Vol. 7 No. 1 Juni 2026 Hal : 153 - 163 E-ISSN: 2723-1089
P-ISSN: 2776-7779

Analisis Sentimen Program Makanan Bergizi Gratis Menggunakan Naïve Bayes dan Random Forest Berbasis CRISP-DM

Ragilia Putri Dinanti[1)], Siti Sarah[2)*], Fiqri Dwi Al Hafiz[3)], Aidil Alfarizi[4)]

[1,2,3,4)]Ilmu Komputer, Fakultas Sains Dan Teknologi,
Universitas Islam Negeri Sumatera Utara, Indonesia

*Coresponding Email: ragiliaputri14@gmail.com, sitisarah210305@gmail.com, fikridwialhafis@gmail.com, aidilalfarizi1101@gmail.com

Abstrak

Program Makanan Bergizi Gratis (MBG) merupakan kebijakan strategis pemerintah Indonesia untuk menangani masalah kekurangan gizi dan menekan angka stunting yang mencapai 14% pada tahun 2024 demi mencapai visi Indonesia Emas 2045. Kebijakan ini memicu diskusi publik yang masif di media sosial YouTube, yang menghadirkan tantangan berupa volume data besar dan keberagaman opini. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap program MBG menggunakan metodologi Cross-Industry Standard Process for Data Mining (CRISP-DM). Dataset yang digunakan bersumber dari Kaggle yang terdiri dari 6.419 komentar YouTube. Data diproses melalui tahapan preprocessing teks dan representasi fitur TF-IDF, kemudian diklasifikasikan menggunakan algoritma Naïve Bayes dan Random Forest. Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan performa yang lebih unggul dengan akurasi sebesar 77,43%, sementara Naïve Bayes mencapai 65,64%. Berdasarkan distribusi data, sentimen Netral mendominasi sebesar 56,40%, diikuti oleh sentimen Positif (24,13%) dan Negatif (19,47%). Dominasi sentimen netral ini mengindikasikan bahwa masyarakat masih bersikap wait-and-see terhadap efektivitas implementasi program MBG di lapangan. Penelitian ini menyimpulkan bahwa pendekatan ensemble learning pada Random Forest lebih efektif dalam menangkap pola bahasa alami yang kompleks dibandingkan metode berbasis probabilitas murni.

Kata Kunci: Analisis Sentimen; Makanan Bergizi Gratis; Naïve Bayes; Random Forest; CRISP-DM

Abstract

The Free Nutritious Food Program (MBG) is a strategic policy of the Indonesian government aimed at addressing malnutrition and reducing the stunting rate—which is projected to reach 14% by 2024—in order to achieve the Indonesia Emas 2045 vision. This policy has sparked massive public discussion on YouTube, presenting challenges in the form of large volumes of data and a diversity of opinions. This study aims to analyze public sentiment toward the MBG program using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The dataset used is sourced from Kaggle and consists of 6,419 YouTube comments. The data was processed through text preprocessing and TF-IDF feature representation stages, then classified using the Naïve Bayes and Random Forest algorithms. The results show that the Random Forest algorithm outperformed the Naïve Bayes algorithm, achieving an accuracy of 77.43% compared to 65.64%. Based on the data distribution, Neutral sentiment dominated at 56.40%, followed by Positive (24.13%) and Negative (19.47%). This dominance of neutral sentiment indicates that the public remains in a wait-and-see stance regarding the effectiveness of the MBG program's implementation on the ground. This study concludes that the

ensemble learning approach using Random Forest is more effective at capturing complex natural language patterns compared to purely probability-based methods.

Keywords: Sentiment Analysis; Free Nutritious Food; Naive Bayes; Random Forest; CRISP-DM

PENDAHULUAN

Program Makan Bergizi Gratis (MBG) merupakan kebijakan strategis yang diluncurkan oleh pemerintah Indonesia dengan tujuan utama menangani permasalahan kekurangan gizi dan meningkatkan kualitas Sumber Daya Manusia (SDM). Urgensi program ini didasarkan pada angka *stunting* di Indonesia tahun 2024 yang mencapai 14%, sehingga pemenuhan hak dasar anak atas pangan yang aman, sehat, dan bergizi menjadi prioritas nasional demi mencapai visi Indonesia Emas 2045 [1]. Kebijakan ini merupakan bentuk transformasi pendidikan dan kesehatan yang menyasar anak-anak usia sekolah guna meningkatkan produktivitas jangka panjang individu [2]. Sebagai salah satu program unggulan yang menyentuh hajat hidup orang banyak, peluncuran MBG memicu diskursus publik yang sangat masif di berbagai platform media online [3]. Media sosial saat ini telah menjadi sarana komunikasi yang paling efektif, transparan, dan efisien bagi masyarakat untuk menyampaikan dukungan maupun kritik terhadap kebijakan pemerintah [4].

Namun, volume data yang sangat besar di platform digital tersebut menghadirkan tantangan berupa "Data Sampah" (*noise*) seperti spam dan akun bot yang tidak memiliki substansi informasi [5]. Keberadaan data yang tidak relevan ini berpotensi menyebabkan bias dalam analisis opini publik jika tidak ditangani secara sistematis. Oleh karena itu, pendekatan analisis sentimen diperlukan sebagai solusi untuk mengevaluasi kualitas kebijakan berdasarkan persepsi masyarakat. Analisis sentimen sendiri merupakan gabungan dari pemrosesan bahasa alami dan penambangan teks yang mengkaji pandangan, perasaan, serta emosi masyarakat terhadap suatu entitas atau isu [6]. Penggunaan teknologi kecerdasan buatan dalam memproses data teks ini terbukti mampu menangkap aspek emosional yang sering kali tidak terjangkau oleh metode evaluasi kuantitatif tradisional [7].

Dalam bidang pengolahan bahasa alami, implementasi *machine learning* berfokus pada pembangunan sistem yang dapat mempelajari dan meningkatkan kinerja secara mandiri berdasarkan data yang tersedia [8]. Algoritma Random Forest sering menjadi pilihan utama dalam menangani data berdimensi tinggi karena kemampuannya dalam mengurangi *overfitting* melalui agregasi banyak pohon keputusan [9]. Di sisi lain, algoritma Naïve Bayes sangat populer karena kesederhanaan dan efisiensi komputasinya, terutama dalam mengklasifikasikan dokumen berdasarkan frekuensi kemunculan kata [10]. Masalah utama dalam analisis opini masyarakat di Indonesia adalah keberagaman kosakata dan penggunaan bahasa informal yang kompleks, sehingga diperlukan perbandingan model klasifikasi untuk menentukan algoritma mana yang paling *robust* dalam menangkap pola bahasa alami tersebut [7].

Untuk menjamin kualitas penelitian yang terstruktur, metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) diterapkan sebagai kerangka kerja utama. CRISP-DM merupakan standar industri yang menyediakan tahapan sistematis mulai dari *Business Understanding* hingga *Evaluation* guna memastikan hasil penambangan data sesuai dengan tujuan organisasi [11]. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap program Makan Bergizi Gratis dengan membandingkan performa algoritma Naïve Bayes dan Random Forest berbasis pendekatan CRISP-DM. Melalui pemanfaatan data komentar dari platform YouTube, penelitian ini berusaha mengidentifikasi pola persepsi publik serta memberikan wawasan berbasis data yang berharga bagi para pengambil kebijakan dalam mengevaluasi efektivitas implementasi program di lapangan. Hasil akhir dari penelitian ini diharapkan dapat menjadi landasan bagi pemerintah dalam merancang strategi komunikasi dan pelayanan yang lebih responsif terhadap kebutuhan masyarakat.

METODE PENELITIAN

1. Business Understanding

Tahap ini bertujuan untuk mendefinisikan tujuan penelitian dari perspektif analisis data. Tujuan utama yang ditetapkan adalah mengklasifikasikan sentimen publik terhadap program MBG menjadi tiga kategori (Positif, Negatif, Netral) berdasarkan komentar YouTube, serta membandingkan kinerja algoritma Naïve Bayes dan Random Forest dalam menyelesaikan tugas klasifikasi tersebut.

2. Data Understanding

Data yang digunakan dalam penelitian ini adalah komentar YouTube yang dikumpulkan dari berbagai video yang membahas program MBG. Total data yang berhasil dikumpulkan berjumlah 6.419 komentar. Proses pelabelan sentimen dilakukan secara manual dengan bantuan pendekatan leksikon sentimen berbahasa Indonesia.

3. Data Preparation

Tahap persiapan data mencakup serangkaian proses pra-pemrosesan teks yang terdiri dari:

1) Case folding - mengubah seluruh teks menjadi huruf kecil;
2) Cleansing – menghapus karakter khusus, URL, dan angka;
3) Tokenisasi - memecah teks menjadi satuan kata;
4) Stopword removal - menghapus kata tidak bermakna dalam bahasa Indonesia; dan
5) Stemming - mengubah kata ke bentuk dasarnya menggunakan algoritma Sastrawi.

Representasi fitur teks dilakukan menggunakan TF-IDF dengan maksimum 5.000 fitur dan kombinasi unigram-bigram.

4. Modeling

Data dibagi menjadi data latih (80%) dan data uji (20%) menggunakan stratified splitting. Dua algoritma diterapkan: (1) Naïve Bayes Multinomial yang cocok untuk data teks berbasis frekuensi kata, dan (2) Random Forest dengan 100 pohon keputusan (n_estimators=100)

sebagai metode ensemble learning berbasis bagging.

5. Evaluation

Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Validasi tambahan menggunakan 5-Fold Cross Validation untuk memastikan konsistensi dan generalisasi model.

HASIL DAN PEMBAHASAN

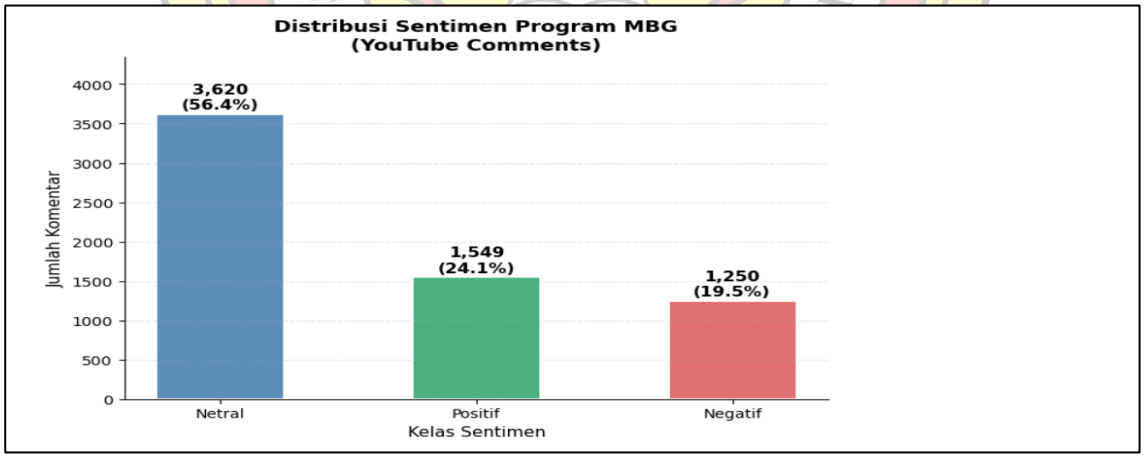
Distribusi Data Sentimen

Dataset penelitian terdiri dari 6.419 komentar YouTube yang telah melalui proses pelabelan sentimen. Distribusi kelas sentimen dapat dilihat pada Tabel 1 dan divisualisasikan pada

Tabel 1. Distribusi Data Sentimen Program MBG

Kelas Sentimen	Jumlah Data	Persentase (%)
Netral	3.620	56,40
Positif	1.549	24,13
Negatif	1.250	19,47
Total	6.419	100,00

Sumber Tabel Hasil Pengumpilan dan Pelabelan data (2026)

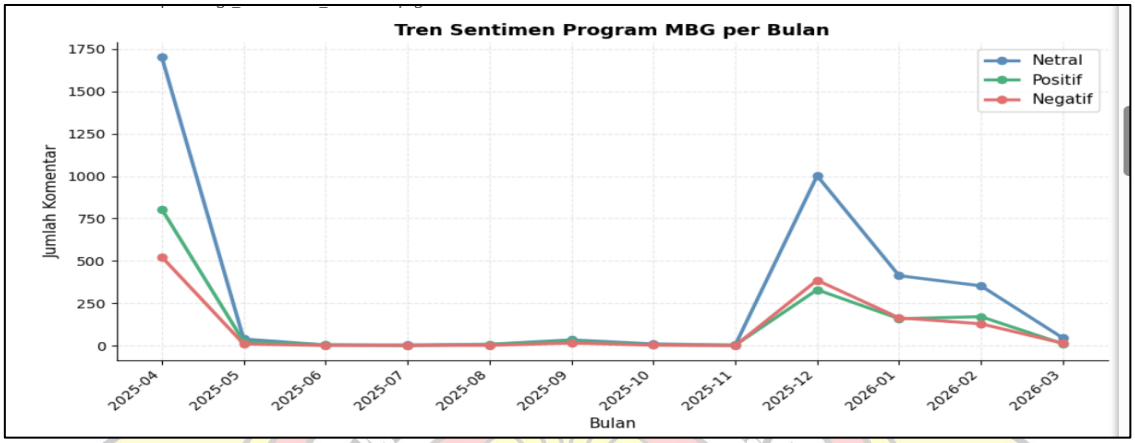


Gambar 1. Dsitribusi Sentime Program MBG pada Komentar Youtobe

Gambar 1 dan Tabel 1 menunjukkan bahwa sentimen Netral mendominasi dataset dengan proporsi 56,4%. Hal ini mengindikasikan bahwa sebagian besar masyarakat belum mengambil sikap tegas terhadap program MBG, kemungkinan disebabkan oleh program yang masih dalam tahap implementasi awal. Sentimen

Positif (24,1%) mencerminkan kelompok yang mengapresiasi inisiatif pemerintah, sedangkan Negatif (19,5%) mencerminkan kelompok yang memiliki kekhawatiran atau kritik.

Tren Sentimen per Bulan



Gambar 2. Tren Sentimen Program MBG per Bulan

Grafik tren pada Gambar 2 menunjukkan bahwa sentimen Netral secara konsisten mendominasi sepanjang periode pengamatan. Lonjakan komentar pada bulan tertentu berkorelasi dengan momen peluncuran dan pemberitaan besar terkait program MBG, yang memicu peningkatan diskusi publik di YouTube.

Hasil Pemodelan Naïve Bayes

Model Naïve Bayes dilatih menggunakan 5.052 sampel data latih dan dievaluasi pada 1.263 sampel data uji.

Tabel 2. Hasil Evaluasi Naïve Bayes per Kelas

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,86	0,23	0,36	250
Netral	0,63	0,95	0,76	704
Positif	0,79	0,33	0,46	309
Weighted Avg	0,71	0,66	0,61	1263

Model Naïve Bayes memperoleh akurasi keseluruhan sebesar 65,49% dengan 5-Fold Cross Validatio. Kelas Netral memiliki recall tertinggi (0,95), namun kelas Negatif dan Positif menunjukkan recall rendah (0,23 dan 0,33). Hal ini

mengindikasikan Naïve Bayes cenderung bias mengklasifikasikan data ke kelas Netral yang merupakan kelas mayoritas.

Hasil Pemodelan Random Forest

Model Random Forest dengan 100 pohon keputusan menghasilkan kinerja yang lebih baik. Hasil evaluasi per kelas disajikan pada Tabel 3

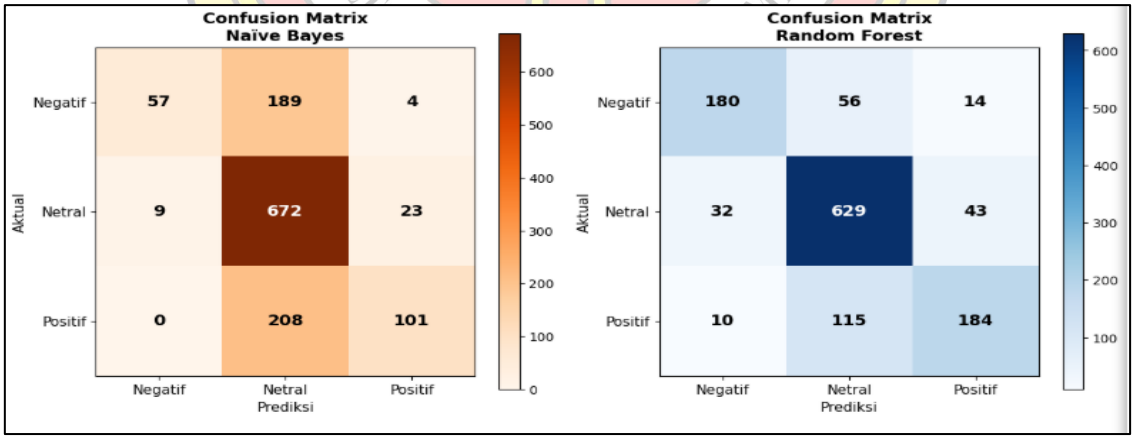
Tabel 3. Hasil Evaluasi Random Forest per Kelas

Kelas	Precision	Recall	F1-Score	Support
Negatif	0,81	0,72	0,76	250
Netral	0,79	0,89	0,84	704
Positif	0,76	0,60	0,67	309
Weighted Avg	0,79	0,79	0,78	1263

Random Forest menghasilkan akurasi 76,74% dengan 5-Fold CV. Dibandingkan Naïve Bayes, distribusi kinerja antar kelas jauh lebih seimbang, dengan recall Negatif 0,72 dan Positif 0,60 yang jauh lebih baik. Hal ini disebabkan kemampuan ensemble Random Forest menangkap pola kompleks melalui averaging 100 pohon keputusan.

Confusion Matrix

Gambar 3 menyajikan confusion matrix dari kedua model secara berdampingan untuk memudahkan perbandingan distribusi kesalahan klasifikasi.

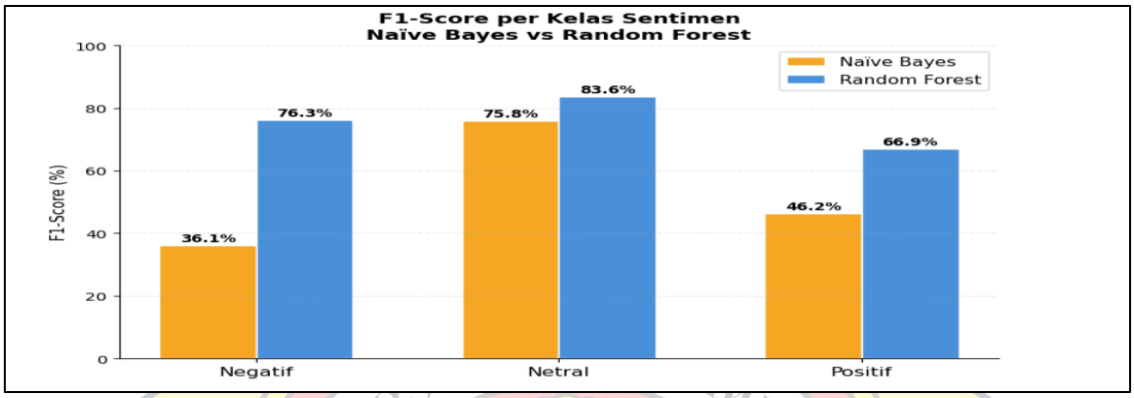


Gambar 3. Confusion Matrix Naïve Bayes dan Randon Forest

Confusion matrix pada Gambar 3 memperlihatkan kelemahan signifikan Naive Bayes: 190 sampel Negatif dan 208 sampel Positif salah diklasifikasikan sebagai

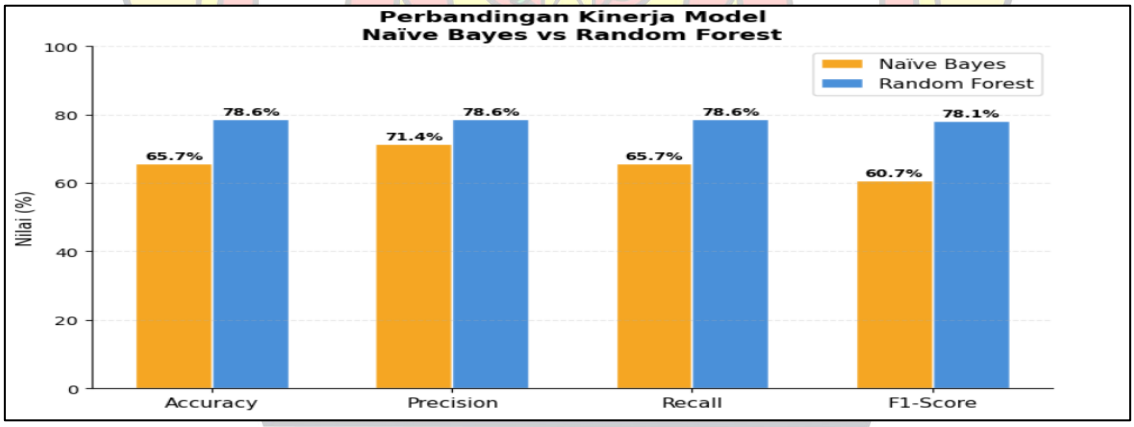
Netral. Random Forest menunjukkan distribusi kesalahan yang lebih merata, dengan hanya 66 sampel Negatif dan 120 sampel Positif yang salah diklasifikasikan sebagai Netral.

F1-Score per Kelas



Gambar 4. Perbandingan F1-Score per Kelas Sentime

Perbandingan Kinerja Model



Gambar 5. Perbandingan Matriks Evaluasi Naïve Bayes vs Random Forest

Tabel 4. Perbandingan Kinerja Model Naïve Bayes dan Random Forest

Algoritma	Akurasi	Presisi	Recall	F1-Score	CV (-5Fold)	Cv std
Naïve Bayes	65,72%	71,44%	65,72%	60,70%	65.49%	0,93%
Randon Forest	78,62%	78,55%	78,62%	78,09%	76,74%	1.40%

Tabel 4 dan Gambar 5 menunjukkan secara jelas keunggulan Random Forest pada seluruh metrik evaluasi. Perbedaan akurasi dan F1-Score merupakan gap yang signifikan. Keunggulan ini disebabkan kemampuan Random Forest sebagai metode ensemble dalam menangkap hubungan non-linear antar fitur. Hasil 5-Fold Cross Validation yang konsisten mengkonfirmasi kedua model tidak mengalami overfitting yang signifikan.

Temuan ini sejalan dengan penelitian Hidayat et al. (2024) yang juga menunjukkan superioritas Random Forest dalam klasifikasi sentimen teks berbahasa Indonesia. Dominasi sentimen Netral juga konsisten dengan temuan Pratama & Wulandari (2023) bahwa masyarakat cenderung bersikap netral terhadap program pemerintah yang baru berjalan.

Implikasi praktis penelitian ini adalah model Random Forest dapat dimanfaatkan untuk memantau sentimen publik terhadap program MBG secara real-time, membantu pemerintah mengidentifikasi isu-isu yang memicu sentimen negatif dan merancang strategi komunikasi yang lebih efektif.

SIMPULAN

Penelitian ini berhasil membangun model analisis sentimen terhadap program Makanan Bergizi Gratis (MBG) berbasis komentar YouTube dengan menerapkan kerangka kerja CRISP-DM. Berdasarkan hasil analisis terhadap 6.419 data komentar, ditemukan bahwa sentimen Netral mendominasi dengan proporsi 56,40%, diikuti oleh sentimen Positif (24,13%) dan Negatif (19,47%). Tingginya angka sentimen netral mencerminkan persepsi masyarakat yang masih berada pada tahap observasi atau menunggu bukti nyata efektivitas program yang baru diimplementasikan.

Dalam aspek performa model, algoritma Random Forest secara konsisten menunjukkan keunggulan dibandingkan Naïve Bayes. Random Forest mencapai tingkat akurasi sebesar 78,62% dan F1-score 78,09%, sedangkan Naïve Bayes hanya memperoleh akurasi 65,72% dan F1-score 60,70%. Superioritas Random Forest disebabkan oleh kemampuannya sebagai metode ensemble dalam menangani

hubungan non-linear antar fitur dan mengurangi bias terhadap kelas mayoritas dibandingkan dengan Naïve Bayes yang sangat bergantung pada frekuensi kemunculan kata.

Implikasi praktis dari penelitian ini adalah pemerintah dapat memanfaatkan model Random Forest untuk memantau dinamika opini publik secara real-time guna merancang strategi komunikasi yang lebih responsif. Untuk pengembangan selanjutnya, disarankan untuk mengeksplorasi teknik penyeimbangan data seperti SMOTE guna menangani ketidakseimbangan kelas sentimen, serta mengintegrasikan model pembelajaran mendalam berbasis IndoBERT untuk meningkatkan pemahaman konteks pada bahasa informal media sosial yang lebih kompleks.

DAFTAR PUSTAKA

[1] A. Kiftiyah, F. A. Palestina, F. U. Abshar, and K. Rofiah, "Program makan bergizi gratis (MBG) dalam perspektif keadilan sosial dan dinamika sosial–politik," *Pancasila J. Keindonesiaan*, vol. 5, no. 1, pp. 101–112, 2025.

[2] W. Tanuwijaya, C. E. Setiawan, H. Irsyad, and A. Rahman, "Implementasi TF-IDF dan Cosine Similarity untuk Penyaringan Dokumen Berita Program Makan Siang Gratis Pemerintah Indonesia," *Device J. Inf. Syst. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 322–334, 2025.

[3] M. E. D. Vanti, V. Octaviani, and Maryaningsih, "Analisis Framing Pemberitaan Program Makan Gratis Prabowo Subianto Di Media Online," *Prof. J. Komun. dan Adm. Publik*, 2024.

[4] D. Kurniawan and M. Yasir, "Optimization Sentiment Analysis Using CRISP-DM and Naïve Bayes Methods Implemented on Social Media," *Cybersp. J. Pendidik. Teknol. Inf.*, vol. 6, no. 2, pp. 74–85, 2022.

[5] L. S. Syabilla, M. I. Natasyah, Fathoni, and J. A. Siahaan, "Klasifikasi Opini Tidak Informatif Pada Program Makan Bergizi Gratis (MBG) Menggunakan Random Forest," *Indones. J. Comput. Sci.*, vol. 5, no. 1, pp. 56–64, 2026.

[6] Hajaroh, T. Suprapti, and R. Narasati, "Implementasi Algoritma Naive Bayes Untuk Analisis Sentimen Ulasan Produk Makanan Dan Minuman Di Tokopedia," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 111–118, 2024.

[7] R. Hidayat and D. J. Ratnaningsih, "Analisis Sentimen Program Makanan

Bergizi Gratis Menggunakan Algoritma Random Forest dan Naive Bayes," *J. Comput. Informatics Res.*, vol. 5, no. 1, pp. 395–400, 2025, doi: 10.47065/comforch.v5i1.2355.

[8] W. A. Rahmat, S. M. Ladjamuddin, and D. T. Awaludin, "Perbandingan Algoritma Decision Tree, Random Forest dan Naive Bayes pada Prediksi Penilaian Kepuasan Penumpang Maskapai Pesawat Menggunakan Dataset Kaggle," *J. Rekayasa Inf.*, vol. 12, no. 2, pp. 150–159, 2023.

[9] M. R. U. Pulungan, D. E. Ratnawati, and B. Rahayudi, "Analisis Sentimen Ulasan Aplikasi PeduliLindungi dengan Metode Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4378–4385, 2022.

[10] Kevin, M. Enjeli, and A. Wijaya, "Analisis Sentimen Pengunaaan Aplikasi Kinemaster Menggunakan Metode Naive Bayes," *J. Ilm. Comput. Sci.*, vol. 2, no. 2, pp. 89–98, 2024, doi: 10.58602/jics.v2i2.24.

[11] Saikin, M. T. A. Zaen, S. Fadli, and H. Fahmi, "Penerapan Algoritma BERT dalam Analisis Sentimen Opini Publik terhadap Destinasi Wisata dengan Metode CRISP-DM," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 4, pp. 5382–5392, 2025, doi: 10.31004/riggs.v4i4.4373.