Computer Based Information System Journal
ISSN (Print): 2337-8794 | E- ISSN : 2621-5292
web jurnal : http://ejournal.upbatam.ac.id/index.php/cbis

PERBANDINGAN KINERJA SVM, KNN, DAN NAÏVE BAYES PADA ANALISIS SENTIMEN TWEET MBG

Ibnu Abdul Ghofur[1], Anggia Dasa Putri[2]
[1,2]*Universitas Putera Batam, Indonesia.*

INFORMASI ARTIKEL

Diterima Redaksi: Desember 2025
Diterbitkan *Online*: Maret 2026

KATA KUNCI

Analisis Sentimen, SVM, KNN, *Naïve Bayes*, Twitter, Program Makan Bergizi Gratis.

KORESPONDENSI

E-mail:
pb220210054@upbatam.ac.id[1]
Anggia.Dasa@upbatam.ac.id[2]

A B S T R A C T

This study explores the effectiveness of three machine learning classifiers—Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes—in analyzing public sentiment toward the Free Nutritious Meal Program using data from the X (Twitter) platform. Tweets were collected via Tweet Harvester by applying keyword-based filtering over the August–October 2025 period. Prior to model implementation, the textual dataset underwent comprehensive preprocessing, including data cleaning, case normalization, lexical standardization, tokenization, stopword elimination, and stemming. Sentiment labels were generated using a lexicon-based approach to distinguish between positive and negative opinions. The processed data were divided into training and testing subsets for classification. Model performance was evaluated using accuracy metrics derived from the confusion matrix. The results show that SVM outperformed the other models with an accuracy of 91.7%, followed by Naïve Bayes at 79.6% and KNN at 79.3%, indicating the strong capability of SVM in handling complex textual representations in social media sentiment analysis.

I. Latar Belakang

Perkembangan media sosial telah mengubah cara masyarakat menyampaikan opini terhadap isu sosial dan kebijakan publik. salah satu media yang paling aktif digunakan yang sifatnya terbuka, *real-time*, dan berbasis teks pendek. Karakteristik ini menjadikan platform X sebagai sumber data yang potensial namun menantang untuk dianalisis otomatis [1].

Program Makan Bergizi Gratis adalah upaya jangka panjang pemerintah untuk meningkatkan status gizi terutama anak-anak. Besarnya anggaran dan cakupan program tersebut menimbulkan respons publik yang beragam, baik berupa dukungan maupun kritik. Oleh karena itu, diperlukan metode yang dapat secara sistematis dan objektif memetakan opini publik.

Natural Language Processing yang dipadukan dengan metode pembelajaran mesin telah menjadi pendekatan utama dalam klasifikasi opini publik pada data berbentuk teks. [2]. Setiap algoritma klasifikasi memiliki karakteristik yang berbeda sehingga

menghasilkan tingkat performa yang bervariasi. Berdasarkan hal tersebut, penelitian ini berfokus pada evaluasi beberapa algoritma untuk mengidentifikasi metode yang paling efektif dalam analisis sentimen opini publik terkait Program Makan Bergizi Gratis.

II. Kajian Literatur

2.1 Data Mining

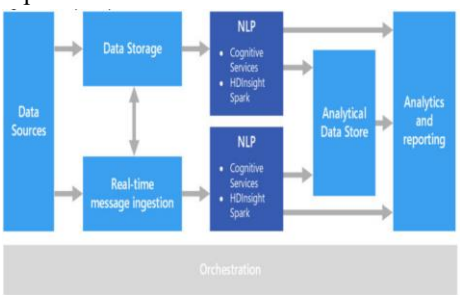
Pendekatan analitis untuk menggali pengetahuan dan informasi bernilai dari kumpulan data berskala besar melalui penerapan teknik statistik, pembelajaran mesin, serta metode pengenalan pola. Dalam bidang ilmu komputer, data mining dikenal sebagai bagian dari proses *Knowledge Discovery in Databases*, yang mencakup serangkaian tahapan sistematis mulai dari pembersihan data, integrasi, seleksi, transformasi, hingga proses penemuan pola dan evaluasi hasil. Tujuan utama dari data mining adalah mengidentifikasi hubungan, kecenderungan, maupun pola tersembunyi dalam data yang sulit ditemukan melalui analisis manual atau konvensional.

Teknik yang sering diterapkan dalam data mining antara lain klasifikasi, klasterisasi, asosiasi, dan regresi, di mana setiap metode memiliki peran yang berbeda dalam proses analisis serta pemahaman karakteristik data. Dalam konteks pengolahan bahasa alami *Natural Language Processing*, *data mining* berperan sebagai fondasi utama bagi sistem yang melakukan analisis dan interpretasi terhadap data berbasis teks.

Dalam analisis sentimen, data mining dimanfaatkan untuk mengekstraksi pola linguistik, tren, serta keterkaitan implisit dalam data teks guna mengelompokkan opini pengguna ke dalam kategori sentimen positif, negatif, maupun netral. Selain berfungsi untuk menemukan informasi tersembunyi, data mining juga berperan dalam pembangunan model prediktif yang mampu memahami konteks bahasa. [3] menjelaskan bahwa data mining menjadi dasar dalam penerapan algoritma *machine learning* seperti SVM dan *Naïve Bayes*, yang bekerja dengan cara melatih model menggunakan dataset ulasan pengguna sebelum diterapkan pada data baru. roses pelatihan ini memungkinkan sistem mempelajari pola kata, struktur kalimat, serta konteks sentimen sehingga klasifikasi opini pengguna pada aplikasi daring dapat dilakukan secara otomatis.

2.2 Natural Language Processing

Salah satu bidang dalam kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Melalui NLP, sistem komputer dapat memproses, memahami, serta menghasilkan bahasa alami secara otomatis sehingga mendekati kemampuan linguistik manusia. NLP tidak hanya mengolah teks sebagai kumpulan kata, tetapi juga menilai makna kontekstualnya melalui tahapan seperti *tokenisasi*, *part-of-speech tagging*, dan analisis semantik untuk memperoleh pemahaman terhadap maksud suatu teks.



Gambar 1. Pentingnya NLP dalam Pengembangan Data

Terlihat pada Gambar 2.1, alur proses penambangan data berbasis NLP dimulai dari tahap data *sources* yang dikumpulkan ke dalam data storage atau *real-time* message ingestion. Selanjutnya, modul NLP (melalui *Cognitive Services*, *HDInsight*, dan *Spark*) melakukan proses pemrosesan bahasa alami untuk menginterpretasikan makna dari data tersebut sebelum dikirimkan ke *Analytical Data Store*. Dari tahap ini, data hasil analisis kemudian digunakan dalam proses *analytics and reporting* untuk menghasilkan wawasan *insight* yang dapat mendukung pengambilan keputusan.

Dalam konteks analisis sentimen modern, NLP memainkan peran sentral dalam

mengekstraksi makna dan emosi dari data besar *big data*, terutama dari platform media sosial. [4] mengungkapkan bahwa penerapan NLP mampu memproses ribuan komentar YouTube secara otomatis dan efisien, sehingga opini publik dapat dikelompokkan ke dalam sentimen positif, negatif, maupun netral dengan akurasi yang tinggi. Selain itu, penelitian tersebut menekankan bahwa pengolahan bahasa natural lebih baik dalam mengidentifikasi nuansa emosional yang rumit, seperti sarkasme dan ambiguitas bahasa, yang sering kali tidak mampu diungkapkan secara optimal menggunakan metode statistik yang lazim digunakan. Keunggulan ini menjadikan NLP sebagai dasar penting dalam pengembangan berbagai model analisis sentimen berbasis *machine learning*, seperti SVM dan *Naïve Bayes*, yang banyak digunakan untuk memahami dinamika opini publik pada konteks sosial, politik, maupun perilaku pengguna. Temuan ini menunjukkan bahwa integrasi NLP dan *machine learning* berperan penting dalam pemetaan persepsi masyarakat yang cepat dan akurat.

2.3 Analisis Sentimen

Analisis sentimen adalah bagian dari pemrosesan bahasa alam NLP yang bertujuan untuk memahami, menafsirkan, dan mengklasifikasikan emosi, pendapat, dan sikap manusia yang diungkapkan dalam teks. [5] Proses ini mengevaluasi polaritas pernyataan terhadap entitas tertentu ke dalam kategori positif, negatif, atau netral. Dalam konteks digital, analisis sentimen menjadi alat penting untuk mengekstraksi informasi subjektif dari data besar seperti ulasan produk, komentar media sosial, dan forum daring.

Gambar 2. Tahapan proses analisis sentimen

Secara umum, analisis sentimen dapat dibedakan menjadi tiga jenis utama, yaitu analisis sentimen tingkat dokumen, kalimat, dan aspek. Analisis tingkat dokumen mengkaji keseluruhan teks untuk menentukan polaritas sentimen, sedangkan analisis tingkat kalimat fokus pada satu kalimat untuk mendeteksi emosi atau opini yang terkandung di dalamnya. Adapun analisis berbasis aspek *aspect-based sentiment analysis* meneliti opini terhadap atribut tertentu dari suatu entitas, misalnya fitur produk atau layanan spesifik [5] Selain itu, bentuk lanjutan dari analisis sentimen meliputi emotion detection yang berfokus pada identifikasi emosi seperti marah, bahagia, atau sedih, serta intention analysis yang menilai niat atau tujuan di balik suatu ekspresi teks.

2.4 *Scraping data.*

Scraping data merupakan proses pengambilan data secara otomatis dari sumber menggunakan program yang mengekstraksi informasi spesifik dari struktur halaman web atau memanfaatkan API sebagai sarana penghubung antarmuka pemrograman aplikasi.teknik ini memperoleh data dalam jumlah besar secara efisien dan sistematis tanpa melakukan penyalinan. Hasil *scraping* ini kemudian menjadi dataset terstruktur dan digunakan pada tahap pra-pemrosesan. *Scraping* semacam ini terbukti efektif untuk memperoleh data yang relevan, terstandar, dan siap digunakan untuk analisis sentimen karena mampu menghemat waktu dan meminimalkan kesalahan manusia dalam pengumpulan data secara manual.

Scraping data banyak diterapkan pada platform X (Twitter) karena karakteristik datanya

yang bersifat *real-time* dan representatif terhadap opini publik [6] proses scraping bisa dilakukan dengan menggunakan *library tweppy* pada bahasa Python untuk mengumpulkan tweet berdasarkan kata kunci tertentu seperti "olahraga", "bersepeda", "bulu tangkis", dan "lari". Data yang dikumpulkan mencakup isi teks, tanggal unggah, nama pengguna, serta metadata lainnya seperti jumlah retweet dan *likes*. Penggunaan *scraping* data di platfrom X memungkinkan peneliti mendapatkan data autentik dan kontekstual dalam jumlah besar yang mencerminkan opini masyarakat secara langsung.

2.5 Pra-pemrosesan data

Tahap pra-pemrosesan data dilakukan untuk memperbaiki kualitas data teks agar siap digunakan pada proses analisis berikutnya. Data teks yang diperoleh dari media sosial umumnya bersifat tidak terstruktur, mengandung *noise*, serta variasi penulisan yang tidak baku. Oleh karena itu, diperlukan serangkaian tahapan pra-pemrosesan untuk memastikan data siap digunakan oleh algoritma pembelajaran mesin. [7] dijelaskan bahwa pra-pemrosesan berfungsi mengubah data teks mentah ke dalam bentuk yang terstruktur dan bermakna, sedangkan [8] menegaskan tahapan ini berperan penting dalam menjaga konsistensi dan kualitas data sebelum proses ekstraksi fitur dan analisis sentimen.

1. Cleaning

Cleaning pembersihan data teks dari unsur-unsur yang tidak memiliki kontribusi terhadap makna, seperti URL, tanda baca, angka, emoji, serta simbol khusus lainnya. Proses ini bertujuan menghilangkan *noise* yang dapat mengganggu kinerja algoritma pada tahap analisis selanjutnya. Data teks yang berasal dari media sosial umumnya mengandung banyak karakter *non-alfabetik*, sehingga perlu dilakukan pembersihan agar kualitas data meningkat. [7] menyatakan bahwa proses cleaning sangat penting untuk mengurangi gangguan dalam data dan meningkatkan performa analisis.

2. Case Folding

Case folding dilakukan untuk menyeragamkan teks dengan mengubah semua huruf menjadi huruf kecil. Tahapan ini dilakukan untuk menghindari perbedaan representasi kata yang disebabkan oleh variasi penggunaan huruf kapital, sehingga kata seperti "Data", "data", dan "DATA" diperlakukan sebagai entitas yang sama oleh sistem. [8] *case folding* bagian dari normalisasi leksikal yang bertujuan untuk menstandarkan format teks [3] *case folding* harus dilakukan sebelum *tokenisasi* dan pembobotan fitur agar representasi teks menjadi konsisten.

3. Normalisasi

Normalisasi bertujuan untuk mengonversi kata tidak baku, singkatan, maupun bahasa informal ke dalam bentuk kata baku sesuai kaidah bahasa. Pada data media sosial, normalisasi menjadi tahap penting karena sering ditemui variasi penulisan yang tidak standar. Proses ini umumnya dilakukan menggunakan pendekatan pemetaan leksikon (*lexicon mapping*), yaitu mencocokkan kata dengan kamus normalisas. [9] Normalisasi diterapkan setelah case folding agar proses pencocokan kata dapat berjalan optimal tanpa dipengaruhi perbedaan kapitalisasi.

4. Tokenisasi

Tokenisasi merupakan tahap pemisahan teks ke dalam bentuk token atau kata-kata individual sehingga memungkinkan analisis pada tingkat kata. Proses ini dilakukan setelah normalisasi guna menjamin keseragaman bentuk token. [6] tokenisasi berperan penting dalam menyiapkan data sebelum proses klasifikasi, karena hasil token inilah yang akan digunakan sebagai representasi fitur dalam model *machine learning*.

5. Stopword Removal

Stopword removal merupakan proses penghapusan kata-kata umum yang sering muncul namun memiliki kontribusi informasi yang rendah, seperti "dan", "yang", dan "di". Tahap ini bertujuan untuk memusatkan analisis pada kata-kata yang memiliki makna semantik lebih signifikan. [10] penghapusan *stopword* membantu meningkatkan performa analisis

sentimen dengan mengurangi dimensi fitur yang tidak relevan. Proses ini dilakukan setelah tokenisasi agar sistem dapat mencocokkan token dengan daftar *stopword* secara efektif.

6. Stemming

Stemming merupakan proses mengonversi kata ke bentuk dasarnya dengan menghilangkan berbagai imbuhan, seperti awalan, akhiran, dan sisipan. Dalam penelitian ini, stemming diterapkan untuk menyatukan beragam variasi kata yang memiliki akar makna yang sama. [6] menyatakan bahwa *stemming* dapat mengurangi kompleksitas fitur dan meningkatkan akurasi model klasifikasi karena sistem hanya memproses kata dasar sebagai representasi fitur.

2.6 Algoritma *Support Vector Machine* (SVM)

Support Vector Machine (SVM) diterapkan dalam penelitian ini karena kemampuannya mengoptimalkan margin pemisah antar kelas, sehingga dapat memberikan performa generalisasi yang stabil pada data teks dengan dimensi tinggi. SVM bekerja dengan memetakan data ke ruang fitur dan menentukan fungsi keputusan yang dinyatakan dengan persamaan $f(x) = w \cdot x + b$, dimana x merupakan vektor fitur teks, w adalah bobot yang merepresentasikan kontribusi setiap fitur, dan b adalah bias yang mengatur posisi *hyperplane* pemisah. Dalam konteks analisis sentimen tweet, fungsi keputusan ini digunakan untuk menentukan kelas sentimen berdasarkan representasi fitur teks, seperti TF-IDF atau n-gram. Sejumlah penelitian menunjukkan bahwa SVM memiliki performa yang stabil dan efisien dalam klasifikasi sentimen berbasis teks.

2.7 Algoritma *K-Nearest Neighbors* (KNN)

K-Nearest Neighbors (KNN) merupakan metode *supervised learning* non-parametrik yang menentukan kelas suatu data berdasarkan kesamaan dengan data latih berlabel, di mana keputusan klasifikasi diperoleh dari mayoritas kelas k tetangga terdekat yang dihitung menggunakan metrik jarak, termasuk jarak *Euclidean*. Secara matematis, jarak *Euclidean* antara dua vektor fitur x dan y dirumuskan $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ Algoritma ini berasumsi bahwa data dengan karakteristik serupa cenderung berada pada kelas yang sama. Nilai k menjadi parameter penting dalam KNN karena pemilihan nilai yang terlalu rendah dapat memicu pengaruh *noise*, sedangkan nilai yang terlalu tinggi berisiko menurunkan performa klasifikasi. Dalam analisis sentimen, KNN umumnya digunakan pada dataset berukuran kecil hingga menengah dan sering dijadikan metode pembanding karena konsepnya yang sederhana.

2.8 Algoritma *Naïve Bayes*

Naïve Bayes mengklasifikasikan teks dengan menghitung probabilitas kelas menggunakan *Teorema Bayes* dan asumsi independensi antar fitur, sehingga setiap kata dianalisis secara terpisah. Pendekatan ini efektif dalam klasifikasi dokumen karena memanfaatkan distribusi frekuensi kata pada data pelatihan dengan proses perhitungan yang ringan Dalam penerapan analisis sentimen, *Naïve Bayes* mampu mengelompokkan opini publik ke dalam kategori sentimen utama, terutama pada dataset berskala kecil hingga menengah. Walaupun asumsi independensi antar fitur sering kali tidak sepenuhnya sesuai dengan bahasa alami, berbagai penelitian tetap menunjukkan kinerja algoritma ini yang kompetitif. Hubungan probabilistik antara kelas C dan fitur X dinyatakan dalam suatu persamaan $P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$ di mana $P(C|X)$ merupakan probabilitas dokumen termasuk ke dalam kelas C, $P(X|C)$ adalah probabilitas kemunculan fitur pada kelas tertentu, $P(C)$ menyatakan probabilitas awal kelas, dan $P(X)$ merupakan probabilitas keseluruhan fitur.

2.9 Studi Komparatif

Studi komparatif digunakan untuk membandingkan kinerja beberapa algoritma klasifikasi dalam menyelesaikan permasalahan

analisis sentimen, dengan tujuan menentukan metode yang paling akurat dan sesuai dengan karakteristik dataset. Penelitian ini, algoritma yang dibandingkan *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes*, yang masing-masing merepresentasikan pendekatan berbasis margin, kedekatan jarak, dan probabilitas. Kinerja masing-masing algoritma dibandingkan menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score* yang dihitung dari confusion *matrix*, sehingga dapat memberikan gambaran objektif mengenai ketepatan klasifikasi sentiment [9]. Pendekatan komparatif ini penting karena setiap algoritma memiliki keunggulan dan keterbatasan tersendiri, sehingga hasil perbandingan dapat menjadi dasar dalam menentukan algoritma yang paling efektif untuk analisis sentimen opini publik pada media sosial.

2.10 Media Sosial X (Twitter)

Media sosial X (Twitter) sebagai sumber data utama dalam penelitian ini karena mampu merepresentasikan opini publik secara *real-time*, terbuka, dan dinamis terhadap isu sosial maupun kebijakan pemerintah. Platform ini memungkinkan pengguna menyampaikan pesan singkat dalam bentuk tweet yang bersifat interaktif dan mudah diakses, sehingga efektif dalam menangkap persepsi serta respons masyarakat terhadap suatu peristiwa. [1]. Melalui Twitter API, data mentah seperti teks tweet, waktu unggahan, dan metadata pengguna dapat dikumpulkan secara sistematis untuk keperluan analisis sentimen. Tweet, yang dibatasi hingga 280 karakter, mengandung ekspresi emosional dan pandangan spontan pengguna, sehingga memiliki nilai informatif yang tinggi dalam kajian opini publik digital [11] Sejumlah penelitian menunjukkan bahwa pemanfaatan data Twitter dalam analisis sentimen sangat bergantung pada karakteristik data dan metode klasifikasi yang digunakan, di mana pemilihan algoritma yang tepat berpengaruh langsung terhadap akurasi dan kualitas hasil analisis [9].

2.11 *Tools*

Penelitian analisis sentimen ini memanfaatkan beberapa perangkat lunak pendukung untuk menunjang proses pengumpulan, pengolahan, dan analisis data secara efisien. Bahasa pemrograman Python digunakan sebagai alat utama karena didukung oleh pustaka *Natural Language Processing* dan *machine learning* seperti *NLTK*, *Sastrawi*, dan *scikit-learn* yang berperan dalam tahapan pra-pemrosesan teks, pemodelan klasifikasi, serta evaluasi kinerja algoritma. Proses komputasi dilakukan menggunakan Google Colab sebagai lingkungan berbasis *cloud* yang memungkinkan eksekusi kode Python tanpa instalasi lokal serta mendukung pengolahan data berskala besar secara efisien. Sementara itu, pengumpulan data tweet dilakukan menggunakan Tweet Harvest yang memanfaatkan Twitter API untuk mengekstraksi data berdasarkan kata kunci tertentu dan menyimpannya dalam format terstruktur. Kombinasi penggunaan ketiga tools tersebut memungkinkan proses analisis sentimen dilakukan secara sistematis, efisien, dan dapat direplikasi sesuai kebutuhan penelitian.

2.12 Keranga Pemikiran

Penelitian ini disusun untuk menganalisis opini publik terhadap Program Makan Bergizi Gratis melalui studi komparatif algoritma *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes* menggunakan data tweet dari platform X (Twitter). Proses penelitian diawali dengan pengumpulan data tweet menggunakan teknik crawling sebagai sumber data utama, kemudian dilakukan tahap pra-pemrosesan untuk membersihkan dan menstandarkan teks agar siap dianalisis. Data yang telah diproses selanjutnya diklasifikasikan menggunakan ketiga algoritma pembelajaran mesin tersebut, lalu dievaluasi berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil evaluasi digunakan untuk menentukan algoritma dengan performa terbaik dan paling konsisten dalam mengklasifikasikan sentimen tweet terkait Program Makan Bergizi Gratis, sehingga dapat menjadi dasar pemetaan opini publik secara objektif dan sistematis.

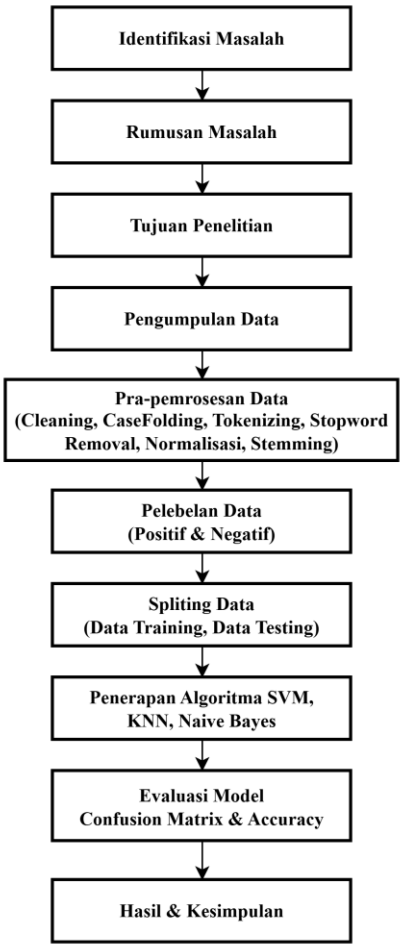
Gambar 3. Kerangka Pemikiran

III. Metodologi

Data penelitian diperoleh dari tweet pada platform X (Twitter) yang berkaitan dengan Program Makan Bergizi Gratis.. Data yang diperoleh kemudian melalui tahap pra-pemrosesan meliputi *cleaning*, *case folding*, *tokenizing*, *stopword removal*, normalisasi, dan stemming untuk menghasilkan data teks yang bersih dan siap diolah. Selanjutnya, data diberi label sentimen positif dan negatif, kemudian dibagi menjadi data latih dan data uji. Proses klasifikasi dilakukan menggunakan algoritma *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan Naïve Bayes. Evaluasi kinerja model dilakukan menggunakan *confusion matrix* dengan fokus utama pada metrik akurasi untuk menentukan algoritma dengan performa terbaik dalam mengklasifikasikan sentimen tweet.

3.1 Desain Penelitrain

Desain penelitian ini memiliki beberapa tahapan yaitu, (1) identifikasi masalah terkait opini publik terhadap Program Makan Bergizi Gratis di platform X (Twitter), (2) perumusan masalah penelitian, (3) penentuan tujuan penelitian, (4) pengumpulan data berupa tweet menggunakan teknik *crawling*, (5) pra-pemrosesan data yang meliputi *cleaning, case folding*, tokenisasi, *stopword removal*, normalisasi, dan *stemming*, (6) pelabelan data sentimen ke dalam kategori positif dan negatif, (7) pembagian data menjadi data latih dan data uji, (8) penerapan algoritma *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes* untuk proses klasifikasi, serta (9) evaluasi hasil klasifikasi menggunakan metrik akurasi untuk menentukan algoritma dengan kinerja terbaik dan menarik kesimpulan penelitian.



Gambar 4. Desain Penelitian

3.2 Metode Pengumpulan Data

Proses pengumpulan data pada penelitian menggunakan Tweet Harvest sebagai alat bantu scraping data dari platform X (Twitter). Pemilihan Tweet Harvest didasarkan pada kemampuannya dalam mengekstraksi tweet secara otomatis berdasarkan kata kunci, rentang waktu unggahan, bahasa, serta jumlah data yang dapat disesuaikan dengan kebutuhan penelitian. Data yang dikumpulkan berupa tweet publik yang membahas Program Makan Bergizi Gratis, baik yang menunjukkan sentimen positif maupun negatif terhadap kebijakan tersebut. Tweet dikumpulkan melalui kata kunci relevan dan disimpan dalam format .CSV untuk keperluan pra-pemrosesan dan analisis sentimen.

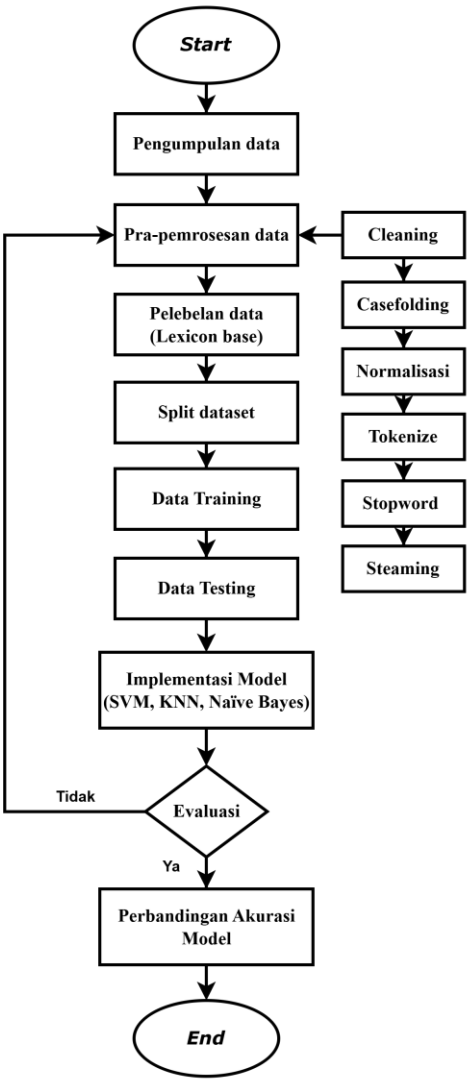
Table 5. Hasil Pengumpulan data

No	Bulan	Jumlah Tweet	Format File
1	Agustus	798	.CSV
2	September	3323	.CSV
3	Oktober	1843	.CSV
	Total	5964	.CSV

Berdasarkan Tabel 5. total data yang berhasil dikumpulkan selama periode Agustus hingga Oktober 2025 sebanyak 5.964 tweet berbahasa Indonesia, yang selanjutnya digunakan sebagai dataset pada tahap pra-pemrosesan dan analisis sentimen.

3.3 Metode Perancangan

Metode perancangan dalam penelitian ini meliputi pengumpulan data tweet dari platform X (Twitter) kemudian diproses melalui tahapan pra-pemrosesan teks. Data yang telah diproses selanjutnya diberi label sentimen positif dan negatif, lalu dibagi menjadi data latih dan data uji. Tahap berikutnya adalah penerapan algoritma *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes* untuk klasifikasi sentimen. Kinerja masing-masing algoritma kemudian dievaluasi berdasarkan nilai akurasi untuk menentukan metode yang paling efektif.



Gambar 6. Alur Proses Analisis Sentimen

3.3.1 Penumpulan Data

Data diperoleh dari platform X (Twitter) dengan melakukan scraping menggunakan Tweet Harvest. Pengambilan data dilakukan berdasarkan kata kunci relevan seperti "Makan Bergizi Gratis" dalam periode waktu tertentu agar mencerminkan opini publik yang aktual. Tweet yang berhasil dikumpulkan disimpan dalam format CSV sebagai data teks mentah untuk diproses lebih lanjut.

3.3.2 Pra-pemrosesan Data

Data teks yang terkumpul diproses melalui tahapan pra-pemrosesan untuk meningkatkan kualitas dan konsistensi data. Tahapan ini meliputi *cleaning* untuk menghilangkan elemen tidak relevan, *case folding* untuk menyeragamkan huruf, normalisasi untuk mengubah kata tidak baku menjadi bentuk standar, *tokenisasi* untuk memecah teks menjadi kata, *stopword removal* untuk menghapus kata umum yang tidak bermakna, serta *stemming* untuk mengembalikan kata ke bentuk dasarnya. Hasil pra-pemrosesan menghasilkan data teks yang siap dianalisis.

3.3.3 Pelebelan Data

Pendekatan *lexicon-based* digunakan dalam pelabelan sentimen dengan memanfaatkan kamus sentimen bahasa Indonesia yang terdiri atas kosakata positif dan negatif dari InSet. Setiap teks hasil pra-pemrosesan diberi skor sentimen berdasarkan selisih jumlah kata positif dan negatif. Skor tersebut digunakan untuk menentukan label sentimen, yaitu positif atau negatif, yang selanjutnya berfungsi sebagai *ground truth* pada tahap pelatihan dan pengujian model.

3.3.4 Split Dataset

Dataset yang telah berlabel dipisahkan menjadi data latih dan data uji menggunakan fungsi *train_test_split* dari pustaka *scikit-learn* dengan proporsi 80% data latih dan 20% data uji. Parameter *stratify* diterapkan guna menjaga keseimbangan distribusi kelas, sedangkan *random_state* digunakan untuk memastikan hasil pemisahan konsisten dan dapat direproduksi. Sebelum pemisahan, data yang tidak valid dihapus untuk menjaga kualitas dataset.

3.3.5 Implementasi Model

Data teks direpresentasikan ke dalam bentuk vektor numerik menggunakan *CountVectorizer*. Representasi ini digunakan sebagai masukan bagi tiga algoritma klasifikasi, *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes*, yang diimplementasikan menggunakan pustaka *scikit-learn*. SVM diterapkan dengan kernel linear, KNN menggunakan pendekatan kedekatan antar data, sedangkan *Naïve Bayes* menggunakan pendekatan probabilistik berbasis frekuensi kata. Ketiga model dilatih menggunakan data latih dan diuji pada data uji.

3.3.6 Evaluasi

Tahapan evaluasi dilakukan melalui perbandingan antara hasil prediksi model dan label sentimen pada data uji yang telah ditetapkan sebelumnya. Proses ini bertujuan untuk menilai tingkat ketepatan masing-masing model dalam mengklasifikasikan sentimen tweet. Evaluasi kinerja didasarkan pada hasil *confusion matrix*, yang digunakan sebagai dasar perhitungan nilai akurasi. Dalam penelitian ini, akurasi dipilih sebagai metrik utama karena selaras dengan tujuan penelitian, yaitu menentukan algoritma dengan tingkat prediksi paling tepat.

Nilai akurasi dihitung berdasarkan proporsi prediksi yang sesuai dengan label sebenarnya pada data uji. Hasil perhitungan akurasi dari setiap algoritma kemudian digunakan sebagai indikator utama untuk menilai kinerja model dan menjadi dasar dalam proses perbandingan efektivitas klasifikasi sentimen tweet terkait Program Makan Bergizi Gratis.

3.3.7 Perbandingan Akurasi Model

Perbandingan akurasi dilakukan terhadap tiga algoritma klasifikasi yang digunakan, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Naïve Bayes*, untuk mengidentifikasi metode dengan tingkat ketepatan tertinggi dalam memprediksi sentimen pada data uji. Nilai akurasi diperoleh dari hasil evaluasi masing-masing model dan dihitung berdasarkan proporsi jumlah prediksi yang benar terhadap keseluruhan data uji. Pendekatan yang berfokus pada akurasi memberikan gambaran yang sederhana dan langsung mengenai efektivitas setiap algoritma dalam menangani data teks media sosial. Model dengan nilai akurasi tertinggi ditetapkan sebagai

metode paling efektif dalam penelitian ini. Untuk mendukung interpretasi hasil secara sistematis dan objektif, nilai akurasi dikelompokkan ke dalam beberapa kategori performa, yaitu kurang baik (<70%), cukup baik (70–79%), baik (≥80%), dan sangat baik (≥90%), sebagaimana umum digunakan dalam penelitian analisis sentimen dan *Natural Language Processing*.

IV. Pembahasan

4.1 Metode Perancangan

Penelitian ini berhasil melakukan analisis sentimen terhadap tweet di platform X (Twitter) yang membahas Program Makan Bergizi Gratis menggunakan tiga algoritma klasifikasi, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes*. Berdasarkan hasil evaluasi menggunakan metrik akurasi, diperoleh perbedaan tingkat kinerja pada masing-masing algoritma, sehingga dapat diketahui model yang paling efektif dalam mengklasifikasikan sentimen opini publik terhadap kebijakan tersebut.

4.1.1 Hasil Pengumpulan Data

Data penelitian diperoleh melalui proses pengumpulan tweet memanfaatkan tools Tweet Harvest dari platform X (Twitter) pada periode Agustus hingga Oktober 2025. Berdasarkan hasil crawling awal, terkumpul sebanyak 5.966 tweet yang berkaitan dengan topik penelitian. Setelah dilakukan seleksi data untuk menghilangkan tweet duplikat, tidak relevan, dan tidak bermakna, jumlah data akhir yang digunakan sebanyak 5.420 tweet. Seluruh data disimpan dalam bentuk dataframe dan tidak mengandung nilai kosong (*non-null*), sehingga dinyatakan layak dan siap digunakan pada tahap pra-pemrosesan dan analisis sentimen selanjutnya.

```
1 df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 5420 entries, 0 to 5965
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   full_text  5420 non-null   object
dtypes: object(1)
memory usage: 84.7+ KB
```

Gambar 7. Hasil Pengumpulan Data

4.1.2 Hasil Pra-pemrosesan Data

Pra-pemrosesan data bertujuan menyiapkan teks agar layak dianalisis melalui tahapan *cleaning*, *case folding*, normalisasi, tokenisasi, *stopword removal*, dan *stemming*.

1. Cleaning

Table 1. Hasil *Cleaning*

Tweet	Hasil *Cleaning*
Makan bergizi gratis menciptakan generasi emas di masa yang akan datang ayo bersama sama gotong royong merealisasikan program ini #MakanBergiziGratis https://t.co/yQDzu7N2ww	Makan bergizi gratis menciptakan generasi emas di masa yang akan datang ayo bersama sama gotong royong merealisasikan program ini MakanBergiziGratis
@grok @Bunga62496 @RickyKardjono MBG itu makan bergizi gratis atau makan beracun gratis sih udah banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung	MBG itu makan bergizi gratis atau makan beracun gratis sih udah banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung

Tabel ini menunjukkan proses penghapusan tanda baca, angka, URL, mention, hashtag, serta simbol lain yang tidak memiliki kontribusi terhadap analisis sentimen, sehingga data yang dihasilkan menjadi lebih bersih dan terfokus pada informasi utama yang relevan.

2. *Case Folding*

Table 2. Hasil *Case Folding*

Hasil *Cleaning*	Hasil *Case folding*
Makan bergizi gratis menciptakan generasi emas di masa yang akan datang ayo bersama sama gotong royong merealisasikan program ini MakanBergiziGratis	makan bergizi gratis menciptakan generasi emas di masa yang akan datang ayo bersama sama gotong royong merealisasikan program ini makanbergizigratis
MBG itu makan bergizi gratis atau makan beracun gratis sih udah banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung	mbg itu makan bergizi gratis atau makan beracun gratis sih udah banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung

Tabel ini ditunjukkan bahwa proses *case folding* dilakukan dengan mengonversi seluruh teks menjadi huruf kecil (*lowercase*) untuk menghindari perbedaan makna akibat variasi penggunaan huruf kapital dalam data teks.

3. Normalisasi

Table 3. Hasil Noramalisasi

Hasil Case folding	Hasil Normalisasi
k inilah yg disebut makan bergizi gratis semoga gak bosen bacanya karena adegannya very slow burn	**ke** inilah yang disebut makan bergizi gratis semoga **tidak bosan** bacanya karena adegannya very slow burn
mbg itu makan bergizi gratis atau makan beracun gratis sih udah banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung	mbak itu makan bergizi gratis atau makan beracun gratis sih **sudah** banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung

Tabel ini menunjukkan proses normalisasi teks, yaitu perubahan kata-kata tidak baku, singkatan, serta ungkapan bahasa informal menjadi bentuk kata baku sesuai kaidah Bahasa Indonesia. Transformasi dilakukan dengan mengganti kata seperti “gak” menjadi “tidak”, “udah” menjadi “sudah”, serta penyesuaian istilah lain yang tidak sesuai dengan kosakata standar.

4. Tokenizing

Table 4. Hasil Tokenizing

Hasil Normalisasi	Hasil Tokenizing
ke inilah yang disebut makan bergizi gratis semoga **tidak bosan** bacanya karena adegannya very slow burn	['ke', 'inilah', 'yang', 'disebut', 'makan', 'bergizi', 'gratis', 'semoga', 'tidak', 'bosan', 'bacanya', 'karena', 'adegannya', 'very', 'slow', 'burn']
mbak itu makan bergizi gratis atau makan beracun gratis sih **sudah** banyak kasus keracunan di berbagai tempat bahkan makanan tak layak makan sampai berbelatung	['mbak', 'itu', 'makan', 'bergizi', 'gratis', 'atau', 'makan', 'beracun', 'gratis', 'sih', 'sudah', 'banyak', 'kasus', 'keracunan', 'di', 'berbagai', 'tempat', 'bahkan', 'makanan', 'tak', 'layak', 'makan', 'sampai', 'berbelatung']

Tabel ini menunjukkan hasil tokenisasi, yaitu proses pemecahan teks menjadi unit-unit kata (*token*). Tokenisasi bertujuan mengubah kalimat menjadi kumpulan kata yang terstruktur sehingga dapat diproses secara efektif oleh algoritma klasifikasi sentimen. Hasil tokenisasi berupa daftar kata dari setiap teks yang merepresentasikan isi tweet secara lebih terperinci.

5. *Stopword Removal*

Table 5. Hasil *Stopword Removal*

Hasil *Tokenizing*	Hasil *Stopword Removal*
['ke', 'inilah', 'yang', 'disebut', 'makan', 'bergizi', 'gratis', 'semoga', 'tidak', 'bosan', 'bacanya', 'karena', 'adegannya', 'very', 'slow', 'burn']	['makan', 'bergizi', 'gratis', 'semoga', 'bosan', 'bacanya', 'adegannya', 'very', 'slow', 'burn']
['mbak', 'itu', 'makan', 'bergizi', 'gratis', 'atau', 'makan', 'beracun', 'gratis', 'sih', 'sudah', 'banyak', 'kasus', 'keracunan', 'di', 'berbagai', 'tempat', 'bahkan', 'makanan', 'tak', 'layak', 'makan', 'sampai', 'berbelatung']	['mbak', 'makan', 'bergizi', 'gratis', 'makan', 'beracun', 'gratis', 'sih', 'keracunan', 'makanan', 'layak', 'makan', 'berbelatung']

Tabel ini menunjukkan proses *stopword removal* yang bertujuan menghapus kata-kata umum yang tidak berkontribusi signifikan terhadap penentuan sentimen, seperti kata hubung, kata depan, dan kata ganti. Melalui proses ini, teks menjadi lebih ringkas dan hanya mempertahankan kata-kata bermakna yang relevan untuk analisis sentimen.

6. Steaming

Table 6. Hasil Steaming

Hasil *Stopword Removal*	Hasil *Stemming*
['makan', 'bergizi', 'gratis', 'semoga', 'bosan', 'bacanya', 'adegannya', 'very', 'slow', 'burn']	makan gizi gratis moga bosan baca adegan very slow burn
['mbak', 'makan', 'bergizi', 'gratis', 'makan', 'beracun', 'gratis', 'sih', 'keracunan', 'makanan', 'layak', 'makan', 'berbelatung']	mbak makan gizi gratis makan racun gratis sih racun makan layak makan belatung

Tabel ini menunjukkan proses *stemming*, yaitu tahap pengubahan kata menjadi bentuk dasar atau kata akar. Bertujuan menyatukan variasi kata yang memiliki makna sama sehingga mengurangi kompleksitas data dan membantu algoritma mengenali pola sentimen secara lebih efektif. Hasil stemming berupa kata dasar yang selanjutnya digunakan pada tahap pelabelan dan klasifikasi sentimen.

4.1.3 Analisi Data

Analisis data teks dilakukan untuk mengamati karakteristik dataset sebelum dan sesudah tahap pra-pemrosesan. Analisis ini bertujuan mengidentifikasi perubahan pola kata, distribusi frekuensi, serta kata-kata dominan yang muncul dalam data. Metode yang digunakan meliputi visualisasi *wordcloud*, analisis frekuensi kata, dan pengamatan n-gram untuk memperoleh gambaran umum kualitas data teks.

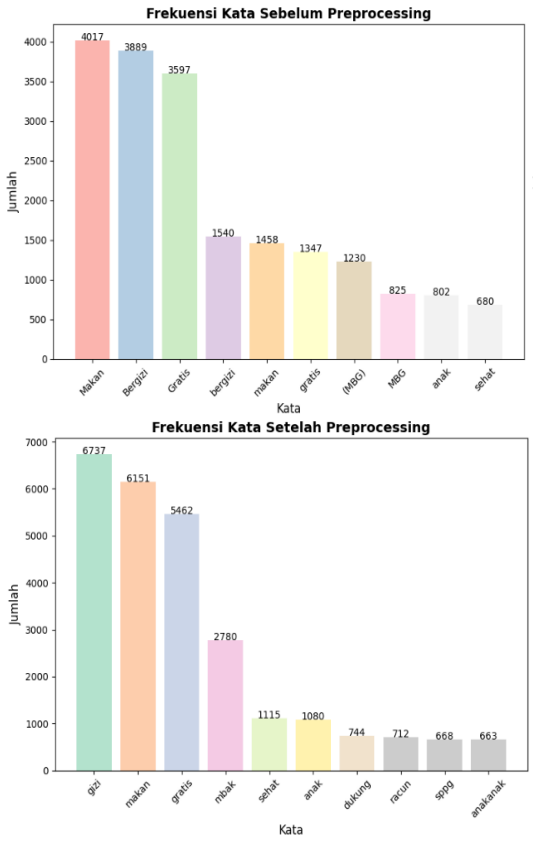
1. *Wordcloud* Sebelum dan Sesudah Pra-Pemrosesan

Gambar 8. *Wordcloud* Sebelum dan Sesudah Pra-Pemrosesan Data

Gambar 8. menampilkan perbandingan *wordcloud* sebelum dan sesudah pra-pemrosesan data, di mana pada kondisi awal masih terlihat dominasi kata-kata seperti “bergizi”, “gratis”, “makan”, dan “anak” yang disertai variasi penulisan, singkatan, perbedaan kapitalisasi, serta kata tidak relevan yang menunjukkan adanya *noise* pada data mentah. Setelah pra-pemrosesan, *wordcloud* menjadi lebih bersih dan

terfokus dengan kata-kata dominan yang telah seragam, seperti "gizi", "gratis", "makan", "anak", dan "layanan", yang menandakan keberhasilan proses pembersihan, normalisasi, dan stemming dalam meningkatkan kualitas teks sehingga data lebih representatif terhadap konteks penelitian dan siap digunakan pada tahap analisis sentimen berikutnya.

2. Frekuensi Kata Sebelum dan Sesudah Pra-Pemrosesan Data

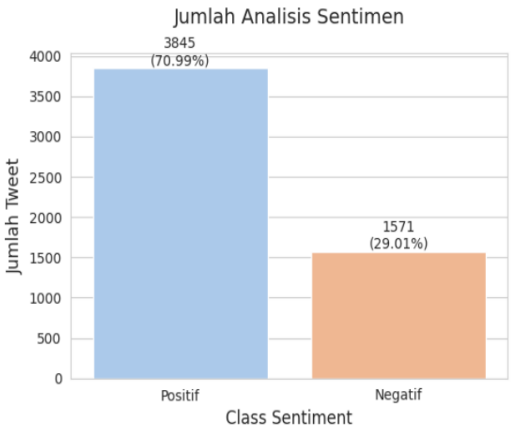


Gambar 9. Frekuensi Kata Sebelum dan Sesudah Pra-Pemrosesan

Analisis frekuensi kata pada Gambar 9. menunjukkan perbedaan yang jelas antara kondisi data sebelum dan sesudah pra-pemrosesan. Pada data awal, kata-kata seperti "makan", "bergizi", dan "gratis" muncul dengan frekuensi tinggi, namun masih terdapat variasi penulisan akibat perbedaan kapitalisasi, penggunaan singkatan seperti "MBG", serta kemunculan kata yang kurang relevan sehingga menimbulkan redundansi dan noise linguistik. Setelah pra-pemrosesan diterapkan, distribusi kata menjadi lebih seragam dan terfokus, dengan kata "gizi", "makan", dan "gratis" mendominasi secara konsisten, disertai kemunculan istilah yang lebih kontekstual terhadap topik penelitian. Perubahan ini menunjukkan bahwa pra-pemrosesan berhasil meningkatkan kualitas representasi teks, sehingga data menjadi lebih relevan, bersih, dan siap digunakan pada tahap analisis lanjutan seperti ekstraksi fitur dan klasifikasi sentimen.

4.1.4 Hasil Pelebelan Data

Metode *lexicon-based* diterapkan dalam pelabelan data dengan memanfaatkan kamus sentimen untuk menetapkan skor positif dan negatif pada setiap kata dalam teks yang telah dipra-pemrosesan. Skor total dari setiap teks kemudian digunakan sebagai dasar penentuan label sentimen, di mana teks dengan skor lebih dari nol dikategorikan sebagai sentimen positif, sedangkan teks dengan skor nol atau kurang diklasifikasikan sebagai sentimen negatif.



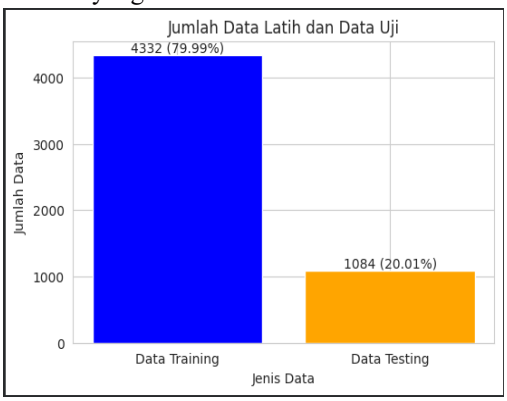
Gambar 10. Distribusi Sentimen Positif dan Negatif Hasil Pelebelan Data

Berdasarkan hasil pelabelan, setiap tweet berhasil diberi label sentimen sesuai nilai skor yang diperoleh. Distribusi hasil pelabelan selanjutnya divisualisasikan dalam bentuk diagram batang sebagaimana ditunjukkan pada

Gambar 10. Hasil visualisasi menunjukkan bahwa sentimen positif mendominasi dengan jumlah 3.845 tweet atau sebesar 70,99%, sedangkan sentimen negatif berjumlah 1.571 tweet atau sebesar 29,01%. Dominasi sentimen positif ini mengindikasikan bahwa mayoritas opini publik di platform X memberikan respons yang mendukung terhadap Program Makan Bergizi Gratis, meskipun masih terdapat sejumlah respons negatif yang mencerminkan kritik atau ketidakpuasan terhadap kebijakan tersebut.

4.1.5 Hasil Spliting Data

Tahap splitting data dilakukan untuk memisahkan dataset hasil pelabelan sentimen menjadi data latih dan data uji agar proses pelatihan dan evaluasi model dapat dilakukan secara objektif tanpa terjadinya *data leakage*. Pemisahan data menggunakan fungsi *train_test_split* dari pustaka *scikit-learn* setelah terlebih dahulu menghapus data yang memiliki nilai kosong pada kolom *stemming_data*. Dataset akhir berjumlah 5.416 data dengan label sentimen yang valid.



Gambar 11. Pembagian *Data Training* dan *Testing*

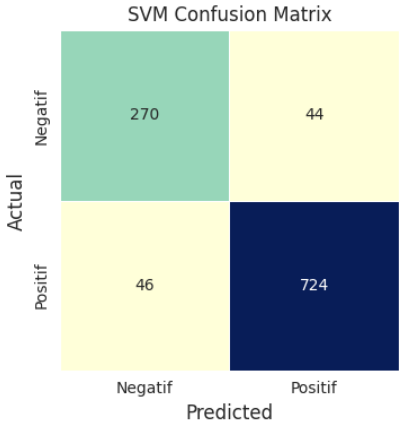
Pada proses ini, data teks hasil stemming digunakan sebagai variabel independen (X), sedangkan label sentimen positif dan negatif digunakan sebagai variabel dependen (y). Pembagian data dilakukan dengan proporsi 80% sebagai data latih dan 20% sebagai data uji, menggunakan *random_state = 42* untuk menjaga konsistensi hasil serta *stratify = y* guna memastikan distribusi kelas tetap seimbang. Hasil pembagian menunjukkan sebanyak 4.332 data digunakan sebagai data latih dan 1.084 data sebagai data uji, dengan representasi fitur hasil *CountVectorizer* berdimensi 5.565 kata unik. Distribusi pembagian data ditampilkan pada Gambar 11, yang menunjukkan bahwa mayoritas data dialokasikan untuk pelatihan model, sementara sebagian lainnya digunakan untuk pengujian performa secara objektif.

4.1.6 Hasil Implementasi Model

Pada tahap pengujian, model klasifikasi sentimen diuji menggunakan algoritma SVM, KNN, dan *Naïve Bayes*. Nilai akurasi digunakan sebagai dasar evaluasi kinerja untuk menentukan algoritma yang paling tepat dalam mengklasifikasikan sentimen terkait Program Makan Bergizi Gratis.

1. *Support Vector Machine* (SVM)

Model *Support Vector Machine* (SVM) diterapkan untuk mengklasifikasikan data sentimen ke dalam dua kelas dengan menentukan hyperplane optimal sebagai batas pemisah antar kelas. Hasil pengujian kinerja model SVM ditampilkan dalam bentuk *confusion matrix* sebagaimana ditunjukkan pada Gambar 12.



Gambar 12. *Confusion Matrix* Model SVM

Berdasarkan *confusion matrix* tersebut, diperoleh nilai evaluasi sebagai berikut:

Table 7. Nilai Confusion Matrix Model SVM

No	Kategori	Nilai
1.	*True Positive* (TP)	724
2.	*True Negative* (TN)	270
3.	*False Positive* (FP)	44
4.	*False Negative* (FN)	46

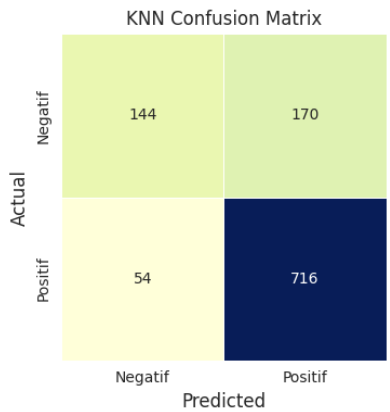
Nilai-nilai tersebut digunakan untuk menghitung tingkat akurasi model SVM dengan rumus berikut:

$$Accuracy = \frac{(724 + 270)}{(724 + 270 + 44 + 46)} = 0{,}916\ (91{,}6\%)$$

Hasil perhitungan menunjukkan bahwa model SVM memperoleh tingkat akurasi sebesar 91,6%, yang menandakan performa klasifikasi yang sangat baik. Tingginya nilai True Positive juga mengindikasikan bahwa model mampu mengenali tweet dengan sentimen positif secara efektif, serta memiliki tingkat kesalahan klasifikasi yang relatif rendah.

2. *K-Nearest Neighbor* (KNN)

Model *K-Nearest Neighbor* (KNN) melakukan klasifikasi sentimen berdasarkan prinsip kedekatan jarak antar data, di mana data uji akan dikelompokkan ke dalam kelas yang paling dominan di antara sejumlah tetangga terdekatnya. Pada penelitian ini, nilai k ditentukan melalui penyesuaian terhadap data latih untuk memperoleh hasil klasifikasi yang optimal. Hasil evaluasi model KNN dalam bentuk *Confusion Matrix* pada Gambar 13.



Gambar 13. *Confusion Matrix* Model KNN

Berdasarkan *Confusion Matrix* tersebut, diperoleh nilai sebagai berikut:

Table 8. Nilai Confusion Matrix Model KNN

No	Kategori	Nilai
1.	True Positive (TP)	716
2.	True Negative (TN)	144
3.	False Positive (FP)	170
4.	False Negative (FN)	54

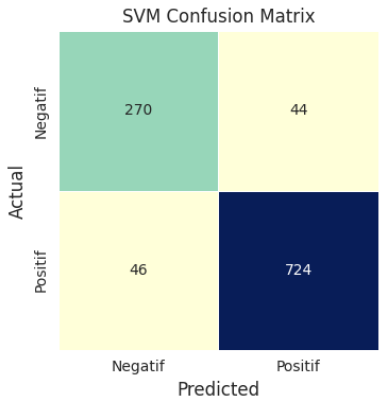
Perhitungan akurasi model KNN dilakukan menggunakan rumus berikut:

$$Accuracy = \frac{(716 + 144)}{(716 + 144 + 170 + 54)} = 0{,}791\ (79{,}1\%)$$

Hasil pengujian menunjukkan bahwa model KNN memperoleh nilai akurasi sebesar 79,1%, yang termasuk dalam kategori cukup baik. Meskipun demikian, performa KNN masih berada di bawah model SVM. Hal ini disebabkan oleh karakteristik data teks yang memiliki dimensi fitur tinggi, sehingga metode berbasis jarak seperti KNN cenderung kurang optimal dibandingkan metode berbasis margin dalam menangani klasifikasi sentimen.

3. Naïve Bayes

Model *Naïve Bayes* digunakan karena kemampuannya dalam mengklasifikasikan data teks secara efisien melalui pendekatan probabilistik dengan asumsi independensi antar fitur. Evaluasi performa model dilakukan menggunakan *confusion matrix* pada Gambar 14.



Gambar 14. *Confusion Matrix* Model Naïve Bayes

Berdasarkan confusion matrix tersebut, diperoleh nilai sebagai berikut:

Table 9. Nilai Confusion Matrix Model Naïve Bayes

No	Kategori	Nilai
1.	*True Positive* (TP)	704
2.	*True Negative* (TN)	156
3.	*False Positive* (FP)	158
4.	*False Negative* (FN)	66

Nilai akurasi model Naïve Bayes dihitung menggunakan persamaan berikut:

$$Accuracy = \frac{(704 + 156)}{(704 + 156 + 158 + 66)} = 0{,}796\ (79{,}6\%)$$

Hasil pengujian menunjukkan bahwa model *Naïve Bayes* memperoleh akurasi 79,6%, sedikit lebih tinggi dibandingkan KNN. Temuan ini mengindikasikan bahwa pendekatan probabilistik Naïve Bayes masih cukup efektif dalam melakukan klasifikasi sentimen teks dua kelas, meskipun performanya belum mampu melampaui model SVM.

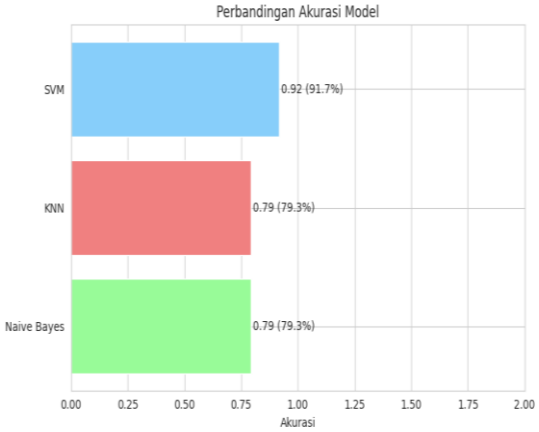
4.2 PembahasanMetode Perancangan

4.2.1 Perbandingan Kinerja Model

Perbandingan kinerja dilakukan terhadap tiga algoritma klasifikasi *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Naïve Bayes*, dengan tujuan menentukan model yang memiliki tingkat akurasi tertinggi dalam mengklasifikasikan sentimen positif dan negatif terhadap Program Makan Bergizi Gratis. Evaluasi kinerja difokuskan pada metrik akurasi yang diperoleh dari hasil *confusion matrix*, karena selaras dengan tujuan penelitian yang menitikberatkan pada ketepatan klasifikasi model.

Table 10 Perbandingan Nilai Confusion Matrik Model Klasifikasi Sentimen

Model	TP	TN	FP	FN	Akurasi (%)
SVM	724	270	44	46	91.7
KNN	716	144	170	54	79.3
Naïve Bayes	704	156	158	66	79.6



Gambar 15. Perbandingan Akurasi Model Klasifikasi Sentimen

Hasil pengujian menunjukkan bahwa algoritma SVM menghasilkan akurasi tertinggi sebesar 91,7%, diikuti oleh *Naïve Bayes* sebesar 79,6%, dan KNN sebesar 79,3%. Perbandingan akurasi ketiga model tersebut divisualisasikan pada Gambar 15. Temuan ini mengindikasikan bahwa SVM memiliki kemampuan terbaik dalam memisahkan kelas sentimen pada data teks yang digunakan, khususnya karena efektivitasnya dalam menangani fitur berdimensi tinggi hasil proses vektorisasi.

4.2.2 Analisis Performa Algoritma

Analisis performa dilakukan untuk menjelaskan perbedaan tingkat akurasi yang dihasilkan oleh masing-masing algoritma serta faktor-faktor yang memengaruhinya.

1. Algoritma *Support Vector Machine* (SVM) menunjukkan performa paling optimal karena menerapkan prinsip *maximum-margin classifier*, yang memungkinkan pemisahan kelas secara lebih tegas dan stabil. Pendekatan ini membuat SVM lebih tahan terhadap *overfitting* dan sangat efektif untuk data teks dengan jumlah fitur yang besar.
2. Algoritma *K-Nearest Neighbor* (KNN) memperoleh akurasi yang lebih rendah karena bekerja berdasarkan perhitungan

jarak antar data. Pada data teks berdimensi tinggi, fenomena curse of dimensionality menyebabkan jarak antar vektor menjadi kurang representatif, sehingga menurunkan kemampuan model dalam membedakan kelas sentimen.

3. Algoritma *Naïve Bayes* menggunakan pendekatan probabilistik dengan asumsi independensi antar fitur. Meskipun asumsi tidak sepenuhnya terpenuhi, *Naïve Bayes* tetap menunjukkan kinerja yang efektif dengan keunggulan pada efisiensi komputasi dan kecepatan klasifikasi.

Secara keseluruhan, hasil penelitian menunjukkan bahwa SVM merupakan algoritma yang paling akurat dan stabil dalam analisis sentimen berbasis teks, sedangkan *Naïve Bayes* dan KNN tetap memberikan performa yang cukup baik dengan keterbatasan masing-masing.

V. Kesimpulan

Penelitian ini menganalisis sentimen opini publik terkait Program Makan Bergizi Gratis menggunakan data tweet dari X (Twitter) berbasis *machine learning*. Proses penelitian meliputi pengumpulan data, pra-pemrosesan teks, pelabelan sentimen lexicon, serta klasifikasi sentimen menggunakan tiga algoritma, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Naïve Bayes*. Evaluasi kinerja model dilakukan dengan berfokus pada metrik akurasi untuk menilai tingkat ketepatan klasifikasi sentimen positif dan negatif.

Hasil pengujian menunjukkan bahwa algoritma SVM memberikan performa terbaik dengan tingkat akurasi sebesar 91,7%, lebih tinggi dibandingkan *Naïve Bayes* sebesar 79,6% dan KNN sebesar 79,3%. Keunggulan SVM disebabkan oleh kemampuannya dalam menangani data teks berdimensi tinggi hasil proses vektorisasi, serta kemampuannya membentuk hyperplane pemisah yang optimal antar kelas sentimen. Sementara itu, Naïve Bayes dan KNN masih menunjukkan performa yang cukup baik, namun memiliki keterbatasan akibat asumsi independensi fitur dan sensitivitas terhadap ruang fitur berdimensi tinggi.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa algoritma SVM merupakan metode yang paling efektif untuk analisis sentimen tweet terkait Program Makan Bergizi Gratis pada penelitian ini. Temuan ini diharapkan dapat menjadi acuan bagi penelitian selanjutnya dalam pemilihan algoritma klasifikasi sentimen berbasis teks, serta sebagai dasar pengembangan sistem pemantauan opini publik berbasis media sosial untuk mendukung evaluasi kebijakan pemerintah.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa, orang tua, dan keluarga atas dukungan yang diberikan, serta kepada Ibu Anggia Dasa Putri selaku dosen pembimbing atas bimbingan dan arahan selama proses penelitian.

Daftar Pustaka

[1] H. R. Budiana, A. Koswara, F. A. A. Prastowo, and E. Ratnasari, "Public Opinion Dynamics on Twitter: A Preliminary Analysis of Conversations Related to the 2024 General Election in Indonesia," *J. Law Sustain. Dev.*, vol. 12, no. 1, p. e2132, 2024, doi: 10.55908/sdgs.v12i1.2132.

[2] F. Fatkhurrohman, B. I. Nugroho, and N. Fadillah, "Analisis Sentimen Program Makan Bergizi Gratis Pemerintah RI Melalui Twitter Menggunakan Metode SVM," vol. 4, no. 3, pp. 3906–3917, 2025.

[3] R. R. S. Dwi Amelia, Pratomo Setiaji, "Perbandingan Algoritma SVM dan Naive Bayes dalam Klasifikasi Sentimen pada Ulasan Aplikasi Traveloka dan Agoda," vol. 11, no. 2, pp. 263–270, 2025.

[4] D. Prastyo, D. Irawan, and I. H. Mursyidin, "Klasifikasi Sentimen Komentar YouTube dengan NLP pada Debat Pilkada Banten 2024," *bit-Tech*, vol. 7, no. 2, pp. 413–421, 2024, doi: 10.32877/bt.v7i2.1833.

[5] K. Alahmadi, S. Alharbi, J. Chen, and X. Wang, "Generalizing sentiment analysis : a review of progress , challenges , and emerging directions," *Soc. Netw. Anal. Min.*, 2025, doi: 10.1007/s13278-025-01461-8.

[6] S. Mulya, H. Sujaini, and T. Tursina, "Analisis Sentimen Tren Olahraga di Masa Pandemi COVID-19 pada Twitter dengan Metode Naïve Bayes Classifier (NBC)," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 2, p. 284, 2022, doi: 10.26418/jp.v8i2.52815.

[7] A. M. Ndapamuri, D. Manongga, and A. Iriani, "Analisis Sentimen Ulasan Aplikasi Tripadvisor Dengan Metode Support Vector Machine, K-Nearest Neighbor, Dan Naive Bayes," *INOVTEK Polbeng - Seri Inform.*, vol. 8, no. 1, p. 127, 2023, doi: 10.35314/isi.v8i1.3260.

[8] Malikhatul Ibriza, Maya Rini Handayani, Wenty Dwi Yuniarti, and Khothibul Umam, "Modeling Political Discourse in Indonesia's 2024 Election Using Unsupervised Machine Learning," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 14, no. 2, pp. 174–182, 2025, doi: 10.32736/sisfokom.v14i2.2359.

[9] T. A. Putra and I. Permana, "Klasifikasi Penerima Bantuan Program Indonesia Pintar (PIP) Pada Siswa SMK Menggunakan Algoritma KNN, NBC dan C4.5," *Technol. Sci.*, vol. 6, no. 4, pp. 2131–2138, 2025, doi: 10.47065/bits.v6i4.6395.

[10] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1120–1126, 2021, doi: 10.29207/resti.v5i6.3588.

[11] J. Katalinić and I. Dunđer, "Neural Network-Based Sentiment Analysis and Anomaly Detection in Crisis-Related Tweets," *Electron.*, vol. 14, no. 11, 2025, doi: 10.3390/electronics14112273.