapkan.

Dalam konteks penelitian ini, tahapan terakhir (*Deployment*) tidak dilakukan, karena fokus penelitian hanya sampai tahap evaluasi performa model.

Alur metodologi penelitian yang digunakan ditunjukkan pada Gambar 1




Gambar 1. Alur Metodologi Penelitian

P-ISSN : 2089-3353
E-ISSN : 2808-9162

2.1. Pengumpulan Data

Penelitian ini menggunakan data dari media sosial X (sebelumnya *Twitter*) yang dikumpulkan melalui proses data *crawling*. Data dikumpulkan menggunakan perangkat lunak *Tweet Harvest*, sebuah scraper pihak ketiga yang dikembangkan oleh Helmi Satria [14]. Alat ini dipilih karena kemampuannya untuk melakukan pengambilan data dalam jumlah besar (*mass scraping*) secara fleksibel, tanpa bergantung pada API resmi X yang seringkali memiliki keterbatasan kuota dan akses.

Proses *crawling* dilakukan dengan memanfaatkan token otentikasi (*auth_token*). Berbeda dengan API resmi yang terbatas, *Tweet Harvest* menggunakan *auth_token* dari akun-akun X untuk meniru sesi pengguna, sehingga dapat mengakses data publik secara lebih luas. Untuk menghindari pembatasan rate limit dari *platform*, proses ini dilakukan dengan menggunakan lima *auth_token* dari akun yang berbeda secara bergantian. *Tweet* yang berhasil dihimpun adalah tweet berbahasa Indonesia yang mengandung salah satu dari kata kunci "sampah plastik", "kurangi plastik", atau "polusi plastik" dan dipublikasikan antara November 2024 hingga April 2025. Dari proses ini, sebanyak 4.351 *tweet* berhasil dikumpulkan dan digunakan sebagai dataset utama.

2.2. *Business Understanding*

Permasalahan utama dalam penelitian ini adalah perlunya memahami persepsi masyarakat terhadap kampanye pengurangan sampah plastik di media sosial. Dengan adanya volume data yang sangat besar dan beragam di *platform* seperti X, pengolahan informasi secara manual menjadi tidak efisien. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis sentimen, sebuah pendekatan berbasis machine learning untuk mengklasifikasikan opini publik ke dalam tiga kategori: positif, negatif, dan netral, secara otomatis.

Namun, penelitian ini juga mengidentifikasi dan menangani tantangan signifikan yang sering terjadi pada data sentimen, yaitu ketidakseimbangan kelas (*class imbalance*). Ketidakseimbangan ini dapat menyebabkan model menjadi bias dan gagal mengenali sentimen minoritas. Untuk mengatasi masalah ini, diterapkan teknik *Random Over Sampling* (ROS). Dengan demikian, penelitian ini tidak hanya menganalisis sentimen, tetapi juga mengeksplorasi efektivitas ROS dalam menghasilkan wawasan yang lebih akurat dan seimbang mengenai persepsi publik.

2.3. Data *Understanding*

Data yang telah dikumpulkan kemudian dianalisis secara eksploratif untuk memahami struktur dan kualitasnya. Salah satu langkah penting dalam tahap ini adalah pengecekan terhadap data duplikat untuk memastikan tidak terdapat tweet yang sama muncul lebih dari satu kali. Hal ini bertujuan untuk menjaga validitas analisis serta menghindari bias dalam proses klasifikasi yang disebabkan oleh pengulangan data karena duplikasi data, jika tidak ditangani, dapat memengaruhi distribusi kelas dan menyebabkan bias dalam proses pelatihan model klasifikasi [15]. Selain itu, dilakukan peninjauan umum terhadap jumlah data, distribusi berdasarkan kata kunci, serta karakteristik dasar lainnya sebelum masuk ke tahap persiapan data lebih lanjut.

2.4. Data *Preparation*

Data *preparation* melibatkan proses *preprocessing* teks dan transformasi ke bentuk numerik yang dapat diolah oleh model klasifikasi. Tahapan *preprocessing* yang dilakukan mencakup:

a. *Cleaning*: Proses ini menghilangkan elemen-elemen yang tidak relevan dan dapat mengganggu analisis sentimen, seperti URL, *mention* (@), *hashtag* (#), angka, dan tanda baca. Langkah ini memastikan bahwa model hanya fokus pada kata-kata yang mengandung makna sentimen.
b. *Case Folding*: Semua huruf diubah menjadi huruf kecil (*lowercase*). Tujuannya adalah untuk menyeragamkan kata, sehingga model dapat mengenali "Sangat Bagus" dan "sangat bagus" sebagai entitas yang sama.
c. *Tokenizing*: Kalimat dipecah menjadi unit-unit terkecil, yaitu *token* atau kata. Proses ini esensial sebagai langkah awal untuk analisis lebih lanjut, seperti penghitungan frekuensi kata.
d. *Stopword Removal*: Kata-kata umum yang tidak memiliki makna penting dalam menentukan sentimen (misalnya, "yang", "dan", "di") dihapus. Langkah ini sangat penting untuk mengurangi dimensi data dan menghilangkan noise, sehingga model dapat lebih fokus pada kata-kata kunci yang benar-benar memengaruhi sentimen (misalnya, "baik", "buruk", "sedih").
e. *Stemming*: Kata diubah ke bentuk dasarnya (*root word*). Tujuan stemming adalah untuk menyatukan kata-kata dengan makna serupa, meskipun memiliki akhiran yang berbeda, seperti "mengurangi", "pengurangan", dan "dikurangi" menjadi "kurang". Ini membantu mengurangi redundansi kata dan membuat model lebih efisien dalam mengenali tema sentimen
f. Normalisasi: Proses ini menstandarisasi variasi kata tidak baku atau singkatan (*slang*) menjadi bentuk bakunya, seperti "gk" dan "ga" menjadi "tidak". Langkah ini memastikan konsistensi data sehingga model dapat memproses setiap kata dengan benar.

Tahapan ini mengacu pada praktik umum dalam pemrosesan bahasa alami [16].

Setelah *preprocessing*, dilakukan *labeling* otomatis terhadap tweet menggunakan model *pre-trained* IndoBERT *Sentiment Classification (mdhugol/indonesia-bert-sentiment-classification)* untuk memberikan label sentimen positif, negatif, atau netral. Selanjutnya, dilakukan ekstraksi fitur menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengubah data teks menjadi representasi numerik [17].

P-ISSN : 2089-3353
E-ISSN : 2808-9162

2.5. *Modeling*

Pemodelan dilakukan menggunakan algoritma Naïve Bayes, salah satu metode klasifikasi berbasis probabilistik yang bekerja dengan menerapkan *Teorema Bayes* dengan asumsi independensi antar fitur [18]. Naïve Bayes merupakan metode klasifikasi berbasis probabilitas dan statistik yang dikembangkan berdasarkan *Teorema Bayes* yang dicetuskan oleh ilmuwan Inggris Thomas Bayes. Algotritma ini memprediksi probabilitas suatu kelas berdasarkan informasi dari data historis, sehingga efektif untuk tugas klasifikasi seperti analisis sentimen [19]. Pemilihan Naïve Bayes dalam penelitian ini didasarkan pada karakteristik data teks dari media sosial yang umumnya berupa teks pendek, berdimensi tinggi, dan mengandung kata-kata yang sering berulang [20]. Algoritma ini mampu menangani kondisi tersebut secara efisien karena proses perhitungannya sederhana, tidak memerlukan sumber daya komputasi yang besar, dan memiliki kinerja yang baik meskipun asumsi independensi antar fitur tidak sepenuhnya terpenuhi [21]. Dibandingkan dengan metode lain seperti *Support Vector Machine* (SVM) atau *Decision Tree* yang memerlukan waktu pelatihan lebih lama dan penyesuaian parameter yang kompleks, Naïve Bayes menawarkan kecepatan dan kemudahan dalam implementasinya.

Untuk mengevaluasi performa model, dilakukan tiga skenario pembagian data latih dan uji:

a. Skenario 1: 60% data latih, 40% data uji
b. Skenario 2: 70% data latih, 30% data uji
c. Skenario 3: 80% data latih, 20% data uji

Selain itu, diterapkan teknik *Random Over Sampling* (ROS) pada data latih untuk menangani masalah ketidakseimbangan kelas. Masalah ini sering terjadi pada dataset sentimen di media sosial, di mana jumlah tweet dari satu sentimen (misalnya, netral atau positif) jauh lebih banyak daripada sentimen lainnya. ROS bekerja dengan mendistribusikan sampel dari kelas minoritas secara acak hingga jumlahnya seimbang dengan kelas mayoritas [22]. Teknik ini dipilih karena kemudahannya dalam implementasi dan efisiensi komputasi dibandingkan metode lain seperti SMOTE (*Synthetic Minority Over-sampling Technique*) yang membutuhkan perhitungan lebih kompleks.

2.6. *Evaluation*

Evaluasi model dilakukan menggunakan *Confusion Matrix* dan metrik turunannya. *Confusion Matrix* digunakan untuk memberikan gambaran yang lebih rinci tentang kinerja model, dengan menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas, sehingga membantu memahami jenis kesalahan yang terjadi *(false positives* dan *false negatives*), Metrik-metrik evaluasi yang digunakan adalah:

a. Akurasi (*Accuracy*): Persentase prediksi benar dari total prediksi
b. Presisi (*Precision*): Proporsi prediksi positif yang benar.
c. *Recall*: Kemampuan model dalam mendeteksi semua data yang benar untuk satu kelas.
d. *F1-Score*: Rata-rata harmonik antara presisi dan recall.

Rumus evaluasi metrik disajikan sebagai berikut [23]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP+FP}{TP} \quad (2)$$

$$Recall = \frac{TP+FN}{TP} \quad (3)$$

$$F1\,Score = 2 \times \frac{Precision+Recall}{Precision \times Recall} \quad (4)$$

Model dievaluasi pada dua skenario, yaitu dengan dan tanpa penerapan ROS. Perbandingan ini dilakukan untuk melihat dampak ROS terhadap kinerja klasifikasi sentimen, terutama pada peningkatan nilai *recall* pada kelas minoritas. Hasilnya diharapkan dapat menunjukkan secara eksplisit bahwa penanganan ketidakseimbangan kelas mampu meningkatkan kemampuan model untuk mendeteksi sentimen yang jarang muncul.

## 3. Hasil dan Pembahasan

3.1 *Business Understanding*

Berdasarkan latar belakang penelitian, permasalahan utama yang diangkat adalah volume data yang sangat besar dan beragam terkait isu pengurangan sampah plastik di media sosial X. Dengan adanya 4.351 *tweet* yang berhasil dikumpulkan, pengolahan informasi secara manual menjadi tidak efisien. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sentimen analisis berbasis machine learning untuk menganalisis komentar tersebut secara otomatis. Ini sangat dibutuhkan karena dapat membantu pelaku kampanye dan pengambil kebijakan dalam memahami sentimen publik secara cepat dan akurat, sehingga mereka bisa merancang strategi komunikasi yang lebih efektif.

Untuk mencapai tujuan tersebut, model yang digunakan adalah algoritma Naive Bayes. Algoritma ini dipilih karena efektivitasnya yang tinggi dalam klasifikasi teks dan efisiensi komputasinya, menjadikannya ideal untuk mengolah dataset yang besar. Selain itu, teknik *Random Over Sampling* (ROS) diterapkan untuk mengatasi ketidakseimbangan kelas yang umumnya terjadi pada data sentimen, di mana satu sentimen tertentu (misalnya, sentimen negatif) lebih dominan. Penerapan ROS memastikan model tidak bias pada kelas mayoritas dan mampu mengidentifikasi sentimen minoritas (misalnya, sentimen positif) dengan lebih baik. Penelitian ini menggunakan Python sebagai bahasa pemrograman dengan pustaka pendukung seperti Sastrawi, NLTK, *scikit-learn* dan pandas serta dijalankan pada platform *Google Colaboratory*.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

3.2 Data *Understanding*

Pada tahap ini, kami melakukan eksplorasi awal terhadap dataset *tweet* yang berhasil dikumpulkan dari media sosial X. Dataset ini difokuskan pada isu kampanye pengurangan penggunaan plastik. Pemilihan kata kunci "sampah plastik", "kurangi plastik", dan "polusi plastik" dilakukan untuk memastikan data yang terkumpul relevan dan mewakili berbagai aspek dari isu tersebut, mulai dari deskripsi masalah (sampah plastik, polusi plastik) hingga seruan aksi (kurangi plastik). Data mencakup rentang waktu dari November 2024 hingga April 2025.

3.2.1 Memeriksa Struktur Data dan Properti Awal

Tahap ini berfokus pada pemeriksaan struktur dasar dataset yang telah digabungkan dari berbagai file CSV. Pemeriksaan ini mencakup verifikasi data untuk memastikan tidak ada duplikasi tweet dan setiap tweet memiliki struktur data yang valid. Selain itu, kami juga melakukan identifikasi nama-nama kolom, tipe data masing-masing kolom, serta total jumlah baris dan kolom.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4662 entries, 0 to 4661
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   conversation_id_str     4662 non-null   float64
 1   created_at              4662 non-null   object
 2   favorite_count          4662 non-null   int64
 3   full_text               4662 non-null   object
 4   id_str                  4662 non-null   float64
 5   image_url               1017 non-null   object
 6   in_reply_to_screen_name 2778 non-null   object
 7   lang                    4662 non-null   object
 8   location                0 non-null      float64
 9   quote_count             4662 non-null   int64
 10  reply_count             4662 non-null   int64
 11  retweet_count           4662 non-null   int64
 12  tweet_url               4662 non-null   object
 13  user_id_str             4662 non-null   float64
 14  username                0 non-null      float64
dtypes: float64(5), int64(4), object(6)
memory usage: 546.5+ KB
```

Gambar 2. Hasil Periksa Data dan Properti

*Dataset* awal terdiri dari 4.662 baris dan 15 kolom. Kolom-kolom utama antara lain:
a. *full_text*: Berisi isi lengkap tweet (tipe data: string)
b. *created_at*: Waktu tweet dibuat (tipe data: string)
c. *id_str*: ID unik tiap tweet (tipe data: float64)
d. *favorite_count, reply_count, retweet_count, dan quote_count*: Data interaksi
e. *location* dan username: Kedua kolom ini tidak memiliki nilai yang lengkap (0 *non-null*), sehingga tidak dapat memberikan informasi yang relevan untuk analisis sentimen yang berbasis teks. Selain itu, fokus penelitian ini adalah pada isi teks tweet itu sendiri, bukan pada atribut demografi atau identitas pengguna, sehingga kolom ini tidak diperlukan.

3.2.2 Identifikasi Duplikasi Data

Identifikasi duplikasi dilakukan pada kolom full_text untuk menghindari bias dalam analisis sentimen.

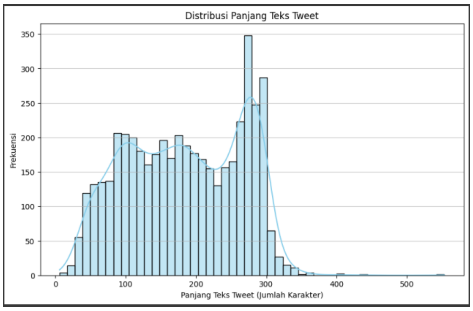
```
Total baris dalam dataset awal: 4662
Jumlah duplikat yang teridentifikasi (berdasarkan 'full_text'): 311
Jumlah baris unik yang teridentifikasi: 4351
```

Gambar 3. Hasil Identifikasi Duplikasi Data

Dari 4.662 *tweet*, ditemukan 311 *tweet* duplikat. Artinya terdapat 4.351 *tweet* unik yang akan digunakan dalam proses selanjutnya.

3.2.3 Analisis Distribusi Panjang *Teks Tweet*

Analisis ini membantu memahami variasi panjang teks dan dampaknya pada proses *preprocessing* dan ekstraksi fitur.




Gambar 4. Visualisasi Distribusi Panjang *Teks Tweet*

Pada Gambar 4, Distribusi memiliki dua puncak:
a. Puncak pertama (90-110 karakter): Menunjukkan kelompok pengguna yang menyampaikan pesan singkat
b. Puncak kedua (270-300 karakter): Menunjukkan penggunaan teks panjang, mendekati batas maksimal *tweet.*

*Tweet* sangat pendek (<50 karakter) dan sangat panjang (>350 karakter) memiliki frekuensi yang rendah.

3.3 Data *Preparation*

Setelah melalui tahapan data *understanding*, data mentah hasil *crawling* perlu diproses lebih lanjut agar dapat digunakan dalam analisis sentimen. Seperti yang telah dijelaskan pada bagian sebelumnya, seluruh proses *preprocessing* dilakukan untuk membersihkan data dari *noise*, menstandarisasi penulisan, dan menghilangkan elemen non-teks yang tidak relevan. Tahapan *preprocessing* meliputi penghapusan duplikat, *cleansing*, *case folding*, tokenisasi, normalisasi, *stopword removal*, dan stemming. Untuk memberikan gambaran menyeluruh terhadap perubahan data dalam setiap tahapan, Tabel 1 menyajikan gambaran yang jelas mengenai perubahan data dari satu *tweet* mentah setelah melewati setiap tahapan *preprocessing* tersebut.

Tabel 1. Hasil Data setelah dilakukan beberapa tahapan

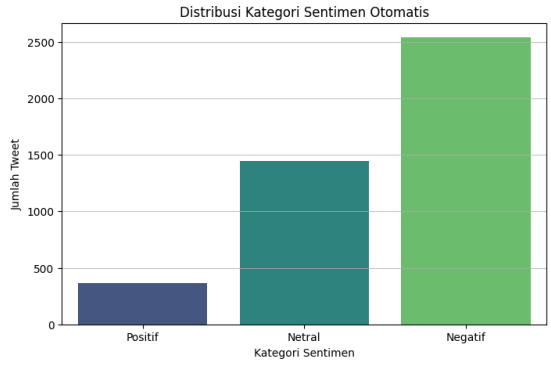
| Tahap | Hasil |
|---|---|
| Tweet Awal | Polusi Plastik: Ancaman Nyata untuk Masa Depan Polusi plastik di Indonesia diperkirakan melonjak hingga 27 juta ton pada tahun 2040. #SelamatkanBumi #KurangiPlastik #*GoGreen* #PolusiPlastik #IndonesiaBersih #Ayobacanews https://t.co/hfpA9ZKgEG |
| Data *Cleaning* | Polusi Plastik Ancaman Nyata untuk Masa Depan Polusi plastik di Indonesia diperkirakan melonjak hingga juta ton pada tahun SelamatkanBumi KurangiPlastik *GoGreen* PolusiPlastik IndonesiaBersih Ayobacanews |
| *Case Folding* | polusi plastik ancaman nyata untuk masa depan polusi plastik di indonesia diperkirakan |

P-ISSN : 2089-3353
E-ISSN : 2808-9162

| | |
|---|---|
| Tokenisasi | melonjak hingga juta ton pada tahun selamatkanbumi kurangiplastik *gogreen* polusiplastik indonesiabersih ayobacanews [polusi, plastik, ancaman, nyata, untuk, masa, depan, polusi, plastik, di, indonesia, diperkirakan, melonjak, hingga, juta, ton, pada, tahun, selamatkanbumi, kurangiplastik, *gogreen*, polusiplastik, indonesiabersih, ayobacanews] |
| Normalisasi | [polusi, plastik, ancaman, nyata, untuk, masa, depan, polusi, plastik, di, indonesia, diperkirakan, melonjak, hingga, juta, ton, pada, tahun, selamatkanbumi, kurangiplastik, *gogreen*, polusiplastik, indonesiabersih, ayobacanews] |
| *Stopword Removal* | [polusi, plastik, ancaman, nyata, masa, depan, polusi, plastik, indonesia, diperkirakan, melonjak, hingga, juta, ton, tahun, selamatkanbumi, kurangiplastik, *gogreen*, polusiplastik, indonesiabersih, ayobacanews] |
| *Stemming* | [polusi, plastik, ancam, nyata, masa, depan, polusi, plastik, indonesia, kira, lonjak, hingga, juta, ton, tahun, selamatkanbumi, kurangiplastik, *gogreen*, polusiplastik, indonesiabersih, ayobacanews] |

Setelah seluruh proses *preprocessing* selesai, dilakukan tahap pelabelan sentimen otomatis untuk mengklasifikasikan *tweet* ke dalam tiga kategori: Positif, Netral, dan Negatif. Berbeda dengan pelabelan manual yang memakan waktu, penelitian ini mengadopsi pendekatan otomatis dengan memanfaatkan model pra-terlatih dari *Hugging Face*, yaitu *mdhugol/indonesia-bert-sentiment-classification*.

Model ini dirancang khusus untuk klasifikasi sentimen dalam bahasa Indonesia dan dibangun berdasarkan arsitektur BERT (*Bidirectional Encoder Representations from Transformers*), yang memungkinkan pemahaman konteks dua arah dalam teks. Penggunaan model ini memungkinkan proses pelabelan yang efisien, objektif, serta dapat diskalakan untuk seluruh dataset yang terdiri dari ribuan tweet. Setiap data tweet yang telah diproses akan dilengkapi dua kolom baru, yaitu: Label, yang menunjukkan kategori sentimen (Positif, Netral, atau Negatif), dan *Sentiment Score*, yang menunjukkan tingkat kepercayaan model terhadap label tersebut.




Gambar 5. Distribusi Kategori Sentimen

Hasil pelabelan otomatis ini menunjukkan ketidakseimbangan kelas (*class imbalance*) pada Gambar 5. Kategori sentimen negatif mendominasi, mencakup sekitar 2.500 *tweet* (57,45%) dari total *tweet*. *Tweet-tweet* dalam kategori ini umumnya berisi kritik, kekhawatiran, atau kekecewaan terhadap masalah polusi plastik. Sementara itu, kategori netral sekitar 1.405 *tweet* (33,32%) banyak berisi informasi faktual atau pernyataan tanpa emosi eksplisit, dan kategori positif hanya mencakup sekitar 350 *tweet* (8,05%), yang mencerminkan dukungan atau optimisme terhadap kampanye pengurangan plastik.

Ketidakseimbangan ini menjadi masalah krusial karena dapat menyebabkan model menjadi bias terhadap kelas mayoritas (negatif) dan sulit mengidentifikasi sentimen pada kelas minoritas (positif). Untuk mengatasi masalah tersebut, digunakan teknik *Random Over Sampling* (ROS) pada data latih. ROS bekerja dengan menggandakan sampel dari kelas minoritas secara acak hingga jumlahnya setara dengan kelas mayoritas, sehingga distribusi kelas menjadi seimbang. Tabel 2 menunjukkan jumlah data tiap kelas sebelum dan sesudah ROS untuk setiap skenario pembagian data (60:40, 70:30, dan 80:20). Penerapan ROS menghasilkan distribusi yang merata pada setiap kelas, yang membantu model belajar secara lebih seimbang dan meningkatkan potensi *recall* pada kelas positif. Dampak penerapan ROS terhadap performa model akan dibahas lebih lanjut pada bagian hasil evaluasi menggunakan *Confusion Matrix* di subbab selanjutnya.

Tabel 2. Hasil Penerapan ROS

| Skenario | Kelas | Sebelum ROS | Setelah ROS |
|---|---|---|---|
| **60:40** | Positif | 221 | 1525 |
| | Netral | 864 | 1525 |
| | Negatif | 1525 | 1525 |
| **70:30** | Positif | 258 | 1779 |
| | Netral | 1008 | 1779 |
| | Negatif | 1779 | 1779 |
| **80:20** | Positif | 285 | 2033 |
| | Netral | 1034 | 2033 |
| | Negatif | 2033 | 2033 |

Langkah berikutnya adalah ekstraksi fitur menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*). Proses ini mengubah representasi teks menjadi bentuk numerik, dengan memberikan bobot lebih besar pada kata-kata yang dianggap penting dalam membedakan dokumen. Representasi ini membentuk matriks sparse, di mana sebagian besar elemen bernilai nol karena tidak semua kata muncul di setiap dokumen. Hasil TF-IDF inilah yang digunakan sebagai input dalam pelatihan model klasifikasi pada tahap selanjutnya.

### 3.4 *Modeling*

Tahap pemodelan merupakan inti dari proses klasifikasi sentimen dalam penelitian ini. Setelah fitur diekstraksi menggunakan metode TF-IDF (*Term Frequency–Inverse Document Frequency*), hasil representasi numerik tersebut digunakan sebagai input dalam pelatihan model. Proses pemodelan dalam penelitian ini menggunakan algoritma *Multinomial* Naïve Bayes, yang dinilai sesuai untuk tugas klasifikasi teks seperti analisis sentimen karena kesederhanaan,

P-ISSN : 2089-3353
E-ISSN : 2808-9162

efisiensi, serta kemampuannya menangani data dalam bentuk frekuensi kata atau bobot.

Untuk menangani masalah ketidakseimbangan data yang ditemukan sebelumnya, dilakukan penerapan *Random Over Sampling* (ROS) secara khusus pada data latih. Teknik ini bertujuan untuk memperbanyak jumlah data pada kelas minoritas sehingga distribusi kelas menjadi seimbang, memungkinkan model belajar secara adil terhadap semua kelas sentimen (positif, netral, dan negatif) [24].

Penelitian ini menggunakan tiga skenario pembagian data untuk pelatihan dan pengujian, yaitu 60% data latih dan 40% data uji (60:40), 70% data latih dan 30% data uji (70:30), serta 80% data latih dan 20% data uji (80:20). Setiap skenario dijalankan dalam dua eksperimen: tanpa penerapan ROS pada data latih (*baseline model*) dan dengan penerapan ROS pada data latih (*balanced model*). Distribusi data sebelum dan sesudah ROS dapat dilihat pada Tabel 2. Sebagai contoh, pada skenario 70:30, jumlah data latih awal terdiri dari 1.779 *tweet* negatif, 1.008 *tweet* netral, dan hanya 258 *tweet* positif; setelah ROS diterapkan, masing-masing kelas memiliki 1.779 data. Perlu dicatat bahwa data uji tetap menggunakan distribusi asli (tanpa ROS) agar performa model dapat diuji dalam kondisi nyata. Model dilatih dengan data latih (yang telah diseimbangkan jika menggunakan ROS) dan matriks TF-IDF, kemudian diuji menggunakan data uji untuk mengukur akurasi dan efektivitas klasifikasi.

### 3.5 *Evaluation*

Tahapan ini bertujuan untuk mengevaluasi kinerja model algoritma Naïve Bayes sebelum dan setelah penanganan ketidakseimbangan kelas (*imbalanced class*) menggunakan *Random Over Sampling* (ROS). Evaluasi dilakukan berdasarkan beberapa metrik performa utama: Akurasi, Presisi, Recall, dan *F1-Score*, serta visualisasi melalui *Confusion Matrix*. Metrik-metrik ini dipilih karena masing-masing memberikan perspektif yang berbeda dan sangat relevan dalam konteks analisis sentimen, terutama pada data yang tidak seimbang.

1. Akurasi digunakan untuk mengukur persentase prediksi benar dari total data, namun pada data yang tidak seimbang, akurasi dapat menyesatkan karena model tetap bisa menghasilkan akurasi tinggi meski gagal mengenali kelas minoritas.
2. Presisi penting untuk memastikan bahwa prediksi positif benar-benar relevan, sehingga kesalahan klasifikasi tidak mengganggu interpretasi sentimen.
3. *Recall* mengukur kemampuan model menemukan seluruh data positif, yang sangat krusial pada analisis sentimen karena setiap sinyal dukungan atau kritik penting untuk diidentifikasi.
4. *F1-score* memberikan rata-rata harmonik antara presisi dan recall, sehingga menjadi ukuran yang seimbang saat keduanya sama-sama penting.

#### 3.5.1 Evaluasi Awal (Sebelum ROS)

Sebelum dilakukan penanganan terhadap ketidakseimbangan kelas (imbalanced class), dilakukan evaluasi awal terhadap performa model Naïve Bayes pada tiga skenario pembagian data, yaitu 60:40, 70:30, dan 80:20. Evaluasi ini bertujuan untuk mengetahui sejauh mana kemampuan model dalam mengklasifikasikan sentimen pada data dengan distribusi kelas yang belum diseimbangkan.

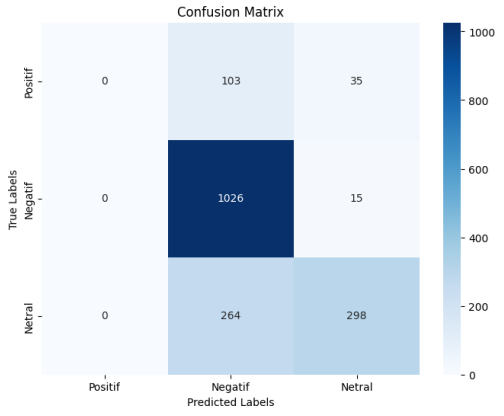
Tabel 3. Ringkasan Metrik Evaluasi Awal Model *Naïve Bayes*

| Skenario | Akurasi | Presisi | Recall | *F1-Score* |
|---|---|---|---|---|
| **60:40** | 76.05% | 0.5310 | 0.5053 | 0.4993 |
| **70:30** | 75.04% | 0.5239 | 0.4988 | 0.4910 |
| **80:20** | 75.55% | 0.5247 | 0.5008 | 0.4937 |

Berdasarkan Tabel 3, skenario 60:40 memberikan kinerja terbaik pada seluruh metrik. Hasil ini kemungkinan disebabkan oleh keseimbangan optimal antara jumlah data latih dan data uji, memungkinkan model belajar dari sampel yang memadai tanpa terlalu banyak terpapar bias dari data latih yang tidak seimbang.

Namun, hasil pada ketiga skenario menunjukkan kelemahan signifikan dalam mengklasifikasikan kelas minoritas, khususnya kelas Positif. Ketidakmampuan model untuk mengenali kelas positif ini terjadi karena model cenderung bias terhadap kelas mayoritas (negatif) selama pelatihan. Akibatnya, model mengabaikan karakteristik unik dari kelas minoritas. Hal ini terlihat dari nilai *recall* yang konsisten berada di angka nol pada seluruh skenario.

Untuk memperjelas pola kesalahan model, khususnya terhadap kelas minoritas, dilakukan analisis lebih lanjut melalui visualisasi *Confusion Matrix* berikut ini:

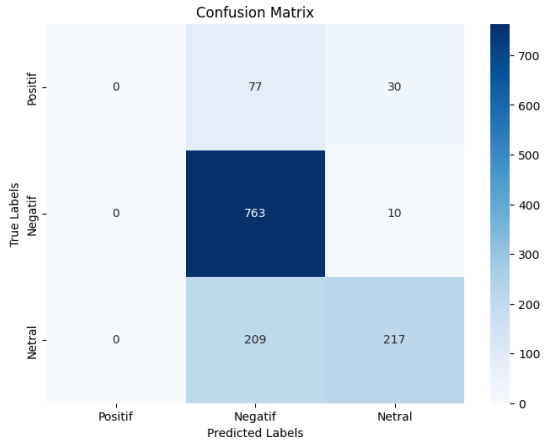



Gambar 6. *Confusion Matrix* (skenario 60:40)

Pada Gambar 6, model tidak berhasil mengenali satu pun *tweet* berlabel Positif dari total 138 *tweet* yang ada. Seluruh data tersebut diprediksi sebagai kelas lain, yaitu Negatif atau Netral, sehingga nilai recall untuk kelas Positif adalah nol. Hal ini terjadi karena model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Sementara itu, kelas Negatif dikenali dengan sangat baik oleh model, di mana sebagian besar tweet dengan label ini berhasil diklasifikasikan dengan benar. Untuk kelas Netral, sebagian data berhasil dikenali, namun masih terjadi

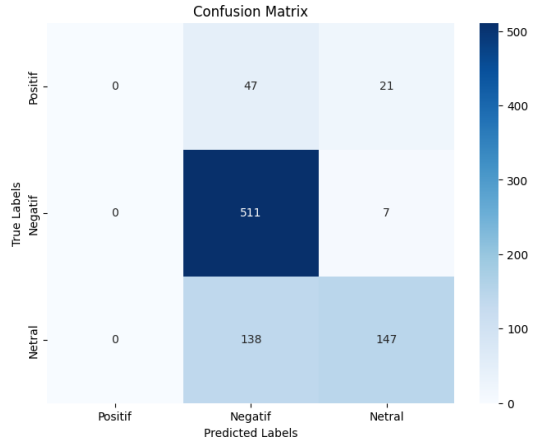
P-ISSN : 2089-3353
E-ISSN : 2808-9162

kesalahan klasifikasi, terutama pengelompokan yang salah ke kelas Negatif.




Gambar 7. *Confusion Matrix* (skenario 70:30)

Pada Gambar 7, pola kesalahan yang terjadi sangat mirip dengan skenario sebelumnya. Dari total 107 tweet berlabel Positif, tidak ada satu pun yang berhasil dikenali oleh model, yang sekali lagi menghasilkan nilai recall nol. Hasil ini sekali lagi menegaskan bias model terhadap kelas dominan. Kinerja model terhadap kelas Negatif tetap tinggi, dan kelas Netral menunjukkan hasil yang relatif serupa dengan skenario 60:40.




Gambar 8. *Confusion Matrix* (skenario 80:20)

Kemudian pada Gambar 8, meskipun proporsi data latih lebih besar, hasil yang diperoleh tidak menunjukkan peningkatan dalam mengenali kelas minoritas. Dari 68 *tweet* yang seharusnya diklasifikasikan sebagai Positif, tidak ada satu pun yang dikenali dengan benar, sehingga *recall* tetap berada pada angka nol. Model cenderung kembali mengklasifikasikan data secara dominan ke kelas mayoritas, yaitu Negatif, dan hanya sebagian dari data Netral yang dapat dikenali dengan tepat.

Berdasarkan ketiga skenario tersebut, dapat disimpulkan bahwa model Naïve Bayes menunjukkan kelemahan yang sangat jelas dalam mengenali kelas Positif yang termasuk dalam kategori minoritas. Ketidakmampuan model untuk mengklasifikasikan satu pun data dari kelas ini secara benar, sebagaimana diperlihatkan oleh nilai recall yang konsisten berada di angka nol, menunjukkan bahwa distribusi data yang tidak seimbang memberikan dampak negatif signifikan terhadap performa model. *Confusion Matrix* dalam hal ini berperan penting dalam memperlihatkan bahwa tanpa penanganan terhadap ketidakseimbangan data, model cenderung bias terhadap kelas mayoritas dan mengabaikan keberadaan kelas minoritas secara keseluruhan.

### 3.5.2 Evaluasi Setelah Penanganan *Imbalanced Class* (ROS)

Setelah dilakukan evaluasi awal terhadap model *Naïve Bayes* tanpa penanganan ketidakseimbangan kelas, langkah selanjutnya adalah menerapkan teknik *Random Over Sampling* (ROS) guna meningkatkan kemampuan model dalam mengenali kelas minoritas, khususnya sentimen Positif yang sebelumnya tidak terprediksi sama sekali. Teknik ROS diterapkan secara khusus pada data latih dalam setiap skenario pembagian data, yaitu 60:40, 70:30, dan 80:20. Tujuan utama dari penerapan teknik ini adalah untuk menyamakan jumlah data pada masing-masing kelas, sehingga model memperoleh representasi data yang seimbang selama proses pelatihan.

Evaluasi dilakukan kembali menggunakan metrik yang sama seperti sebelumnya, yaitu Akurasi, Presisi, *Recall*, dan *F1-Score*. Hasil evaluasi yang diperoleh setelah penerapan ROS ditampilkan pada Tabel 3.5 berikut:

Tabel 4. Ringkasan Metrik Evaluasi setelah ROS Pada Model *Naïve Bayes*

| Skenario | Akurasi | Presisi | Recall | *F1-Score* |
|---|---|---|---|---|
| **60:40** | 75.36% | 0.6585 | 0.7017 | 0.6622 |
| **70:30** | 73.43% | 0.6415 | 0.6794 | 0.6407 |
| **80:20** | 75.89% | 0.6619 | 0.7058 | 0.6673 |

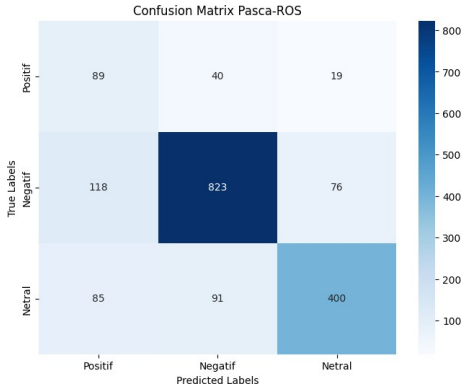
Berdasarkan hasil pada Tabel 4, terlihat bahwa seluruh metrik performa mengalami peningkatan setelah dilakukan penyeimbangan kelas menggunakan ROS. Meskipun terjadi sedikit penurunan akurasi pada skenario 70:30, namun peningkatan nilai *Recall* dan *F1-Score* menunjukkan bahwa model menjadi lebih adil dalam mengklasifikasikan seluruh kelas sentimen, terutama kelas Positif yang sebelumnya diabaikan.

Setelah dilakukan evaluasi awal terhadap model Naïve Bayes tanpa penanganan ketidakseimbangan kelas, langkah selanjutnya adalah menerapkan teknik *Random Over Sampling* (ROS) untuk memperbaiki performa klasifikasi, khususnya terhadap kelas minoritas. Teknik ROS diterapkan hanya pada data latih untuk setiap skenario, yaitu 60:40, 70:30, dan 80:20. Tujuan utamanya adalah agar model memiliki representasi yang seimbang dari masing-masing kelas selama proses pembelajaran, sehingga mampu mengenali kelas minoritas seperti Positif dengan lebih baik.

Evaluasi dilakukan kembali pada ketiga skenario menggunakan metrik yang sama seperti sebelumnya,

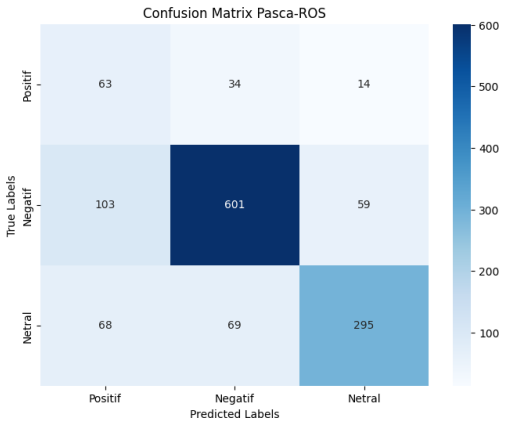
P-ISSN : 2089-3353
E-ISSN : 2808-9162

yaitu akurasi, presisi, *recall*, dan *F1-score*. Hasil evaluasi menunjukkan bahwa penerapan ROS berhasil meningkatkan performa model dalam mengenali semua kelas sentimen, terutama kelas Positif yang sebelumnya memiliki nilai *recall* nol.



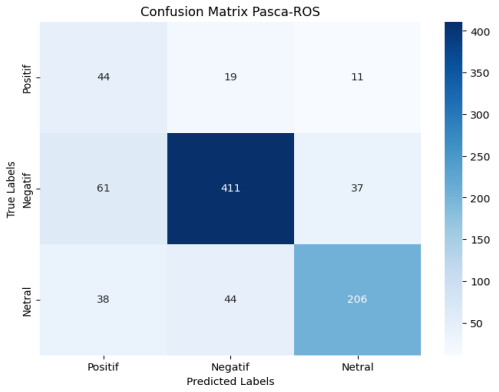

Gambar 9. *Confusion Matrix* Setelah ROS (skenario 60:40)

Pada Gambar 9, model *Naïve Bayes* yang dilatih menggunakan data hasil ROS menunjukkan peningkatan signifikan pada metrik presisi, recall, dan *F1-score*. Presisi untuk kelas Positif mencapai sekitar 0.2927, dengan *recall* sebesar 0.6014 dan *F1-score* sebesar 0.3922. Untuk kelas Negatif, presisi dan *recall* masing-masing mencapai 0.8617 dan 0.7965, sementara kelas Netral memiliki presisi sebesar 0.8032 dan recall 0.6875. Secara keseluruhan, model mencapai akurasi sebesar 75.36%. Meskipun nilai akurasi ini sedikit menurun dibandingkan sebelum ROS, yaitu 76.05%, peningkatan signifikan pada kemampuan model dalam mengenali kelas minoritas menunjukkan bahwa ROS berhasil memperbaiki distribusi prediksi dan meningkatkan keseimbangan klasifikasi.




Gambar 10. *Confusion Matrix* setelah ROS (skenario 70:30)

Hasil serupa juga terlihat pada Gambar 10. Setelah penerapan ROS, model mampu mengenali kelas Positif dengan jauh lebih baik dibandingkan sebelumnya. Presisi untuk kelas ini tercatat sekitar 0.2692, dengan *recall* sebesar 0.5676 dan *F1-score* sebesar 0.3671. Kelas Negatif memiliki presisi sekitar 0.8537 dan recall 0.7874, sedangkan kelas Netral mencatatkan presisi sebesar 0.8011 dan *recall* 0.6829. Akurasi keseluruhan model tercatat sebesar 73.43%, sedikit menurun dari akurasi sebelum ROS yaitu 75.04%. Namun peningkatan signifikan pada metrik *recall* dan *F1-score*, khususnya untuk kelas Positif dan Netral, menunjukkan bahwa model menjadi lebih adil dan tidak lagi terlalu mendominasi kelas mayoritas.




Gambar 11. *Confusion Matrix* setelah ROS (skenario 80:20)

Pada skenario 80:20, model menunjukkan performa terbaik setelah penerapan ROS. Nilai presisi untuk kelas Positif mencapai sekitar 0.3070, dengan recall sebesar 0.5946 dan *F1-score* sebesar 0.4052. Kelas Negatif mencatatkan presisi sekitar 0.8665 dan *recall* 0.8075, sementara kelas Netral memiliki presisi 0.8110 dan recall 0.7153. Model secara keseluruhan mencatatkan akurasi sebesar 75.89%, yang sedikit lebih tinggi dibandingkan dengan skenario 60:40 dan 70:30. Peningkatan pada nilai recall dan *F1-score* untuk kelas Positif sangat menonjol dibandingkan dengan evaluasi awal sebelum ROS, di mana kelas ini sama sekali tidak terklasifikasi dengan benar.

Dari ketiga skenario evaluasi, dapat disimpulkan bahwa penerapan ROS memberikan dampak yang positif terhadap kemampuan model dalam mengenali ketiga kelas sentimen secara lebih seimbang. Meskipun terjadi sedikit penurunan pada nilai akurasi, peningkatan pada nilai presisi, recall, dan *F1-score*, terutama untuk kelas Positif yang sebelumnya tidak dikenali sama sekali, menunjukkan bahwa ROS mampu memperbaiki kelemahan utama model Naïve Bayes dalam menangani ketidakseimbangan kelas.

## 4. Kesimpulan

Penelitian ini menunjukkan bahwa algoritma Naïve Bayes efektif digunakan untuk melakukan analisis sentimen terhadap opini masyarakat terkait isu pengurangan sampah plastik di media sosial X. Namun, temuan kunci dari penelitian ini adalah pentingnya penanganan ketidakseimbangan kelas (*class imbalance*). Tanpa penanganan tersebut, model cenderung bias terhadap kelas mayoritas dan gagal total dalam mengenali kelas minoritas seperti sentimen positif. Hal ini terbukti dari nilai recall untuk kelas positif yang konsisten berada di angka nol pada semua skenario awal.

Untuk mengatasi masalah tersebut, diterapkan teknik *Random Over Sampling* (ROS) pada data latih. Penerapan ROS terbukti berhasil meningkatkan performa model secara signifikan. Meskipun terjadi

P-ISSN : 2089-3353
E-ISSN : 2808-9162

sedikit penurunan pada Akurasi secara keseluruhan, peningkatan substansial pada metrik Recall dan F1-Score menunjukkan bahwa model menjadi lebih seimbang dan adil dalam mengenali sentimen. Secara spesifik, ROS berhasil membuat model mampu mengenali sentimen positif, yang sebelumnya diabaikan, sehingga kemampuan model dalam mendeteksi semua kelas sentimen meningkat. Penurunan akurasi ini dapat diterima karena mengorbankan sedikit keakuratan keseluruhan untuk mendapatkan kemampuan deteksi yang jauh lebih baik pada sentimen minoritas yang krusial.

Temuan ini menegaskan pentingnya penanganan data tidak seimbang dalam analisis teks dan dampaknya yang nyata terhadap validitas hasil. Penelitian ini dapat diaplikasikan sebagai dasar dalam merancang strategi komunikasi kampanye lingkungan yang berbasis pada pemahaman publik secara otomatis. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi algoritma lain serta mempertimbangkan pendekatan ensemble dan teknik pelabelan manual guna meningkatkan akurasi dan validitas model.

## Daftar Rujukan

[1] Kementerian Lingkungan Hidup dan Kehutanan, 2025. sistem informasi pengelolaan sampah nasional. [Online]. Tersedia: https://sipsn.menlhk.go.id/sipsn/ [Diakses 28 Apr 2025].

[2] Portal Informasi Indonesia, 2019. menenggelamkan pembuang sampah plastik di laut. [Online]. Tersedia: https://indonesia.go.id/narasi/indonesia-dalam-angka/sosial/menenggelamkan-pembuang-sampah-plastik-di-laut [Diakses 11 Agt 2025].

[3] Y. A. Hidayat, S. Kiranamahsa, and M. A. Zamal, "A study of plastic waste management effectiveness in Indonesia industries," AIMS Energy, vol. 7, no. 3, pp. 350–370, 2019, doi: 10.3934/ENERGY.2019.3.350.

[4] Putu, N. and Arwini, D., 2022. Sampah Plastik Dan Upaya Pengurangan Timbulan Sampah Plastik. Jurnal Ilmiah Vastuwidya, 5(1), 72–82. doi: https://doi.org/10.47532/jiv.v5i1.412

[5] Pilapitiya, P.G.C.N.T. and Ratnayake, A.S., 2024. The world of plastic waste: a review. Cleaner Materials, 11, p.100220. doi: https://doi.org/10.1016/j.clema.2024.100220.

[6] Maheswari, J.S., Hikmah, E.S. and Rizaldhi, M.B., 2023. Penggunaan media Instagram dalam kampanye pengurangan sampah plastik: studi pustaka artikel ilmiah periode 2019-2022. In: Prosiding Seminar Nasional. pp.702–711.

[7] Obidje, B.M. and Pakereng, M.A.I., 2025. Analisis sentimen pemilihan presiden dan wakil presiden tahun 2024 di Twitter menggunakan metode klasifikasi Naive Bayes. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10 (1), pp.424–433. doi: https://doi.org/10.29100/jipi.v10i1.5836

[8] Rakhman, F.R., Ramadhani, R.W. and Kuncoroyakti, Y.A., 2021. Analisis sentimen dan opini digital kampanye 3M di masa Covid-19 melalui media sosial Twitter. Komunikologi: Jurnal Ilmiah Ilmu Komunikasi, 18 (1).

[9] Farhani, A., 2024. Analisis sentimen terhadap kendaraan listrik di Indonesia menggunakan metode klasifikasi Naïve Bayes. Jurnal Indonesia: Manajemen Informatika dan Komunikasi, 5 (3), pp.2680–2690.

[10] Purwanti, Z., 2024. Pemodelan text mining untuk analisis sentimen terhadap program makan siang gratis di media sosial X menggunakan algoritma support vector machine (SVM). [Online] Available at: https://journal.stmiki.ac.id [Accessed 17 August 2025].

[11] Millennianita, F., Athiyah, U. and Muhammad, A.W., 2024. Comparison of Naïve Bayes classifier and support vector machine methods for sentiment classification of responses to bullying cases on Twitter. Journal of Mechatronics and Artificial Intelligence. [Online] Available at: http://ejournal.upi.edu/index.php/jmai/ [Accessed 17 August 2025].

[12] Alfiyani, W., Fatah, D.A. and Irhamni, F., 2025. Penerapan algoritma Naïve Bayes untuk analisis sentimen pada media sosial X terhadap performa tim nasional sepak bola Indonesia di era kepemimpinan Shin Tae-yong. JATI (Jurnal Mahasiswa Teknik Informatika), 9 (3), pp.3969– 3977. doi: https://doi.org/10.36040/jati.v9i3.13451

[13] Wirth, R. and Hipp, J., 2000. CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. Manchester, pp.29–39.

[14] Satria, H., 2023. Crawl data Twitter menggunakan Tweet Harvest – Juli 2023. [Online] Available at: https://helmisatria.com/blog/crawl-data-twitter- menggunakan-tweet-harvest/ [Accessed 30 June 2025].

[15] Sholihah, N.N. and Hermawan, A., 2023. Implementation of random forest and smote methods for economic status classification in Cirebon City. Jurnal Teknik Informatika (Jutif), 4 (6), pp.1387–1397. doi: https://doi.org/10.52436/1.jutif.2023.4.6.1135

[16] Findawati, Y., Indahyanti, U., Rahmawati, Y. and Puspitasari, R., 2023. Sentiment analysis of potential presidential candidates 2024: a Twitter-based study. Academia Open, 8 (1), pp.10–21070. doi: https://doi.org/10.21070/acopen.8.2023.7138

[17] Rabbani, S., Safitri, D., Rahmadhani, N., Sani, A.A.F. and Anam, M.K., 2023. Perbandingan evaluasi kernel SVM untuk klasifikasi sentimen dalam analisis kenaikan harga BBM: comparative evaluation of SVM kernels for sentiment classification in fuel price increase analysis. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3 (2), pp.153–160. https://doi.org/10.57152/malcom.v3i2.897

[18] Widodo, Y.B., Anggraeini, S.A. and Sutabri, T., 2021. Perancangan sistem pakar diagnosis penyakit diabetes berbasis web menggunakan algoritma Naive Bayes. Jurnal Teknologi Informasi dan Komputer, 7 (1), pp.112–123. doi: https://doi.org/10.37012/jtik.v7i1.507

[19] Rina, 2025. Algoritma Naive Bayes: pemahaman, contoh perhitungan manual serta implementasi dengan Python dan Orange Data Mining. Medium. [Online] Available at: https://esairina.medium.com/algoritma-naive-bayes- pemahaman-contoh-perhitungan-manual-dan- implementasi-dengan-python-dan-475091cae835 [Accessed 15 May 2025].

[20] Mualfah, D., Prihatin, A. and Firdaus, R., 2023. Analisis sentimen masyarakat terhadap kasus pembobolan data nasabah Bank BSI pada Twitter menggunakan metode random forest dan Naïve Bayes. Jurnal Fasilkom, 13 (3), pp.614–620. doi: https://doi.org/10.37859/jf.v13i3.6478

[21] Octariadi, B.C., 2025. Penerapan algoritma (Naïve Bayes) untuk memprediksi penyakit diare. Jurnal Fasilkom, 15 (1), pp.49–56. doi: https://doi.org/10.37859/jf.v15i1.8993

[22] Hasanah, U., Soleh, A.M. and Sadik, K., 2024. Effect of random undersampling, oversampling, and SMOTE on the performance of cardiovascular disease prediction models. Jurnal Matematika, Statistika dan Komputasi,

P-ISSN : 2089-3353
E-ISSN : 2808-9162

21 (1), pp.88–102. doi: https://doi.org/10.20956/j.v21i1.35552

[23] Passa, R.S., Nurmaini, S. and Rini, D.P., 2023. Deteksi tumor otak pada magnetic resonance imaging menggunakan YOLOv7. Jurnal Ilmiah Matrik, 25 (2), pp.116–121. doi: https://doi.org/10.33557/jurnalmatrik.v25i2.240

[24] Noorizki, A.Z., Pratikno, H. and Kusumawati, W.I., 2024. Klasifikasi emosional ulasan pelanggan dengan pendekatan NLP menggunakan metode ensemble dan ROS. Techno.com, 23 (4). doi: https://doi.org/10.62411/tc.v23i4.1155