JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

Perbandingan Naive Bayes, SVM, dan IndoBERT untuk Analisis Sentimen Publik terhadap Program Makan Bergizi Gratis pada Komentar YouTube

Stella Maris[1*], Windi Irmayani[2], Muhammad Ifan Rifani Ihsan[3]

[1,2,3]Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Pontianak, Indonesia
Email: [1*]stellamaris3460@email.com, [2]windi.wnr@bsi.ac.id, [3]ifan.mii@bsi.ac.id
(*Email Corresponding Author: stellamaris3460@gmail.com)
Received: 1 Juli 2026 | Revision: 5 Juli 2026 | Accepted: 11 Juli 2026

Abstrak

Pelaksanaan Program Makan Bergizi Gratis (MBG) pada tahun 2025 dialokasikan anggaran sebesar Rp171 triliun dengan target 82,9 juta penerima manfaat, namun realisasi awal yang tercatat baru mencapai Rp3 triliun untuk 3,98 juta orang. Kesenjangan target pelaksanaan dan tantangan di lapangan tersebut memicu berbagai opini publik pada platform digital seperti YouTube. Studi ini menganalisis sentimen publik terhadap Program MBG dengan membandingkan efektivitas algoritma Naive Bayes, Support Vector Machine (SVM), dan IndoBERT menggunakan 6.419 komentar YouTube dari Kaggle. Metodologi penelitian meliputi *preprocessing* teks, pelabelan sentimen sberbasis *Lexicon*, pembagian data, ekstraksi fitur (TF-IDF dan IndoBERT *Tokenizer*), penerapan SMOTE pada Naive Bayes dan SVM, serta evaluasi kinerja model. Hasil pelabelan awal menunjukkan adanya ketidakseimbangan data yang didominasi oleh kelas netral sebesar 72,9%, diikuti kelas negatif sebesar 15,0%, dan kelas positif sebesar 12,1%. Setelah dilakukan penanganan menggunakan metode SMOTE, distribusi kelas pada data *training* menjadi seimbang guna mengoptimalkan proses pembelajaran model Naive Bayes dan SVM. Berdasarkan evaluasi akhir eksperimen, IndoBERT mencatatkan kinerja paling superior dengan nilai *accuracy* tertinggi mencapai 97,35%, mengungguli SVM sebesar 95,72% dan Naive Bayes sebesar 65,11%. Kesimpulannya, pendekatan *context-aware* pada model IndoBERT serta penanganan data sebelum pemodelan sangat efektif dan andal dalam memetakan persepsi serta kekhawatiran masyarakat terhadap efisiensi pelaksanaan Program Makan Bergizi Gratis.

Kata Kunci: Analisis Sentimen, Makan Bergizi Gratis, Naive Bayes, SVM, IndoBERT.

Abstract

The implementation of the Free Nutritious Meal (MBG) Program in 2025 was allocated a budget of IDR 171 trillion targeting 82.9 million beneficiaries, but initial realization only reached IDR 3 trillion for 3.98 million people. This gap between implementation targets and field challenges has triggered various public opinions on digital platforms such as YouTube. This study analyzes public sentiment toward the MBG Program by comparing the effectiveness of Naive Bayes, Support Vector Machine (SVM), and IndoBERT algorithms using 6,419 YouTube comments from Kaggle. The methodology involves text preprocessing, Lexicon-based sentiment labeling, data splitting, feature extraction (TF-IDF and IndoBERT Tokenizer), SMOTE implementation for Naive Bayes and SVM, and model performance evaluation. The initial labeling results show data imbalance dominated by the neutral class at 72.9%, followed by the negative class at 15.0%, and the positive class at 12.1%. After handling this with the SMOTE method, the class distribution in the training data became balanced to optimize the learning process of the Naive Bayes and SVM models. Based on the final experimental evaluation, IndoBERT achieves superior performance with the highest accuracy of 97.35%, outperforming SVM at 95.72% and Naive Bayes at 65.11%. In conclusion, the context-aware approach of the IndoBERT model and data handling prior to modeling are highly effective and reliable in mapping public perceptions and concerns regarding the implementation efficiency of the Free Nutritious Meal Program.

***Keywords**: Sentiment Analysis, Free Nutritious Meal, Naive Bayes, SVM, IndoBERT.*

1. PENDAHULUAN

Implementasi Program Makan Bergizi Gratis (MBG) sebagai agenda strategis nasional saat ini tengah menghadapi tantangan operasional yang signifikan di lapangan. Pada tahun 2025, pemerintah mengalokasikan kapasitas fiskal yang sangat besar mencapai Rp171 triliun dengan target 82,9 juta penerima manfaat, dan memproyeksikan peningkatan anggaran hingga Rp300 triliun pada tahun 2026. Namun, data realisasi hingga 21 Mei 2025 menunjukkan penyerapan anggaran baru mencapai Rp3 triliun dengan jangkauan 3,98 juta penerima manfaat [1]. Kontras yang tajam antara target makro perencanaan dan fakta realisasi riil inilah yang memicu gelombang tanggapan publik di berbagai platform digital, salah satunya YouTube yang menurut data APJII 2026 merupakan media berbasis video nomor satu yang paling intensif diakses pengguna internet domestik dengan persentase 49,3% [2]. Besarnya volume rekam opini

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

pada kolom komentar YouTube tersebut memunculkan kendala baru dari sisi komputasi karena karakteristik datanya bersifat tidak terstruktur (*unstructured data*), didominasi bahasa informal, sarat singkatan, serta mengalami ketidakseimbangan kelas (*class imbalance*) yang menyebabkan model klasifikasi standar kerap gagal menangkap konteks semantik secara presisi.

Guna mengatasi hambatan tersebut, penelitian ini mengajukan solusi berupa penerapan kerangka pemodelan analisis sentimen (*sentiment analysis*) komparatif yang andal menggunakan algoritma *context-aware* berbasis *Deep Learning* yaitu IndoBERT, yang disandingkan dengan algoritma berbasis probabilistik (Naive Bayes) serta model berbasis *hyperplane* pembatas optimal (Support Vector Machine). Melalui pendekatan komparatif ini, riset ini diharapkan mampu memetakan polaritas persepsi masyarakat terhadap dinamika Program MBG secara objektif dan akurat. Untuk memosisikan kebaruan ilmiah (*novelty*) dari pendekatan yang diajukan, perlu dilakukan penelaahan mendalam terhadap sejumlah penelitian sejenis yang relevan dalam kurun waktu beberapa tahun terakhir.

Pertama, studi yang dilakukan oleh Vebrian dan Kustiyono (2025) mengeksplorasi opini publik mengenai rencana program makan gratis di media sosial X menggunakan metode Naive Bayes dan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Riset dengan sampel 501 data tersebut menghasilkan akurasi sebesar 69,3%, namun mencatat keterbatasan mendasar dalam mengenali dokumen berpolaritas negatif secara optimal [3]. Kedua, Ghofur dan Putri (2026) menyajikan analisis komparatif menggunakan algoritma Support Vector Machine (SVM), K-Nearest Neighbor (KNN), dan Naive Bayes pada 5.420 data *tweet* di platform X terkait program MBG. Hasil pengujian menunjukkan keunggulan superior model SVM dengan tingkat akurasi mencapai 91,7% dibandingkan Naive Bayes yang hanya meraih 79,6% [4].

Ketiga, Muntaza (2026) melakukan ekspansi riset pada platform YouTube menggunakan algoritma SVM untuk mendeteksi tanggapan masyarakat terhadap isu spesifik terkait keamanan pangan pada program MBG. Melalui ekstraksi komentar dan pelabelan menggunakan model bahasa RoBERTa, penelitian tersebut menghasilkan akurasi sebesar 84% dan menemukan dominasi sentimen negatif dari publik terkait kualitas pelaksanaan program [5]. Keempat, Rahmatullah dkk. (2025) mendayagunakan algoritma Naive Bayes untuk mengklasifikasikan 1.470 komentar pada akun YouTube resmi Sekretariat Negara mengenai kebijakan MBG. Eksperimen tersebut memperoleh tingkat akurasi sebesar 84,69%, namun menghasilkan nilai *recall* yang cukup rendah yaitu 53,73% pada kelas minoritas [6]. Kelima, Rohman dan Agung (2025) melakukan studi komparasi tingkat lanjut dengan melibatkan arsitektur *Transformer* IndoBERT, SVM, dan Random Forest terhadap 54.105 data *tweet* di platform X. Hasil akhir riset ini memvalidasi bahwa IndoBERT menyajikan performa paling efektif dengan raihan akurasi tertinggi sebesar 87,60% jika dibandingkan dengan SVM (83,39%) maupun Random Forest (70,28%) [7].

Melalui penelaahan komprehensif terhadap lima studi terdahulu tersebut, ditemukan sebuah kesenjangan riset (*Gap Analysis*) yang menjadi landasan urgensi penelitian ini. Sebagian besar literatur yang ada saat ini masih mendominasi pengujian pada platform media sosial X, sedangkan eksplorasi sentimen berbasis dokumen komentar YouTube terkait Program Makan Bergizi Gratis masih relatif terbatas. Di samping itu, beberapa riset yang sudah memanfaatkan repositori YouTube umumnya hanya menerapkan satu metode klasifikasi tunggal seperti Naive Bayes atau SVM secara parsial, sehingga belum menyajikan gambaran komparasi yang utuh mengenai performa relatif dari masing-masing arsitektur pada satu dataset yang sama. Karakteristik teks komentar YouTube yang memiliki tingkat kebebasan berbahasa sangat tinggi memerlukan pengujian lintas model untuk menguji konsistensi akurasi. Celah ilmiah inilah yang diisi oleh penelitian ini melalui eksperimen komparasi langsung antara metode Naive Bayes, Support Vector Machine (SVM), dan IndoBERT pada satu korpus data yang seragam.

Berdasarkan kesenjangan ilmiah yang telah diidentifikasi, tujuan utama dari penelitian ini adalah untuk mengevaluasi, memvalidasi, dan menentukan algoritma klasifikasi yang menyajikan kinerja performa paling superior serta akurat dalam mengenali kecenderungan sentimen masyarakat mengenai Program MBG pada kolom komentar YouTube. Melalui eksperimen komparatif ini, harapan teoretis yang ingin dicapai adalah hasil penelitian dapat memberikan kontribusi ilmiah yang valid serta menjadi acuan akademis yang kokoh dalam pengembangan aplikasi pemrosesan bahasa alami (*Natural Language Processing*) pada korpus dokumen berbahasa Indonesia yang tidak terstruktur. Secara praktis, riset ini diharapkan mampu menyuguhkan visualisasi peta persepsi publik yang objektif dan berbasis data (*evidence-based*) bagi pihak regulator. Hasil analisis ini diharapkan dapat menjadi referensi strategis bagi pemerintah dalam mengevaluasi tata kelola, sistem distribusi logistik, serta efisiensi pelaksanaan kebijakan Program Makan Bergizi Gratis demi mendukung optimalisasi pembangunan nasional.

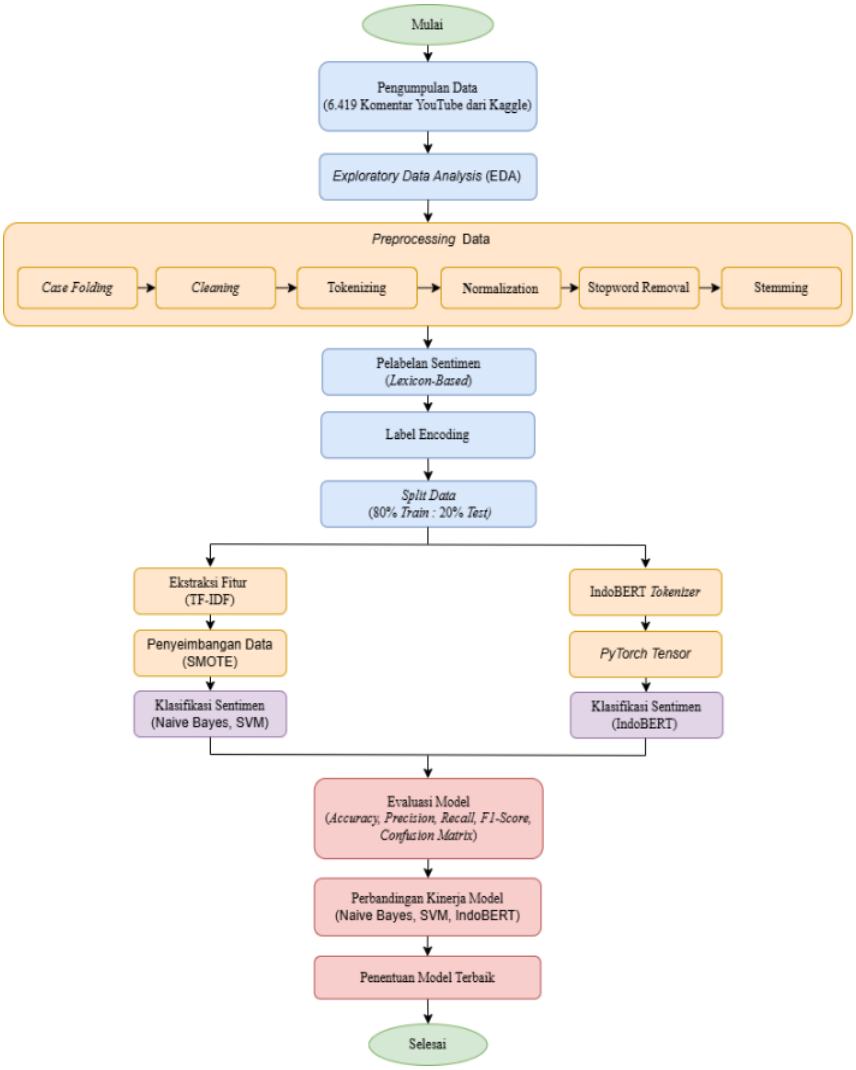
2. METODOLOGI PENELITIAN

Pendekatan kuantitatif diterapkan untuk menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG). Proses pengolahan data dilakukan menggunakan Google Colaboratory dengan bahasa pemrograman Python. Tahapan analisis diawali dengan pemanfaatan data sekunder, dilanjutkan dengan *preprocessing* teks, pelabelan sentimen, pembagian data, ekstraksi fitur, penyeimbangan data menggunakan *Synthetic Minority Oversampling*

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

Technique (SMOTE), serta proses klasifikasi menggunakan algoritma Naive Bayes (NB), Support Vector Machine (SVM), dan IndoBERT. Kinerja setiap algoritma kemudian dievaluasi dan dibandingkan untuk menentukan model yang memberikan hasil klasifikasi terbaik.

Keseluruhan tahapan penelitian disajikan pada Gambar 1.



Gambar 1. Alur Penelitian Analisis Sentimen Komentar *YouTube* Program Makan Bergizi Gratis

2.1 Pengumpulan Data

Data penelitian bersumber dari data sekunder yang diperoleh melalui platform Kaggle. Sumber data tersebut berupa *dataset* dalam format CSV yang memuat 6.419 komentar pengguna YouTube terkait Program Makan Bergizi Gratis (MBG). Teks komentar digunakan sebagai atribut utama dalam proses analisis sentimen, sedangkan atribut lainnya berfungsi sebagai informasi pendukung. Setelah dataset diperoleh, dilakukan pemeriksaan terhadap struktur dan kualitas data untuk memastikan kelengkapan serta kesesuaiannya sebelum memasuki tahap *preprocessing* [8].

2.2 *Exploratory Data Analysis* (EDA)

Sebelum memasuki tahap *preprocessing*, dilakukan pemeriksaan terhadap data untuk memperoleh gambaran mengenai kondisi dataset. Tahapan ini meliputi identifikasi karakteristik data, distribusi, nilai yang hilang (*missing values*), serta kecenderungan pola yang terdapat pada data. Proses tersebut dikenal sebagai *Exploratory Data Analysis* (EDA) dan berperan sebagai dasar dalam menentukan langkah pengolahan data selanjutnya. EDA dilakukan sebelum proses pemodelan sebagai langkah untuk memahami kondisi data yang akan dianalisis [9]. Pada tahap ini dilakukan analisis terhadap informasi dataset, pemeriksaan *missing value*, serta distribusi data sentimen. Hasil dari EDA digunakan untuk mengidentifikasi kondisi awal data, mengetahui kualitas dataset, dan memastikan data siap digunakan pada tahapan *preprocessing* serta proses klasifikasi.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

2.3 *Preprocessing* Data

Data yang telah dikumpulkan selanjutnya melalui proses seleksi dan pembersihan guna memperoleh komentar yang sesuai untuk dianalisis. Seleksi dilakukan dengan memperhatikan relevansi isi komentar, kejelasan bahasa, dan kualitas data sehingga dapat digunakan pada tahapan analisis berikutnya [10]. *Preprocessing* dilakukan menghasilkan data teks yang bersih, konsisten, dan siap digunakan sebagai masukan bagi algoritma klasifikasi. Tahapan *preprocessing* yang dilakukan meliputi:

a. *Case Folding*: Tahap ini dilakukan dengan mengonversi seluruh karakter alfabet pada teks menjadi huruf kecil (*lowercase*) sehingga penulisan setiap kata menjadi seragam.
b. *Cleaning*: Pada tahap ini, data teks dibersihkan dari berbagai elemen yang tidak memiliki kontribusi terhadap proses analisis sentimen, meliputi URL, angka, tanda baca, emoji, simbol, dan karakter khusus lainnya.
c. *Tokenizing*: Memecah kalimat atau dokumen menjadi elemen-elemen kata terpisah (*token*) supaya siap masuk pada tahap selanjutnya.
d. *Normalization*: Tahap ini menghasilkan bentuk penulisan yang baku dengan menyesuaikan kata tidak baku, singkatan, maupun bahasa gaul ke dalam padanan kata sesuai kamus yang digunakan.
e. *Normalization*: Tahap ini menghasilkan bentuk penulisan yang baku dengan menyesuaikan kata tidak baku, singkatan, maupun bahasa gaul ke dalam padanan kata sesuai kamus yang digunakan.
f. *Stemming*: menyederhanakan kata berimbuhan ke bentuk kata dasar sehingga berbagai variasi kata yang memiliki makna serupa dapat direpresentasikan dalam satu bentuk.

2.4 Pelabelan Sentimen

Pendekatan *lexicon-based* merupakan metode analisis sentimen yang menggunakan kamus (*lexicon*) berisi kata beserta nilai sentimennya untuk menentukan polaritas suatu teks [11]. Pada tahap ini, setiap komentar dianalisis berdasarkan kamus *lexicon* untuk memperoleh skor sentimen. Skor Penentuan label sentimen dilakukan berdasarkan hasil yang diperoleh pada tahapan sebelumnya. dikelompokkan ke dalam tiga kategori, yaitu positif, netral, dan negatif. Selanjutnya, hasil pelabelan dimanfaatkan sebagai target (*label*) dalam proses pelatihan (*training*) maupun pengujian (*testing*) model klasifikasi sentimen.

2.5 Pembagian Data (*Data Splitting*)

Melalui *data splitting*, keseluruhan dataset akan dikelompokkan menjadi himpunan latih (*training*) dan himpunan uji *(testing)* [12]. Setelah proses *label encoding* selesai, *dataset* dibagi menjadi dua bagian, yaitu data pelatihan (*training*) sebesar 80% dan data pengujian (*testing*) sebesar 20% menggunakan fungsi *train_test_split*. Pembagian data dilakukan untuk memisahkan data yang digunakan dalam proses pelatihan model dari himpunan data yang dipakai pada tahap validasi [13]. Porsi data latih digunakan guna membangun sistem model Naive Bayes, Support Vector Machine (SVM), dan IndoBERT, sedangkan data pengujian digunakan untuk mengevaluasi kemampuan masing-masing model untuk mengidentifikasi kategori sentimen pada komentar mengenai Program Makan Bergizi Gratis (MBG).

2.6 Ekstraksi Fitur

Tahap ekstraksi fitur dilakukan untuk mengubah data teks hasil *preprocessing* menjadi representasi numerik yang dapat diproses oleh model klasifikasi [14]. *TF* menunjukkan jumlah kemunculan kata pada setiap dokumen, sedangkan IDF digunakan untuk menentukan tingkat kelangkaan istilah tersebut berdasarkan keseluruhan kumpulan dokumen [15]. Pada tahap ekstraksi fitur, TF-IDF digunakan untuk merepresentasikan data teks dalam bentuk fitur numerik yang kemudian diproses oleh algoritma Naive Bayes dan Support Vector Machine (SVM). Pada model IndoBERT, pembentukan representasi data dilakukan melalui IndoBERT Tokenizer yang mengubah komentar menjadi token sesuai kebutuhan masukan model. Selanjutnya, hasil tokenisasi dikonversi menjadi PyTorch Tensor agar dapat diproses oleh model IndoBERT pada tahap pelatihan dan pengujian. Hasil ekstraksi fitur kemudian digunakan sebagai data masukan pada proses pelatihan dan pengujian masing-masing model klasifikasi.

2.7 Penanganan Ketidakseimbangan Data

Penanganan ketidakseimbangan data bertujuan untuk menghasilkan distribusi kelas yang lebih seimbang sebelum proses klasifikasi [16]. Setelah data pelatihan diekstraksi menjadi fitur numerik menggunakan TF-IDF, metode SMOTE diterapkan sebelum data diproses oleh algoritma Naive Bayes dan Support Vector Machine (SVM). Sementara itu, pelatihan model IndoBERT memanfaatkan data yang telah melalui proses tokenisasi oleh IndoBERT Tokenizer tanpa penerapan SMOTE. Selanjutnya dilakukan evaluasi untuk menilai kinerja pada setiap model setelah penanganan ketidakseimbangan data.

2.8 Klasifikasi Sentimen

Tahap klasifikasi sentimen pada penelitian ini bertujuan mengklasifikasikan setiap komentar ke dalam kategori sentimen positif, netral, atau negatif berdasarkan isi teks [17]. Untuk mencapai tujuan tersebut, sebagai model

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

pembanding, penelitian ini memanfaatkan algoritma Naive Bayes, Support Vector Machine (SVM), dan IndoBERT untuk melakukan klasifikasi sentimen. Data yang digunakan pada tahap klasifikasi merupakan hasil dari preprocessing, *label encoding*, pembagian data (*data splitting*), dan ekstraksi fitur. Khusus pada model Naive Bayes dan Support Vector Machine (SVM), metode Synthetic Minority Oversampling Technique (SMOTE) diterapkan pada data latih untuk menyeimbangkan distribusi jumlah sampel sehingga setiap kelas sentimen memiliki jumlah data yang lebih proporsional. Berbeda dengan kedua model tersebut, IndoBERT memanfaatkan token yang dihasilkan oleh IndoBERT Tokenizer sebagai representasi masukan sesuai dengan mekanisme kerja arsitektur Transformer. Selanjutnya, kinerja masing-masing algoritma dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menentukan model yang memberikan hasil terbaik dalam mengklasifikasikan sentimen masyarakat terkait pelaksanaan Program Makan Bergizi Gratis (MBG).

2.9 Evaluasi Model

Tahap akhir penelitian berfokus pada penilaian kemampuan setiap model klasifikasi dalam memprediksi data uji yang tidak terlibat selama proses pelatihan [4]. Kemampuan model dalam melakukan prediksi terhadap data yang belum pernah diproses sebelumnya dinilai melalui tahap evaluasi kinerja [18]. Tahap evaluasi bertujuan menilai performa algoritma Naive Bayes, Support Vector Machine (SVM), dan IndoBERT untuk melakukan klasifikasi sentimen pada komentar mengenai Program Makan Bergizi Gratis (MBG). Penilaian dilakukan menggunakan confusion matrix yang menyajikan hubungan antara hasil prediksi dan label sebenarnya pada setiap kategori sentimen. Berdasarkan hasil tersebut, performa masing-masing algoritma dianalisis menggunakan metrik accuracy, precision, recall, dan F1-score. Seluruh hasil pengukuran kemudian dijadikan dasar untuk membandingkan kemampuan ketiga algoritma sehingga dapat ditentukan model yang memberikan kinerja paling baik pada penelitian ini.

3. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil pengujian komparatif algoritma klasifikasi sentimen publik terhadap Program Makan Bergizi Gratis di YouTube. Analisis dilakukan secara mendalam mencakup evaluasi performa model sebelum dan sesudah optimasi penanganan data.

3.1 Hasil *Preprocessing* Data

Sebelum memasuki proses klasifikasi, kualitas data komentar ditingkatkan melalui serangkaian tahapan *preprocessing*. Tahapan ini berperan dalam menyiapkan data agar memiliki format yang seragam dan layak untuk dianalisis. Proses yang diterapkan terdiri atas *case folding*, *cleaning*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*. Perubahan yang duhasilkan pada tahapan disajikan secara rinci pada Tabel 1.

Tabel 1. Hasil Tahapan *Preprocessing* Datas

Case Folding	penguasa dan pelaksana program mbg menjadi penguasa naga ke 10 (dari penguasa 9 naga)
Cleaning	penguasa dan pelaksana program mbg menjadi penguasa naga ke dari penguasa naga
Tokenizing	[penguasa, dan, pelaksana, program, mbg, menjadi, penguasa, naga, ke, dari, penguasa, naga]
Normalization	[penguasa, dan, pelaksana, program, mbg, menjadi, penguasa, naga, ke, dari, penguasa, naga]
Stopword Removal	[penguasa, pelaksana, program, mbg, menjadi, penguasa, naga, penguasa, naga]
Stemming	[kuasa, laksana, program, mbg, jadi, kuasa, naga, kuasa, naga]
Final Preprocessing Text	kuasa laksana program mbg jadi kuasa naga kuasa naga

Tabel 1 memperlihatkan perubahan teks komentar secara bertahap yang menjadi semakin rapi dan bersih setelah melewati setiap proses *preprocessing*. Eliminasi karakter angka, simbol, serta kata fungsional pada tahap awal berhasil memangkas derau teks, yang kemudian disempurnakan dengan pemotongan imbuhan menjadi kata dasar di tahap akhir. Pola transformasi yang sistematis ini menghasilkan baris *Final Preprocessing Text* yang sangat rapi dan siap diproses oleh algoritma klasifikasi.

3.2 Hasil Pelabelan Sentimen

Setiap komentar yang telah melalui tahap prapemrosesan diberikan kategori sentimen secara otomatis menggunakan pendekatan *lexicon-based*. Penentuan kategori dilakukan berdasarkan skor yang diperoleh dari kamus sentimen (*lexicon*),

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

sehingga setiap data diklasifikasikan ke dalam kelas positif, netral, atau negatif. Label yang dihasilkan pada tahap ini selanjutnya menjadi acuan dalam proses analisis dan pemodelan.

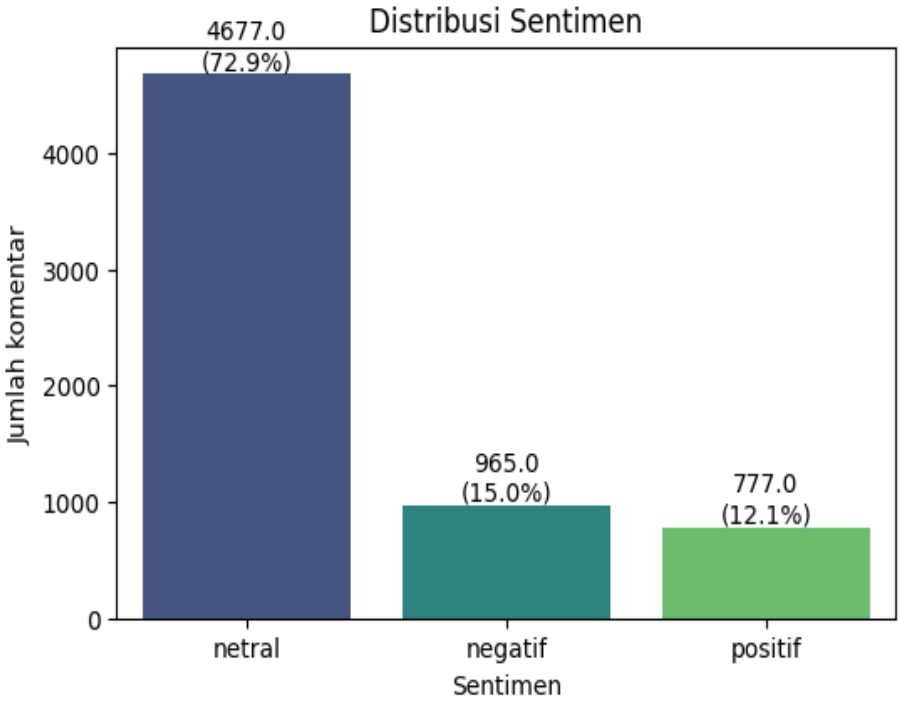
Tabel 2. Label Sentimen Berdasarkan Pendekatan *Lexicon-Based*

Teks Komentar	Hasil Pelabelan Sentimen
konten tarik bangun wawas bangsa indonesia lebih kritis br cuma sayang presentasi jangan terlalu formal depan jabat br coba review youtuber rumah editor kamar film dokumentasi bagus presentasi bagus youtube br tetap semangat kak dukung konten jangan lupa cantum refrensi laku crosschec	Positif
partai yayasan aph layak jadi mitra mbg py ahli mampu alam dlm bidang kuliner	Positif
rusak lah negara kalau nepotisme korupsi makin terang terang pelihara mbg siap langgeng kuasa bangsat kan ancur negara indonesia	Negatif
minimal chanel youtube netizen banyak tolak mbg banding balik tik tok sangat banyak buzzer perintah idiot memang beda kelas netizen youtube netizen tik tok	Negatif
anak makan x bapak penganguran	Netral
maling kedok gizi	Netral

Tabel 2 menyajikan hasil otomatisasi pelabelan dokumen sentimen menggunakan pendekatan *lexicon-based* untuk mengatasi ketiadaan label orisinal pada *dataset* komentar mentah. Melalui mekanisme ini, setiap baris opini publik dapat diidentifikasi secara objektif ke dalam polaritas sentimen positif, negatif, atau netral berdasarkan pembobotan nilai kata yang terkandung di dalamnya. Proses anotasi otomatis ini menjadi landasan krusial untuk memetakan distribusi sebaran opini sekaligus menyediakan data latih (*training data*) yang valid bagi tahapan klasifikasi algoritma tingkat lanjut.

3.3 Visulisasi Distribusi Sentimen

Untuk mengetahui persebaran hasil pelabelan sentimen, dilakukan Visualisasi distribusi data dilakukan untuk menunjukkan persebaran hasil pelabelan pada setiap kategori sentimen yang terdiri atas kategori positif, negatif, dan netral.



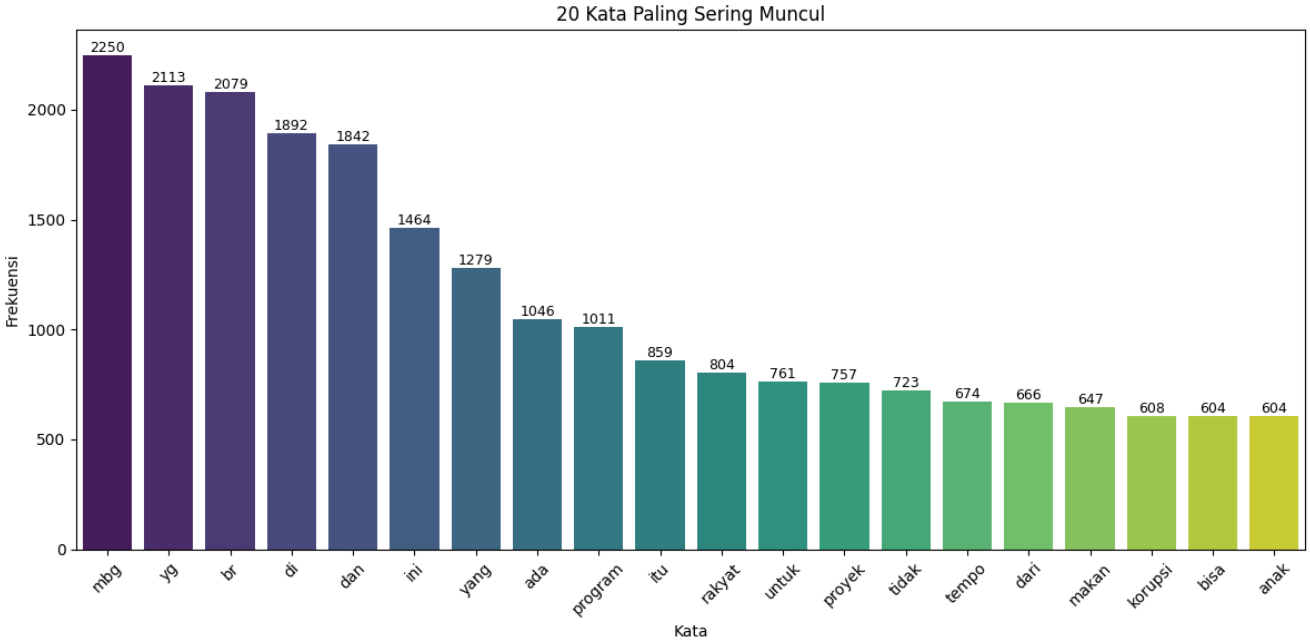
Gambar 2. Grafik Batang Distribusi Persentase Hasil Pelabelan Sentimen

Hasil pelabelan sentimen menunjukkan bahwa kelas sentimen netral memiliki jumlah komentar terbanyak, yaitu sebanyak 4.677 komentar atau 72,9% dari keseluruhan data. Selanjutnya, kelas sentimen negatif berjumlah 965 komentar atau 15,0%, sedangkan kelas sentimen positif berjumlah 777 komentar atau 12,1%. Distribusi tersebut menunjukkan bahwa data didominasi oleh sentimen netral dibandingkan dengan sentimen negatif maupun sentimen positif.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

3.4 Analisis Frekuensi Kata

Setelah melalui tahapan prapemrosesan, data komentar dianalisis untuk mengetahui kata-kata yang paling dominan digunakan dalam pembahasan Program Makan Bergizi Gratis (MBG). Analisis ini menghasilkan daftar 20 kata dengan tingkat kemunculan tertinggi.

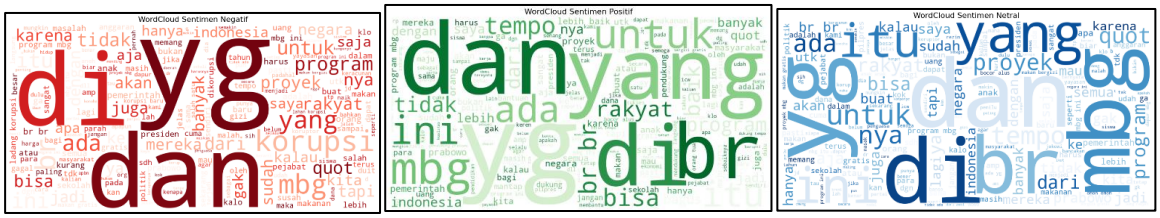


Gambar 3. Diagram Frekuensi Kemunculan Kata Tertinggi pada Komentar Program MBG

Berdasarkan Gambar 3, kata “mbg” mendominasi hasil analisis frekuensi kata dengan jumlah kemunculan tertinggi pada data komentar dengan frekuensi sebanyak 2.250 kali, diikuti oleh kata “yg” sebanyak 2.113 kali, “br” sebanyak 2.079 kali, “di” sebanyak 1.892 kali, dan “dan” sebanyak 1.842 kali. Selain itu, beberapa kata lain, seperti “program”, “rakyat”, “proyek”, “makan”, dan “anak”, juga memiliki frekuensi kemunculan yang tinggi. Hal ini menunjukkan bahwa komentar masyarakat banyak membahas pelaksanaan Program Makan Bergizi Gratis (MBG).

3.5 *Word cloud* Berdasarkan Sentimen

Word cloud dimanfaatkan sebagai media untuk menggambarkan dominasi istilah pada setiap kategori sentimen. Gambar 4 memperlihatkan hasil pemetaan tersebut.



(a) Negatif (b) Positif (c) Netral

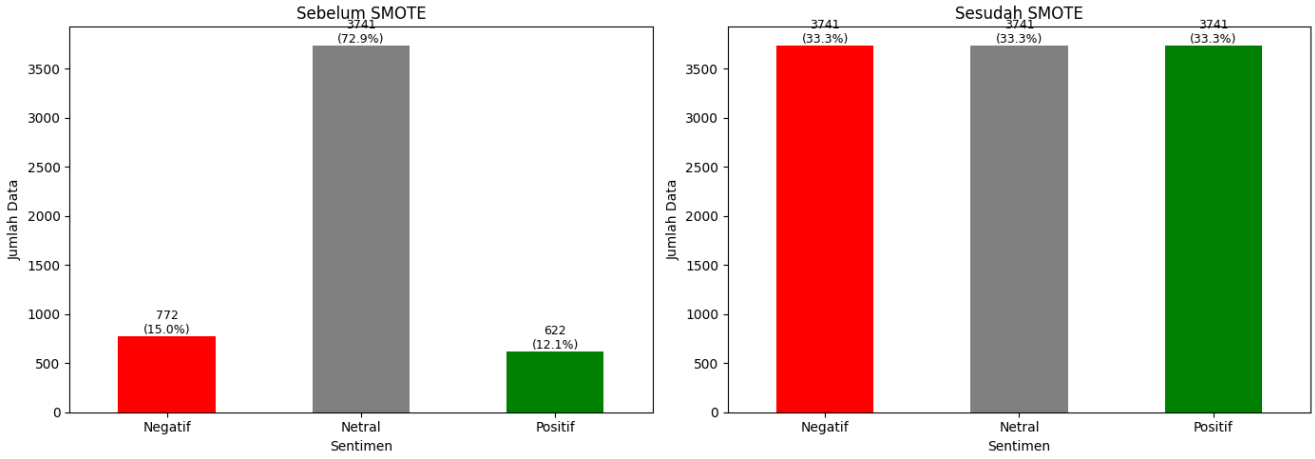
Gambar 4. *Word cloud* Berdasarkan Sentimen

Berdasarkan Gambar 4, Visualisasi *word cloud* memperlihatkan istilah yang paling sering muncul pada masing-masing kategori sentimen. Pada kategori sentimen negatif, istilah “dan”, “di”, dan “ yg” memiliki frekuensi kemunculan yang lebih tinggi dibandingkan istilah lainnya. Pada sentimen positif, kata “dan”, “yg”, "untuk", “br”, “ada”, “rakyat”, ‘tidak”, dan “mbg” menjadi kata yang paling dominan. Sementara itu, pada sentimen netral, kata “yg”, “dan”, “di”, “bo”, “untuk”, “ada”, “proyek”, dan “rakyat” memiliki ukuran yang lebih besar dibandingkan kata lainnya. Ukuran setiap istilah pada *word cloud* merepresentasikan tingkat frekuensi kemunculannya, sehingga istilah yang ditampilkan lebih besar menunjukkan frekuensi yang lebih tinggi pada kategori sentimen terkait.

3.6 Visualisasi Data Sebelum dan Sesudah SMOTE

Ketidakseimbangan distribusi data latih pada setiap kelas sentimen diatasi secara optimal menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE).

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416



(a) Sebelum *SMOTE* (b) Sesudah *SMOTE*

Gambar 5. Distribusi Data Sebelum dan Sesudah *SMOTE*

Distribusi data latih sebelum dilakukan penyeimbangan masih didominasi oleh kelas sentimen netral sebanyak 3.741 data, sedangkan kelas sentimen negatif dan positif masing-masing terdiri atas 772 dan 622 data. Ketidakseimbangan tersebut kemudian ditangani menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). Setelah tahap penyeimbangan selesai, setiap kelas sentimen memiliki jumlah data yang sama, yaitu 3.741 data. Kondisi ini menunjukkan bahwa metode SMOTE mampu menghasilkan distribusi data latih yang lebih proporsional sehingga setiap kelas memperoleh representasi yang setara selama proses pelatihan model.

3.7 Hasil Perbandingan Kinerja Model

Perbandingan performa algoritma Naive Bayes, Support Vector Machine (SVM), dan IndoBERT dilakukan berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Ringkasan hasil pengukuran ketiga algoritma ditampilkan pada Tabel 3.

Tabel 3. Hasil Perbandingan Kinerja Model Klasifikasi

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naive Bayes	65,11%	77,13%	65,11%	67,24%
Support Vector Machine (SVM)	95,72%	95,74%	95,72%	95,72%
IndoBERT	97,35%	97,40%	97,35%	97,32%

Perbandingan pada Tabel 3 menempatkan IndoBERT sebagai algoritma dengan hasil terbaik. Nilai *accuracy* yang diperoleh mencapai 97,35%, disertai *precision* sebesar 97,40%, *recall* 97,35%, dan *F1-score* 97,32%. Nilai tersebut mencerminkan kemampuan IndoBERT dalam melakukan klasifikasi sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG) dengan tingkat ketepatan yang lebih tinggi dibandingkan Naive Bayes dan Support Vector Machine (SVM).

3.8. Hasil Evaluasi Model Terbaik

Tabel 4 menyajikan *classification report* model IndoBERT berdasarkan hasil pengujian pada data uji.

Tabel 4. Hasil Klasifikasi Model IsndoBERT

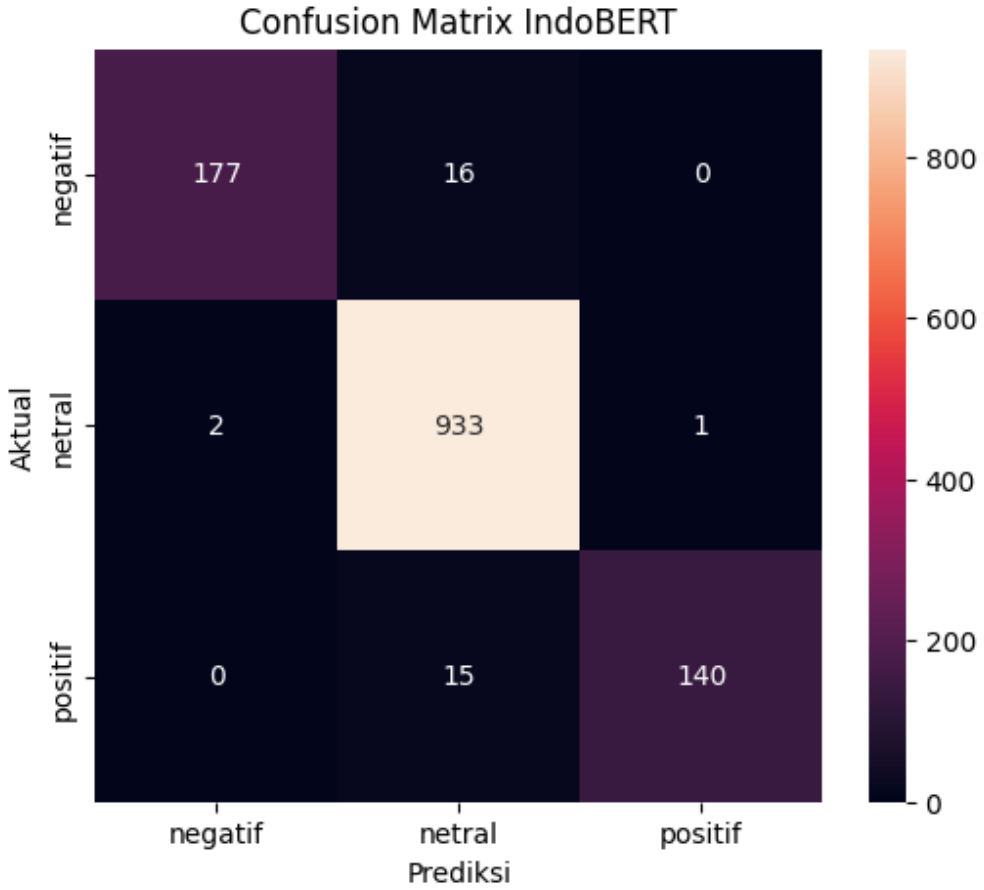
Kelas Sentimen	*Precision*	*Recall*	*F1-Score*	*Support*
Negatif	98,88%	91,71%	95,16%	193
Netral	96,78%	99,68%	98,21%	936
Positif	99,29%	90,32%	94,59%	155
Accuracy			97,35%	1284
Macro avg	98,32%	93,90%	95,99%	1284
Weighted avg	97,40%	97,35%	97,32%	1284

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

Nilai *F1-score* tertinggi pada Tabel 4 dicapai oleh kelas Netral dengan persentase 98,21%. Kelas Negatif dan Positif masing-masing mencatat *F1-score* sebesar 95,16% dan 94,59%. Selain itu, model mencatat nilai *accuracy* keseluruhan sebesar 97,35%.

3.9 Evaluasi Model Terbaik

Evaluasi terhadap model IndoBERT dilakukan menggunakan *confusion matrix* untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kelas sentimen. Hasil *confusion matrix* model IndoBERT disajikan pada Gambar 6.



Gambar 6. *Confusion Matrix* Model IndoBERT Hasil Klasifikasi Sentimen

Confusion matrix pada Gambar 6 digunakan untuk menggambarkan hasil prediksi model IndoBERT terhadap setiap kategori sentimen. Sebagian besar data uji berhasil ditempatkan pada kelas yang ssesuai, yaitu 177 data untuk sentimen negatif, 933 data untuk sentimen netral, dan 140 data untuk sentimen positif. Kesalahan prediksi hanya ditemukan pada sejumlah kecil data, yang sebagian besar berpindah ke kategori sentimen netral. Kondisi tersebut mencerminkan kemampuan IndoBERT dalam mengenali pola sentimen pada komentar mengenai Program Makan Bergizi Gratis (MBG), sejalan dengan nilai *accuracy* yang mencapai 97,35%.

4. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa implementasi rangkaian pengolahan data yang sistematis, mulai dari tahapan *preprocessing* teks, otomatisasi pelabelan dokumen, pembagian data, hingga ekstraksi fitur, mampu meningkatkan kualitas data secara signifikan. Selain itu, penerapan metode *Synthetic Minority Oversampling Technique* (SMOTE) terbukti efektif dalam mengatasi kendala ketidakseimbangan kelas (*class imbalance*) pada model *Naïve Bayes* dan *Support Vector Machine* (SVM) dengan menghasilkan distribusi data latih yang jauh lebih proporsional serta seimbang di setiap kelas sentimen. Melalui pengujian komparatif secara menyeluruh, hasil eksperimen menunjukkan bahwa arsitektur berbasis *Deep Learning* yaitu IndoBERT menyajikan performa klasifikasi yang paling superior dan optimal dibandingkan dengan model *baseline Naïve Bayes* maupun *Support Vector Machine* (SVM) dalam memetakan polaritas analisis sentimen terkait Program Makan Bergizi Gratis (MBG) pada media sosial YouTube. Keunggulan performa yang dicapai oleh IndoBERT ini didorong oleh kapabilitas arsitekturnya yang berbasis *Transformer* dan bersifat *context-aware*, sehingga mampu menangkap esensi semantik, konteks kalimat, serta karakteristik struktur bahasa Indonesia informal

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1703-1712
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1416

secara mendalam dan konsisten. Dengan demikian, penelitian ini menyimpulkan bahwa IndoBERT merupakan model yang sangat efektif, andal, dan layak dipertimbangkan sebagai referensi utama untuk kebutuhan riset analisis sentimen berskala besar pada korpus dokumen berbahasa Indonesia yang tidak terstruktur di masa mendatang.

REFERENCES

[1] W. W. Anshory, "Naik Hampir Dua Kali Lipat, Luhut Sebut Anggaran MBG 2026 Jadi Rp 300 Triliun," Kompas.com. Accessed: Jun. 30, 2026. [Online]. Available: https://www.kompas.com/sumatera-utara/read/2025/06/12/133500688/naik-hampir-dua-kali-lipat-luhut-sebut-anggaran-mbg-2026-jadi-rp?

[2] A. Rendanianti, "Platform Video yang Sering Ditonton Pengguna Internet Indonesia 2026, YouTube Teratas!," GoodStats Data. Accessed: Jul. 08, 2026. [Online]. Available: https://data.goodstats.id/statistic/platform-video-yang-sering-ditonton-pengguna-internet-indonesia-2026-youtube-teratas-gzzii

[3] Y. Z. Vebrian, T. Informatika, and U. N. Waluyo, "A SENTIMENT ANALYSIS OF FREE MEAL PLANS ON SOCIAL MEDIA USING NAÏVE BAYES ALGORITHMS ANALISIS SENTIMEN TERHADAP RENCANA MAKAN GRATIS DI SOSIAL MEDIA X MENGGUNAKAN," vol. 10, no. 1, 2025.

[4] A. D. P. Ibnu Abdul Ghofur, "P a g e | 100 Computer Based Information System Journal PERBANDINGAN KINERJA SVM , KNN , DAN NAÏVE BAYES PADA ANALISIS SENTIMEN TWEET MBG," vol. 01, pp. 100–117, 2026.

[5] W. R. Muntaza and T. Informatika, "Analisis Sentimen Tanggapan Masyarakat Terhadap Program Makan Bergizi Gratis Pada Platform YouTube Menggunakan SVM," vol. 5, pp. 297–302, 2026.

[6] B. Rahmatullah, S. A. Saputra, P. Budiono, and D. P. Wigandi, "Sentimen Analisis Makan Bergizi Gratis Menggunakan Algoritma Naive Bayes," vol. 05, no. 01, 2025.

[7] S. Rohman and I. W. P. Agung, "DALAM ANALISIS SENTIMEN PROGRAM MAKAN BERGIZI GRATIS," vol. 9, no. 6, pp. 9464–9471, 2025.

[8] E. Kurniasari, F. K. Putra, U. Gunadarma, A. Sentimen, P. D. Nasional, and M. Sosial, "Pemanfaatan Metode Naïve Bayes untuk Analisis Sentimen Terkait Insiden Server Down dan Peretasan pada Pusat Data Nasional," vol. 14, pp. 926–933, 2025.

[9] D. A. Arvi Pramudyantoro, Ema Utami, "Penggabungan k-nearest neighbors dan lightgbm untuk prediksi diabetes pada dataset pima indians: menggunakan pendekatan exploratory data analysis 1.," vol. 9, no. 3, pp. 1133–1144, 2024.

[10] F. F. R. Agnia Suci Rizkia, Wufron, "Sentiment Analysis of Coretax : A Comparison of Manual , Transformers-Based , and Lexicon-Based Data Labeling on IndoBERT Performance Analisis Sentimen Coretax : Perbandingan Pelabelan Data Manual ," vol. 5, no. July, pp. 1037–1048, 2025.

[11] Z. Akbar, I. Riadi, and R. Umar, "Analisis Sentimen Program Makan Bergizi Gratis Menggunakan Lexicon-Based dan Support Vector Machine," vol. 16, no. 1, pp. 150–156, 2026.

[12] A. P. Maretta, A. N. Kusuma, R. M. Sasmita, A. F. Rizkyllah, and A. Ibrahim, "PERBANDINGAN METODE NAÏVE BAYES DAN K-NEAREST NEIGHBOR TERHADAP ANALISIS SENTIMEN ULASAN PROGRAM MAKAN SIANG GRATIS DI INDONESIA," vol. 9, no. 4, 2025.

[13] Y. A. Prasetyo, E. Utami, and A. Yaqin, "Pengaruh Komposisi Split Data Terhadap Performa Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM," vol. 6, no. 2, pp. 382–390, 2024, doi: 10.33650/jeecom.v4i2.

[14] A. Firizkiansah, A. Muhammad, I. R. Maulana, U. S. Indonesia, and K. Bekasi, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," vol. 2, no. 1, pp. 29–36, 2025.

[15] I. I. Abdurrahman Arasy, Surya Agustian, Lestari Handayani, "Sentiment Classification Using Multilayer Perceptron Algorithm with TF-IDF Features Klasifikasi Sentimen Menggunakan Metode Multilayer Perceptron dengan Fitur TF-IDF," vol. 5, no. July, pp. 908–919, 2025.

[16] A. A. Qolbu, N. Fitriyati, and N. Inayah, "Performa Naïve Bayes , SVM , dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang," vol. 814, no. 1, pp. 29–44, 2025, doi: 10.14421/fourier.2025.141.29-44.

[17] T. Setyani, K. Sari, H. N. Heidy, and R. R. Suryono, "Sentiment Analysis of Jamsostek Mobile Application Reviews on Google Play Store Using Support Vector Machine and Naive Bayes Algorithms Analisis Sentimen Pengguna Aplikasi Jamsostek Mobile Berdasarkan Ulasan Google Play Store Menggunakan Algoritma Support Vector Machine dan Naive Bayes," vol. 6, no. January, pp. 373–384, 2026.

[18] A. Ahfaz, R. Ramdani, Y. B. Pratama, A. Pramudyantoro, E. Altiarika, and Z. Wahyuzi, "VIDEO PROMOSI BISNIS KULINER DI BANGKA BELITUNG," vol. 3, no. 1, pp. 7–12, 2025.