IMPLEMENTASI SUPPORT VECTOR MACHINE DALAM ANALISIS SENTIMEN NETIZEN TERHADAP PROGRAM MAKAN BERGIZI GRATIS

Muthia Rifky Ananda, Lidya Wati, Desi Wahana
Rekayasa Perangkat Lunak, Politeknik Negeri Bengkalis
Jalan Sungai Alam, Kec. Bengkalis, Kabupaten Bengkalis, Riau, Indonesia
muthiarifkyananda@gmail.com

ABSTRAK

Program Makan Bergizi Gratis (MBG) sebagai kebijakan pemerintah Prabowo-Gibran memicu respons beragam di media sosial. Penelitian ini bertujuan mengimplementasikan algoritma Support Vector Machine (SVM) untuk mengklasifikasi sentimen netizen terhadap program MBG menjadi kategori positif dan negatif. Metode penelitian menggunakan pendekatan Knowledge Discovery in Database (KDD) dengan pembobotan TF-IDF pada 7.354 komentar yang diambil dari platform X dan TikTok. Hasil pengujian menunjukkan akurasi 82% pada platform X dan 88% pada TikTok, yang dipengaruhi ketidakseimbangan distribusi data. Pengujian pada data gabungan dengan distribusi lebih seimbang menghasilkan akurasi 83% dengan kinerja klasifikasi lebih stabil. Analisis word cloud mengungkap dominasi sentimen negatif pada X (70,53%) yang berfokus pada kekhawatiran anggaran dan transparansi, sedangkan TikTok didominasi sentimen positif (85,65%) yang menekankan empati dan apresiasi terhadap manfaat program. Temuan ini membuktikan SVM efektif dalam analisis sentimen dengan performa optimal pada data berdistribusi seimbang.

Kata kunci : *Analisis Sentimen, Support Vector Machine, Makan Bergizi Gratis, Media Sosial, TF-IDF, KDD*

1. PENDAHULUAN

Platform media sosial telah berkembang menjadi ruang diskusi digital yang dinamis dan efektif dalam kehidupan masyarakat modern, berperan signifikan dalam membentuk opini publik terhadap kebijakan pemerintah [1]. X (sebelumnya Twitter) sebagai media untuk menyampaikan gagasan dan informasi melalui cuitan singkat, serta TikTok sebagai aplikasi berbagi video pendek yang populer di kalangan generasi muda, menjadi sumber data yang relevan untuk memahami pandangan netizen terhadap isu-isu nasional. Salah satu kebijakan utama pemerintahan Prabowo-Gibran adalah Program Makan Bergizi Gratis (MBG) yang diluncurkan pada 6 Januari 2025 sebagai upaya mengatasi malnutrisi dan meningkatkan kualitas sumber daya manusia (SDM) Indonesia, dengan sasaran utama anak-anak, pelajar, dan ibu hamil [2]. Berdasarkan data Kementerian Kesehatan dan Kementerian Koordinator Pembangunan Manusia dan Kebudayaan, sekitar 41% siswa di Indonesia mengalami kelaparan yang berdampak langsung terhadap menurunnya kualitas pendidikan [2]. Namun, respons masyarakat terhadap program MBG ini beragam, dengan sebagian mendukung tujuan pemerintah untuk mengurangi stunting, sementara yang lain mempertanyakan efektivitas program, kesiapan logistik, potensi pemborosan anggaran, dan keberlanjutan program [3].

Analisis sentimen merupakan teknik dalam text mining yang bertujuan mengidentifikasi dan mengevaluasi emosi dalam teks, untuk menentukan apakah suatu teks memiliki sifat positif, negatif, atau netral [4]. Teknik ini dapat diterapkan untuk menganalisis unggahan di media sosial guna melihat kecenderungan opini publik terhadap suatu kebijakan melalui algoritma klasifikasi yang telah dilatih menggunakan data teks berlabel. Dalam konteks ini, algoritma *Support Vector Machine* (SVM) dipilih karena memiliki keunggulan dalam proses klasifikasi, yaitu mampu mencari *hyperplane* optimal yang memaksimalkan margin antar kelas sehingga menghasilkan pemisahan data yang lebih jelas. Selain itu, SVM efektif dalam menangani data berdimensi tinggi seperti teks hasil representasi TF-IDF, karena hanya bergantung pada *support vector* (data penting di batas kelas) sehingga lebih tahan terhadap *noise* dan *overfitting* dibandingkan metode lain seperti *Naïve Bayes* yang berbasis probabilistik dan *K-Nearest Neighbor* (KNN) yang bergantung pada kedekatan jarak seluruh data. Keunggulan tersebut terbukti dalam berbagai penelitian terdahulu, seperti penelitian Handayani dan Zufria yang berhasil mengklasifikasikan pendapat netizen tentang calon presiden 2024 di Twitter dengan akurasi 86% [5], serta penelitian Pamungkas dan Kharisudin yang menunjukkan SVM memiliki akurasi rata-rata 90,01% dalam menganalisis tanggapan masyarakat terhadap pandemi COVID-19 dibandingkan algoritma Naïve Bayes (79,20%) dan KNN (62,10%) [6].

Penelitian ini diharapkan dapat memberikan pemahaman mengenai penerapan algoritma SVM dalam mengklasifikasi opini publik terhadap kebijakan pemerintah, sekaligus menjadi dasar pengambilan keputusan bagi pihak terkait dalam perbaikan program dan strategi komunikasi yang lebih efektif.

2. TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen adalah teknik text mining untuk mengidentifikasi emosi dalam teks guna mengklasifikasikannya sebagai positif, negatif, atau netral [4]. Teknik ini umum digunakan untuk menilai kepuasan pelanggan atau memahami opini publik di media sosial, dengan memanfaatkan algoritma klasifikasi yang dilatih menggunakan data teks berlabel secara manual [4].

2.2. Program Makan Bergizi Gratis

Program Makan Bergizi Gratis (MBG) merupakan kebijakan utama pemerintahan Prabowo-Gibran yang diluncurkan pada 6 Januari 2025 untuk mengatasi malnutrisi dan meningkatkan kualitas SDM Indonesia, dengan sasaran anak-anak, pelajar, dan ibu hamil [2]. Berdasarkan data Kementerian Kesehatan dan Kemenko PMK, sekitar 41% siswa di Indonesia mengalami kelaparan yang berdampak pada menurunnya kualitas pendidikan [2]. Meski bertujuan mulia, respons masyarakat beragam: ada yang mendukung upaya pengurangan stunting, namun banyak pula yang mempertanyakan efektivitas, kesiapan logistik, potensi pemborosan anggaran, dan keberlanjutan program [3].

2.3. Media Sosial sebagai Sumber Data Opini Publik

X (sebelumnya Twitter) adalah *Platform* media sosial yang memungkinkan pengguna menyampaikan gagasan, informasi, atau berita melalui cuitan singkat, serta menjadi sarana utama untuk diskusi publik, penyebaran isu terkini, dan aktivisme sosial-politik. Opini di X dianggap akurat dan relevan untuk memahami pandangan netizen terhadap suatu isu [1]. Sementara itu, TikTok merupakan aplikasi berbagi video pendek yang populer di kalangan generasi muda, tidak hanya sebagai hiburan tetapi juga sebagai medium yang membentuk opini dan perilaku publik melalui konten kreatif yang mudah diakses dan dibagikan [7].

2.4. *Knowledge Discovery in Database* (KDD)

Knowledge Discovery in Database (KDD) adalah proses sistematis untuk mengekstraksi pengetahuan bernilai dari data skala besar, terdiri atas lima tahap: selection, pre-processing, transformation, data mining, dan evaluation [8]. Pendekatan ini menyediakan kerangka metodologis komprehensif untuk mengubah data mentah menjadi pengetahuan yang dapat ditindaklanjuti.

2.5. *Text Preprocessing*

Text preprocessing adalah tahapan kritis dalam analisis teks untuk mengubah data tidak terstruktur menjadi format yang siap diproses oleh algoritma [9]. Tahapan umumnya meliputi: *cleaning* (menghapus URL, *mention*, *hashtag*), *case folding* (konversi ke huruf kecil), *tokenizing* (pemecahan teks menjadi token), normalisasi (menstandarkan kata tidak baku), *stopword removal* (menghapus kata umum), dan *stemming* (mengembalikan kata ke bentuk dasar). Proses ini meningkatkan kualitas representasi teks dan mengurangi dimensi data untuk efisiensi komputasi [9].

2.6. *Term Frequency–Inverse Document Frequency* (TF-IDF)

TF-IDF menggabungkan dua komponen: *Term Frequency* (TF) yang menghitung frekuensi kemunculan term dalam dokumen, dan *Inverse Document Frequency* (IDF) yang mengukur kelangkaan term di seluruh koleksi dokumen [9]. Bobot TF-IDF dihitung dengan rumus:

a. *Term Frequency (TF*), adalah frekuensi atau banyaknya kemunculan kata pada suatu dokumen.

$$tf_{t.d} = 1 + log_{10} tf_{t.d}, if tf_{t.d} > \quad (1)$$

b. Inverse Document Frequency (IDF), yaitu jumlah kemunculn term pada keseluruhan dokumen

$$idf_t = log_{10}\left(\frac{N}{df_{(t)}}\right) \quad (2)$$

c. TF-IDF dilakukan dengan

$$tf - idf_{td} = tf_{td} \times idf_t \quad (3)$$

2.7. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) menggunakan himpunan data latih dalam format (Xi, yi), di mana Xi adalah tupel dan yi adalah label kelas dengan i=1....N [4]:

$$f(x_i) = \{\geq\ 0, y_i\ =\ +1\ <0, y_i\} = -1 \quad (4)$$

Dimana terbentuknya *hyperplane* pada penjelasan sebagai berikut:

$$w.x + b = 0 \quad (5)$$

a. (w) adalah vektor bobot yang terdiri dari w1, w2, w3, …, wn, di mana n adalah jumlah atribut.
b. (b) adalah skalar yang juga dikenal sebagai bias.
c. (x) adalah kumpulan data latih atau tupel pelatihan

2.8. Evaluasi Kinerja Model

Confusion matrix merupakan tabel evaluasi yang mempresentasikan performa model klasifikasi dengan membandingkan prediksi terhadap label aktual [10]. Matriks ini terdiri dari empat komponen: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan komponen tersebut, metrik evaluasi dihitung sebagai berikut:

a. Akurasi (*Accuracy*): Menghitung *persentase* keseluruhan prediksi yang benar.

$$\text{Akurasi} = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN} \quad (6)$$

b. Presisi (*Precision*): Menghitung *persentase* prediksi positif yang benar dari semua prediksi positif.

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (7)$$

c. *Recall*: Menghitung *persentase* kasus positif yang sesuai dengan label asli.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

d. *F1-Score*: Menggabungkan presisi dan recall dalam satu matrik melalui rata-rata harmonis untuk mencerminkan keseimbangan antara keduanya.

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (9)$$

2.9. Studi Terdahulu

Penelitian Handayani dan Zufria (2023) menerapkan SVM untuk menganalisis sentimen terhadap bakal calon presiden RI 2024 di Twitter dengan akurasi 86% untuk calon urut 2 [5]. Pamungkas dan Kharisudin (2021) membandingkan tiga algoritma (SVM, Naïve Bayes, KNN) dalam analisis sentimen pandemi COVID-19, menemukan SVM dengan kernel linear mencapai akurasi rata-rata 90,01%, lebih unggul dibanding Naïve Bayes (79,20%) dan KNN (62,10%) [6]. Arsi dan Waluyo (2021) menggunakan SVM untuk menganalisis sentimen pemindahan ibu kota dengan akurasi 96,68% [11]. Rafli dan Heryana menganalisis sentimen terhadap Indonesia Emas 2045 menggunakan SVM dengan skenario pembagian data 80:20 menghasilkan akurasi 94%, presisi 95%, recall 98%, dan F1-Score 97% [12]. Studi-studi tersebut membuktikan efektivitas SVM dalam berbagai konteks analisis sentimen, termasuk isu kebijakan publik yang menjadi fokus penelitian ini.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Database* (KDD) yang terdiri dari lima tahapan utama, yaitu *selection*, *pre-processing*, *transformation*, data *mining*, dan *evaluation* [8]. Tahapan-tahapan tersebut diimplementasikan untuk menganalisis sentimen netizen terhadap Program Makan Bergizi Gratis dari *Platform* X dan TikTok menggunakan algoritma *Support Vector Machine* (SVM).

3.1. Pengumpulan Data

Data penelitian diperoleh melalui proses scraping dari dua *Platform* media sosial, yaitu X (sebelumnya Twitter) dan TikTok. Untuk *Platform* X, pengumpulan data dilakukan menggunakan *Application Programming Interface* (API) dengan kata kunci "program makan siang bergizi gratis", menghasilkan 5.364 komentar awal. Sedangkan untuk *Platform* TikTok, pengumpulan data dilakukan menggunakan *tools* Instant Data *Scraper* yang menghasilkan 2.306 komentar. Setelah proses penggabungan dan pembersihan data (dataset preparation), diperoleh total 7.354 data komentar yang siap diproses [1].

Seluruh data komentar kemudian dilabelkan secara manual menjadi dua kategori sentimen, yaitu positif dan negatif, dengan validasi oleh ahli bahasa Indonesia. Pelabelan dilakukan berdasarkan konten emosional dan opini yang terkandung dalam komentar, di mana komentar yang mengandung dukungan, apresiasi, atau harapan positif terhadap program diklasifikasikan sebagai sentimen positif, sedangkan komentar yang mengandung kritik, kekhawatiran, atau penolakan diklasifikasikan sebagai sentimen negatif [9].

3.2. *Pre-processing*

Tahap *pre-processing* bertujuan mengubah data teks tidak terstruktur menjadi format yang siap diproses oleh algoritma klasifikasi. Tahapan pre-processing yang diterapkan meliputi [9]:

a. *Cleaning*: Menghapus elemen tidak relevan seperti URL, mention (@username), hashtag, retweet (RT), emoji, dan karakter khusus lainnya menggunakan ekspresi reguler (regular expression).
b. *Case folding*: Mengonversi seluruh karakter teks menjadi huruf kecil (lowercase) untuk menghindari duplikasi kata akibat perbedaan kapitalisasi.
c. *Tokenizing*: Memecah teks menjadi unit kata (token) menggunakan *library* NLTK.
d. Normalisasi: Mengubah kata tidak baku, singkatan, atau typo menjadi bentuk baku berdasarkan kamus normalisasi bahasa Indonesia (contoh: "gk" → "tidak", "gua" → "aku").
e. *Stopword removal*: Menghapus kata-kata umum yang tidak membawa makna spesifik seperti "yang", "di", "ke", "dan" menggunakan daftar *stopword* bahasa Indonesia dari *library* NLTK.
f. *Stemming*: Mengubah kata berimbuhan menjadi bentuk dasarnya menggunakan Sastrawi Stemmer (contoh: "makanan" → "makan", "bergizi" → "gizi").

3.3. Transformasi dengan TF-IDF

Setelah tahap *pre-processing*, dilakukan transformasi data teks menjadi representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). TF-IDF mengukur relevansi suatu term terhadap dokumen dengan mempertimbangkan frekuensi kemunculannya dalam dokumen dan kelangkaannya di seluruh korpus [13]. Bobot TF-IDF dihitung dengan rumus:

a. *Term Frequency (TF*), adalah frekuensi atau banyaknya kemunculan kata pada suatu dokumen.

$$tf_{t.d} = 1 + log_{10} tf_{t.d}, if tf_{t.d} > \quad (10)$$

b. Inverse Document Frequency (IDF), yaitu jumlah kemunculn term pada keseluruhan dokumen

$$idf_t = log_{10}\left(\frac{N}{df_{(t)}}\right) \quad (11)$$

c. TF-IDF dilakukan dengan

$$tf - idf_{td} = tf_{td} \times idf_t \quad (12)$$

3.4. Klasifikasi dengan *Support Vector Machine*

Algoritma *Support Vector Machine* (SVM) dengan kernel linear diimplementasikan untuk mengklasifikasikan sentimen komentar menjadi dua kategori (positif dan negatif) [14]. SVM bekerja dengan menemukan *hyperplane* optimal yang memisahkan kedua kelas dengan margin maksimum [15]. Fungsi keputusan SVM dinyatakan sebagai:

$$f(x_i) = \{\geq 0, y_i = +1 < 0, y_i\} = -1 \quad (13)$$

Dimana terbentuknya *hyperplane* pada penjelasan sebagai berikut:

$$w.x + b = 0$$

a. (w) adalah vektor bobot yang terdiri dari w1, w2, w3, …, wn, di mana n adalah jumlah atribut.
b. (b) adalah skalar yang juga dikenal sebagai bias.
c. (x) adalah kumpulan data latih atau tupel pelatihan

Data dibagi dengan rasio 80:20, di mana 80% digunakan sebagai data latih (training set) dan 20% sebagai data uji (*testing set*) dengan metode *stratified sampling* untuk mempertahankan proporsi distribusi kelas pada kedua subset [10].

3.5. Evaluasi Model

Kinerja model dievaluasi menggunakan *confusion matrix* yang terdiri dari empat komponen: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) [10]. Berdasarkan komponen tersebut, dihitung empat metrik evaluasi:

a. Akurasi (*Accuracy*): Menghitung *persentase* keseluruhan prediksi yang benar.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

b. Presisi (*Precision*): Menghitung *persentase* prediksi positif yang benar dari semua prediksi positif.

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (15)$$

c. *Recall*: Menghitung *persentase* kasus positif yang sesuai dengan label asli.

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

d. *F1-Score*: Menggabungkan presisi dan recall dalam satu matrik melalui rata-rata harmonis untuk mencerminkan keseimbangan antara keduanya.

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (17)$$

Evaluasi dilakukan secara terpisah untuk masing-masing *Platform* (X dan TikTok) serta pada data gabungan untuk menganalisis pengaruh distribusi data terhadap kinerja model klasifikasi [10]. Selain itu, dilakukan visualisasi *word cloud* untuk mengidentifikasi kata-kata dominan pada masing-masing kelas sentimen guna memperkaya interpretasi hasil analisis [16].

4. HASIL DAN PEMBAHASAN

Penelitian ini mengimplementasikan algoritma *Support Vector Machine* (SVM) dengan pendekatan *Knowledge Discovery in Database* (KDD) untuk menganalisis sentimen netizen terhadap Program Makan Bergizi Gratis (MBG) yang diluncurkan pemerintahan Prabowo-Gibran pada 6 Januari 2025. Total 7.354 komentar diperoleh dari dua platform media sosial—X (5.089 komentar) dan TikTok (2.265 komentar)—setelah melalui proses dataset preparation yang meliputi validasi oleh ahli bahasa dan pelabelan manual menjadi dua kategori sentimen: positif dan negatif.

4.1. Data *Selection*

Proses seleksi data dimulai dengan pengumpulan komentar melalui scraping dari *Platform* X menggunakan API dengan kata kunci "program makan siang bergizi gratis", menghasilkan 5.364 data mentah. Sedangkan untuk platform TikTok, pengumpulan data dilakukan menggunakan tools Instant Data *Scraper* yang menghasilkan 2.306 komentar. Setelah proses penggabungan dan pembersihan awal (dataset *preparation*), diperoleh total 7.354 data yang kemudian divalidasi secara manual oleh ahli bahasa.

Distribusi data berdasarkan platform dan kelas sentimen disajikan pada Tabel 1

Tabel 1. Distribusi Data

Platform	Total Data	Sentimen Positif	Sentimen Negatif
X	5.089	1.503	3.586
TikTok	2.265	1.940	325
Gabungan	7.354	3.443	3.911

Berdasarkan Tabel 1. terlihat ketimpangan distribusi sentimen yang signifikan antar platform. Platform X didominasi sentimen negatif (70,47%), sedangkan TikTok justru didominasi sentimen positif (85,65%). Perbedaan ini mencerminkan karakteristik pengguna dan konteks komunikasi yang berbeda pada masing-masing *Platform*. Pengguna X cenderung lebih kritis dan analitis dalam menyampaikan opini politik, sementara pengguna TikTok—yang mayoritas generasi muda—lebih mengekspresikan respons emosional dan apresiatif terhadap program sosial.

4.2. *Preprocessing*

Tahap *preprocessing* bertujuan mengubah data teks tidak terstruktur menjadi format terstruktur yang siap diproses oleh algoritma klasifikasi. Seluruh tahapan *preprocessing* diimplementasikan menggunakan *Python* dengan *library* NLTK dan regular expression. Berikut penjabaran setiap sub-tahap beserta contoh hasilnya.

4.3. *Cleaning*

Proses *cleaning* menghapus elemen non-teks yang tidak relevan seperti URL, mention (@username), hashtag, emoji, retweet (RT), dan karakter khusus. Implementasi dilakukan menggunakan regular expression dengan pola khusus untuk setiap elemen yang dihapus.

Tabel 2. *Cleaning*

Sebelum Cleaning	Sesudah Cleaning
"Tak ada anak yang boleh lapar! Bersama kita dukung makan siang bergizi gratis untuk masa depan yang lebih cerah. #PenuhiGiziIndonesia #DukungGiziSeimbang"	"Tak ada anak yang boleh lapar Bersama kita dukung makan siang bergizi gratis untuk masa depan yang lebih cerah"
"Untung ada makan siang gratis yang berganti makan bergizi gratis. @Gerindra @prabowo"	"Untung ada makan siang gratis yang berganti makan bergizi gratis"
"Najis ah. Makanan begini gua bisa beli judulnya makan siang gratis tapi makan siang mana ada budgetnya sekecil ini? Mending makan makanan emaknya wehhhhhhh lebih bergizi!!!!!"	"Najis ah Makanan begini gua bisa beli judulnya makan siang gratis tapi makan siang mana ada budgetnya sekecil ini Mending makan makanan emaknya wehhhhhhh lebih bergizi"

Proses cleaning berhasil menghilangkan 316 data yang tidak memiliki konten teks setelah pembersihan, sehingga total data berkurang dari 7.670 menjadi 7.354 komentar valid.

4.4. *Case Folding*

Case folding dilakukan dengan mengonversi seluruh karakter teks menjadi huruf kecil (lowercase) untuk menghindari duplikasi kata akibat perbedaan kapitalisasi. Proses ini diimplementasikan menggunakan fungsi str.lower() pada Python.

Tabel 3. *Case Folding*

Sebelum Case Folding	Sesudah Case Folding
"Tak ada anak yang boleh lapar Bersama kita dukung makan siang bergizi gratis untuk masa depan yang lebih cerah"	"tak ada anak yang boleh lapar bersama kita dukung makan siang bergizi gratis untuk masa depan yang lebih cerah"
"Makan siang Bergizi Gratis Mimin seneng banget lihat adeknya lahap makan"	"makan siang bergizi gratis mimin seneng banget lihat adeknya lahap makan"

4.5. *Tokenizing*

Tokenizing memecah teks menjadi unit-unit kata (token) menggunakan library NLTK. Proses ini mengidentifikasi batas kata berdasarkan spasi dan tanda baca.

Tabel 4. *Tokenizing*

Hasil Proses Sebelumnya	Hasil *Tokennizing*
tak ada anak yang boleh lapar bersama kita dukung makan siang bergizi gratis untuk masa depan yang lebih cerah	[‘tak’, ‘ada’, ‘anak’, ‘yang’, ‘boleh’, ‘lapar’, ‘bersama’, ‘kita’, ‘dukung’, ‘makan’, ‘siang’, ‘bergizi’, ‘gratis’, ‘untuk’, ‘masa’, ‘depan’, ‘yang’, ‘lebih’, ‘cerah’]
untung ada makan siang gratis yang berganti makan bergizi gratis	[‘untung’, ‘ada’, ‘makan’, ‘siang’, ‘gratis’, ‘yang’, ‘berganti’, ‘makan’, ‘bergizi’, ‘gratis’]

4.6. *Normalization*

Normalisasi mengubah kata tidak baku, singkatan, atau typo menjadi bentuk baku berdasarkan kamus normalisasi bahasa Indonesia yang dikembangkan khusus untuk penelitian ini.

Tabel 5. *Normalization*

Sebelum Normalisasi	Sesudah Normalisasi
['tak', 'ada', 'anak', ...]	['tidak', 'ada', 'anak', ...]
['gua', 'bisa', 'beli', ...]	['aku', 'bisa', 'beli', ...]
['tp', 'buahnya', 'gasuka']	['tapi', 'buahnya', 'tidak', 'suka']
['mimin', 'seneng', 'banget']	['admin', 'senang', 'sekali']

4.7. *Stopwords Removal*

Stopword removal menghapus kata-kata umum yang tidak membawa makna spesifik seperti "yang", "di", "ke", "dan", "untuk", "ini", "itu" menggunakan daftar stopword bahasa Indonesia dari library NLTK yang telah dikustomisasi.

Tabel 6. *Stopwords Removal*

Sebelum Stopword Removal	Sesudah Stopword Removal
['tidak', 'ada', 'anak', 'yang', 'boleh', 'lapar', 'bersama', 'kita', 'dukung', 'makan', 'siang', 'bergizi', 'gratis', 'untuk', 'masa', 'depan', 'yang', 'lebih', 'cerah']	['anak', 'boleh', 'lapar', 'dukung', 'makan', 'siang', 'bergizi', 'gratis', 'masa', 'depan', 'cerah']

Proses ini mengurangi dimensi data secara signifikan dan meningkatkan fokus analisis pada kata-kata bermakna yang relevan dengan sentimen.

4.8. *Stemming*

Stemming mengubah kata berimbuhan menjadi bentuk dasarnya menggunakan Sastrawi Stemmer untuk bahasa Indonesia.

Tabel 7. *Stemming*

Sebelum Stemming	Sesudah Stemming
['anak', 'boleh', 'lapar', 'dukung', 'makan', 'siang', 'bergizi', 'gratis', 'masa', 'depan', 'cerah']	['anak', 'boleh', 'lapar', 'dukung', 'makan', 'siang', 'gizi', 'gratis', 'masa', 'depan', 'cerah']
['makanan', 'begini', 'beli', 'judulnya', 'makan', 'siang', 'gratis']	['makan', 'gini', 'beli', 'judul', 'makan', 'siang', 'gratis']

Setelah seluruh tahap preprocessing, diperoleh representasi teks yang bersih, terstandarisasi, dan optimal untuk proses transformasi numerik selanjutnya.

4.9. *Transformation*

Transformasi data teks menjadi representasi numerik dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF mengukur relevansi suatu term terhadap dokumen dengan mempertimbangkan frekuensi kemunculannya dalam dokumen dan kelangkaannya di seluruh korpus.

Hasil transformasi menghasilkan matriks berdimensi 7.354 × 12.847, di mana setiap baris merepresentasikan vektor fitur numerik dari satu komentar. Dimensi 12.847 merepresentasikan jumlah term unik setelah preprocessing dan seleksi fitur berbasis frekuensi.

Dokumen 1 (Positif): [0.000, 0.699, 1.000, 1.000, 1.000, ..., 0.000]
Dokumen 2 (Positif): [0.000, 0.000, 0.000, 0.000, 0.000, ..., 1.000]
Dokumen 3 (Negatif): [0.000, 0.000, 0.000, 0.000, 0.000, ..., 1.000, 1.000, 1.000]

4.10. *Features Selection*

Data dibagi dengan rasio 80:20 menggunakan metode stratified sampling untuk mempertahankan proporsi distribusi kelas pada kedua subset. Pembagian dilakukan secara terpisah untuk masing-masing platform dan data gabungan:

Tabel 8. *Feature Selection*

Platform	Total Data	Data Latih	Data Uji	Rasio Positif Negatif (latih)	Rasio Positif Negatif (uji)
X	5.089	4.071	1.018	29,5% : 70,5%	29,6% : 70,4%
TikTok	2.265	1.812	453	85,7% : 14,3%	85,6% : 14,4%
Gabungan	7.354	5.883	1.471	46,8% : 53,2%	46,9% : 53,1%

Stratified sampling memastikan bahwa distribusi kelas pada data latih dan uji relatif seimbang sesuai proporsi populasi, sehingga model dapat dilatih dan dievaluasi secara representatif.

4.11. Data *Mining*

Algoritma *Support Vector Machine* (SVM) dengan kernel linear diimplementasikan menggunakan library scikit-learn pada Python.

Model dilatih menggunakan data latih yang telah ditransformasi ke dalam vektor TF-IDF. Proses pelatihan menghasilkan hyperplane optimal yang memisahkan kelas positif dan negatif dengan margin maksimum.

$$f(x) = \text{sign}(\sum \alpha_i y_i K(x_i, x) + b)$$

4.12. *Evaluation*

Evaluasi kinerja model dilakukan menggunakan confusion matrix dan metrik turunannya (akurasi, presisi, recall, F1-score). Pengujian dilakukan dengan membandingkan hasil secara terpisah untuk masing-masing platform serta pada data gabungan untuk menganalisis pengaruh distribusi data terhadap kinerja model.

4.13. Hasil *Evaluation Platform* X

Tabel 9. Hasil Evaluasi *Platform X*

	Prediksi Positif	Prediksi Negatif
Aktual Positif	248 (TP)	53 (FN)
Aktual Negatif	129 (FP)	588 (TN)

Metrik Evaluasi *Platform* X:

a. Akurasi: 82,02%
b. Presisi: 65,78%
c. Recall: 82,39%
d. F1-Score: 73,15%

Berdasarkan hasil pengujian pada Platform X, Nilai recall yang relatif tinggi menunjukkan bahwa model cukup baik dalam mendeteksi data yang termasuk dalam kelas positif. Namun, nilai presisi yang lebih rendah mengindikasikan adanya kesalahan klasifikasi yang cukup signifikan, khususnya pada data negatif yang diprediksi sebagai positif (false positive) sebanyak 129 data.

Distribusi data pada Platform X didominasi oleh kelas negatif sebesar 70,47%, yang menyebabkan model cenderung bias terhadap kelas mayoritas. Hal ini terlihat dari jumlah true negative (588) yang jauh lebih tinggi dibandingkan true positive (248). Ketidakseimbangan data ini mempengaruhi kemampuan model dalam membedakan kedua kelas secara optimal, sehingga meskipun model memiliki sensitivitas yang baik (recall tinggi), tingkat ketepatan prediksi terhadap kelas positif masih rendah (presisi rendah).

4.14. Hasil *Evaluation Platform* Tiktok

Tabel 10. Hasil Evaluasi *Platform* Tiktok

	Prediksi Positif	Prediksi Negatif
Aktual Positif	378 (TP)	11 (FN)
Aktual Negatif	44 (FP)	29 (TN)

Metrik Evaluasi *Platform* Tiktok:

a. Akurasi: 88,08%
b. Presisi: 89,57%
c. Recall: 97,17%
d. F1-Score: 93,22%

Pada *Platform* TikTok, Nilai *recall* yang sangat tinggi menunjukkan bahwa model hampir mampu mendeteksi seluruh data positif dengan sangat baik, yang ditandai dengan jumlah *false negative* yang sangat kecil (11 data).

Namun demikian, distribusi data pada *platform* ini sangat tidak seimbang, dengan dominasi kelas positif sebesar 85,65%. Kondisi ini menyebabkan model cenderung "*overfitting*" terhadap kelas positif, sehingga meskipun performa metrik terlihat sangat tinggi, kemampuan model dalam mengenali kelas negatif relatif kurang optimal. Hal ini terlihat dari jumlah *true negative* yang rendah (29) dibandingkan *false positive* (44), yang menunjukkan bahwa model masih mengalami kesulitan dalam membedakan data negatif secara akurat.

Dengan demikian, tingginya nilai akurasi dan *F1-score* pada *Platform* TikTok tidak sepenuhnya mencerminkan kemampuan generalisasi model, melainkan dipengaruhi oleh dominasi kelas positif yang sangat kuat.

4.15. Hasil *Evaluation Platform* Gabungan

Tabel 11. Hasil Evaluasi *Platform* Gabungan

	Prediksi Positif	**Prediksi Negatif**
Aktual Positif	318 (TP)	133 (FN)
Aktual Negatif	118 (FP)	902 (TN)

Metrik Evaluasi Platform Gabungan:

a. Akurasi: 83,01%
b. Presisi: 72,94%
c. Recall: 70,51%
d. F1-Score: 71,70%

Berbeda dengan pengujian pada masing-masing platform, distribusi data pada dataset gabungan relatif lebih seimbang (46,82% positif dan 53,18% negatif), sehingga menghasilkan performa model yang lebih stabil dan tidak terlalu bias terhadap salah satu kelas.

Nilai presisi dan *recall* yang cukup seimbang menunjukkan bahwa model dapat mengklasifikasikan kedua kelas sentimen dengan lebih stabil. Meskipun akurasi lebih rendah dibandingkan *Platform* TikTok, nilai *F1-score* pada data gabungan lebih mencerminkan kinerja model secara keseluruhan karena mempertimbangkan keseimbangan antara presisi dan *recall*.

4.16. Analisis Perbandingan Antar Platform

Berdasarkan hasil evaluasi pada ketiga skenario pengujian, terlihat bahwa distribusi data memiliki pengaruh yang signifikan terhadap kinerja model. *Platform* TikTok menghasilkan performa tertinggi secara metrik, namun hal ini dipengaruhi oleh dominasi kelas positif yang sangat tinggi, sehingga model menjadi kurang optimal dalam mengenali kelas negatif. Sebaliknya, *Platform* X menunjukkan performa yang lebih rendah, terutama pada presisi, akibat dominasi kelas negatif yang menyebabkan meningkatnya jumlah *false positive*.

Sementara itu, pengujian pada data gabungan menghasilkan performa yang lebih seimbang meskipun tidak mencapai nilai akurasi tertinggi. Hal ini menunjukkan bahwa distribusi data yang lebih proporsional mampu meningkatkan kemampuan generalisasi model dalam mengklasifikasikan kedua kelas sentimen secara lebih konsisten.

Secara keseluruhan, dapat disimpulkan bahwa keseimbangan distribusi data merupakan faktor penting dalam meningkatkan kualitas model klasifikasi, terutama dalam mengurangi bias terhadap kelas tertentu dan meningkatkan stabilitas performa model.

5. KESIMPULAN DAN SARAN

Penelitian ini membuktikan bahwa algoritma Support Vector Machine (SVM) dengan kernel linear efektif dalam mengklasifikasikan sentimen netizen terhadap Program Makan Bergizi Gratis dengan akurasi 82% pada platform X, 88% pada TikTok, dan 83% pada data gabungan. Kinerja model optimal dicapai pada kondisi distribusi data seimbang (46,84% positif : 53,16% negatif), sedangkan ketimpangan ekstrem pada pengujian terpisah memengaruhi stabilitas klasifikasi. Analisis word cloud mengungkap perbedaan karakteristik sentimen dominan antar platform: X didominasi sentimen negatif (70,53%) yang kritis terhadap aspek anggaran, pajak, dan transparansi kebijakan, sedangkan TikTok didominasi sentimen positif (85,65%) yang bersifat emosional dan apresiatif terhadap manfaat langsung program bagi anak-anak. Untuk penelitian selanjutnya disarankan menggunakan dataset seimbang, memperluas evaluasi dengan metrik precision, recall, dan F1-score, membandingkan SVM dengan algoritma lain (Naïve Bayes, Random Forest), serta memperluas sumber data ke platform media sosial tambahan guna memperoleh gambaran persepsi publik yang lebih holistik bagi perbaikan implementasi dan strategi komunikasi program.

DAFTAR PUSTAKA

[1] N. Wear, N. Rosita, and I. Mulia, "Analisis Sentimen Terhadap Hasil Pilpres 2024 Pada Aplikasi Tiktok Dan X Menggunakan Metode Naïve Bayes Classifier," *Jurnal Imliah Informatika Mulia*, vol. 1, no. 2, pp. 15–22, 2024.

[2] D. E. Prasetyo, *Pancasila : Jurnal Keindonesiaan*, 2nd ed., vol. 3. J. Keindonesiaan, 2023.

[3] C. Sentosa, *Program Makan Siang Gratis di Indonesia: Solusi Atau Beban?* Character Building BINUS University, 2025.

[4] W. Amna, *Data Mining*. Sumatra Barat: PT. Global Eksekutif Teknologi, 2023.

[5] A. Handayani and I. Zufria, "Analisis Sentimen Terhadap Bakal Capres RI 2024 di Twitter Menggunakan Algoritma SVM," *Journal of Information System Research (JOSH)*, vol. 5, no. 1, pp. 53–63, Oct. 2023, doi: 10.47065/josh.v5i1.4379.

[6] F. S. Pamungkas and K. Iqbal, "Analisis Sentimen dengan SVM, NAIVE BAYES dan

KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 4, no. 1, pp. 628–634, 2021.

[7] N. F. Ilmi, *Problematika TeoriI & Praktik Komunikasi*. Jakarta Selatan: PT. Mahakarya Citra Utama Group, 2023.

[8] B. Efori, *Data Mining Untuk Perguruan Tinggi*. Yogyakarta: Deepublish, 2020.

[9] H. C. Morama and D. E. Ratnawati, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 4, pp. 1702–1708, 2022.

[10] Y. Nurdin, K. Saddami, and N. Nasaruddin, *Pengenalan Praktis Supervised Machine Learning: Dengan Jupyter Notebook*. Banda Aceh: USK Press, 2025.

[11] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, p. 147, Feb. 2021, doi: 10.25126/jtiik.0813944.

[12] M. Rafli Tanjung, N. Heryana, and S. Siska, "Analisis Sentiment Pengguna Platform X Terhadap Indonesia Emas 2045 Menggunakan Algoritma Suport Vector Machine," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 4, pp. 7657–7665, Aug. 2024, doi: 10.36040/jati.v8i4.10327.

[13] S. U. Masruroh and E. S. Qorina, *Fast Text dan Word2vec Pada Query Kesamaan Semantik Sistem Temu Kembali Informatika*. Yogyakarta: CV. Budi Utama, 2022.

[14] R. Hidayat, M. Haris, and Z. K. Simbolon, "Implementasi Metode Support Vector Machine (SVM) Pada Klasifikasi Drop Out (DO) Mahasiswa," *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, vol. 9, no. 2, pp. 51–59, 2024.

[15] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink," *Teknika*, vol. 13, no. 1, pp. 127–136, Mar. 2024, doi: 10.34148/teknika.v13i1.746.

[16] I. Saputra and A. Dinar, *Machine Learning Untuk Pemula*, vol. 1. Informatika, 2022