Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 4, Nomor 3,November 2025

MACHINE LEARNING OPTIMIZATION USING CORRELATION FEAUTER SELECTION AND SMOTE-ENN FOR EDUCATOR SENTIMENT

[1]Jupriadi*, [2]Anthony Anggarawan, [3]Hairani

[12] S1 Ilmu Komputer, Fakultas Teknik, Universitas Bumigora
Jl. Ismail Marzuki No.22, Cilinaya, Kec. Cakranegara, Kota Mataram, NTB,Indonesia

*e-mail: Jupeerrr@gmail.com[1], Anthony.anggrawan@universitasbumigora.ac.id[2] , Hairani@universitasbumigora.ac.id[3]

Received: 2025-08-22
Revised: 2025-09-30
Accepted: 2025-10-15

Page : 1-17

Abstrak : Tenaga pendidik memiliki peran strategis dalam kemajuan pendidikan nasional, sehingga pemahaman terhadap sentimen mereka penting dalam meningkatkan kesejahteraan dan kualitas layanan pendidikan. Dalam pengolahan data sentimen, seringkali ditemui tantangan seperti ketidakseimbangan data antara sentimen mayoritas dan minoritas, serta tingginya jumlah fitur yang menyebabkan dimensionalitas data menjadi besar.Tujuan penelitian ini adalah menganalisis sentimen attitude tenaga pendidik di Indonesia menggunakan metode klasifikasi Machine Learning, dengan pendekatan seleksi fitur berbasis korelasi dan penyeimbangan data melalui Synthetic Minority Over-sampling Technique–Edited Nearest Neighbours (SMOTE ENN). Model klasifikasi dibangun menggunakan algoritma Naïve Bayes dan Support Vector Machine (SVM).Hasil penelitian menunjukkan bahwa SVM memberikan akurasi lebih tinggi dibandingkan Naïve Bayes, baik pada data asli maupun setelah penerapan SMOTE ENN. Akurasi Naïve Bayes meningkat dari 61% menjadi 89% setelah seleksi fitur berbasis korelasi, sedangkan SVM meningkat dari 69% menjadi 97%. Penelitian ini membuktikan bahwa kombinasi SVM, SMOTE ENN, dan seleksi fitur berbasis korelasi mampu meningkatkan akurasi klasifikasi sentimen tenaga pendidik di Indonesia secara signifikan.

Kata kunci : analisis sentimen, tenaga pendidik, SVM, SMOTE, seleksi fitur

Abstract : *Educators play a strategic role in the progress of national education, so understanding their sentiments is important for improving their well-being and the quality of educational services. In sentiment data processing, challenges are often encountered, such as data imbalance between majority and minority sentiments, and a high number of features leading to high data dimensionality. The purpose of this study is to analyze the sentiment of Indonesian educators' attitudes using Machine Learning classification methods, with a correlation-based feature selection approach and data balancing through Synthetic Minority Over-sampling Technique–Edited Nearest Neighbours (SMOTE ENN). Classification models were built using the Naïve Bayes and Support Vector Machine (SVM) algorithms. The*

research results show that SVM provides higher accuracy compared to Naïve Bayes, both on the original data and after applying SMOTE ENN. Naïve Bayes' accuracy increased from 61% to 89% after correlation-based feature selection, while SVM's increased from 69% to 97%. This study proves that the combination of SVM, SMOTE ENN, and correlation-based feature selection can significantly improve the accuracy of sentiment classification for Indonesian educators

Keywords: *Sentiment analysis, Educators, SMOTE, Feature selection*

DOI: https://doi.org/10.63893/jetcom.v4i3.314

1 Pendahuluan (*or Introduction*)

Tenaga pendidik memiliki peran strategis dalam menentukan arah dan kualitas pendidikan suatu bangsa [1]. Di tengah kompleksitas sistem pendidikan di Indonesia, persepsi dan sikap para pendidik terhadap profesinya menjadi aspek penting yang perlu dipahami secara mendalam [2]. Sikap tersebut mencerminkan sejauh mana mereka merasa dihargai, didukung, dan mampu menjalankan peran mereka secara optimal dalam lingkungan kerja yang dinamis. Oleh karena itu, diperlukan pendekatan yang dapat menggambarkan sentimen mereka secara objektif dan berbasis data [3].

Dengan pesatnya perkembangan teknologi informasi, media sosial dan platform daring telah menjadi sarana bagi para tenaga pendidik untuk mengekspresikan opini, persepsi, dan perasaan mereka [4]. Data sentimen yang tersebar luas di media sosial, survei daring, dan forum diskusi menjadi sumber informasi yang kaya untuk dianalisis secara kuantitatif menggunakan metode analisis sentimen berbasis machine learning [5]. Namun, dalam pengolahan data sentimen, seringkali ditemui beberapa tantangan seperti **ketidakseimbangan data** antara sentimen mayoritas dan minoritas, serta tingginya jumlah fitur yang menyebabkan dimensionalitas data menjadi besar [6].

2 Tinjauan Literatur (*or Literature Review*)

(Beberapa penelitian sebelumnya telah memanfaatkan algoritma machine learning seperti Naïve Bayes dan Support Vector Machine (SVM) untuk mengklasifikasikan sentimen di berbagai bidang, seperti transportasi online, layanan e-commerce, dan opini pelanggan pada layanan publik. Penelitian ini menerapkan metode ensemble dengan SMOTE untuk klasifikasi sentimen pada pendidikan daring, dan berhasil mencapai akurasi tertinggi sebesar 79,62%**,** melebihi performa metode klasifikasi konvensional tanpa penanganan data tidak seimbang [7]. Penelitian yang dilakukan [8] juga menghadapi kendala dalam pengolahan data berdimensi besar yang mempengaruhi akurasi model. Penelitian yang dilakukan [9] studi hanya memanfaatkan teknik balancing data standar seperti SMOTE tanpa mengeliminasi noise, serta mengabaikan pentingnya seleksi fitur untuk mereduksi kompleksitas data. Penelitian [10] membandingkan Naïve Bayes dan SVM pada ulasan aplikasi Mobile JKN, mencatat peningkatan akurasi hingga 90 % setelah menggunakan SMOTE untuk mengatasi ketidakseimbangan data. Penelitian [11] memanfaatkan Naïve Bayes dipadukan dengan SMOTE pada ulasan pengguna by.U, berhasil mengoptimalkan akurasi dari 84,42 % menjadi 85,83 %. Pada penelitian [12] menunjukkan bahwa penerapan

SMOTE pada SVM untuk data media sosial X meningkatkan akurasi dari 71,41% menjadi 83,89%, serta menaikkan precision, recall, dan F1-score menjadi 0,84. Penelitian [13] menerapkan seleksi fitur berdasarkan information gain pada data sentimen produk e-commerce dan menemukan peningkatan performa klasifikasi meskipun belum memanfaatkan korelasi statistik fitur secara langsung

Dari penelitian sebelumnya, masih terdapat beberapa kekurangan yang perlu dilengkapi, khususnya pada teknik balancing data yang diterapkan umumnya masih terbatas pada metode oversampling biasa seperti SMOTE, tanpa melakukan proses kombinasi dengan metode pembersihan data seperti *Edited Nearest Neighbor* (ENN) yang mampu mengurangi noise pada data hasil oversampling [14]. Proses yang dilakukan penelitian sebelumnya jarang melakukan optimasi [15], melalui seleksi fitur berbasis korelasi yang bertujuan menghilangkan fitur tidak relevan dan mengurangi dimensionalitas data, hal ini sangat penting dalam meningkatkan efisiensi dan akurasi model [16]. Penelitian terkait analisis sentimen yang mengkaji attitude tenaga pendidik masih sangat terbatas, padahal data sentimen dari tenaga pendidik sudah cukup melimpah di berbagai media digital [17].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan optimasi metode machine learning berbasis kombinasi seleksi fitur korelasi dan SMOTE ENN untuk meningkatkan kinerja model klasifikasi sentimen. Seleksi fitur korelasi digunakan untuk menyaring fitur-fitur yang memiliki hubungan signifikan dengan target, sehingga dapat meningkatkan efisiensi komputasi dan mengurangi risiko overfitting. Sementara itu, SMOTE ENN diterapkan untuk proses penyeimbangan distribusi kelas sekaligus menghilangkan data yang berpotensi menjadi noise atau outlier, sehingga kualitas dataset menjadi lebih baik dan model menjadi lebih akurat.

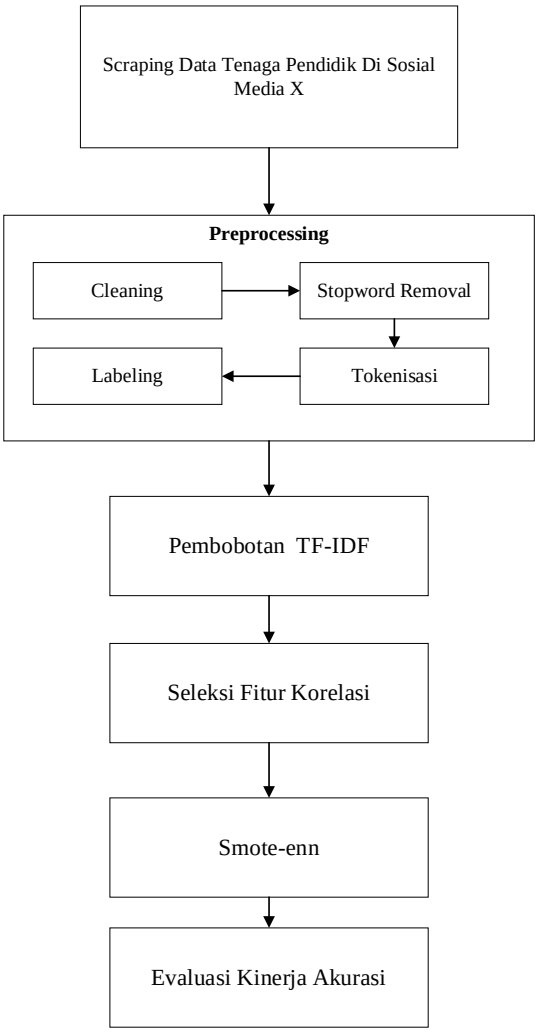
Penelitian ini juga membandingkan performa dua algoritma machine learning, yaitu Naïve Bayes dan SVM, untuk melihat bagaimana efektivitas optimasi dengan seleksi fitur dan balancing data berdampak pada akurasi dan ketahanan model. Dengan menerapkan pendekatan optimasi yang lebih lengkap ini, diharapkan hasil penelitian dapat memberikan kontribusi nyata dalam mengembangkan model analisis sentimen yang lebih akurat, efisien, dan adaptif terhadap karakteristik data yang tidak seimbang. Pada Tabel 1 berikut menyajikan beberapa penelitian terdahulu yang relevan dengan penelitian ini,tabel ini mencakup nama penulis, tahun publikasi, metode yang digunakan dan akurasi yang di capai. Perhatikan bahwa Gambar 1 *caption* di bawah, huruf awal kata huruf besar dan *bold* serta harus diacu dalam teks atau ada narasi yang menjelaskannya, gambar dan tulisan harus jelas atau resolusi tinggi.

Tabel 1 Tabel perbandingan penelitian

Penulis dan Tahun	SVM	Naïve Bayes	Smote	Seleksi fitur	Akurasi
Imran et al.(2022)	✓	✓	✓	-	79%
Gifari et al. (2022)	✓	-	-	-	85%
Septiani et al. (2024)	✓	✓	✓	-	90%
Luthfi et al. (2025)	-	✓	✓	-	85,83%
Andriyani et al. (2025)	✓	-	✓	-	83,89%
Madhumathi et al. (2022)	-	-	-	✓	98,5%
Rahat et al. (2020)	✓	✓	-	-	82,48%
Nurhaliza Agustina et al. (2024)	✓	-	-	-	86%
Purnawan et al. (2024)	-	-	✓	-	92%
Gunawan et al. (2018)	-	✓	-	-	77,78%
Penelitian Saya	✓	✓	✓	✓	97,19 %

3 Metode Penelitian (or Research Method)

Pada penelitian ini untuk menganalisis sentimen masyarakat terhadap tenaga pendidik indonesia dengan menggunakan algoritma Naïve Bayes dan SVM [18]. Seperti yang ditampilkan pada gambar 1.



Gambar 1 Kerangka penelitian.

Gambar 1. Kerangka penelitian menunjukkan tahapan dalam proses analisis sentimen terhadap data tenaga pendidik yang diperoleh dari sosial media X. Alur penelitian ini diawali dengan tahap *scraping data***,** yaitu proses pengambilan data secara otomatis dari platform sosial media yang menjadi sumber utama data penelitian. Data yang dikumpulkan merupakan opini atau sentimen yang berkaitan dengan tenaga pendidik.

Selanjutnya, data yang telah dikumpulkan akan masuk ke dalam tahap *preprocessing*. Pada tahap ini terdapat beberapa proses penting yaitu Cleaning data**,** pembersihan data dari karakter-karakter yang tidak relevan seperti simbol, angka, dan tanda baca yang tidak diperlukan [19]. Setelah itu di lakukan Stopword Removal, merupakan proses menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis seperti "dan", "yang", "di", dan sebagainya [20].Selanjutnya pada tahap tokenisasi, Proses ini memecah kalimat menjadi kata-kata yang akan mempermudah dalam proses analisis lebih lanjut. Tahap terakhir adalah Labeling, Memberikan label pada data sesuai dengan kategori sentimen seperti positif, negatif, atau netral [21].

Setelah data selesai diproses, langkah berikutnya adalah pembobotan TF-IDF (*Term Frequency - Inverse Document Frequency*) [22]**.** Proses ini bertujuan untuk memberikan bobot pada setiap kata berdasarkan seberapa sering kata tersebut muncul dalam dokumen dan seberapa unik kata tersebut dalam keseluruhan korpus data.

Tahapan selanjutnya adalah seleksi fitur korelasi**,** yang berfungsi untuk memilih fitur-fitur (kata) yang memiliki pengaruh signifikan terhadap hasil klasifikasi, dengan tujuan untuk mengurangi dimensi data dan meningkatkan performa model.

Kemudian, dilakukan proses penyeimbangan data menggunakan metode SMOTE ENN [23]. Penggunaan teknik ini mengkombinasikan Teknik *Synthetic Minority Oversampling Technique* (SMOTE) untuk menambah data minoritas secara sintetis dan *Edited Nearest Neighbors* (ENN) untuk menghapus data yang salah klasifikasi, sehingga distribusi data menjadi lebih seimbang [24].

Tahap terakhir adalah evaluasi kinerja akurasi, yaitu mengukur performa model klasifikasi yang dibangun berdasarkan data yang telah diproses, menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk menentukan tingkat keberhasilan dari sistem yang dikembangkan [25].

3.1 Pengambilan Data

Data terkait tenaga pendidik diambil dari platform sosial media tertentu (dalam hal ini disebut Sosial Media X) menggunakan teknik web scraping untuk mengumpulkan informasi [26].

3.2 Preprocessing

Pada tahap ini data akan di proses terlebih dahulu hingga siap digunakan untuk training model nantinya, tujuan dari tahap ini supaya data yang masuk pada model sudah bersih dari data duplikat, data yang kosong, tidak imbang, kata-kata tidak baku, dan kata yang tidak berguna yang bisa mempengaruhi kinerja dari model lebih buruk [27].

Pada tahap ini, data akan diproses terlebih dahulu sehingga siap digunakan untuk model pelatihan [28]. Proses ini meliputi penghapusan data duplikat untuk menghilangkan informasi tidak relevan (seperti tanda baca, angka, atau karakter lainnya) [29], penghapusan kata umum yang tidak membawa makna penting (seperti “dan”, “atau”, “dari”) [20], serta tokenisasi untuk memecah kalimat menjadi satuan kata atau frasa yang lebih kecil agar bisa diolah lebih lanjut [30], pemberian label pada data untuk mengkategorikan teks sesuai kelas yang relevan [21].

3.3 Pembobotan TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode yang digunakan untuk membobot kata-kata dalam dokumen [21]. Kata yang lebih sering muncul dalam satu dokumen tapi jarang muncul di dokumen lain akan diberi bobot lebih tinggi. Rumus TF-IDF ditunjukan pada pesamaan berikut.

$$\mathrm{TF} - \mathrm{IDF}(\mathrm{t},\mathrm{d}) = \mathrm{TF}(\mathrm{t},\mathrm{d}) \times \mathrm{IDF}(\mathrm{t}) \quad (1)$$

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \mathrm{TF}(\mathrm{t},\mathrm{d}) \times \mathrm{IDF}(\mathrm{t}) \quad (2)$$

$$\mathrm{IDF}(\mathrm{t}) \left(\frac{N}{df(t)}\right) \quad (3)$$

Persamaan (1) menunjukkan rumus utama dari metode TF-IDF *(Term Frequency–Inverse Document Frequency).* Persamaan ini mengalikan nilai *TF(t,d),* yaitu frekuensi kemunculan suatu kata *t* dalam dokumen *d*, dengan nilai IDF(t), yaitu pembobotan berdasarkan seberapa jarang kata tersebut muncul di seluruh kumpulan dokumen. Tujuannya adalah untuk menentukan seberapa penting suatu kata dalam dokumen tertentu dibandingkan dengan seluruh korpus [31].

Persamaan (2) mendefinisikan komponen *Term Frequency* (TF). TF dihitung sebagai rasio antara jumlah kemunculan kata *t* dalam dokumen *d* terhadap jumlah total kata dalam dokumen tersebut. Nilai ini merepresentasikan seberapa sering suatu kata muncul dalam satu dokumen [32].

Persamaan (3) adalah rumus untuk menghitung *Inverse Document Frequency* (IDF). IDF dihitung menggunakan logaritma dari perbandingan jumlah total dokumen (*N*) terhadap jumlah

dokumen yang mengandung kata *t* (*df(t)*). Semakin sedikit dokumen yang mengandung kata tersebut, semakin besar nilai IDF-nya, yang berarti kata tersebut lebih unik atau spesifik untuk dokumen tertentu [33].

3.4 Seleksi Fitur Korelasi

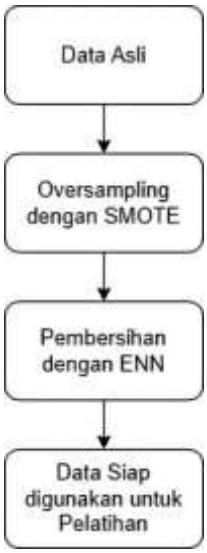
Pada tahap ini, dilakukan pemilihan fitur-fitur yang paling relevan dengan menggunakan korelasi antar fitur untuk mengurangi dimensi data dan meningkatkan kinerja model [34]. Rumus yang digunakan dalam seleksi fitur korelasi menggunakan *Pearson Correlation,* ditunjukan pada persamaan (4).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (4)$$

Persamaan (4) merupakan rumus koefisien korelasi Pearson (r) yang digunakan untuk mengukur tingkat hubungan linear antara dua variable [35]. Dalam konteks seleksi fitur, korelasi ini digunakan untuk mengidentifikasi fitur-fitur yang memiliki hubungan signifikan terhadap variabel target, sehingga fitur yang kurang relevan dapat dieliminasi untuk menyederhanakan model dan meningkatkan performa [36].

3.5 SMOTE ENN

SMOTE ENN (*Synthetic Minority Over-sampling Technique Edited Nearest Neighbour*), adalah gabungan dua teknik yang digunakan untuk mengatasi ketidakseimbangan kelas [37]. SMOTE berfungsi untuk membuat sampel sintetik dari kelas minoritas berdasarkan interpolasi antara sampel yang berdekatan [23]. ENN digunakan untuk membersihkan data dengan menghapus sampel outlier yang salah diklasifikasikan oleh tetangga terdekatnya [38]. Alur SMOTE ENN ditampilkan pada gambar 2.



Gambar 2. Alur Proses SMOTE ENN

Gambar 2. Alur Proses SMOTE ENN menunjukkan tahapan dalam penyeimbangan dan pembersihan data sebelum digunakan untuk pelatihan model klasifikasi.

Adapun penjelasan setiap langkahnya Data Asli Tahap awal berupa data mentah yang diperoleh dari proses pengambilan data, biasanya memiliki ketidakseimbangan distribusi antar kelas (misalnya, data sentimen positif jauh lebih banyak daripada negatif).

Oversampling dengan SMOTE Pada tahap ini, dilakukan penambahan data sintetis untuk kelas minoritas menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE)**.** SMOTE

bekerja dengan membuat sampel baru berdasarkan interpolasi nilai-nilai dari sampel minoritas yang ada, sehingga distribusi kelas menjadi lebih seimbang.

Pembersihan dengan ENN Setelah proses oversampling, data diproses menggunakan *Edited Nearest Neighbour* (ENN) yang bertujuan untuk menghapus data outlier atau data yang tidak konsisten. ENN akan mengeliminasi sampel yang kelasnya berbeda dari mayoritas tetangga terdekatnya, sehingga membantu meningkatkan kualitas data.

Data Siap Digunakan untuk Pelatihan Model data yang telah melalui proses SMOTE dan ENN kini lebih seimbang dan bersih. Data ini kemudian siap digunakan pada tahap pelatihan model klasifikasi, seperti Naïve Bayes atau SVM.

3.6 Evaluasi Kinerja

Pada tahap terakhir, model dievaluasi berdasarkan metrik kinerja seperti akurasi, presisi, dan recall untuk menilai seberapa baik model tersebut dalam melakukan klasifikasi atau prediksi [39]. Berikut rumus confusion matriks yang digunakan ditunjukan pada tabel 2.

Tabel 2 Tabel Confusion Matriks

	Prediksi *Positive*	*Negative*
Positive	*True Positive (TP)*	*False Positive (FP)*
Negative	*False Negative (FN)*	*True Negative (TN)*

Pada Tabel 2 *confusion matrix* di atas merupakan alat yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan cara membandingkan hasil prediksi model terhadap nilai sebenarnya. Tabel ini terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) [40]. True Positive menunjukkan jumlah data yang benar-benar positif dan diprediksi positif oleh model, sedangkan True Negative menunjukkan jumlah data yang benar-benar negatif dan berhasil diprediksi negative [40]. False Positive merupakan kesalahan ketika model memprediksi data sebagai positif padahal sebenarnya negative. Keempat komponen ini menjadi dasar dalam menghitung metrik evaluasi seperti akurasi, presisi, dan recall untuk menilai seberapa baik model dalam melakukan klasifikasi [40].

4 Hasil dan Pembahasan (or Results and Analysis)

Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma Naïve Bayes dan SVM melakukan klasifikasi sentimen terhadap data sosial media X yang membahas tenaga pendidik di Indonesia. Proses klasifikasi didahului oleh tahapan data preparation yang meliputi preprocessing, pembobotan dengan TF-IDF, seleksi fitur berbasis korelasi, dan penyeimbangan data menggunakan SMOTE dan ENN hingga tahapan Evaluasi Kinerja.

4.1 Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 1.508 data mentah yang diperoleh dari platform media sosial X. Kategori sentimen dalam dataset ini dibagi menjadi tiga kelas, yaitu: Positif (1), Netral (0), dan Negatif (-1). Penentuan kelas sentimen dilakukan secara manual melalui proses *labeling* oleh peneliti dengan mempertimbangkan konteks dan makna dari isi teks *tweet*. Proses labeling ini juga mengacu pada pedoman analisis sentimen serta dibantu oleh dua annotator untuk memastikan konsistensi dan validitas klasifikasi. Setiap entri pada dataset mencakup informasi seperti nama pengguna (anonim), tanggal unggahan, isi teks tweet, serta jumlah *likes* dan *retweets* yang diperoleh.

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025

4.2 Tahap Preprocessing

Tahap preprocessing dilakukan untuk mempersiapkan data sebelum masuk ke proses pembobotan dan klasifikasi. Untuk menggambarkan hasil preprocessing, berikut ditampilkan data sebelum dan data sesudah preprocessing,ditunjukan pada tabel 3.

Tabel 3 Data Sebelum Preprocessing

No	Teks Tweet Asli
1	Aplikasi bantu rakyat
2	Banyak tampil beranda donasi tampil beranda
3	Semoga berkah kelola aplikasi donatur
4	guru honorer di tempatku dibayar 300 ribu per bulan itu sangat tidak manusiawi
…	…
1508	Aplikasi crush buka padahal sudah di update

Pada tahap ini, karakter tidak penting seperti emotikon, tanda baca, simbol, dan URL dihapus, gambar 3 menyajikan tampilan *source code* yang digunakan untuk melakukan *data preprocessing*. Hasil *cleaning* data ditunjukan pada tabel 4.

Gambar 3 *Source Code Cleaning* Data

Tabel 4 Hasil Cleaning

No	Hasil Cleaning
1	aplikasi bantu rakyat
2	banyak tampil beranda donasi tampil beranda
3	semoga berkah kelola aplikasi donatur
4	guru honorer di tempatku dibayar 300 ribu per bulan itu sangat tidak manusiawi
…	…
1508	aplikasi crush buka padahal sudah di update

Pada tahap ini, kata-kata umum yang tidak memiliki pengaruh signifikan dihilangkan, gambar 4 menyajikan tampilan *source code* yang digunakan untuk penerapan *Stopword Removal*. Hasil *Stopword Removal* ditunjukan pada tabel 5.

Gambar 4 *Source Code Stopword Removal*

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025

Tabel 5 Hasil *Stopword Removal*

No	Hasil Stopword Removal
1	aplikasi bantu rakyat
2	banyak tampil beranda donasi tampil beranda
3	berkah kelola aplikasi donatur
4	guru honorer tempat bayar ribu bulan manusiawi
…	…
1508	aplikasi crush buka update

Teks yang telah dibersihkan dari stopword dipecah menjadi token (kata-kata terpisah), gambar 5 menunjukkan s*ource code* yang menjalankan proses pemotongan atau *tokenizing*. Hasil Tokenisasi di tunjukan pada tabel 6.

```
import nltk
from nltk.tokenize import word_tokenize
def tokenizing(text):
    """
    Memecah string teks menjadi daftar kata.
    Args:
        text (str): String teks yang ingin di-tokenisasi.
    Returns:
        list: Daftar kata (tokens) dari teks yang diinput.
    """
    text = nltk.tokenize.word_tokenize(text)
    return text
```

Gambar 5 Source Code Tokenisasi

Tabel 6 Hasil Tokenisasi

No	Hasil Tokenisasi
1	['aplikasi', 'bantu', 'rakyat']
2	['banyak', 'tampil', 'beranda', 'donasi', 'tampil', 'beranda']
3	['berkah', 'kelola', 'aplikasi', 'donatur']
4	['guru', 'honorer', 'tempat', 'bayar', 'ribu', 'bulan', 'manusiawi']
…	…
1508	['aplikasi', 'crush', 'buka', 'update']

Setelah data dibersihkan, setiap tweet diberikan label sentimen berdasarkan isi teks. Gambar 6 menunjukkan *source code* untuk proses *labeling* data, yang mengategorikan *tweet* sebagai positif, negatif, atau netral.Hasil Labeling data ditunjukan pada tabel 7.

```
import pandas as pd
import re
positive_data = pd.read_csv('positive.txt', delimiter="\t", header=None)
negative_data = pd.read_csv('negative.txt', delimiter="\t", header=None)
def check_positive_negative(stemming_words):
    """
    Memeriksa sentimen dari kata-kata yang sudah di-stemming berdasarkan kamus kata.
    Args:
        stemming_words (str): Teks yang sudah di-stemming.
    Returns:
        str: Label sentimen ("Positive", "Negative", atau "Neutral").
    """
    words = stemming_words.split()
    positive_count = 0
    negative_count = 0
    for word in words:
        if word in positive_data[0].values:
            positive_count += 1
        if word in negative_data[0].values:
            negative_count += 1
    if positive_count > negative_count:
        return "Positive"
    elif positive_count == negative_count:
        return "Neutral"
    else:
        return "Negative"
```

Gambar 6 *Source Code* Labeling Data

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025

Tabel 7 Hasil Labeling Data

No	Teks Cleaning	Label
1	aplikasi bantu rakyat	Positif
2	banyak tampil beranda donasi tampil beranda	Negatif
3	semoga berkah kelola aplikasi donatur	Positif
4	guru honorer di tempatku dibayar 300 ribu per bulan itu sangat tidak manusiawi	Negatif
…	…	…
1508	aplikasi crush buka padahal sudah di update	Negatif

Dari hasil tersebut terlihat bahwa data telah dibersihkan dan disederhanakan hanya menjadi kata-kata inti yang memiliki makna penting. Proses ini bertujuan untuk mengurangi kompleksitas data dan memfokuskan pada fitur-fitur penting yang akan digunakan dalam pembobotan TF-IDF dan klasifikasi.

4.3 Pembobotan TF-IDF

TF-IDF digunakan untuk mengkonversi data teks menjadi representasi numerik yang dapat diproses oleh model pembelajaran mesin. Proses ini menghasilkan bobot numerik untuk setiap kata berdasarkan frekuensi kemunculannya dan signifikansinya dalam seluruh dokumen. Visualisasi hasil pembobotan ditunjukkan pada tabel 8.

Tabel 8 Hasil TF-IDF

No	Term	TF				DF	IDF = log n/DF			
		A	B	C	D		A	B	C	D
1	aplikasi	1	0	1	1	3	0.124	0	0.124	0.124
2	bantu	1	0	0	0	1	0.602	0	0	0
3	rakyat	1	1	0	1	3	0.124	0.124	0	0.124
4	banyak	0	1	0	0	1	0	0.602	0	0
5	tampil	0	1	0	0	1	0	0.602	0	0
6	beranda	0	1	0	0	1	0	0.602	0	0
7	donasi	0	1	0	1	2	0	0.301	0	0.301
8	moga	0	0	1	0	1	0	0	0.602	0
9	berkah	0	1	1	0	2	0	0.301	0.301	0
10	kelola	0	0	1	0	1	0	0	0.602	0
11	donatur	0	0	1	1	2	0	0	0.301	0.301
12	crush	0	0	0	1	1	0	0	0	0.602
13	buka	0	0	0	1	1	0	0	0	0.602
14	padahal	0	0	0	1	1	0	0	0	0.602
15	update	0	0	0	1	1	0	0	0	0.602

4.4 Seleksi Fitur Korelasi

Seleksi fitur dilakukan menggunakan metode korelasi untuk menyaring fitur-fitur yang memiliki hubungan signifikan dengan kelas target. Dari 4676 fitur awal yang dihasilkan oleh proses ekstraksi TF-IDF, dan hanya 3000 fitur yang dipertahankan berdasarkan nilai korelasi tertinggi

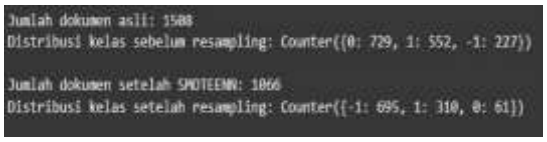
terhadap target. Proses ini bertujuan untuk mengurangi dimensionalitas data, meningkatkan efisiensi komputasi, serta mengoptimalkan akurasi model klasifikasi. Visualisasi 15 hasil seleksi fitur terbaik yang diperoleh ditampilkan pada Tabel 9.

Tabel 9 Hasil Seleksi Fitur Korelasi

No	*Feature*	*Chi2_score*
1	kurang	34.545245
2	salah	20.970373
3	selamat	17.893703
4	sehat	15.497975
5	terima	15.115159
6	lengkap	14.928489
7	seksual	13.394112
8	malang	12.625925
9	semangat	12.111629
10	serius	11.515996
11	subsidi	10.252700
12	prestasi	10.104908
13	lu	9.879105
14	lowong	9.833992
15	keras	8.795974

4.5 Penyeimbangan Data Menggunakan SMOTE dan ENN

Untuk mengatasi ketidakseimbangan kelas, digunakan metode Synthetic Minority Oversampling Technique (SMOTE) yang diikuti oleh Edited Nearest Neighbor (ENN). SMOTE menghasilkan sampel sintetik untuk kelas minoritas, sementara ENN menghapus data yang berpotensi sebagai noise berdasarkan kedekatan tetangga. Kombinasi keduanya menghasilkan data latih yang lebih seimbang dan bersih. Pembagian data dilakukan menggunakan *train_test_split* dengan rasio 80:20.Pada tahapan SMOTE ENN data awal yaitu 1508 data menjadi 1066 data di karenakan proses oversampling dan undersampling,hasil SMOTE ENN di tampilkan pada Gambar 7.



Gambar 7 Hasil SMOTE ENN

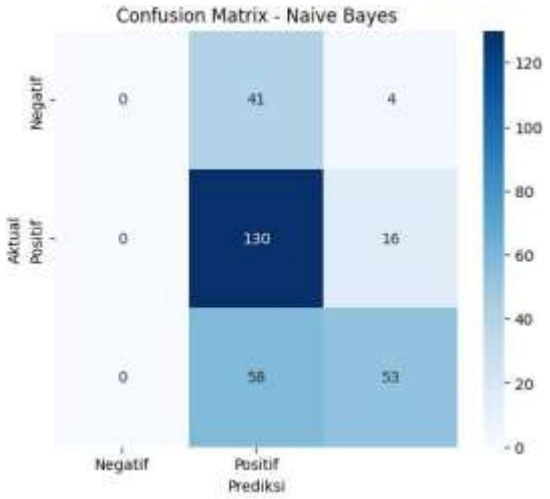
4.6 Evaluasi Model

Model dievaluasi berdasarkan metrik akurasi, presisi, dan recall dari confusion matrix. Tiga skenario evaluasi dilakukan: menggunakan data original tanpa seleksi fitur dan ENN, dan menggunakan data hasil SMOTE ENN tanpa seleksi fitur, dan terakhir menggunakan data hasil SMOTE ENN dengan Seleksi Fitur.

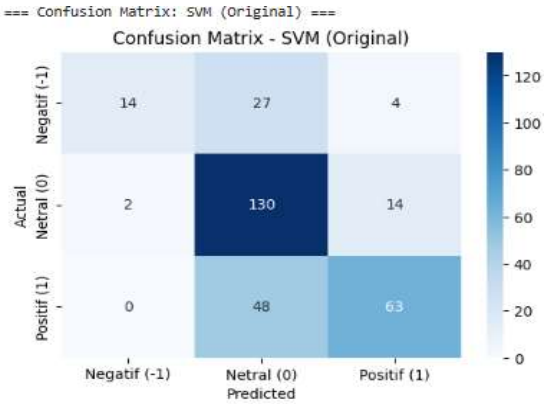
Evaluasi dengan Data Original

Pengujian pertama dilakukan dengan menggunakan data asli tanpa melalui proses seleksi fitur dan tanpa penerapan metode Edited Nearest Neighbor (ENN). Hasil evaluasinya ditampilkan dalam gambar 8 dan gambar 9 berikut:

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025



Gambar 8 Confusion Matriks dengan data Original Naive Bayes



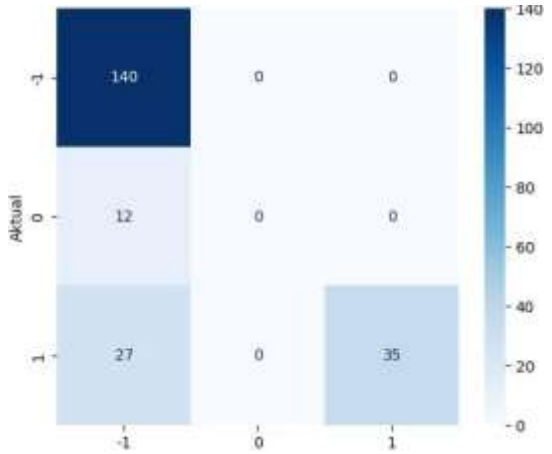
Gambar 9 Confusion Matriks dengan data Original SVM

Pada skenario pertama, data diuji tanpa proses seleksi fitur maupun balancing. Hasil menunjukkan bahwa algoritma SVM menghasilkan akurasi sebesar 68,54%, lebih tinggi dibandingkan Naïve Bayes sebesar 60,59%. Presisi tertinggi pada SVM berada di kelas negatif dengan nilai 87,50%, sementara Naïve Bayes tidak mampu mengklasifikasikan kelas negatif (presisi 0%). SVM juga unggul dalam *recall* pada seluruh kelas, khususnya kelas positif dengan nilai 77,78%, dibandingkan Naïve Bayes yang hanya mencapai 72,60%.

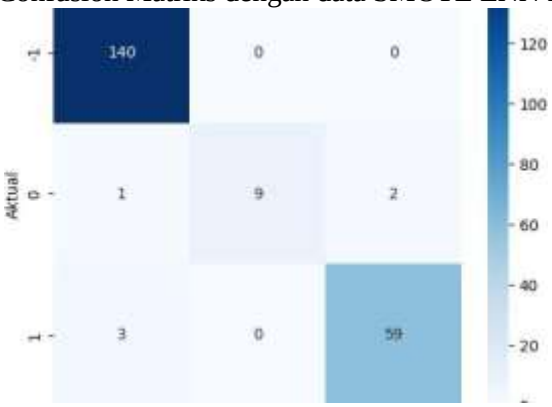
Hasil menunjukkan bahwa SVM memiliki performa lebih baik dibandingkan Naïve Bayes, khususnya dalam mengenali kelas negatif yang sebelumnya tidak terdeteksi oleh Naïve Bayes.

Evaluasi dengan Data SMOTE + ENN (Tanpa Seleksi Fitur) Pada tahap ini, data diuji dengan metode SMOTE ENN tetapi tanpa seleksi fitur. Berikut adalah hasil evaluasi,Berikut adalah hasil evaluasi pada gambar 10 dan gambar 11:

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025



Gambar 10 Confusion Matriks dengan data SMOTE ENN Naive Bayes

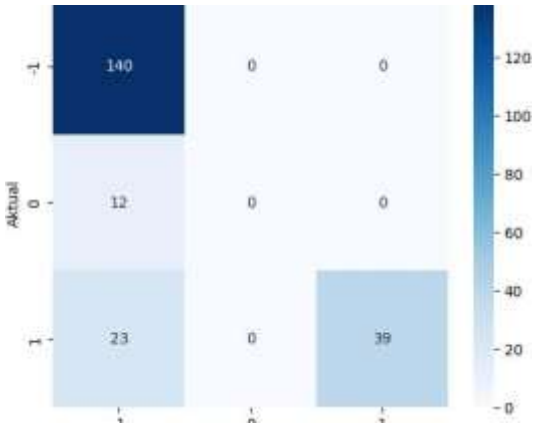


Gambar 11 Confusion Matriks dengan data SMOTE ENN SVM

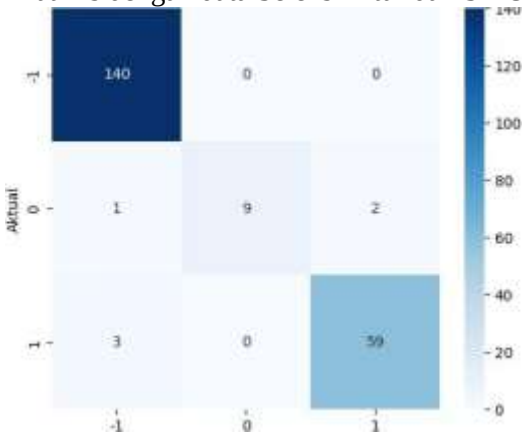
Setelah penerapan metode *balancing* SMOTE ENN, performa kedua model mengalami peningkatan signifikan. Akurasi Naïve Bayes meningkat drastis menjadi 81,87%, sementara SVM mencatat akurasi tertinggi sebesar 97,19%. Pada model SVM, presisi dan *recall* untuk seluruh kelas meningkat secara signifikan, dengan presisi kelas positif mencapai 97,22% dan *recall* kelas negatif sebesar 100%. Naïve Bayes juga menunjukkan peningkatan pada kelas negatif dengan *recall* 100%, namun tetap mengalami kendala dalam proses mengklasifikasikan kelas netral (presisi dan *recall* 0%). Hasil menunjukan bahwa balancing data mampu mengatasi ketidakseimbangan kelas dan meningkatkan generalisasi model.

Evaluasi dengan Data SMOTE ENN + Seleksi Fitur Skenario terakhir menggabungkan seleksi fitur dan metode *balancing* SMOTE ENN. Hasil menunjukkan bahwa pendekatan ini memberikan kinerja terbaik untuk kedua model.Berikut adalah hasil evaluasi pada gambar 12 dan gambar 13

Journal of Engineering, Technology and Computing (JETCom) E-ISSN: 2828-0512
Volume 1, Nomor 1, September 2025



Gambar 12 Confusion Matriks dengan data Seleksi Fitur dan SMOTE ENN Naive Bayes



Gambar 13 Confusion Matriks dengan data Seleksi Fitur dan SMOTE ENN SVM

Naïve Bayes mencapai akurasi 84%, sedangkan SVM mempertahankan performa unggul dengan akurasi 97,19%. Pada model SVM, seluruh metrik evaluasi menunjukkan hasil sangat baik, dengan presisi kelas positif dan negatif mencapai 100% dan 97,22%, serta *recall* kelas negatif tetap 100%. Naïve Bayes menunjukkan peningkatan pada presisi dan *recall* kelas positif, namun tetap gagal mengklasifikasikan kelas netral secara efektif. Berdasarkan hasil pengujian pada algoritma *Naïve Bayes* dan *Support Vector Machine* dengan menggunakan seleksi fitur berbasis korelasi dan metode SMOTE ENN untuk menyeimbang data serta melakukan evaluasi berdasarkan akurasi, presisi,dan recall,didapatkan perbandingan kedua algoritma dengan mengambil percobaan terbaik yaitu SVM sebagai berikut.

Gambar 14 Perbandingan Akurasi Naïve Bayes dan SVM

Berdasarkan gambar 14 perbandingan akurasi antara model *Naïve Bayes* dan SVM, dapat disimpulkan bahwa model SVM memiliki performa yang lebih baik dalam hal akurasi dibandingkan dengan *Naïve Bayes*. Pada data original, akurasi SVM sebesar 69%, sedangkan *Naïve Bayes* hanya mencapai 61%. Setelah diterapkan metode SMOTE ENN, akurasi kedua model meningkat secara signifikan, dengan Naïve Bayes mencapai 82% dan SVM meningkat drastis hingga 97,19%. Penggunaan kombinasi Feature Selection, SMOTE ENN semakin meningkatkan akurasi Naïve Bayes menjadi 84%, dan metode SVM meningkat menjadi 97,19%.

Hal ini menunjukkan bahwa teknik pengolahan data seperti SMOTE ENN berperan penting dalam meningkatkan performa model, terutama bagi *Naïve Bayes*. Namun, secara keseluruhan, model SVM tetap menunjukkan akurasi tertinggi, menjadikannya pilihan yang lebih efektif dalam mengklasifikasikan data dengan benar pada dataset yang digunakan.

5 Kesimpulan (*or Conclusion*)

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa tujuan utama penelitian, yaitu mengevaluasi performa model klasifikasi dalam analisis sentimen terhadap tenaga pendidik di Indonesia, telah tercapai dengan baik. Model Support Vector Machine (SVM) menunjukkan performa yang lebih unggul dibandingkan dengan Naïve Bayes, baik pada data asli maupun setelah diterapkan teknik penyeimbangan data seperti SMOTE ENN. Hal ini menunjukkan bahwa pemilihan algoritma yang tepat sangat berpengaruh terhadap hasil klasifikasi. Selain itu, metode pemrosesan data seperti seleksi fitur dan teknik penyeimbangan data terbukti meningkatkan akurasi, khususnya pada algoritma *Naïve Bayes* yang lebih sensitif terhadap pemrosesan fitur. Analisis sentimen juga berhasil mengidentifikasi faktor-faktor utama yang memengaruhi opini publik terhadap tenaga pendidik di Indonesia, seperti kebijakan pendidikan, kesejahteraan guru, sistem penggajian, serta dukungan pemerintah. Penelitian ini tidak hanya menyoroti efektivitas teknik klasifikasi dalam pengolahan data teks, tetapi juga memberikan pemahaman yang lebih mendalam mengenai isu-isu yang dihadapi tenaga pendidik.

Penelitian ini memiliki kelemahan dalam mengeksplorasi algoritma lain seperti Random Forest, XGBoost, LSTM, atau BERT guna meningkatkan akurasi klasifikasi dan pemahaman konteks dalam analisis sentimen.

Referensi (*Reference*)

[1] [1] M. Munir, “The Role Of The Teacher Determines The Quality Of Education In Indonesia,” pp. 1–8, 2021.
[2] A. Iskandar, I. Rusydi, H. Amin, M. N. Hakim, and H. A. Haqq, “Strategic Management in Improving the Quality of Education in Boarding School,” AL-ISHLAH J. Pendidik., vol. 14, no. 4, pp. 7229–7238, Dec. 2022, doi: 10.35445/alishlah.v14i4.2075.
[3] S. Nurhayati and S. Musa, “Teaching With Purpose: Indonesian Educators’ Response to The Challenges of Society 5.0,” Proc. Int. Conf. Res. Educ. Sci., vol. 10, no. 1, pp. 360–372, 2024.
[4] M. Sofi-Karim, A. O. Bali, and K. Rached, “Online education via media platforms and applications as an innovative teaching method,” Educ. Inf. Technol., vol. 28, no. 1, pp. 507–523, Jan. 2023, doi: 10.1007/S10639-022-11188-0/METRICS.
[5] F. Aftab et al., “A Comprehensive Survey on Sentiment Analysis Techniques,” Int. J. Technol., vol. 14, no. 6, pp. 1288–1298, 2023, doi: 10.14716/ijtech.v14i6.6632.
[6] R. Obiedat et al., “Sentiment Analysis of Customers’ Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution,” IEEE Access, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.
[7] M. Imran, S. Hina, and M. M. Baig, “Analysis of Learner’s Sentiments to Evaluate Sustainability of Online Education System during COVID-19 Pandemic,” Sustain., vol. 14, no. 8, pp. 1–18, 2022, doi: 10.3390/su14084529.

[8] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," J. Inf. Technol., vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.
[9] P. P. Putra, M. K. Anam, A. S. Chan, A. Hadi, N. Hendri, and A. Masnur, "Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques," J. Appl. Data Sci., vol. 6, no. 2, pp. 936–951, 2025, doi: 10.47738/jads.v6i2.618.
[10] E. Septiani, T. M. Akhriza, and M. Husni, "Comparison of the Accuracy Between Naive Bayes Classifier and Support Vector Machine Algorithms for Sentiment Analysis in Mobile JKN Application Reviews," vol. 1, no. 1, pp. 21–32, 2024.
[11] A. H. Luthfi, A. Faqih, and G. Dwilestari, "Enhancing Model Accuracy in Sentiment Analysis of the by . U Application Using Naïve Bayes and SMOTE Techniques," vol. 4, no. 2, 2025.
[12] D. Andriyani, Ahmad Faqih, and Sandy Eka Permana, "The Effect of SMOTE Application on Support Vector Machine Performance in Sentiment Classification on Imbalanced Datasets," J. Artif. Intell. Eng. Appl., vol. 4, no. 2, pp. 752–757, 2025, doi: 10.59934/jaiea.v4i2.742.
[13] R. Madhumathi, A. M. Kowshalya, and R. Shruthi, "Assessment of Sentiment Analysis Using Information Gain Based Feature Selection Approach," Comput. Syst. Sci. Eng., vol. 43, no. 2, pp. 849–860, 2022, doi: 10.32604/csse.2022.023568.
[14] M. Mukherjee and M. Khushi, "Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features," Appl. Syst. Innov., vol. 4, no. 1, pp. 1–15, 2021, doi: 10.3390/asi4010018.
[15] I. K. A. Purnawan, A. D. Wibawa, A. Kurniawati, and M. H. Purnomo, "Optimizing Diabetic Neuropathy Severity Classification Using Electromyography Signals Through Synthetic Oversampling Techniques," J. Nas. Pendidik. Tek. Inform., vol. 13, no. 3, pp. 681–690, 2024, doi: 10.23887/janapati.v13i3.85675.
[16] M. Muthukrishnan, S. Andavar, and R. S. P. Raj, "A Fused Feature Selection Technique for Enhanced Sentiment Analysis Using Deep Learning," Brazilian Arch. Biol. Technol., vol. 67, pp. 1–16, 2024, doi: 10.1590/1678-4324-2024240183.
[17] J. Zhou and J. min Ye, "Sentiment analysis in education research: a review of journal publications," Interact. Learn. Environ., vol. 31, no. 3, pp. 1252–1264, Apr. 2023, doi: 10.1080/10494820.2020.1826985;PAGE:STRING:ARTICLE/CHAPTER.
[18] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019, pp. 266–270, 2020, doi: 10.1109/SMART46866.2019.9117512.
[19] P. Edastama, A. S. Bist, and A. Prambudi, "Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm," Int. J. Cyber IT Serv. Manag., vol. 1, no. 2, pp. 159–172, 2021, doi: 10.34306/ijcitsm.v1i2.46.
[20] A. Sukarno Hatta, "Clustering Pada Data Sentimen Penggunaan Transportasi Online Menggunakan Algoritma Spectral Clustering," e-Proceeding Eng., vol. 8, no. 6, p. 11945, 2021.
[21] M. Soleimani, A. Intezari, and D. J. Pauleen, "Mitigating cognitive biases in developing ai-assisted recruitment systems: A knowledge-sharing approach," Int. J. Knowl. Manag., vol. 18, no. 1, pp. 1–18, 2022, doi: 10.4018/IJKM.290022.
[22] B. M. Iqbal, K. M. Lhaksmana, and E. B. Setiawan, "2024 Presidential Election Sentiment Analysis in News Media Using Support Vector Machine," J. Comput. Syst. Informatics, vol. 4, no. 2, pp. 397–404, 2023, doi: 10.47065/josyc.v4i2.3051.
[23] M. A. Latief, L. R. Nabila, W. Miftakhurrahman, S. Ma'rufatullah, and H. Tantyoko, "Handling Imbalance Data using Hybrid Sampling SMOTE ENN in Lung Cancer Classification," Int. J. Eng. Comput. Sci. Appl., vol. 3, no. 1, pp. 11–18, 2024, doi: 10.30812/ijecsa.v3i1.3758.
[24] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique

(SMOTE) Strategies," vol. 8, no. September, pp. 1310–1318, 2024.
[25] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," Int. J. Electr. Comput. Eng., vol. 11, no. 3, pp. 2407–2413, 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.
[26] S. Kumar, J. Thakur, D. Ekka, and I. Sahu, "Web Scraping Using Python," Int. J. Adv. Eng. Manag., vol. 4, no. 9, p. 235, 2022, doi: 10.35629/5252-0409235237.
[27] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," IEEE Access, vol. 12, no. May, pp. 80698–80730, 2024, doi: 10.1109/ACCESS.2024.3406748.
[28] A. Rahman and M. G. Muktadir, "SPSS: An Imperative Quantitative Data Analysis Tool for Social Science Research," Int. J. Res. Innov. Soc. Sci., vol. 05, no. 10, pp. 300–302, 2021, doi: 10.47772/ijriss.2021.51012.
[29] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," Pamukkale Univ. J. Eng. Sci., vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.
[30] N. Ebrahimiyan, M. Lotfi Ghahroud, S. Bastani, A. Abadi, and F. Jafari, "Tokenization and its application in different countries," J. FinTech Artif. Intell., vol. 2021, no. 1, pp. 14–019, 2021, doi: 10.47277/JFAI/1(1)019.
[31] M. T. Mohammed and O. F. Rashid, "Document retrieval using term frequency inverse sentence frequency weighting scheme," Indones. J. Electr. Eng. Comput. Sci., vol. 31, no. 3, pp. 1478–1485, 2023, doi: 10.11591/ijeecs.v31.i3.pp1478-1485.
[32] I. Arroyo-Fernández, C. F. Méndez-Cruz, G. Sierra, J. M. Torres-Moreno, and G. Sidorov, "Unsupervised sentence representations as word information series: Revisiting TF–IDF," Comput. Speech Lang., vol. 56, pp. 107–129, 2021, doi: 10.1016/j.csl.2019.01.005.
[33] C. A. Nurhaliza Agustina, R. Novita, Mustakim, and N. E. Rozanda, "The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm," Procedia Comput. Sci., vol. 234, pp. 156–163, 2024, doi: 10.1016/j.procs.2024.02.162.
[34] C. P. Vandana and A. A. Chikkamannur, "Feature selection: An empirical study," Int. J. Eng. Trends Technol., vol. 69, no. 2, pp. 165–170, 2021, doi: 10.14445/22315381/IJETT-V69I2P223.
[35] P. Rani, R. Kumar, and A. Jain, "A Hybrid Approach for Feature Selection Based on Correlation Feature Selection and Genetic Algorithm," Int. J. Softw. Innov., vol. 10, no. 1, pp. 1–17, 2022, doi: 10.4018/IJSI.292028.
[36] H. K. Bhuyan, C. Chakraborty, S. K. Pani, and V. Ravi, "Feature and Subfeature Selection for Classification Using Correlation Coefficient and Fuzzy Model," IEEE Trans. Eng. Manag., vol. 70, no. 5, pp. 1655–1669, 2023, doi: 10.1109/TEM.2021.3065699.
[37] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 8, pp. 585–590, 2023, doi: 10.14569/IJACSA.2023.0140864.
[38] S. K. Umar, S. Kumari, P. Samui, and D. Kumar, "A Liquefaction Study Using ENN, CA, and Biogeography Optimized-Based ANFIS Technique," Int. J. Appl. Metaheuristic Comput., vol. 13, no. 1, pp. 1–23, 2021, doi: 10.4018/ijamc.290535.
[39] M. Heydarian, T. E. Doyle, and R. Samavi, "Multi-Label Confusion Matrix," IEEE Access, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.
[40] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," Int. J. Educ. Integr., vol. 19, no. 1, pp. 1–16, 2023, doi: 10.1007/s40979-023-00140-5.