

Analisis Sentimen Terhadap Program Makan Bergizi Gratis Menggunakan Metode *Logistic Regression*

DOI: <http://dx.doi.org/10.35889/progresif.v22i2.3610>

Creative Commons License 4.0 (CC BY –NC)



Ari Budi Indrawan^{1*}, Donny Maulana², M. Zubair Abdurrohmah³

Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia

*e-mail *Corresponding Author*: Aribudi17@mhs.pelitabangsa.ac.id

Abstract

This study aims to analyze public sentiment toward the Free Nutritious Meal Program (MBG), a policy implemented by the Indonesian government to improve the nutritional quality of students. The data used consist of 1,440 tweets collected through a scraping process from the X/Twitter platform. The data processing stages include preprocessing steps such as case folding, cleaning, tokenizing, stopword removal, and stemming using the Sastrawi library. Furthermore, the text data are transformed into numerical representations using the TF-IDF method and classified using the Logistic Regression algorithm. To enhance the model's performance, the SMOTE technique is applied to address data imbalance, along with GridSearchCV for parameter optimization. The results indicate that the Logistic Regression model achieves excellent performance, with an Accuracy of 98.96%, Precision of 99.14%, Recall of 98.10%, and an F1-Score of 98.61%. This study is expected to provide an objective overview of public perception and serve as a reference for policy evaluation and decision-making.

Keywords: *Sentiment Analysis; Free Nutritious Meal Program; Logistic Regression; Text mining; NLP.*

Abstrak

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG) yang merupakan kebijakan pemerintah Indonesia dalam meningkatkan kualitas gizi peserta didik. Data yang digunakan berupa 1.440 tweet yang diperoleh melalui proses scraping dari platform X/Twitter. Tahapan pengolahan data meliputi *preprocessing* yang terdiri dari *case folding*, *cleaning*, *tokenizing*, *stopword removal*, dan *stemming* menggunakan library Sastrawi. Selanjutnya, data teks diubah menjadi representasi numerik menggunakan metode *TF-IDF* dan diklasifikasikan menggunakan algoritma *Logistic Regression*. Untuk meningkatkan performa model, diterapkan teknik *SMOTE* dalam mengatasi ketidakseimbangan data serta *GridSearchCV* untuk optimasi parameter. Hasil penelitian menunjukkan bahwa model *Logistic Regression* memiliki kinerja yang sangat baik dengan akurasi sebesar 98,96%, *Precision* 99,14%, *Recall* 98,10%, dan *F1-Score* 98,61%. Penelitian ini diharapkan dapat memberikan gambaran objektif mengenai persepsi masyarakat serta menjadi bahan evaluasi dalam pengambilan kebijakan.

Kata kunci: *Analisis Sentimen; Makan Bergizi Gratis; Logistic Regression; Text mining; NLP.*

1. Pendahuluan

Program Makan Bergizi Gratis (MBG) merupakan salah satu kebijakan strategis pemerintah Indonesia yang bertujuan untuk meningkatkan kualitas sumber daya manusia melalui pemenuhan kebutuhan gizi peserta didik. Program ini memiliki peran penting dalam meningkatkan konsentrasi belajar, kesehatan, serta kehadiran siswa di sekolah. Selain itu, program MBG juga berkontribusi dalam mendorong pertumbuhan ekonomi lokal melalui keterlibatan UMKM sebagai penyedia makanan bergizi [1]. Oleh karena itu, keberhasilan implementasi program ini menjadi aspek penting dalam pembangunan nasional menuju Indonesia Emas 2045. Namun demikian, implementasi program MBG masih menghadapi

berbagai tantangan, seperti distribusi yang belum merata, kualitas makanan yang tidak konsisten, serta lemahnya pengawasan pelaksanaan program [2]. Kondisi ini memunculkan beragam opini masyarakat yang banyak disampaikan melalui media sosial, khususnya platform X (*Twitter*). Volume data yang besar dan tidak terstruktur menyebabkan analisis manual menjadi tidak efektif dan berpotensi menghasilkan bias subjektif. Oleh karena itu, diperlukan pendekatan berbasis teknologi seperti *Natural Language Processing* (NLP) dan *text mining* untuk menganalisis opini publik secara otomatis dan objektif [3].

Penelitian terkait analisis sentimen terhadap kebijakan pemerintah telah banyak dilakukan untuk memahami persepsi publik berbasis data media sosial. Fitrianti dan Yudhistira [4] menganalisis sentimen masyarakat terhadap dinamika politik dalam konteks pemilihan kepala daerah menggunakan metode *Naïve Bayes*. Hasil penelitian menunjukkan bahwa media sosial dapat menjadi indikator opini publik terhadap kebijakan dan aktor pemerintah. Selanjutnya, Agustina dan Herliana [5] mengkaji sentimen terhadap kebijakan efisiensi anggaran pemerintah menggunakan pendekatan *text mining* dan NLP, yang mampu memberikan gambaran kuantitatif mengenai respons masyarakat terhadap kebijakan pemerintah. Penelitian lain oleh Setyawan et al. [6] menunjukkan bahwa metode *machine learning* seperti *Support Vector Machine* dan *Naïve Bayes* efektif dalam mengklasifikasikan opini publik terhadap isu-isu pemerintahan di media sosial.

Selain itu, beberapa penelitian juga telah menerapkan metode *deep learning* dalam analisis sentimen kebijakan publik. Penelitian oleh Febri Haerani [7] menggunakan metode *Gated Recurrent Unit* (GRU) dan menunjukkan bahwa mayoritas masyarakat memberikan sentimen positif terhadap program MBG dengan nilai *F1-Score* mencapai 91% untuk sentimen positif dan 88% untuk sentimen negatif. Penelitian lain menggunakan metode *Long Short-Term Memory* (LSTM) menunjukkan bahwa model mampu mencapai akurasi sebesar 91,71% dengan dominasi sentimen positif sebesar 73,4% [8]. Hasil ini menunjukkan bahwa pendekatan *deep learning* memiliki kemampuan yang baik dalam memahami konteks teks, namun membutuhkan sumber daya komputasi yang lebih besar dan kompleksitas yang tinggi. Meskipun berbagai metode telah digunakan, penelitian-penelitian sebelumnya umumnya belum mengintegrasikan teknik penyeimbangan data seperti *SMOTE* serta optimasi parameter menggunakan *GridSearchCV* dalam satu kerangka kerja yang sistematis. Selain itu, penelitian yang secara khusus mengkaji sentimen terhadap kebijakan Program Makan Bergizi Gratis dengan pendekatan yang teroptimasi masih terbatas. Hal ini menunjukkan adanya gap dalam pengembangan model klasifikasi yang tidak hanya akurat, tetapi juga stabil dan robust.

Berdasarkan gap tersebut, penelitian ini mengusulkan pendekatan analisis sentimen menggunakan algoritma *Logistic Regression* yang dikombinasikan dengan metode *TF-IDF*, *SMOTE*, dan *GridSearchCV*. *Logistic Regression* dipilih karena memiliki kemampuan yang baik dalam memodelkan hubungan antar fitur pada data teks. Penggunaan *SMOTE* bertujuan untuk mengatasi ketidakseimbangan data, sedangkan *GridSearchCV* digunakan untuk memperoleh parameter terbaik secara sistematis. Kebaruan penelitian ini terletak pada integrasi metode tersebut dalam analisis sentimen terhadap kebijakan Program Makan Bergizi Gratis, sehingga diharapkan mampu menghasilkan model dengan tingkat akurasi yang lebih tinggi dan stabil.

3. Metodologi

Penelitian ini menggunakan pendekatan *text mining* dan *machine learning* untuk menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG). Data penelitian diperoleh dari platform media sosial X/*Twitter* melalui proses scraping menggunakan kata kunci yang berkaitan dengan MBG [4]. *Dataset* yang digunakan dalam penelitian ini berjumlah 1.440 data yang kemudian dibagi menjadi data latih sebesar 80% (1.152 data) dan data uji sebesar 20% (288 data). Proses *training* dilakukan menggunakan data latih dengan penerapan teknik *SMOTE* untuk menyeimbangkan distribusi kelas serta optimasi parameter menggunakan *GridSearchCV*. Model dilatih untuk mempelajari pola hubungan antara fitur *TF-IDF* dan label sentimen.

3.1 Pengumpulan Data

Data penelitian diperoleh dari platform media sosial X/*Twitter* melalui proses scraping menggunakan kata kunci yang berkaitan dengan Program Makan Bergizi Gratis. Data yang dikumpulkan berupa komentar atau tweet masyarakat yang memuat opini mengenai program tersebut. *Dataset* yang diperoleh kemudian diseleksi untuk memastikan relevansi dengan topik

penelitian. Data yang tidak relevan, duplikat, atau mengandung karakter yang tidak diperlukan akan dihapus pada tahap *preprocessing*, hal ini peneliti memberikan label nama yaitu *DatasetMBG.csv* untuk memudahkan pemanggilan saat melakukan pengolahan data.

3.2 Preprocessing Data

Tahap *preprocessing* bertujuan untuk membersihkan dan menyiapkan data teks agar dapat diproses oleh algoritma *machine learning*[5]. Tahapan *preprocessing* meliputi: *Case folding* digunakan untuk Mengubah seluruh teks menjadi huruf kecil, *Cleaning* digunakan untuk Menghapus URL, simbol, angka, dan karakter khusus kemudian *Tokenizing* digunakan untuk Memecah teks menjadi kata-kata individual, *Stopword removal* digunakan untuk Menghapus kata umum yang tidak memiliki makna penting dan *Stemming* digunakan untuk Mengubah kata berimbuhan menjadi kata dasar menggunakan *library Sastrawi*[6]. Berikut contoh hasil perubahan teks sebelum dan sesudah *preprocessing* dapat dilihat pada Tabel 1.

Tabel 1 Hasil Pengolahan Data *Preprocessing*

| No | Teks Asli | Setelah <i>Preprocessing</i> | Keterangan Perubahan |
|----|-----------------------------------------------------------------------|------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | "MBG ini bagus banget! Pelayanannya cepat, harganya juga terjangkau." | mbg bagus layanan cepat harga terjangkau | - Lowercase semua huruf - Hapus tanda baca (!, .) - Hapus <i>stopword</i> : ini, juga- <i>Stemming</i> : pelayanannya → layanan, terjangkau → terjangkau |
| 2 | "Jelek banget, adminnya ga respon sama sekali!" | jelek admin respon kali | - Hapus tanda baca (!, .,) - Hapus <i>stopword</i> : ga, sama- <i>Stemming</i> : adminnya → admin, sekali → kali |
| 3 | "Saya puas banget belanja di sini, recommended pokoknya." | puas belanja sini recommend pokok | - Lowercase- Hapus tanda baca (, .) - Hapus <i>stopword</i> : saya, di - <i>Stemming</i> : recommended → recommend, pokoknya → pokok |
| 4 | "Kurirnya lambat, padahal sudah bayar ongkir mahal." | kurir lambat padahal bayar ongkir mahal | - Hapus tanda baca (, .) - Hapus <i>stopword</i> : sudah- <i>Stemming</i> : kurirnya → kurir |
| 5 | "@mbgofficial mantap produknya!! bikin nagih" | mantap produk bikin nagih | - Hapus <i>mention</i> : @mbgofficial - Hapus tanda baca (!!)- <i>Stemming</i> : produknya → produk |

3.3 Ekstraksi Fitur

Setelah proses *preprocessing* selesai, data teks diubah menjadi representasi numerik menggunakan metode *TF-IDF* (*Term Frequency–Inverse Document Frequency*). Metode ini memberikan bobot pada setiap kata berdasarkan frekuensi kemunculan dalam dokumen serta tingkat kepentingannya dalam keseluruhan *dataset*[9].

3.4 Model Klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini adalah *Logistic Regression*. Algoritma ini digunakan untuk mengklasifikasikan sentimen masyarakat menjadi dua kategori yaitu: 1. Sentimen Positif, 2. Sentimen Negatif. Dalam penelitian ini, pembagian data dilakukan dengan rasio 80% sebagai data latih dan 20% sebagai data uji. Rasio tersebut dipilih karena secara empiris dianggap mampu memberikan keseimbangan yang optimal antara kebutuhan model untuk mempelajari pola sentimen secara mendalam dan kebutuhan evaluasi performa yang objektif [10].

Model *Logistic Regression* secara matematis dapat dirumuskan sebagai berikut [16]:

$$P(y = 1|x) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n)}} \quad (1)$$

Keterangan:

- P(y=1|x) = probabilitas kelas positif
- b₀ = intercept
- b_i = koefisien regresi
- x_i = fitur hasil *TF-IDF*
- e = konstanta eksponensial

3.5 Evaluasi Model

Tahap evaluasi model merupakan langkah penting setelah model *Logistic Regression* selesai dibangun untuk mengukur tingkat akurasi dan performa prediksi terhadap sentimen MBG (positif atau negatif). Evaluasi ini penting untuk mengetahui apakah model bekerja dengan baik dan dapat diandalkan dalam mengklasifikasikan opini pengguna terhadap MBG.

Model *Logistic Regression* menggunakan fungsi:

$$p = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots)}} \tag{2}$$

Keterangan:

- P = probabilitas suatu kejadian (misal: sentimen positif, lulus/tidak)
- b₀ = intercept
- b₁, b₂, ... = koefisien regresi
- X₁, X₂, ... = variabel input
- e = konstanta 2.71828...

Pada penelitian ini, evaluasi dilakukan menggunakan beberapa metrik utama: *Confusion Matrix*, *ROC Curve*, dan *Area Under the Curve (AUC)*. *Confusion Matrix* digunakan untuk melihat distribusi prediksi benar (benar positif/benar negatif) maupun salah (*False Positive* dan *False Negative*) yang dilakukan oleh model terhadap sentimen MBG[9]. Sementara itu, ROC dan AUC digunakan untuk menilai kemampuan model dalam membedakan antara sentimen positif dan negatif secara keseluruhan.

Metrik Evaluasi *Logistic Regression* untuk Sentimen MB digunakan untuk memperjelas performa model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision untuk mengukur seberapa banyak prediksi positif yang benar dari semua prediksi positif.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall untuk mengukur seberapa banyak data yang sebenarnya positif berhasil diprediksi dengan benar.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Dan *F1-Score* untuk menggabungkan *Precision* dan *Recall* dalam satu metrik, menghasilkan skor untuk mencapai keseimbangan antar keduanya.

$$F1 - score = 2x \frac{precision \times recall}{precision + recall} \tag{6}$$

Keterangan:

- TP (*True Positive*) = Data sentimen positif yang berhasil diprediksi positif
- TN (*True Negative*) = Data sentimen negatif yang berhasil diprediksi negatif
- FP (*False Positive*) = Data negatif, tetapi model memprediksi positif
- FN (*False Negative*) = Data positif, tetapi model memprediksi negatif

4. Hasil dan Pembahasan

Dataset yang digunakan dalam penelitian ini berjumlah 1.440 data yang kemudian dibagi menjadi data latih sebesar 80% (1.152 data) dan data uji sebesar 20% (288 data). Proses *training* dilakukan menggunakan data latih dengan penerapan teknik *SMOTE* untuk mengatasi ketidakseimbangan data serta optimasi parameter menggunakan *GridSearchCV*. Model *Logistic Regression* dilatih untuk mempelajari hubungan antara fitur teks hasil representasi *TF-IDF* dengan label sentimen.

4.1 Hasil *Preprocessing*

Setelah proses *preprocessing* dilakukan, data teks menjadi lebih bersih dan siap digunakan dalam proses klasifikasi. Tahapan ini berhasil menghapus berbagai karakter yang tidak relevan seperti URL, simbol, dan kata-kata yang tidak memiliki makna penting. Implementasi proses *preprocessing* menggunakan bahasa pemrograman *Python* ditunjukkan pada Gambar 1.

```
def clean_text(text):
    text = text.lower()
    text = re.sub(r"http[s+|www\|s+|https\|s+", '', text) # hapus URL
    text = re.sub(r'@\w+|#\w+', '', text) # hapus mention & hashtag
    text = re.sub(r'^a-z\s]', '', text) # hapus tanda baca & angka
    text = re.sub(r'\s+', '', text).strip() # hapus spasi ganda
    text = ' '.join([w for w in text.split() if w not in stopwords]) # hapus stopword
    text = stemmer.stem(text) # stemming
    return text
```

Gambar 1 Source code Untuk *Preprocessing* Teks

Gambar tersebut memperlihatkan proses penghapusan URL, tanda baca, serta proses *stemming* menggunakan *library Sastrawi*. Tahapan *preprocessing* meliputi *case folding* (mengubah seluruh huruf menjadi *lowercase*), penghapusan URL, penghapusan *mention* dan *hashtag*, penghapusan angka serta tanda baca, penghapusan *stopword* menggunakan *library Sastrawi*, serta proses *stemming* untuk mengembalikan kata ke bentuk dasarnya. Kombinasi tahapan ini bertujuan untuk menyeragamkan struktur teks dan meningkatkan konsistensi representasi kata [11]. Contoh perubahan teks sebelum dan sesudah proses *preprocessing* pada *dataset* penelitian ditunjukkan pada Tabel 2.

Tabel 2 Proses *Preprocessing* Teks pada *Dataset*

| No | Teks Sebelum <i>Preprocessing</i> | Teks Setelah <i>Preprocessing</i> |
|----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | @prabowo Klo program makan bergizi gratis ini benar-benar dijalankan dengan baik dan merata ke seluruh daerah, saya yakin dampaknya akan sangat besar bagi generasi muda Indonesia. | klo program makan bergizi gratis ini benar benar dijalankan dengan baik dan merata ke seluruh daerah saya yakin dampaknya akan sangat besar bagi generasi muda indonesia |
| 2 | Program makan bergizi gratis ini sebenarnya bagus, tapi pelaksanaannya harus diawasi ketat supaya tidak terjadi penyimpangan anggaran seperti program-program sebelumnya. | program makan bergizi gratis ini sebenarnya bagus tapi pelaksanaannya harus diawasi ketat supaya tidak terjadi penyimpangan anggaran seperti program program sebelumnya |
| 3 | Saya mendukung program makan gratis ini karena bisa membantu anak-anak dari keluarga kurang mampu mendapatkan asupan gizi yang lebih baik setiap harinya. | saya asupan gizi yang lebih baik setiap harinya mendukung program makan gratis ini karena bisa membantu anak anak dari keluarga kurang mampu mendapatkan |

4.2 Proses Pelabelan Sentimen

Pelabelan dilakukan menggunakan pendekatan *lexicon-based*, yaitu dengan membandingkan kata dalam teks terhadap daftar kata positif dan negatif yang telah ditentukan sebelumnya [12]. Tabel 3 berikut merupakan contoh hasil pelabelan yang dilakukan peneliti dapat dilihat pada tabel 3.

Tabel 3 Proses Labeling Sentimen

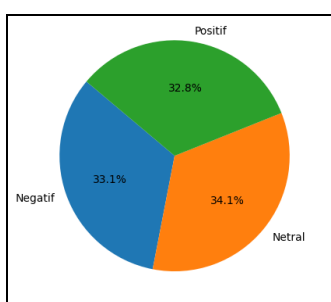
| No | Teks Setelah <i>Preprocessing</i> | Jumlah Kata Positif | Jumlah Kata Negatif | Label |
|----|-------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------|------------------------------|---------|
| 1 | saya mendukung program makan gratis ini karena bisa membantu anak anak dari keluarga kurang mampu mendapatkan asupan gizi yang lebih baik | 3 (<i>mendukung, membantu, baik</i>) | 0 | Positif |
| 2 | program makan bergizi gratis ini sebenarnya bagus tapi pelaksanaannya lambat dan banyak masalah di lapangan | 1 (<i>bagus</i>) | 2 (<i>lambat, masalah</i>) | Negatif |
| 3 | program makan bergizi gratis ini sudah berjalan di beberapa daerah dan sedang dalam tahap evaluasi pemerintah | 0 | 0 | Netral |

Penentuan label sentimen pada tabel di atas dilakukan berdasarkan pendekatan *lexicon-based*, yaitu dengan mengidentifikasi kemunculan kata-kata yang termasuk dalam

daftar kata positif dan negatif. Kalimat dikategorikan sebagai positif apabila mengandung dominasi kata dengan konotasi mendukung, apresiatif, atau menunjukkan kepuasan. Sebaliknya, kalimat dikategorikan sebagai negatif apabila mengandung kata dengan konotasi kritik, kekecewaan, kegagalan, atau penilaian buruk. Sementara itu, kalimat yang hanya bersifat informatif tanpa mengandung unsur evaluatif dikategorikan sebagai netral[13]. Melalui mekanisme ini, setiap teks pada *dataset* diberikan label berdasarkan dominasi jumlah kata yang terdeteksi dalam kamus sentimen. Proses ini menghasilkan variabel target berupa kelas sentimen (positif, negatif, atau netral) yang kemudian digunakan dalam tahap pelatihan model *machine learning*.

4.3 Distribusi Sentimen *Dataset*

Berdasarkan hasil pengolahan pada *dataset DatasetMBG_Labeled.csv*, diperoleh total 1.440 data komentar yang terdiri dari 473 sentimen positif, 491 sentimen netral, dan 476 sentimen negatif. Jika dikonversi dalam bentuk persentase, distribusi data adalah 32,8% positif, 34,1% netral, dan 33,1% negatif. Visualisasi distribusi sentimen pada *dataset* penelitian dapat dilihat pada Gambar 2.



Gambar 2 Distribusi hasil Klasifikasi sentimen

Hasil ini menunjukkan bahwa distribusi data relatif seimbang dengan selisih proporsi kurang dari 2% antar kelas. Keseimbangan kelas merupakan faktor penting dalam proses klasifikasi karena *dataset* yang tidak seimbang dapat menyebabkan model mengalami bias terhadap kelas mayoritas. Dengan distribusi yang relatif merata, model diharapkan mampu mempelajari pola sentimen secara lebih optimal dan menghasilkan performa evaluasi yang lebih stabil.

4.4 Transformasi Fitur dengan *TF-IDF*

Sebelum dilakukan proses klasifikasi, data teks perlu diubah menjadi representasi numerik agar dapat diproses oleh algoritma *machine learning*. Dalam penelitian ini digunakan metode *Term Frequency–Inverse Document Frequency (TF-IDF)* untuk mengubah teks menjadi vektor fitur. *TF-IDF* bekerja dengan menggabungkan dua komponen utama, yaitu *Term Frequency (TF)* yang mengukur frekuensi kemunculan kata dalam suatu dokumen dan *Inverse Document Frequency (IDF)* yang mengukur tingkat kelangkaan kata dalam seluruh dokumen. Kombinasi kedua komponen ini menghasilkan bobot kata yang lebih informatif. Implementasi metode *TF-IDF* pada penelitian ini menggunakan *library TfidfVectorizer* dari *Python*. *Source code* penerapan metode *TF-IDF* dapat dilihat pada Gambar 3.

```
tfidf = TfidfVectorizer(ngram_range=(1,2), max_features=5000)
x = tfidf.fit_transform(df["clean"])
y = df["label_refined"]
```

Gambar 3 *Source code* Penerapan *TF-IDF*

Pada penelitian ini digunakan parameter *max_features = 5000*, *min_df = 2*, serta *ngram_range = (1,2)*. Parameter tersebut memungkinkan model untuk mempertimbangkan kata tunggal (*unigram*) maupun pasangan kata (*bigram*) sehingga dapat menangkap konteks kata dengan lebih baik. Hasil transformasi *TF-IDF* menghasilkan matriks fitur yang merepresentasikan setiap dokumen sebagai vektor numerik. Matriks inilah yang kemudian digunakan sebagai input dalam proses pelatihan model klasifikasi.

4.5 Penanganan Ketidakseimbangan Data dengan SMOTE

Dalam klasifikasi teks, ketidakseimbangan jumlah data antar kelas dapat memengaruhi performa model. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik *Synthetic Minority Oversampling Technique (SMOTE)*. Implementasi metode *SMOTE* dalam proses pemrograman dapat dilihat pada Gambar 4.

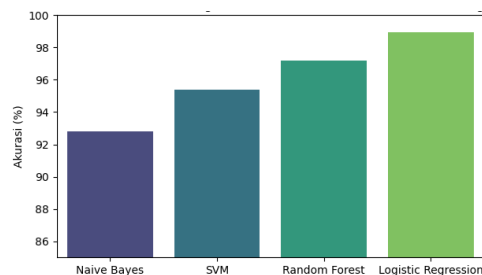
```
sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_resample(X, y)
```

Gambar 4 Source code SMOTE untuk menyeimbangkan Kelas

SMOTE merupakan metode *oversampling* yang menghasilkan data sintesis pada kelas minoritas berdasarkan kedekatan antar sampel pada ruang fitur. Teknik ini bertujuan untuk meningkatkan representasi kelas yang jumlah datanya lebih sedikit sehingga model dapat mempelajari distribusi data secara lebih merata. Proses *SMOTE* dilakukan setelah tahap transformasi fitur *TF-IDF* dan sebelum proses pelatihan model. Dengan pendekatan ini, distribusi data pada proses pelatihan menjadi lebih seimbang sehingga dapat mengurangi potensi bias model terhadap kelas tertentu [14].

4.6 Hasil Klasifikasi Logistic Regression

Model yang digunakan dalam penelitian ini adalah *Logistic Regression*, yang merupakan metode klasifikasi diskriminatif yang banyak digunakan dalam analisis teks dan pengolahan bahasa alami. Berbeda dengan *Naive Bayes* yang berbasis probabilistik sederhana, *Logistic Regression* mampu memodelkan hubungan antara fitur dan kelas secara lebih fleksibel. Untuk mengevaluasi performa model, dilakukan perbandingan akurasi beberapa algoritma *machine learning*. Hasil perbandingan akurasi antar model dapat dilihat pada Gambar 5.



Gambar 5 Perbandingan Akurasi Antar Model *Machine Learning*

Berdasarkan Gambar 5, model *Logistic Regression* menunjukkan performa yang kompetitif dalam proses klasifikasi sentimen dibandingkan dengan model lainnya. Hal ini menunjukkan bahwa *Logistic Regression* mampu memanfaatkan representasi fitur hasil transformasi *TF-IDF* secara efektif dalam membedakan kelas sentimen. Untuk meningkatkan performa model, dilakukan proses optimasi menggunakan *GridSearchCV* yang bertujuan untuk mencari kombinasi parameter terbaik secara sistematis. Parameter yang diuji dalam penelitian ini meliputi nilai *C*, *class_weight*, dan *max_iter*. Proses tuning parameter dilakukan menggunakan teknik *cross-validation* sehingga model dapat diuji pada beberapa subset data secara bergantian. Pendekatan ini bertujuan untuk memperoleh konfigurasi model yang memiliki performa paling stabil. Implementasi proses tuning menggunakan *GridSearchCV* pada model *Logistic Regression* ditunjukkan pada Gambar 6.

```
params = {
    'C': [0.5, 1, 2],
    'max_iter': [300],
    'class_weight': ['balanced']
}
grid = GridSearchCV(LogisticRegression(), param_grid=params, cv=3)
grid.fit(X_train, y_train)

best_model = grid.best_estimator_
print("Best Params:", grid.best_params_)
```

Gambar 6 Source code Model *Logistic Regression* tuning

Hasil optimasi menunjukkan bahwa model *Logistic Regression* mampu mencapai tingkat akurasi sebesar 98,96%, dengan nilai *Precision*, *Recall*, dan *F1-Score* yang mendekati 99%. Nilai ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengklasifikasikan sentimen opini masyarakat. Ringkasan hasil evaluasi model tersebut dapat dilihat pada Gambar 7.

```
Best Params: {'C': 4, 'class_weight': 'balanced', 'max_iter': 300}
◆ Akurasi Model: 98.96 %
```

Gambar 7 Hasil Akurasi *Logistic Regression*

Untuk memvalidasi hasil sistem, dilakukan perhitungan manual menggunakan *Confusion Matrix* pada data uji. Berdasarkan *Confusion Matrix*, jumlah prediksi benar pada diagonal utama adalah 1425 data.

1). Akurasi (*Accuracy*)

Perhitungan nilai akurasi dilakukan menggunakan Persamaan (3) yang telah dijelaskan pada bagian metode. Berdasarkan *Confusion Matrix*, jumlah prediksi benar yang berada pada diagonal utama adalah:

$$476 + 485 + 464 = 1425$$

Sehingga nilai akurasi dihitung sebagai berikut:

$$Accuracy = \frac{\sum \text{Diagonal Utama}}{n} = \frac{(476 + 485 + 464)}{1440} = \frac{1425}{1440} \approx 0,9896 / (98,96\%)$$

Nilai tersebut menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan dengan benar oleh model.

2). *Precision*, *Recall*, dan *F1-Score* (Studi Kasus: Kelas Positif)

Evaluasi juga dilakukan pada tingkat kelas untuk mengetahui performa model secara lebih spesifik. Berdasarkan hasil *Confusion Matrix* diperoleh nilai:

$$TP = 464, FP = 4, FN = 9$$

Perhitungan *Precision* dilakukan menggunakan Persamaan (4).

$$Precision = \frac{TP}{(TP + FP)} = \frac{464}{(464 + 4)} = \frac{464}{468} = 0,9914 / (99,14\%)$$

Perhitungan *Recall* dilakukan menggunakan Persamaan (5).

$$Recall = \frac{TP}{(TP + FN)} = \frac{464}{(464 + 9)} = \frac{464}{473} = 0,9810 / (98,10\%)$$

Selanjutnya, nilai *F1-Score* dihitung menggunakan Persamaan (6).

$$F1-Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} = 2 \times \frac{(0,9914 \times 0,9810)}{(0,9914 + 0,9810)} = 0,9861 / (98,61\%)$$

Sebagai ilustrasi penerapan model *Logistic Regression*, digunakan contoh kalimat “Anak-anak jadi senang dan sehat berkat program ini”. Setelah dilakukan *preprocessing* dan transformasi *TF-IDF*, diperoleh nilai fitur sebagai berikut:

| Fitur | Nilai | Parameter | Nilai |
|--------------|-------|-----------|-------|
| x1 (senang) | 0.65 | b0 | -0.85 |
| x2 (sehat) | 0.71 | b1 | 1.20 |
| x3 (program) | 0.33 | b2 | 1.35 |

Selanjutnya dilakukan substitusi nilai fitur ke dalam Persamaan (2) untuk memperoleh nilai fungsi linear:

$$p = \frac{1}{1 + e^{-(1.036)}} = e^{-1.036} = 0.355 = p = \frac{1}{1 + 0.355} = 0.738$$

Hasil perhitungan menunjukkan bahwa probabilitas kelas positif sebesar 0,738 (73,8%). Karena nilai probabilitas terbesar berada pada kelas positif, maka kalimat tersebut diklasifikasikan sebagai sentimen positif oleh model *Logistic Regression*. Hasil perhitungan ini menunjukkan bahwa model *Logistic Regression* memiliki performa klasifikasi yang sangat baik dalam mengidentifikasi sentimen.

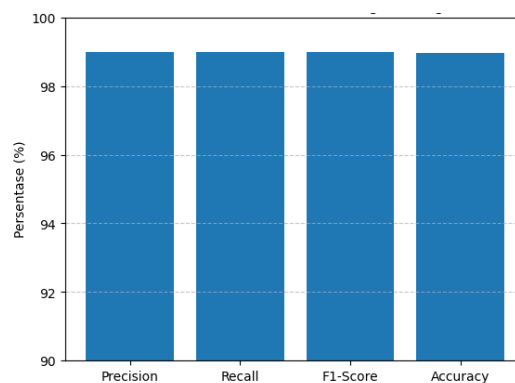
3. Analisis Karakteristik Data Berhasil dan Gagal

Meskipun model memiliki akurasi tinggi, terdapat selisih error sebesar 1,04% (15 data). Berdasarkan observasi pada *dataset* "*DatasetMBG.csv*", dilakukan analisis terhadap karakteristik data untuk memenuhi validasi performa kinerja akhir sebagaimana tersaji pada Tabel 4.

Tabel 4 Analisis Sampel Data Berhasil dan Gagal

| Teks Asli (<i>dataset</i>) | Prediksi | Status | Analisis Penyebab |
|-------------------------------------------------------------------|----------|----------|------------------------------------------------------------------------------------------|
| "Program makan bergizi gratis sangat membantu gizi anak sekolah." | Positif | Berhasil | Memiliki kata kunci formal ("membantu", "gizi") yang dikenali baik oleh <i>TF-IDF</i> . |
| "Menunggu hasil laporan.. Kalau ketemu mohon dibantu yak." | Negatif | Gagal | Penggunaan kata gaul ("yak") dan kalimat menggantung membuat bobot fitur menjadi ambigu. |
| "Anakmu makan yg beginian jatah makan gratis..???" | Netral | Gagal | Unsur sarkasme dan tanda baca berulang sulit diproses secara linear oleh model. |

Selain itu, performa model juga divisualisasikan dalam bentuk grafik untuk memperlihatkan perbandingan nilai evaluasi secara lebih jelas. Grafik performa model *Logistic Regression* ditampilkan pada Gambar 8.

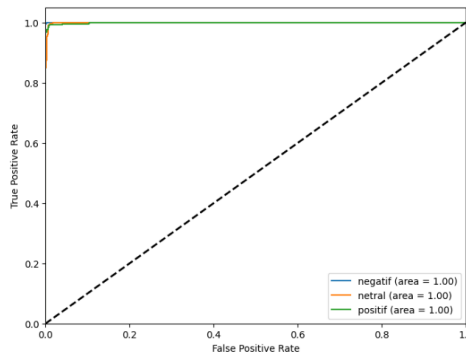


Gambar 8 Grafik Akurasi Model Logistic Regression

Berdasarkan Gambar 7 dan 8, model *Logistic Regression* mencapai performa evaluasi mendekati 99% pada seluruh metrik. Hal ini membuktikan efektivitas model dalam menangkap pola hubungan fitur teks hasil transformasi *TF-IDF*, sehingga lebih unggul dibandingkan metode probabilistik seperti Naive Bayes. Meskipun akurasi sangat tinggi memicu potensi overfitting, penerapan *cross-validation* melalui *GridSearchCV* dalam riset ini berhasil meminimalisir risiko tersebut dan menjamin generalisasi model yang baik pada data baru.

4.7 Evaluasi Model dengan ROC Curve

Evaluasi tambahan dilakukan menggunakan *Receiver Operating Characteristic* (ROC) Curve untuk mengukur kemampuan model dalam membedakan setiap kelas sentimen. Kurva ROC menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang klasifikasi. Selain itu, performa model juga diukur menggunakan nilai *Area Under Curve* (AUC). Visualisasi kurva ROC dari model *Logistic Regression* pada penelitian ini ditunjukkan pada Gambar 9.



Gambar 9 ROC Curve Model Logistic Regression

Berdasarkan Gambar 9, hasil pengujian menunjukkan bahwa nilai AUC mendekati 1.0, yang menandakan bahwa model *Logistic Regression* memiliki kemampuan diskriminasi yang sangat baik dalam membedakan sentimen positif, netral, dan negatif.

4.8 Analisis Word Cloud

Selain analisis kuantitatif, penelitian ini juga melakukan analisis kualitatif menggunakan visualisasi word cloud untuk melihat kata-kata yang paling dominan pada masing-masing kategori sentimen. Visualisasi word cloud untuk sentimen positif dapat dilihat pada Gambar 10.



Gambar 10 Word Cloud Sentimen Positif

Berdasarkan Gambar 10, kata-kata yang sering muncul pada sentimen positif antara lain bagus, baik, membantu, sukses, dan sehat. Kata-kata tersebut mencerminkan adanya persepsi positif masyarakat terhadap Program Makan Bergizi Gratis. Sementara itu, visualisasi kata yang dominan pada sentimen negatif ditunjukkan pada Gambar 11.



Gambar 11 Word Cloud Sentimen Negatif

Berdasarkan Gambar 11, kata-kata seperti gagal, buruk, lambat, kurang, dan tidak menunjukkan adanya kritik atau ketidakpuasan terhadap implementasi program. Visualisasi ini memberikan gambaran tambahan mengenai pola opini masyarakat serta mendukung hasil analisis klasifikasi yang telah dilakukan sebelumnya. Metode word cloud dapat memberikan gambaran cepat mengenai topik atau kata yang paling dominan dalam suatu kategori sentimen [15].

4.10 Pembahasan

Integrasi algoritma *Logistic Regression* dengan teknik *TF-IDF*, *SMOTE*, dan *GridSearchCV* dalam penelitian ini terbukti efektif dengan capaian akurasi sebesar 98,96%. Tingginya performa ini menunjukkan bahwa model mampu menangkap pola sentimen masyarakat terhadap Program Makan Bergizi Gratis secara optimal. Penerapan teknik *SMOTE* berperan penting dalam menyeimbangkan distribusi data yang cenderung tidak merata, sehingga model tidak bias terhadap kelas tertentu. Secara teknis, penyeimbangan data sintesis ini berkontribusi terhadap peningkatan nilai *Recall* hingga mencapai 98,10%, yang menunjukkan bahwa model memiliki sensitivitas tinggi dalam mendeteksi sentimen negatif. Temuan ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa penanganan data tidak seimbang merupakan faktor penting dalam meningkatkan stabilitas dan performa model klasifikasi pada data media sosial.

Jika ditinjau berdasarkan distribusi kelas sentimen, model menunjukkan kemampuan yang relatif seimbang dalam mengklasifikasikan sentimen positif, negatif, dan netral. Hal ini didukung oleh distribusi *dataset* yang hampir merata, yaitu 32,8% positif, 34,1% netral, dan 33,1% negatif. Pada sentimen positif, model mampu mengenali kata-kata dengan konotasi apresiatif seperti “baik”, “membantu”, dan “sehat”. Sementara itu, pada sentimen negatif, model secara efektif mengidentifikasi kata-kata seperti “buruk”, “lambat”, dan “gagal”. Namun, pada sentimen netral, model cenderung mengalami kesulitan karena karakteristik teks yang bersifat informatif tanpa ekspresi emosional yang kuat, sehingga membatasi kemampuan diskriminatif fitur *TF-IDF*.

Keunggulan model dalam penelitian ini juga dipengaruhi oleh penggunaan metode *TF-IDF* yang mampu memberikan bobot lebih tinggi pada kata-kata penting dalam dokumen. Fitur seperti “gizi”, “gratis”, dan “bantu” memiliki kontribusi besar dalam proses klasifikasi karena sering muncul dalam konteks opini publik terhadap program MBG. Dibandingkan dengan metode *Naïve Bayes* yang mengasumsikan independensi antar fitur [17], *Logistic Regression* mampu memodelkan hubungan linear antar fitur dalam ruang berdimensi tinggi secara lebih fleksibel. Hal ini memungkinkan model untuk menangkap pola hubungan antar kata secara lebih akurat. Selain itu, optimasi parameter menggunakan *GridSearchCV* memastikan bahwa model berada pada konfigurasi terbaik, sehingga mampu menghasilkan performa yang stabil dan meminimalkan risiko overfitting.

Jika dibandingkan dengan penelitian sebelumnya, performa model dalam penelitian ini tergolong lebih tinggi. Penelitian oleh Setyawan et al. [6] menunjukkan bahwa metode SVM dan *Naïve Bayes* menghasilkan akurasi pada kisaran 90–92%. Sementara itu, penelitian berbasis *deep learning* oleh Haerani [7] menggunakan GRU menghasilkan *F1-Score* sebesar 91%, dan penelitian oleh Novfuja et al. [8] menggunakan LSTM mencapai akurasi sebesar 91,71%. Dibandingkan dengan hasil tersebut, model *Logistic Regression* dalam penelitian ini menunjukkan peningkatan performa yang signifikan dengan akurasi mencapai 98,96%. Hal ini menunjukkan bahwa kombinasi *TF-IDF*, *SMOTE*, dan *GridSearchCV* mampu meningkatkan kualitas klasifikasi secara optimal meskipun menggunakan pendekatan yang relatif lebih sederhana dibandingkan *deep learning*.

Meskipun model menunjukkan performa yang sangat tinggi, masih terdapat selisih error sebesar 1,04% yang berasal dari data yang gagal diklasifikasikan secara tepat. Berdasarkan analisis pada bagian sebelumnya, kesalahan klasifikasi umumnya disebabkan oleh karakteristik teks yang mengandung unsur sarkasme, penggunaan bahasa gaul (slang), serta ambiguitas linguistik. Kondisi ini menunjukkan bahwa pendekatan berbasis *TF-IDF* masih memiliki keterbatasan dalam memahami konteks semantik yang kompleks. Oleh karena itu, pengembangan model di masa mendatang dapat mempertimbangkan penggunaan pendekatan berbasis *deep learning* atau word embedding untuk meningkatkan kemampuan pemahaman konteks.

Secara praktis, penelitian ini memberikan kontribusi dalam bentuk kerangka kerja sistematis yang mengintegrasikan *Logistic Regression*, *TF-IDF*, *SMOTE*, dan *GridSearchCV* dalam analisis sentimen kebijakan publik. Pendekatan ini tidak hanya meningkatkan akurasi model, tetapi juga menyediakan alat analisis yang dapat digunakan oleh pemerintah untuk memantau persepsi masyarakat secara *real-time*. Dengan demikian, hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam pengambilan keputusan berbasis data (*data-driven decision making*), khususnya dalam evaluasi dan pengembangan kebijakan Program Makan Bergizi Gratis.

5. Simpulan

Berdasarkan hasil penelitian, pendekatan *text mining* dan *machine learning* dapat digunakan secara efektif untuk menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG). Penelitian ini menggunakan 1.440 data komentar dari platform X/Twitter yang diproses melalui tahap *preprocessing*, ekstraksi fitur menggunakan *TF-IDF*, serta klasifikasi dengan algoritma *Logistic Regression*. Hasil evaluasi menunjukkan bahwa model *Logistic Regression* memiliki performa yang sangat baik dengan akurasi 98,96%, *Precision* 99,14%, *Recall* 98,10%, dan *F1-Score* 98,61%. Hasil tersebut menunjukkan bahwa model mampu mengklasifikasikan sentimen masyarakat secara akurat. Dengan demikian, *Logistic Regression* terbukti efektif digunakan dalam analisis sentimen terhadap opini publik di media sosial, serta dapat memberikan gambaran mengenai persepsi masyarakat terhadap kebijakan Program Makan Bergizi Gratis.

Daftar Referensi

- [1]. S. Y. Widyasari, A. Larasati, and W. Y. Alam, "Evaluasi Kebijakan Makan Bergizi Gratis di Sekolah Dasar: Implikasi Terhadap Kesehatan Anak dan Pemberdayaan Ekonomi Lokal," *Innovative: Journal of Social Science Research*, vol. 5, no. 4, pp. 1727–1736, Jul. 2025.
- [2]. D. Oktawila, H. E. Bagijo, and Tanudjaja, "Kedudukan Lembaga Negara Dalam Makan Bergizi Gratis," *Jurnal Media Informatika (JUMIN)*, vol. 6, no. 3, pp. 1595–1602, May–Aug. 2025.
- [3]. M. E. Ga'a, Y. F. Kasi, and Y. D. A. Epu, "Program Makan Bergizi Gratis di Kabupaten Nagekeo: Tujuan dan Tahapannya," *Jurnal Ilmiah Peternakan*, vol. 5, no. 2, pp. 84–90, Jun. 2025.
- [4]. S. Fitrianti and A. Yudhistira, "Analisis Sentimen Media Sosial Terhadap Calon Pilkada 2024 Dengan Metode *Naïve Bayes*," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 1, pp. 167–176, Jan. 2025, doi: 10.52436/1.jpti.610.
- [5]. V. Agustina and A. Herliana, "Analisis Sentimen Publik atas Kebijakan Efisiensi Anggaran 2025 dengan *Text mining* dan *Natural Language Processing*," *Jurnal Media Informatika (JUMIN)*, vol. 6, no. 3, pp. 2182–2194, May–Aug. 2025, doi: 10.37424/jumin.v6i3.
- [6]. H. Setyawan, L. M. Azizah, and A. Y. Pradani, "Sentiment Analysis of Public Responses on Indonesia Government Using *Naïve Bayes* and *Support Vector Machine*," *Emerging Information Science and Technology*, vol. 4, no. 1, pp. 1–7, Feb. 2023, doi: 10.26623/eist.v4i1.
- [7]. F. Haerani, "Analisis Sentimen Program Makan Bergizi Gratis Berdasarkan Media Sosial X Menggunakan Metode Gated Recurrent Unit (GRU)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, no. 1, Jan. 2025, doi: 10.23960/jitet.v14i1.8754.
- [8]. E. Novfuja, E. Ali, S. Susanti, and Lusiana, "Prediksi Dukungan Publik Terhadap Program akan Bergizi Gratis (MBG) Menggunakan Analisis Sentimen Berbasis Long Short-Term Memory (LSTM)," *Jurnal Algoritma*, vol. 22, no. 2, pp. 1704–1715, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2690.
- [9]. G. S. Rasyad and W. Maharani, "*Logistic Regression* and *Naïve Bayes* Comparison in Classifying Emotions on Indonesian X Social Media," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 31–40, Apr. 2025, doi: 10.29408/edumatic.v9i1.29120.
- [10]. S. Khoerunnisa, D. F. Shiddieq, and D. Nurhayati, "Penerapan Algoritma *Naïve Bayes* dengan Teknik *TF-IDF* dan Cross Validation untuk Analisis Sentimen Terhadap Starlink," *MALCOM: Indonesian Journal of Machine learning and Computer Science*, vol. 5, no. 2, pp. 566–577, Mar. 2025, doi: 10.57152/malcom.v5i2.1852.
- [11]. M. U. Albab, Y. K. P., and M. N. Fawaiq, "Optimization of the *Stemming* Technique on Text *Preprocessing* President 3 Periods Topic," *Jurnal Transformatika*, vol. 20, no. 2, pp. 1–12, Jan. 2023, doi: 10.26623/transformatika.v20i2.5374.
- [12]. J. E. B. Sinulingga and H. C. K. Sitorus, "Analisis Sentimen Masyarakat terhadap Film Horor Indonesia Menggunakan Metode SVM dan *TF-IDF*," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 14, no. 1, pp. 45–55, Apr. 2024.
- [13]. A. B. Putra Negara, "The Influence of Applying *Stopword removal* and *SMOTE* on Indonesian Sentiment Classification," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 3, pp. 172–180, Dec. 2023, doi: 10.24843/lkjiti.2023.v14.i03.p05.

-
- [14]. M. N. Fahmi, "Implementasi *Machine learning* menggunakan *Python* Library: Scikit-Learn (Supervised dan Unsupervised Learning)," *Sains Data: Jurnal Studi Matematika dan Teknologi*, vol. 1, no. 2, pp. 87–96, Dec. 2023, doi: 10.52620/sainsdata.v1i2.31.
- [15]. L. Rhomaningtias, A. Khairunisa, S. S. M. Wara, and K. M. Hindrayani, "Analisis Sentimen Ulasan Aplikasi Smile Indonesia Menggunakan Metode *Naïve Bayes* dan *Support Vector Machine* (SVM)," *HOAQ: Jurnal Teknologi Informasi*, vol. 16, no. 1, pp. 79–91, May 2025.
- [16]. I. Ristiana and Mutmainah, "Analisis Sentimen Bencana Banjir Sumatera Menggunakan *TF-IDF* dan Logistic Regression," *Jurnal Sistem Informasi dan Teknologi (SINTEK)*, vol. 6, no. 1, pp. 57-65, 2024. [Online]. Available: <https://sintek.stmikku.ac.id/index.php/home>
- [17]. A. Riviani, & C.B. Santoso, "Perbandingan *Naïve Bayes* dan SVM untuk Analisis Sentimen Pengguna Aplikasi X Terhadap Danantara. *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 14, no. 3, pp. 1843-1855, 2025.