

BAB III METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Sebagai penelitian kuantitatif dengan pendekatan eksperimental (*experimental research*), penelitian ini bertujuan membandingkan kinerja dua algoritma *machine learning* *Naïve Bayes* dan *Support Vector Machine* (SVM) berfitur *Word N-gram* dan *Character N-gram* dalam mengklasifikasikan sentimen opini publik. Pendekatan eksperimental ini dipilih agar metrik evaluasi seperti akurasi, presisi, *recall*, dan *f1-score* dari kedua model dapat diukur secara objektif (Lestari & Hartono, 2024).

3.2 Fokus Penelitian

Objek yang dikaji dalam penelitian ini adalah opini masyarakat mengenai "Program Makan Bergizi Gratis (MBG)" yang bersumber dari dataset publik di Kaggle, dengan fokus pada penerapan teknik *Text Mining* dan *Natural Language Processing* (NLP) untuk analisis sentimen berbahasa Indonesia. Cakupannya dibatasi pada tahap pra-pemrosesan data teks, pembobotan kata dengan TF-IDF (*Term Frequency-Inverse Document Frequency*), serta klasifikasi sentimen ke dalam tiga kelas: Positif, Negatif, dan Netral.

3.3 Sumber Data

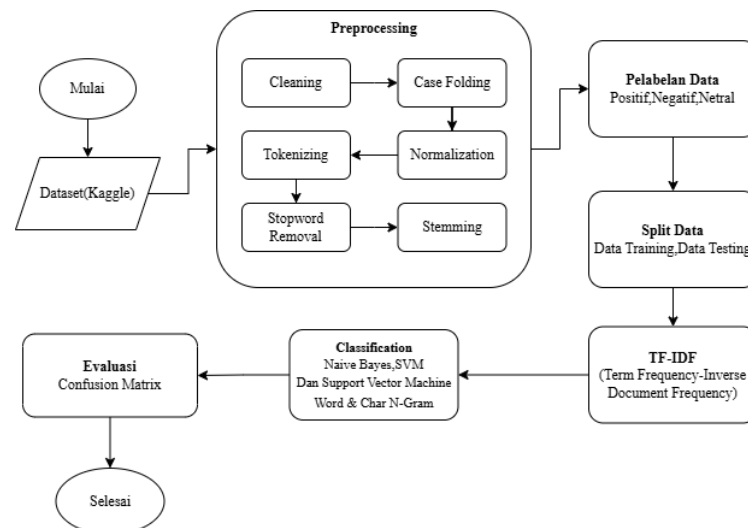
Data penelitian ini diperoleh dari Kaggle, sebuah komunitas *data science* dan *machine learning* yang menyediakan aneka dataset terbuka bagi penggunanya. Dua dataset dipakai sekaligus: pertama berjudul "program_makan_bergizi_gratis_10k data.csv" buatan Almuhammad (2025) yang berisi kumpulan *tweet* komentar masyarakat mengenai Program Makan Bergizi Gratis pada masa pemerintahan Presiden Prabowo Subianto, dan kedua berjudul Dataset.csv dari Ravhi Haris Wibowo (2025). Kedua dataset ini dipilih karena selaras dengan tujuan penelitian, yakni menganalisis sentimen terhadap Program Makan Bergizi Gratis menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM).

3.4 Alat dan Bahan Penelitian

Kelancaran implementasi sistem dan analisis data pada penelitian ini ditunjang oleh perangkat keras dan perangkat lunak dengan spesifikasi berikut:

1. Perangkat Keras (*Hardware*): laptop HP berprosesor Intel Core i3, RAM 12 GB, dan penyimpanan SSD 256 GB.
2. Perangkat Lunak (*Software*): bahasa pemrograman Python, IDE Visual Studio Code/Google Colab, serta pustaka pendukung seperti Pandas, Sastrawi, Scikit-learn, dan Seaborn.

3.5 Desain Sistem Penelitian



Gambar 3.1 Desain Sistem

3.5.1 Dataset Kaggle

Dataset penelitian ini merupakan data sekunder dari Kaggle yang mencakup dua sumber utama seputar opini masyarakat terhadap Program Makan Bergizi Gratis (MBG): kontribusi Almuhammad (2025) sejumlah 10.533 data, dan kontribusi Ravhi Haris Wibowo (2025) sebanyak 11.185 data. Setelah digabungkan, terbentuklah satu kesatuan data penelitian berjumlah 21.718 dokumen teks berisi komentar atau tweet masyarakat berbahasa Indonesia, yang mencerminkan beragam persepsi baik pro maupun kontra terhadap kebijakan tersebut. Pemilihan dataset dari Kaggle ini ditujukan untuk menjamin ketersediaan volume data yang

besar dan relevan, guna menguji keakuratan algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM).

3.5.2 Pra-pemrosesan Data

Karena data mentah dari media sosial umumnya tidak terstruktur dan sarat *noise*, tahapan *preprocessing* berikut perlu dilalui agar data siap diolah (Mauliza & Sipayung, 2024):

1. *Cleaning*, yaitu menyaring data dari elemen yang tidak relevan seperti URL, *mention*, *hashtag*, angka, tanda baca, dan emoji.
2. *Case Folding*, yang menyeragamkan seluruh huruf kapital menjadi huruf kecil (*lowercase*).
3. *Normalization (Slang Removal)*, yakni mengonversi kata tidak baku, singkatan, atau bahasa gaul menjadi kata baku sesuai KBBI. Mengingat maraknya bahasa tidak baku di media sosial, normalisasi kata ini sangat berpengaruh terhadap akurasi akhir (Cahyani & Abdillah, 2024).
4. *Tokenizing*, yaitu memecah kalimat menjadi satuan kata tunggal (*token*).
5. *Stopword Removal*, yang membuang kata hubung umum tanpa makna sentimen berarti.
6. *Stemming*, yaitu mengembalikan kata berimbuhan ke bentuk dasarnya. Optimalisasi teknik ini pada teks Bahasa Indonesia terbukti mampu mendongkrak performa model klasifikasi (Albab, Yuniar, & Fawaiq, 2023).

3.5.3 Pelabelan Data (*Data Labeling*)

Setelah data bersih, proses pelabelan sentimen dilanjutkan secara semi-otomatis menggunakan pendekatan leksikon (*lexicon-based*), yang kemudian diverifikasi manual oleh peneliti untuk menjaga validitasnya (Ryandi, Pratiwi, & Sari, 2025). Data pun dikelompokkan menjadi tiga kelas: Positif, Negatif, dan Netral.

3.5.4 Split Data

Setelah melalui tahap *text preprocessing* dan pembobotan kata, dataset gabungan dibagi menjadi dua bagian melalui proses *split data*, dengan rasio 80% untuk *training data* (data latih) dan 20% untuk *testing data* (data uji). Rasio 80:20 ini lazim dipakai dalam riset *machine learning* karena mampu memberi keseimbangan optimal antara pembelajaran pola dan validasi model (Sari & Nugroho, 2024).

1. Data Latih (80%): porsi ini dipakai model *machine learning* pada fase pelatihan, tempat algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) mempelajari pola kata serta karakteristik setiap kelas sentimen (Positif, Negatif, dan Netral).
2. Data Uji (20%): sisa porsi ini dipakai untuk menilai performa model yang telah dilatih, berperan sebagai data baru yang belum pernah dikenali model sebelumnya, sehingga metrik evaluasi akurasi, presisi, recall, dan f1-score dapat diperoleh secara objektif.

3.5.5 Pembobotan Kata (*Term Weighting*)

Setelah berlabel, data teks dikonversi menjadi representasi numerik memakai metode TF-IDF (*Term Frequency-Inverse Document Frequency*), yang memberi bobot lebih besar pada kata-kata ciri khas sentimen tertentu sekaligus menekan dominasi kata-kata umum sehingga model lebih mampu membedakan antar kelas sentimen (Siregar et al., 2023).

3.5.6 Ekstraksi Fitur N-gram

Sebagai representasi masukan bagi model SVM, ekstraksi fitur N-gram dilakukan setelah proses pembobotan tahapan yang tidak diberlakukan pada model *Naïve Bayes* karena tetap memakai representasi *bag-of-words* standar. Fungsi *TfidfVectorizer* dari pustaka *Scikit-learn* dipakai untuk menjalankan ekstraksi ini, dengan konfigurasi berikut (Iriananda et al., 2024):

1. *Word N-gram*: parameter `analyzer='word'` dipadukan `ngram_range=(1,3)`, guna menghasilkan kombinasi Unigram (1-gram), Bigram (2-gram), dan Trigram (3-gram) berbasis kata pendekatan yang menangkap makna kata tunggal sekaligus konteks frasa bermuatan sentimen.
2. *Character N-gram*: parameter `analyzer='char_wb'` dipadukan `ngram_range=(3,5)`, guna menghasilkan rangkaian karakter sepanjang 3 hingga 5 huruf pendekatan yang ampuh menangani kata tidak baku, singkatan, dan variasi penulisan informal yang lazim pada teks media sosial berbahasa Indonesia (Nugroho, 2021).
3. Penggabungan Fitur (*Feature Concatenation*): sebelum dimasukkan ke model SVM untuk pelatihan dan klasifikasi, matriks fitur Word N-gram dan Character N-gram digabungkan (*concatenated*) memakai fungsi `scipy.sparse.hstack` (Iriananda et al., 2022).

Melalui strategi tersebut, model SVM memperoleh ruang fitur yang kaya dan menyeluruh, memadukan informasi semantik tingkat kata dengan informasi morfologis tingkat karakter secara bersamaan.

3.5.7 Pembentukan Model Klasifikasi

Pada tahap ini, dataset dibagi menjadi Data Latih (80%) untuk melatih model dan Data Uji (20%) untuk pengujian rasio yang lazim dipakai demi menjaga keseimbangan antara pembelajaran pola dan validasi model (Sari & Nugroho, 2024).

1. Algoritma *Naïve Bayes*, yang dilatih memakai varian *Multinomial Naïve Bayes* karena unggul dalam kecepatan komputasi dan efektif untuk dataset teks berukuran besar (Pratama & Wibowo, 2024).
2. Algoritma SVM, yang dilatih dengan kernel Linear atau RBF; dipilih karena kestabilan serta akurasi kompetitifnya pada data berdimensi tinggi (Susanto, 2024).

3. Algoritma SVM dengan *Word & Char N-gram*, yang dilatih menggunakan kernel Linear dengan masukan berupa matriks fitur gabungan Word N-gram dan Character N-gram hasil tahap 3.5.6. Berbeda dari SVM standar, model ini memanfaatkan sekaligus informasi konteks frasa dari Word N-gram dan informasi morfologis sub-kata dari Character N-gram, sehingga diharapkan lebih peka membedakan nuansa sentimen pada teks informal berbahasa Indonesia yang penuh kata tidak baku dan singkatan (Iriananda et al., 2024).

3.5.8 Evaluasi dan Validasi Model

Kinerja kedua model diuji memakai Data Uji yang belum pernah dilihat sebelumnya, dan hasil prediksinya dibandingkan dengan label asli (*Ground Truth*) melalui tabel *Confusion Matrix*. Metrik yang dihitung mencakup Akurasi, Presisi, *Recall*, dan *F1-Score*, untuk menyimpulkan metode mana yang lebih unggul (Ramadhan, 2024).