Router : Jurnal Teknik Informatika dan Terapan
Volume. 3, Nomor. 4 Desember 2025
e-ISSN: 3032-3312, p-ISSN: 3026-3611, Hal 01-10
DOI: https://doi.org/10.62951/router.v3i4.704
Tersedia : https://journal.aptii.or.id/index.php/Router

Analisis Sentimen Kebijakan Pemberlakuan Cukai pada Minuman Berpemanis dalam Kemasan Menggunakan Metode Multinomial Naive Bayes

Muhammad Firdaus[1], Ulya Anisatur Rosyidah[2*], Luluk Handayani[3]
[1-3]Program Studi Teknik Infomatika, Universitas Muhammadiyah Jember, Indonesia
[*]*Penulis korespondensi: ulyaanisatur@unmuhjember.ac.id[2]*

Abstract. *Sugar consumption in Indonesia remains high, with diabetes affecting 20.4 million people. This condition has prompted the government to introduce an excise policy on Minuman Berpemanis Dalam Kemasan (MBDK) to reduce sugar intake. Social media, particularly the X platform, serves as a medium for the public to express their opinions regarding this policy. This study aims to analyze public sentiment toward the MBDK excise policy using a lexicon-based approach for data labeling and the Multinomial Naive Bayes algorithm with unigram and bigram feature extraction. The initial results show that the highest performance was achieved using 5-Fold Cross Validation, with an average accuracy of 83%, precision of 84%, recall of 75%, and an F1-Score of 77%. After applying data balancing using Stratified Cross Validation combined with Borderline-SMOTE and limiting the features to the 700 most frequent terms, the model's performance improved. The best results were obtained with 10-Fold Cross Validation, achieving 86% accuracy, 84% precision, 83% recall, and an F1-Score of 83%. These findings indicate that the Multinomial Naive Bayes model can effectively classify public sentiment regarding the MBDK excise policy after the data balancing process.*

Keywords: *Borderline-SMOTE; MBDK; Multinomial Naive Bayes; N-gram; Stratified Cross Validation*

Abstrak. Konsumsi gula di Indonesia tergolong tinggi dengan penderita diabetes mencapai 20,4 juta jiwa. Kondisi ini menjadi dasar bagi pemerintah untuk memberlakukan kebijakan cukai pada minuman berpemanis dalam kemasan (MBDK) sebagai langkah pengendalian konsumsi. Platform media sosial, terutama X, berperan sebagai media untuk mengemukakan opini terhadap kebijakan tersebut. Tujuan dari penelitian ini untuk menganalisis sentimen masyarakat terhadap kebijakan cukai MBDK menggunakan pendekatan berbasis leksikon untuk pelabelan data dan *Multinomial Naive Bayes* dikombinasikan dengan ekstraksi fitur *unigram* dan *bigram*. Hasil awal menunjukkan bahwa akurasi tertinggi didapatkan pada pengujian 5-*Fold Cross Validation* dengan rata-rata akurasi sebesar 83%, presisi 84%, *recall* 75%, dan *F1-Score* berada pada 77%. Setelah dilakukan *balancing* menggunakan *Stratified Cross Validation* yang dikombinasikan dengan *Borderline-SMOTE* serta pembatasan pada 700 fitur dengan frekuensi tertinggi, didapatkan peningkatan performa model. Hasil tertinggi didapatkan pada pengujian 10-*Fold Cross Validation* dengan menghasilkan akurasi mencapai 86%, presisi 84%, recall 83%, serta *F1-Score* 83%. Hasil tersebut menunjukan bahwa model *Multinomial Naive Bayes* mampu mengklasifikasikan sentimen masyarakat terhadap kebijakan cukai MBDK dengan kinerja yang baik setelah dilakukan proses penyeimbangan data.

Kata kunci: Bayes naif multinomial; *Borderline-SMOTE;* MBDK; N-gram; Validasi Silang Bertingkat

1. LATAR BELAKANG

Indonesia menempati peringkat ke-6 dunia dalam konsumsi gula, dengan total konsumsi mencapai 7,7 juta metrik ton pada periode 2024/2025 (USDA Foreign Agricultural & Service, 2025). Tingginya tingkat konsumsi tersebut tidak terlepas dari pola hidup masyarakat Indonesia yang cenderung menyukai makanan dan minuman manis, terutama di kalangan anak muda (Emiliana & Setiarini, 2024). Konsumsi gula berlebih berpotensi meningkatkan risiko berbagai penyakit tidak menular seperti diabetes, obesitas, penyakit jantung, dan kanker (Sinaga et al., 2024). International Diabetes Federation (2024) juga mencatat bahwa jumlah penderita diabetes di Indonesia mencapai 20,4 juta jiwa dan terus meningkat setiap tahun.

Naskah Masuk: 17 Oktober 2025; Revisi: 21 November 2025; Diterima: 13 Desember 2025;
Terbit: 16 Desember 2025

Sebagai respons atas tingginya konsumsi gula, pemerintah menerapkan kebijakan cukai pada minuman berpemanis dalam kemasan (MBDK) sebagai langkah preventif untuk menekan konsumsi gula di masyarakat (Direktorat Jenderal Perbendaharaan, 2025). Kebijakan tersebut memunculkan beragam respons di kalangan masyarakat. Sebagian pihak mendukung karena dinilai sebagai langkah preventif terhadap penyakit, sementara pihak lainnya menolak karena dianggap menurunkan daya beli konsumen menurun serta memengaruhi pertumbuhan industri minuman. Perbedaan pandangan tersebut menunjukkan perlunya pemetaan opini publik melalui analisis sentimen terutama dari media sosial yang menjadi ruang ekspresi masyarakat.

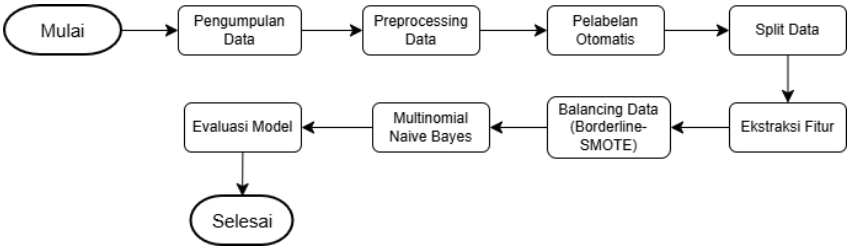
Analisis sentimen merupakan metode pemrosesan dalam bahasa alami yang bertujuan menentukan arah atau kecenderungan opini pada suatu teks, apakah menunjukkan sikap positif, negatif, maupun netral (Eni Tri & Ari, 2021). Salah satu metode yang banyak digunakan adalah Multinomial Naive Bayes, yaitu algoritma klasifikasi berbasis frekuensi kata yang cepat dan stabil untuk data teks skala besar, meski memiliki keterbatasan dalam menangkap konteks kata (Ningsih et al., 2024). Untuk mengatasi keterbatasan tersebut, fitur *N-Gram* digunakan agar model dapat mengenali pola urutan kata yang lebih bermakna.

Berdasarkan kondisi tersebut, penelitian ini dilakukan untuk menganalisis sentimen masarakat terhadap kebijakan cukai minuman berpemanis dalam kemasan (MBDK) dengan menggunakan data dari platform X. Penelitian ini diharapkan mampu menghasilkan pemetaan persepsi masyarakat sebagai bahan pertimbangan bagi pemerintah. Dalam pelaksanaannya, penelitian ini mengombinasikan pendekatan Lexicon-Based dan Multinomial Naive Bayes dengan fitur N-gram untuk menghasilkan analisis sentimen yang lebih akurat dan kontekstual terhadap isu cukai MBDK.

2. METODE PENELITIAN

Tahapan Penelitian

Proses klasifikasi sentimen terkait kebijakan kebijakan cukai minuman berpemanis dalam kemasan menggunakan *Multinomial Naive Bayes* dilakukan melalui sejumlah tahap. Penelitian dimulai dari tahap pengumpulan data, yang selanjutnya dilakukan *preprocessing* untuk membersihkan dan menyiapkan data teks. Selanjutnya data dilakukan pelabelan secara otomatis, lalu dibagi menjadi data latih dan uji sebelum melalui proses ekstraksi fitur. Karena jumlah data tidak seimbang, dilakukan proses *balancing data* menggunakan *Borderline-SMOTE,* kemudian model *Multinomial Naive Bayes* dibangun dan dilatih. Terakhir, dilakukan evaluasi model untuk menilai kinerjanya. Rangkaian tahapan penelitian dalam studi ini disajikan pada Gambar 1.



Gambar 1. Rangkaian Tahapan Penelitian.

Dataset

Dataset pada penelitian ini terdiri atas 500 tweet yang diperoleh dari platform X menggunakan kata kunci terkait kebijakan cukai minuman berpemanis dalam kemasan (MBDK), seperti "cukai MBDK", "cukai minuman manis", dan "minuman berpemanis". Pengumpulan dataset dilakukan mengunakan *tools snscrape* dengan rentang waktu 1 Februari 2024 hingga 30 April 2025, kemudian seluruh komentar tersebut disimpan dalam format CSV untuk keperluan analisis lebih lanjut.

Preprocessing Data

Pada tahap preprocessing, seluruh data dibersihkan agar siap diolah secara optimal, sekaligus untuk meningkatkan akurasi dan kinerja model (Harsemadi et al., 2023). Langkah pertama berupa proses pembersihan teks, di mana berbagai komponen yang tidak relevan seperti tautan, angka, simbol, dan karakter lain yang tidak diperlukan dalam analisis. Setelah itu, dilakukan penyeragaman huruf bentuk huruf melalui *case folding* sehingga seluruh teks dikonversi menjadi huruf kecil. Kemudia teks dinormalisasi agar kata-kata yang tidak baku atau slang disesuaikan dengan padanan yang sesuai KBBI.

Teks yang sudah bersih kemudian diproses melalui tokenisasi untuk memisahkan kalimat menjadi deretan kata. Berikutnya dilakukan stopword removal menggunakan Sastrawi untuk menghilangkan kata-kata yang tidak memiliki nilai informatif terhadap analisis. Tahap terkahir adalah stemming menggunakan library Sastrawi, untuk mengembalikan setiap kata ke bentuk dasarnya, sehingga model dapat membaca makna kata dengan lebih konsisten (Putu et al., 2025).

Pelabelan Otomatis

Dalam penelitian ini, proses pelabelan dilakukan secara otomatis memanfaatkan pendekatan leksikon. Setiap komentar dianalisis berdasarkan skor sentimen yang terdapat dalam kamus sentimen, yaitu Senti_Strenght yang telah disesuaikan dengan kebutuhan penelitian ini. Sistem kemudian menghitung total skor sentimen dari setiap komentar dan menentukan label akhir sebagai positif, negatif, atau netral (Khaira et al., 2020). Dalam

penelitian ini memfokuskan anlisis pada dua kategori sentimen, yaitu positif dan negatif. Berdasarkan proses pelabelan tersebut, diperoleh 346 komentar berlabel positif dan 154 komentar berlabel negatif. Data yang sudah terlabeli tersebut selanjutnya digunakan sebagai dasar dalam proses pelatihan dan pengujian model klasifikasi sentimen. Hasil dari proses pelabelan ditampilkan pada tabel 1.

Tabel 1. Hasil Pelabelan Otomatis.

No	Hasil_preprocessing	label	Skor Total	Rincian_kata
1	bakal naik harga air mineral	Negatif	-1	bakal naik [-1] harga air mineral
2	semua semua saja dipajakin pusing [-1] kepala	Negatif	-1	semua semua saja dipajakin pusing [-1] kepala
3	pajak tinggi	Negatif	-1	pajak tinggi [-1]

Ekstraksi Fitur

Pada tahap Ekstarksi fitur, teks yang telah dibersihkan kemudian diproses menggunakan teknik *N-gram* (Unigram dan Bigram) untuk menangkap konteks berdasarkan satu kata maupun dua kata yang muncul berurutan, sehinggah hubungan antar kata lebih tertangkap. Selanjutnya, hasil *N-Gram* dikonversi menjadi representasi numerik menggunakan *CountVectorizer* dari *Scikit-learn* agar dapat diolah oleh *Multinomial Naive Bayes.* Vektor numerik tersebut kemudian dimanfaatkan sebagai input dalam proses pelatihan dan perhitungan model *Multinomial Naive Bayes* (Vincent et al., 2024). Hasil pengolahannya ditampilkan pada Tabel 2 dan Tabel 3.

Tabel 2. Distribusil Hasil *Unigram.*

No	Unigram	Frekuensi
1	minum	82
2	tidak	72
3	tuju	72
...	...	...
764	utama	1
765	untuk	1
766	umkn	1

Tabel 3. Distribusi Hasil Bigram.

No	Unigram	Frekuensi
1	air putih	28
2	minum manis	27
3	minum air	16
...	...	...
2113	indonesia kena	1
2114	indonesia darurat	1
2115	iya benar	1

Split Data

Pada tahap ini, *K-Fold Cross Validation* diterapkan untuk memastikan hasil evaluasi model lebih stabil dan tidak bergantung pada satu skema pembagian dataset saja (Ridwansyah, 2022). Namun, dataset yang digunakan mengalami ketidakseimbangan kelas. Hal ini mengakibatkan model lebih condong memprediksi kelas mayoritas sehingga akurasi model dalam mengenali kelas minoritas menjadi menurun (Galih & Santi, 2025). Sebagai upaya menangani permasalah tersebut, diterapkan teknik *Borderline-SMOTE,* yaitu metode yang menghasilkan data sintetik dari kelas minoritas yang berada di sekitar daerah batas dengan kelas mayoritas, sehingga distribusi kelas yang dihasilkan menjadi lebih relevan dan representatif (Ridwan et al., 2024).

Multinomial Naive Bayes

Multinomial Naive Bayes merupakan salah satu algoritma yang berlandaskan pada teori bayes, yang bekerja dengan mengasumsikan bahwa setiap kata dalam dokumen bersifat independen serta melakukan proses klasifikasi berdasarkan perhitungan probabilitas kemunculan kata dalam suatu kategori (Hadaina & Budiyanto, 2022). Pendekatan ini efektif untuk data berteks besar karena sederhana, cepat, dan memberikan performa yang stabil (Alfandi Safira & Hasan, 2023).

$$P(c|d) = P(c) \times \prod_{i=1}^{n} P(w_i|c)^{count(w_i,d)} \quad (1)$$

Evaluasi

Dalam menilai efektivitas model dalam melakukan klasifikasi sentimen, evaluasi dilakukan menggunakan *Confusion Matrix*. Teknik evaluasi ini membandingkan label prediksi yang dihasilkan model dengan label sebenarnya, sehingga dapat mengidentifikasi kesalahan klasifikasi (Radhitya et al., n.d.). Confusion Matrix mencakup empat variabel utama: True Positive (TP),yaitu ketika model berhasil memprediksi kelas positif dengan benar; True Negative (TN), yaitu ketika model berhasil memprediksi kelas negatif dengan benar; False Positive (FP), yaitu ketika model salah memprediksi data sebagai kelas positif, padahal sebenarnya termasuk kelas negatif; dan False Negative (FN), yaitu ketika model salah memprediksi data sebagai kelas negatif, padahal sebenarnya termasuk kelas positif (Fikri et al., 2020). Keempat variabel tersebut berperan sebagai acuan dalam perhitungan metrik evaluasi, yaitu presisi, akurasi, recall, dan F1-Score, sehingga kinerja model dapat dinilai secara lebih komprehensif.

Akurasi menunjukkan sejauh mana model melakukan klasifikasi dengan tepat, yang dihitung dengan membandingkan jumlah prediksi yang benar terhadap total data uji.

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \quad (2)$$

Presisi menggambarkan seberapa tepat model dalam mengklasifikasikan data sebagai kelas positif, yang dihitung dari perbandingan antara prediksi positif yang benar dengan seluruh prediksi positif yang dihasilkan.

$$Presisi = \frac{TP}{(TP + FP)} \quad (3)$$

Recall menunjukkan tingkat kemampuan model dalam mengidentifikasi data positid dari seluh data positif yang terdapat dalam dataset.

$$Recall = \frac{TP}{(TP + FN)} \quad (4)$$

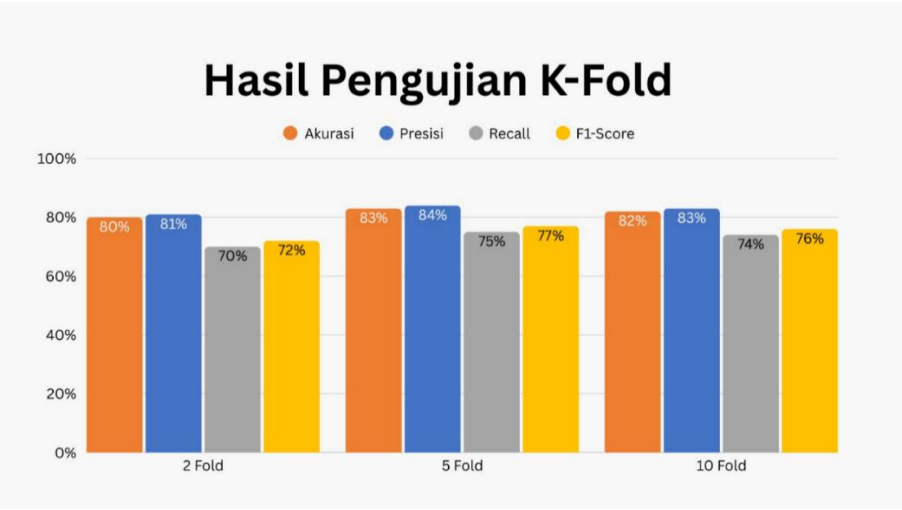
F1-Score digunakan untuk menilai keseimbangan antara presisi dan *recall* dengan menggunakan rata-rata harmonik.

$$F1 - Score = \frac{2 \times Presisi \times Recall}{Presisi + Recall} \quad (5)$$

3. HASIL DAN PEMBAHASAN

Hasil Klasifikasi

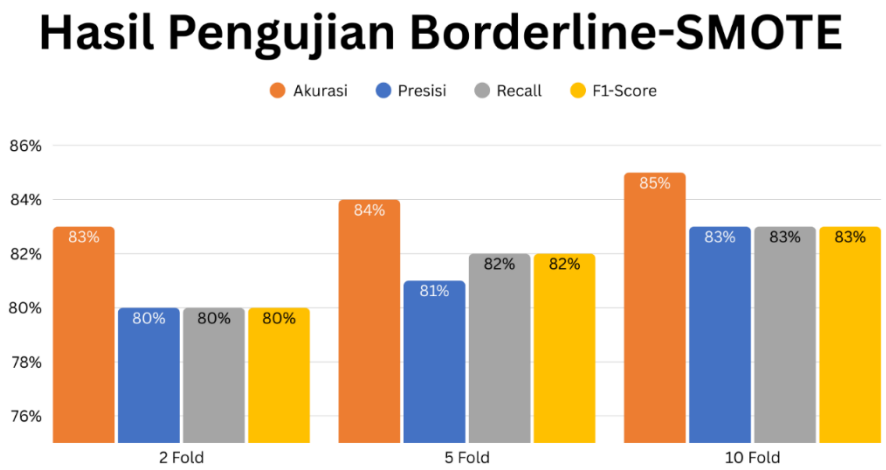
Berikut hasil pengujian model *Multinomial Naive Bayes* menggunakan metode *K-Fold Cross* Validation dengan variasi 4, 5, dan 10 fold. Pegujian ini dilakukan untuk membandingkan performa model pada masing-masing skema pembagian data, dan hasilnya dapat dilihat pada Gambar 2.



Gambar 2. Rata-Rata Hasil Evaluasi K-Fold.

Berdasarkan hasil pengujian dengan skema validasi *K-Fold* 4, 5, dan 10 didapatkan bahwa pada pengujian 5-Fold memberikan performa terbaik dengan akurasi sebesar 83%, presisi 84%, *recall* 75%, dan *F1-Score* 77%. Adapun pada skema 2-fold, model hanya mencapai akurasi 80%, *precision* 81%, *recall* 70%, dan *F1-Score* 72%. Sementara itu, pada skema 10-fold, akurasi berada pada angka 82%, dengan presisi 83%, *recall* 74%, dan *F1-Score* 76%. Meskipun akurasi pada beberapa skema tergolong cukup baik, nilai *recall* dan *F1-Score* yang berada pada rentang 70% hingga 77% mengindikasikan bahwa model masih belum mampu mengidentifikasi kelas monoritas secara maksimal. Kondisi ini terjadi akibat ketidakseimbangan distribusi data, di mana model cenderung lebih sering mengklasifikasikan data ke dalam kelas mayoritas.

Untuk mengatasi permasalahan tersebut, dilakukan proses balancing data menggunakan *Borderline*-SMOTE, dengan tujuan memperkuat representasi kelas minoritas dengan membuat sampel sintetis pada area yang berdekatan dengan batas keputusan (decision boundary). Setelah proses *balancing* diterapkan, model dievaluasi kembali dan hasil pengujian ditampilkan pada Gambar 3.



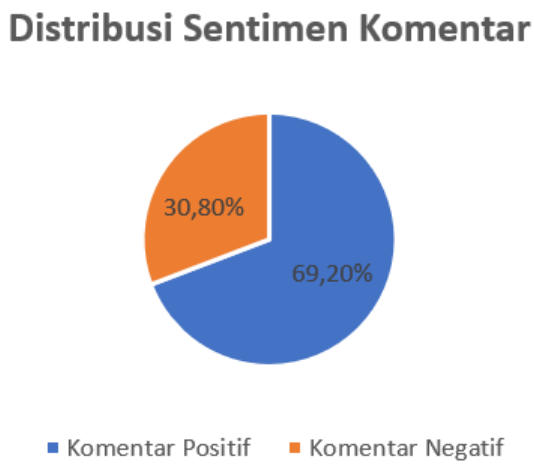
Gambar 3. Rata-Rata hasil *Borderline-SMOTE.*

Setelah dilakukan *balancing dataset* menggunakan *Borderline-SMOTE,* terjadi peningkatan peforma model pada seluruh skema fold 4, 5, dan 10. Pada skema 4-fold, model mencapai akurasi 83% dengan presisi, *recall,* dan *F1-Score* masing-maisng berda pada angka 80%. Performa meningkat pada skema 5-fold dengan akurasi 84%, presisi 81%, serta *recall* dan *F1-Score* sebesar 82%. Adapun performa tertinggi dicapai pada skema 10-fold, Dimana akurasi mencapai 85% dan seluruh metrik lainnya presisi, *recall,* dan *F1-Score* konsisten berada di angka 83%. Peningkatan metrik yang paling menonjol terlihat pada nilai *recall* dan *F1-Score* yang mencapai 83% pada seluruh skema pengujian. Peningkatan performa ini terjadi

karena hasil data minoritas yang dihasilkan *Borderline-SMOTE* memberikan representasi tambahan pada area batas kelas, sehingga membantu model *Multinomial Naïve Bayes* belajar pola klasifikasi secara lebih efektif dan seimbang.

Distribusi Sentimen Komentar

Bagian ini menyajikan distribusi sentiment dari komentar pengguna twitter twitter terkait kebijakan cukai MBDK, dapat dilihat pada gambar 4.



Gambar 4. Persentase Sentimen Komentar.

Berdasarkan hasil pelabelan menggunakan, distribusi sentiment pada komentar pengguna Twitter menunjukkan bahwa mayoritas komentar bersentimen positif yaitu sebanyak 346 komentar atau 69,20%. Sementara itu, komentar bersentimen negatif berjumlah 154 komentar atau 30,80%. Hasil ini memberikan gambaran mengenai kecenderungan respons publik terhadap kebijakan cukai MBDK.

4. KESIMPULAN DAN SARAN

Berdasarkan keseluruhan hasil analisis yang telah dipaparkan, dapat disimpulkan bahwa mayoritas pengguna media sosial X memberikan tanggapan positif terhadap kebijakan cukai minuman berpemanis dalam kemasan (MBDK). Hal ini terlihat dari proporsi sentimen positif mencapai 346 komentar atau 69,2% dan sentimen negatif berjumlah 154 komentar atau 30,8%. Dari sisi performa model, algoritma *Multinomial Naive Bayes* menunjukkan kinerja terbaik setelah dilakukan penyeimbangan data menggunakan *Borderline-SMOTE*, dengan rata-rata akurasi 86%, presisi 84%, recall 83%, dan F1-score 83%. Hasil tersebut menunjukkan bahwa penerapan *Borderline-SMOTE* mampu meningkatkan kemampuan model dalam menangani distribusi kelas yang tidak seimbang, sehingga hasil klasifikasinya menjadi lebih presisi serta representatif terhadap opini publik.

Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar dan rentang waktu pengambilan data yang lebih panjang agar model dapat melakukan generalisasi secara lebih optimal.

DAFTAR REFERENSI

Alfandi Safira, & Hasan, F. N. (2023). Analisis sentimen masyarakat terhadap paylater menggunakan metode Naive Bayes classifier. *ZONAsi: Jurnal Sistem Informasi, 5*(1), 59–70. https://doi.org/10.31849/zn.v5i1.12856

Emiliana, N., & Setiarini, A. (2024). Hubungan konsumsi minuman berpemanis dengan kejadian obesitas pada anak dan remaja: A systematic literature review. *Holistik Jurnal Kesehatan, 18*(4), 509–517.

Eni Tri, H., & Ari, S. (2021). Analisis sentimen respon masyarakat terhadap kabar harian COVID-19 pada Twitter Kementerian Kesehatan. *Jurnal Teknologi dan Sistem Informasi (JTSI), 2*(3), 32–37. http://repository.teknokrat.ac.id/3224/

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Comparison of naïve Bayes and support vector machine methods in Twitter sentiment analysis. *Smatika Jurnal, 10*(2), 71–76.

Galih, I. K., & Santi, I. G. (2025). Penerapan support vector machine untuk klasifikasi tingkat risiko kebakaran hutan. *Jurnal Ilmiah*, *3*, 763–774.

Hadaina, F., & Budiyanto, U. (2022). Implementasi metode multinomial naïve Bayes untuk sentiment analysis terhadap data ulasan produk Colearn pada Google Play Store. Dalam *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)* (pp. 660–666). https://senafti.budiluhur.ac.id/index.php

Harsemadi, I. G., Dharmendra, I. K., & Wijaya, I. M. P. P. (2023). Klasifikasi emosi pada tweet berbahasa Indonesia. *Prosiding Seminar*, 86.

Khaira, U., Johanda, R., Utomo, P. E., Suratno, T., & Info, A. (2020). Sentiment analysis of cyberbullying on Twitter using SentiStrength. *Jurnal Ilmiah*, *3*(1), 21–27.

Ningsih, W., Alfianda, B., Rahmaddeni, R., & Wulandari, D. (2024). Perbandingan algoritma SVM dan naïve Bayes dalam analisis sentimen Twitter pada penggunaan mobil listrik di Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(2), 556–562. https://doi.org/10.57152/malcom.v4i2.1253

Noviyanti, N. P., Ngurah, I. G., & Cahyadi, A. (2025). Penerapan multinomial naïve Bayes dan chi-square pada analisis sentimen makan bergizi gratis. *Jurnal Ilmiah*, *3*, 785–794.

Radhitya, M. L. (n.d.). *Klasifikasi genre musik menggunakan text mining* (pp. 1–9).

Ridwan, R., Hermaliani, E. H., & Ernawati, M. (2024). Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian. *Computer Science (CO-SCIENCE), 4*(1), 80–88. https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990

Ridwansyah, T. (2022). Implementasi text mining terhadap analisis sentimen masyarakat dunia di Twitter terhadap Kota Medan menggunakan k-fold cross validation dan naïve Bayes classifier. *KLIK: Kajian Ilmiah Informatika dan Komputer, 2*(5), 178–185. https://doi.org/10.30865/klik.v2i5.362

Sinaga, J., Sinambela, J. L., Purba, B. C., & Pelawi, S. (2024). Gula dan kesehatan: Kajian terhadap dampak kesehatan akibat konsumsi gula berlebih. *Mutiara: Jurnal Ilmiah Multidisiplin Indonesia, 2*(1), 54–68. https://doi.org/10.61404/jimi.v2i1.84

Vincent, R., Maulana, I., & Komarudin, O. (2024). Perbandingan klasifikasi naïve Bayes dan support vector machine dalam analisis sentimen dengan multiclass di Twitter. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(4), 2496–2505. https://doi.org/10.36040/jati.v7i4.7152