BAB II
LANDASAN TEORI

1.4 Teori dan Konsep Analisis Sentimen

Analisis sentimen publik digunakan untuk memahami bagaimana masyarakat menanggapi kebijakan Program Makan Bergizi Gratis melalui opini dan reaksi yang dipublikasikan di media sosial X. Analisis ini juga membantu dalam mengevaluasi efektivitas komunikasi pemerintah, menilai tingkat kepuasan masyarakat, dan memberikan masukan yang berharga untuk perbaikan program di masa mendatang. Landasan teoretis penelitian ini terletak pada tiga dimensi utama: linguistik, psikologi sosial, dan kecerdasan buatan (*artificial intelligence*). Pendekatan linguistik, psikologi sosial, dan artificial intelligence berfungsi sebagai dasar ilmiah dalam menganalisis pola bahasa dan persepsi publik terhadap kebijakan Program Makan Bergizi Gratis. Ketiga disiplin tersebut berperan dalam membentuk pemahaman terhadap cara bahasa, emosi, dan interaksi sosial dikonversi menjadi data yang dapat diolah secara komputasional. Menurut Rodríguez-Ibáñez et al, analisis sentimen modern memanfaatkan teknik *natural language processing (NLP)* untuk memproses bahasa alami dan mengekstraksi makna yang terkandung di dalamnya(Cas, 2023). Dengan bantuan algoritma *machine learning*, sistem dapat mengidentifikasi polaritas opini (positif, negatif, atau netral) dari teks yang bersifat dinamis dan kompleks.

Perkembangan teknologi informasi telah mengubah pola komunikasi publik, menjadikan media sosial X sebagai data penting karena menampung respons spontan publik mengenai implementasi Program Makan Bergizi Gratis.

Fenomena ini menuntut penggunaan metode ilmiah yang mampu mengolah volume data besar (*big data*) secara efisien dan akurat. Analisis sentimen digunakan untuk menilai respons publik terhadap kebijakan seperti *Program Makan Bergizi Gratis* dengan mengukur distribusi opini masyarakat di media sosial. Hasilnya membantu memahami persepsi sosial terhadap program pemerintah serta arah dukungan atau kritik yang muncul di ruang digital. Analisis berbasis sentimen telah terbukti efektif dalam mendukung pengambilan keputusan kebijakan dengan mempertimbangkan aspirasi masyarakat(Ndubuisi & Olufunke, 2025).

2.1.1 Analisis Sentimen

Sentiment analysis atau analisis sentimen merupakan subbidang dari *natural language processing (NLP)* yang berfokus pada identifikasi dan klasifikasi emosi atau opini yang terkandung dalam teks. Analisis sentimen digunakan dalam penelitian ini untuk mengidentifikasi polaritas opini masyarakat terkait efektivitas Program Makan Bergizi Gratis. Pendekatan berbasis machine learning, terutama *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB), digunakan karena keduanya terbukti efektif menangani klasifikasi sentimen kebijakan publik. *Support Vector Machine* (SVM), dengan kemampuannya dalam menangani data non-linear, mampu memisahkan kategori sentimen dengan margin yang besar, sementara Naive Bayes menggunakan prinsip probabilistik untuk menentukan sentimen berdasarkan distribusi kata dalam teks. Kedua algoritma ini memiliki keunggulan dalam kecepatan proses, akurasi yang tinggi, serta kemampuannya untuk menangani dataset yang besar dan beragam. Dalam konteks penelitian ini, pembelajaran mesin

dianggap lebih efektif karena dapat menyesuaikan diri dengan konteks bahasa lokal yang dinamis di media sosial (Das & Kumar, 2022).

Kokab et al. (2022) menekankan bahwa pendekatan berbasis *machine learning* seperti *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) memiliki kemampuan tinggi dalam mengklasifikasikan teks besar secara otomatis(Kokab et al., 2022). Proses ini dimulai dengan tahap prapemrosesan data, di mana teks dibersihkan dari *noise* seperti emotikon, singkatan, dan bahasa tidak baku. Variasi bahasa informal, singkatan, dan campuran yang umum muncul dalam opini publik pada media sosial X menjadi fokus utama dalam tahap prapemrosesan teks. Pengguna media sosial sering kali menggunakan bahasa yang lebih santai, emotikon, singkatan, serta campuran antara bahasa Indonesia dan bahasa Inggris, yang bisa mengaburkan makna asli dan mempengaruhi akurasi analisis sentimen. Oleh karena itu, analisis sentimen memerlukan teknik yang adaptif terhadap dinamika linguistik agar hasilnya tetap akurat dan representatif terhadap opini publik sebenarnya (Xu et al., 2025).

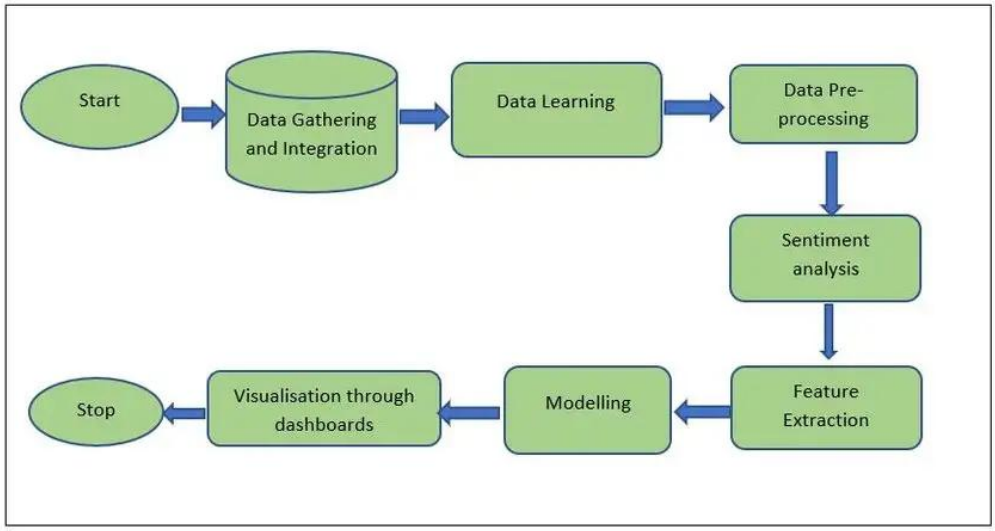
2.1.2 Hubungan Analisis Sentimen dengan Kebijakan Publik

Analisis sentimen memberikan gambaran mengenai tingkat penerimaan publik terhadap Program Makan Bergizi Gratis, sehingga dapat digunakan sebagai dasar evaluasi kebijakan. Dalam konteks *Program Makan Bergizi Gratis*, analisis ini membantu menilai apakah kebijakan tersebut diterima dengan baik atau menimbulkan reaksi negatif. Zhao et al. (2022) menjelaskan bahwa pola sentimen publik dapat menjadi indikator efektivitas komunikasi kebijakan pemerintah. Data yang diperoleh dari analisis sentimen membantu lembaga pemerintahan memahami

isu prioritas, memperbaiki strategi komunikasi, dan meningkatkan transparansi kebijakan(Zhao et al., 2022).

Opini publik di media sosial terbentuk melalui interaksi emosional, bias kognitif, dan dinamika kelompok sosial. Ketika individu mengekspresikan pandangan terhadap kebijakan pemerintah, mereka tidak hanya menyampaikan fakta tetapi juga mengekspresikan emosi dan persepsi terhadap legitimasi kebijakan tersebut(*Relation of Public Opinion with Public Policy*, n.d.). Oleh karena itu, analisis sentimen dapat digunakan sebagai alat ukur psikologis dan sosial yang mencerminkan kondisi emosional masyarakat terhadap kebijakan yang sedang berlangsung.

Analisis polaritas sentimen (positif, negatif, netral) memiliki peran penting dalam menilai tingkat penerimaan publik. Sentimen positif menunjukkan dukungan yang kuat dari masyarakat terhadap program, mencerminkan kepuasan publik atas manfaat yang diterima serta apresiasi terhadap upaya pemerintah dalam menyediakan makanan bergizi bagi warga. Sentimen ini dapat mencakup pujian terhadap efektivitas distribusi, kualitas makanan, atau dampak positif yang dirasakan oleh penerima manfaat. Sebaliknya, sentimen negatif dapat menjadi indikasi adanya ketidakpuasan atau masalah dalam implementasi kebijakan, seperti kesulitan dalam akses, ketidakmerataan distribusi, atau kualitas layanan yang tidak memenuhi harapan. Dengan memahami kedua sisi sentimen ini, pembuat kebijakan dapat merancang intervensi yang lebih tepat, mengoptimalkan program untuk memenuhi kebutuhan masyarakat, dan meningkatkan partisipasi serta kepercayaan publik terhadap kebijakan yang ada.



Gambar 2.1 Evolusi Analisis Sentimen di Media Sosial

Sumber : Wikimedia Commons

Gambar ini menggambarkan evolusi proses analisis sentimen di media sosial dari tahap tradisional berbasis leksikon menuju pendekatan modern yang menggunakan *machine learning* dan *deep learning*. Pada bagian awal, diagram memperlihatkan tahapan dasar seperti pengumpulan data (*data collection*), pembersihan teks (*text preprocessing*), dan ekstraksi fitur (*feature extraction*) yang membentuk fondasi awal analisis opini publik. Selanjutnya, gambar ini menampilkan bagaimana perkembangan teknologi telah memungkinkan penerapan algoritma canggih seperti *Support Vector Machine (SVM)*, *Naïve Bayes (NB)*, dan model *transformer* seperti BERT untuk memahami konteks semantik yang lebih mendalam. Evolusi ini menunjukkan bahwa analisis sentimen kini tidak hanya berfungsi mengklasifikasi opini positif, negatif, atau netral, tetapi juga mampu menangkap emosi dan konteks sosial yang kompleks dalam komunikasi publik.

2.2 Algoritma Yang Digunakan

Analisis sentimen dalam penelitian ini menggunakan dua algoritma utama, yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB). Kedua algoritma ini dipilih karena performanya yang kompetitif dalam mengklasifikasikan opini kebijakan publik yang bersifat dinamis dan berdimensi tinggi, terutama ketika menghadapi data teks yang sangat beragam dan kompleks. SVM bekerja berdasarkan prinsip optimasi geometris untuk memisahkan data ke dalam kelas yang berbeda dengan margin maksimal, sementara NB menggunakan pendekatan probabilistik dengan asumsi independensi antar fitur (Berrar, n.d.)(Uddin et al., n.d.)(Suasnawa et al., 2021). Dalam konteks analisis sentimen publik terhadap *Program Makan Bergizi Gratis*, kedua algoritma tersebut saling melengkapi antara efisiensi perhitungan *Naïve Bayes* (NB) dan akurasi klasifikasi *Support Vector Machine* (SVM). *Naive Bayes* (NB) menawarkan efisiensi luar biasa dalam memproses opini singkat, seperti komentar atau cuitan di media sosial, yang sering kali terdiri dari kalimat pendek, singkatan, dan bahasa tidak baku. Keunggulan utama NB terletak pada kemampuannya untuk bekerja dengan cepat meskipun jumlah data yang dianalisis besar, serta kemudahan dalam mengimplementasikan model pada dataset teks yang tidak terlalu kompleks. Sementara itu, *Support Vector Machine* (SVM) memiliki kemampuan untuk memberikan akurasi yang sangat tinggi dalam mengklasifikasikan opini publik, terutama ketika representasi fitur menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency), yang dapat menangani ribuan dimensi fitur dalam data teks.

Perbandingan keduanya menjadi penting karena karakteristik data media sosial yang cenderung tidak terstruktur, bervariasi, dan memiliki ambiguitas

linguistik. NB mampu memberikan hasil cepat pada data besar karena kesederhanaan modelnya, sedangkan SVM lebih kuat dalam menangani data yang kompleks dengan fitur saling bergantung. Rahmadzani et al. (2025) menekankan bahwa dalam klasifikasi teks yang bersifat multikelas, SVM sering kali menunjukkan akurasi lebih tinggi dibanding NB, terutama ketika data memiliki *noise* atau ketidakseimbangan antar kelas(Studies, 2025).

2.2.1 Prinsip Dasar Support Vector Machine

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin (*machine learning*) yang efektif untuk menganalisis data opini publik yang memiliki fitur tinggi, seperti data teks yang berisi berbagai variabel, kata kunci, dan konteks yang kompleks. Tujuan utama SVM adalah memaksimalkan margin, yaitu jarak antara *hyperplane* dan titik data terdekat dari setiap kelas(Guido et al., 2024). Secara matematis, SVM memecahkan masalah optimasi dengan fungsi objektif untuk meminimalkan kesalahan klasifikasi sambil mempertahankan jarak antar kelas sebesar mungkin. Model ini menggunakan fungsi *kernel* seperti *linear*, *polynomial*, atau *radial basis function (RBF)* untuk mentransformasikan data non-linear ke ruang berdimensi lebih tinggi agar dapat dipisahkan secara linear (Du et al., 2024).

Dalam analisis sentimen, SVM terbukti unggul untuk data berdimensi tinggi seperti teks karena setiap kata dianggap sebagai fitur yang independen. Kemampuan SVM dalam menangani data dengan korelasi tinggi antar fitur menjadikannya sangat efektif dalam memproses bahasa alami yang kompleks(Al-harbi, 2019). Millennianita et al. (2024) menambahkan bahwa SVM memberikan

stabilitas performa yang baik pada data dengan distribusi tidak seimbang karena algoritma ini berfokus pada titik-titik data kritis (*support vectors*) yang berkontribusi signifikan terhadap pembentukan *hyperplane*(Millennianita et al., 2024).

2.2.2 Prinsip Dasar *Naïve Bayes*

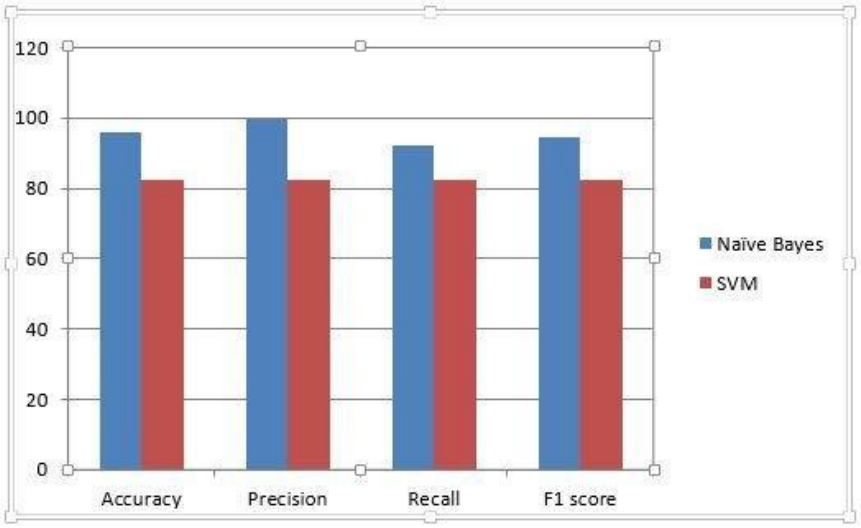
Naïve Bayes (NB) adalah algoritma probabilistik yang bekerja berdasarkan Teorema Bayes untuk menghitung probabilitas suatu kelas berdasarkan fitur yang diamati. Asumsi utamanya adalah bahwa setiap fitur bersifat independen terhadap fitur lainnya, yang membuat proses perhitungan menjadi lebih sederhana dan efisien (Berrar, 2019). Rumus dasar NB adalah:

$$\boldsymbol{P(C|X)} = \frac{\boldsymbol{P(X|C)P(C)}}{\boldsymbol{P(X)}}$$

di mana (C) adalah kelas sentimen (positif, negatif, netral) dan (X) adalah kumpulan fitur teks. Meskipun asumsi independensi ini jarang terpenuhi secara sempurna dalam teks alami, NB tetap memberikan hasil kompetitif, terutama dalam skenario dengan data besar dan berdimensi tinggi (Phatcharathada & Srisuradetchai, 2025).

NB efisien dalam komputasi karena hanya membutuhkan perhitungan probabilitas sederhana berdasarkan frekuensi kemunculan kata(Falasari & Muslim, 2022). Kelebihan utama NB terletak pada kemampuannya menangani *sparse data* dengan sangat baik, menjadikannya cocok untuk menganalisis teks opini singkat yang bersumber dari media sosial X, terutama terkait kebijakan publik pada Program Makan Bergizi Gratis. Dalam konteks kebijakan, data yang diperoleh sering kali berupa komentar, tweet, atau cuitan yang singkat dan mengandung

banyak variasi dalam bahasa, singkatan, serta penggunaan emotikon atau hashtag. Meskipun teks-teks tersebut relatif pendek dan tidak terstruktur, Naive Bayes dapat dengan efektif mengklasifikasikan opini masyarakat berdasarkan distribusi probabilistik kata-kata yang muncul dalam teks. Namun, kelemahannya adalah sensitivitas terhadap korelasi antar fitur, yang dapat menyebabkan penurunan akurasi jika kata-kata dalam teks memiliki hubungan semantik yang kuat (Blanquero et al., 2021).



Gambar 2.2 Perbandingan Teoritis dan Praktis Support Vector Machine dan Naïve Bayes

Sumber :Researchgate

Gambar ini menampilkan perbandingan teoretis dan praktis antara dua algoritma klasifikasi populer dalam analisis sentimen, yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB). Visualisasi ini menggambarkan bagaimana SVM bekerja dengan prinsip geometris untuk menemukan *hyperplane* optimal yang memisahkan data ke dalam kelas dengan margin maksimum, sedangkan NB menggunakan pendekatan probabilistik berdasarkan Teorema Bayes untuk

menghitung kemungkinan sebuah teks termasuk ke dalam kelas sentimen tertentu. Dalam konteks penelitian, gambar ini membantu menunjukkan keunggulan dan keterbatasan masing-masing algoritma: SVM unggul dalam akurasi dan kemampuan menangani data kompleks berdimensi tinggi, sementara NB unggul dalam kecepatan dan efisiensi komputasi.

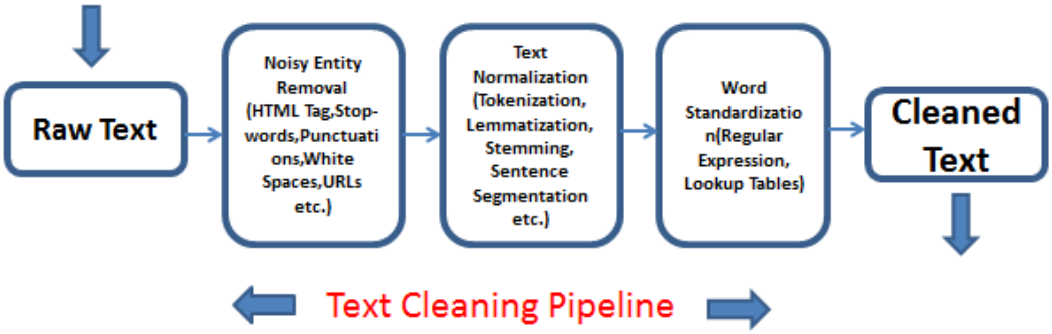
2.3 Langkah – Langkah Dalam Machine Learning

Tahapan dalam *machine learning* berfungsi memastikan proses analisis sentimen berjalan secara sistematis, terstruktur, dan dapat direplikasi. Seluruh langkah mulai dari pengumpulan data, pembersihan teks, ekstraksi fitur, pelatihan model, hingga validasi hasil harus dilakukan secara konsisten agar algoritma mampu menghasilkan prediksi yang akurat. Proses ini sangat penting dalam analisis sentimen berbasis teks, terutama ketika data bersumber dari media sosial yang penuh variasi bahasa, singkatan, serta *noise*. Menurut Krouska et al. (2017), kualitas proses awal akan menentukan kualitas model sehingga tahapan pra-pemrosesan data menjadi komponen yang tidak bisa diabaikan.(Krouska et al., 2016). Dalam penelitian ini, pipeline disusun mengikuti praktik terbaik dalam analisis data berbasis *machine learning* , dengan memastikan setiap langkah mengikuti standar ilmiah dan sesuai konteks kebijakan *Program Makan Bergizi Gratis*(Z. Li et al., 2025).

2.3.1 Tahap Pengumpulan dan Prapemrosesan

Pengumpulan data dilakukan menggunakan kata kunci yang berkaitan langsung dengan Program Makan Bergizi Gratis agar opini yang dianalisis benar-

benar relevan. Tahap ini harus memastikan bahwa hanya cuitan yang benar-benar terkait dengan topik kebijakan Program Makan Bergizi Gratis yang masuk dalam dataset. Dalam konteks analisis sentimen kebijakan publik, teks yang bersumber dari media sosial sering kali mengandung banyak kebisingan, seperti kata-kata yang tidak relevan, singkatan, atau penggunaan bahasa informal yang dapat mengaburkan makna. Setelah data terkumpul, langkah berikutnya adalah *text preprocessing*, dengan melakukan prapemrosesan yang tepat seperti pembersihan data, penghapusan kata-kata umum (stop words), standarisasi singkatan, dan normalisasi teks kualitas data yang digunakan dalam pelatihan model dapat ditingkatkan secara signifikan. Proses ini mengurangi variasi yang tidak perlu dan memastikan bahwa hanya informasi yang relevan yang dipertimbangkan dalam analisis. Hasilnya, model klasifikasi menjadi lebih efisien dalam mengidentifikasi pola sentimen, sehingga mampu menghasilkan prediksi yang lebih akurat dan andal. Prapemrosesan yang baik tidak hanya meningkatkan performa model dalam hal akurasi, tetapi juga memungkinkan model untuk lebih cepat mengenali nuansa dalam opini publik, memberikan wawasan yang lebih dalam mengenai reaksi masyarakat terhadap kebijakan yang dianalisis. yaitu proses membersihkan dan menyiapkan teks agar siap diolah oleh algoritma. *Case folding* digunakan untuk menyeragamkan huruf menjadi kecil, *tokenizing* memecah teks menjadi kata, *stopword removal* menghilangkan kata-kata umum yang tidak bermakna penting, dan *stemming* mengembalikan kata ke bentuk dasarnya. Tahapan pra-pemrosesan ini mampu mengurangi *noise* dalam teks sehingga meningkatkan performa klasifikasi model (Id et al., 2020).



Gambar 2.3 Alur Pra-Pemrosesan Teks

Sumber: linkedin.com

Gambar ini menampilkan tahapan dasar pra-pemrosesan teks yang meliputi pembersihan data, normalisasi, tokenisasi, dan ekstraksi fitur. Visualisasi ini memberikan ilustrasi runtut bagaimana teks mentah diproses menjadi representasi numerik yang siap dimasukkan ke pipeline *machine learning*.

2.3.2 Ektraksi Fitur, Pelatihan, dan Validasi Model

Setelah teks dibersihkan, tahap berikutnya adalah mengubah teks menjadi representasi numerik melalui *Term Frequency–Inverse Document Frequency (TF–IDF)*. *TF–IDF* memberi bobot pada kata yang jarang muncul secara global tetapi sering muncul di dokumen tertentu, sehingga membantu algoritma mengenali kata-kata bermakna (Harmandini & L, 2024). TF–IDF digunakan untuk menonjolkan kata-kata yang sering muncul dalam diskusi mengenai Program Makan Bergizi Gratis sehingga membantu model dalam memahami konteks kebijakan. Dalam diskusi tentang kebijakan seperti Program Makan Bergizi Gratis, terdapat istilah-istilah kunci yang lebih relevan dan spesifik, seperti "gizi", "penerima manfaat", atau "distribusi makanan", yang perlu diberi bobot lebih agar model dapat

menangkap makna penting dari opini publik. Dengan menggunakan TF–IDF, model dapat lebih baik dalam membedakan kata-kata yang signifikan dari yang kurang relevan, sehingga meningkatkan akurasi analisis sentimen dan memungkinkan model untuk mengenali nuansa dalam berbagai pendapat yang disampaikan oleh masyarakat, baik yang mendukung maupun yang mengkritik kebijakan tersebut. Representasi ini memungkinkan model seperti *Support Vector Machine* dan *Naïve Bayes* mempelajari pola sentimen secara akurat. Proses pelatihan model dilakukan dengan pendekatan *supervised learning*, tahap pelatihan dilakukan untuk membandingkan performa SVM (Support Vector Machine) dan NB (Naive Bayes) secara empiris berdasarkan opini publik yang tersebar di media sosial X, dengan tujuan untuk menilai keunggulan dan kelemahan masing-masing algoritma dalam mengklasifikasikan sentimen terhadap Program Makan Bergizi Gratis. Dalam proses ini, model SVM dan NB akan dilatih menggunakan dataset yang terdiri dari berbagai teks opini, seperti komentar, tweet, dan cuitan, yang menggambarkan beragam sentimen publik baik positif, negatif, maupun netral. Tujuan utama dari komparasi ini adalah untuk mengidentifikasi algoritma mana yang lebih efektif dalam menangani data teks yang memiliki dimensi tinggi, variasi bahasa, serta struktur yang tidak teratur, yang sering ditemukan dalam diskusi kebijakan di media sosial. Kombinasi *TF–IDF* dengan SVM dan NB menghasilkan akurasi tinggi pada tugas klasifikasi sentimen media sosial yang semakin menegaskan efektivitas kedua algoritma dalam menganalisis opini publik (Marganingsih et al., 2025).

Validasi hasil dilakukan menggunakan *k-fold cross-validation* yang membagi dataset menjadi beberapa bagian untuk diuji secara bergantian. Teknik ini

mencegah *overfitting* dan memastikan model mampu melakukan generalisasi terhadap data baru. Menurut Xu et al. (2022), validasi silang memberikan ukuran performa yang lebih stabil untuk data media sosial yang penuh variasi(Ariel et al., 2022). Evaluasi model kemudian dilanjutkan menggunakan metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Evaluasi yang komprehensif membantu peneliti memahami kekuatan dan kelemahan model secara lebih mendalam(Zafar et al., 2024).

2.4 Teknik Evaluasi Algoritma dalam *Machine Learning*

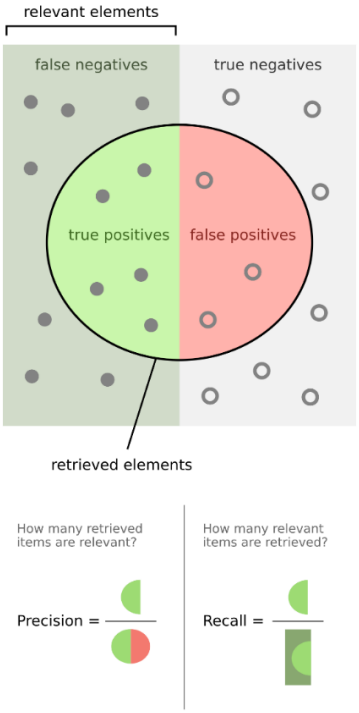
Evaluasi algoritma merupakan langkah penting dalam memastikan bahwa model *machine learning* mampu menghasilkan prediksi yang akurat dan dapat diandalkan, khususnya dalam analisis sentimen berbasis teks. Evaluasi tidak hanya berfungsi untuk mengukur performa model, melainkan juga untuk memahami bagaimana model menangani data yang kompleks, tidak seimbang, dan penuh variasi seperti data dari media sosial. Menurut Riyanto et al. (2023), pemilihan metrik evaluasi harus mempertimbangkan struktur data serta tujuan penelitian agar hasil yang diperoleh tidak bias dan dapat mencerminkan kondisi sebenarnya(Riyanto et al., 2023). Pada penelitian analisis sentimen terhadap *Program Makan Bergizi Gratis*, penggunaan metrik yang tepat menjadi krusial karena opini masyarakat pada media sosial sering kali menampilkan distribusi tidak seimbang antara sentimen positif, negatif, dan netral.

2.4.1 Metrik Evaluasi Kinerja Model

Evaluasi dilakukan untuk mengukur seberapa baik kedua algoritma SVM (Support Vector Machine) dan Naive Bayes (NB) mampu mengklasifikasikan sentimen terkait kebijakan Program Makan Bergizi Gratis, dengan mempertimbangkan berbagai faktor seperti akurasi, presisi, recall, dan F1-score. Precision, recall, F1-score, accuracy, dan MCC (Matthews Correlation Coefficient) adalah metrik yang umum digunakan untuk menilai efektivitas model dalam membaca kecenderungan opini publik terhadap kebijakan Program Makan Bergizi Gratis. Dalam konteks kebijakan ini, penting untuk mengukur tidak hanya seberapa akurat model dalam mengklasifikasikan sentimen, tetapi juga seberapa baik model dapat menangkap opini yang positif maupun negatif dari publik. *Accuracy* mengukur proporsi prediksi yang benar dibanding total prediksi, namun metrik ini tidak ideal untuk data tidak seimbang karena cenderung menguntungkan kelas mayoritas. Vanacore et al. (2022) menjelaskan bahwa *accuracy* harus digunakan dengan hati-hati pada skenario multikelas karena dapat memberikan gambaran semu terhadap performa model(Vanacore et al., 2024). Sebaliknya, *precision* dan *recall* lebih representatif pada data tidak seimbang karena menekankan ketepatan prediksi untuk kelas tertentu (Q. Li et al., 2022).

Precision mengukur ketepatan prediksi positif, sedangkan *recall* mengukur kemampuan model untuk mengidentifikasi seluruh data positif. Keduanya sangat relevan dalam analisis sentimen, terutama ketika kesalahan dalam mengidentifikasi opini negatif dapat berdampak signifikan terhadap evaluasi kebijakan publik. *F1-score*, rata-rata harmonis antara *precision* dan *recall*, memberikan gambaran yang

lebih seimbang dan banyak digunakan dalam penelitian modern. Takahashi et al. (2021) menyebutkan bahwa *F1-score* memberikan stabilitas evaluasi yang tinggi pada data yang memiliki ketidakseimbangan kelas, sedangkan Sitarz (2022) menekankan bahwa metrik ini lebih mampu menangkap performa klasifikasi secara keseluruhan(Takahashi et al., 2022) (Sitarz, 2022).



Gambar 2.4 Contoh Visualisai Metrik Evaluasi

Sumber**:** Wikimedia Commons

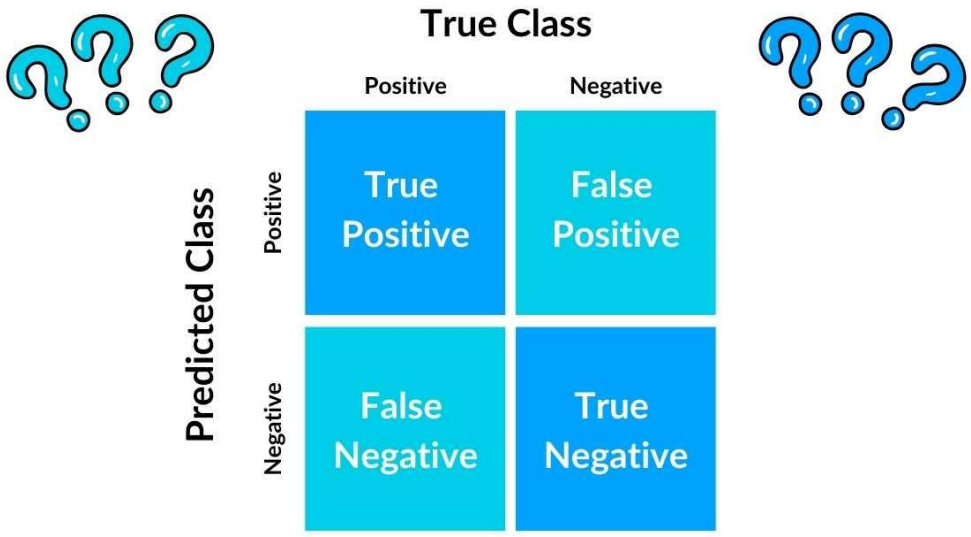
Gambar ini menampilkan hubungan antara *precision*, *recall*, dan *F1-score* dalam evaluasi model klasifikasi. Visualisasi ini menunjukkan bagaimana perubahan dalam nilai *precision* dan *recall* memengaruhi *F1-score*.

2.4.1 Validasi Model dan Analisis Kesalahan

Selain metrik dasar, penelitian ini juga menggunakan teknik *stratified k-fold cross-validation* untuk mengevaluasi kemampuan generalisasi model. Validasi *stratified k-fold* memastikan bahwa model mampu membaca variasi opini publik secara stabil tanpa bias kelas, karena setiap lipatan data tetap memuat proporsi sentimen yang seimbang. Dalam konteks analisis opini terhadap kebijakan Program Makan Bergizi Gratis, teknik validasi ini membantu model menangkap dinamika persepsi masyarakat baik dukungan, kritik, maupun sentimen netral secara lebih representatif, sehingga hasil evaluasi kinerja algoritma menjadi lebih akurat dan mencerminkan kondisi nyata di media sosial X. Teknik ini membagi dataset menjadi beberapa bagian dengan proporsi kelas yang tetap, sehingga setiap *fold* merepresentasikan keseluruhan distribusi kelas. Menurut Bahrami et al. (2024), teknik ini dianggap lebih efektif untuk menghindari bias dan menghasilkan evaluasi yang stabil pada data tidak seimbang. *Stratified cross-validation* memberikan hasil yang lebih akurat dibandingkan *k-fold* biasa karena mempertahankan proporsi kelas minoritas dalam setiap iterasi pelatihan(Lumumba et al., 2024).

Analisis *confusion matrix* digunakan untuk mengidentifikasi berbagai bentuk kesalahan prediksi yang dapat memengaruhi ketepatan interpretasi terhadap sentimen masyarakat mengenai Program Makan Bergizi Gratis. Melalui matriks ini, peneliti dapat mengetahui apakah model cenderung salah mengklasifikasikan sentimen negatif sebagai positif, atau sebaliknya, yang berpotensi menghasilkan pemahaman yang keliru terhadap opini publik. Riyanto et al. (2023) menekankan bahwa *confusion matrix* sangat berguna untuk mengidentifikasi pola kesalahan seperti *false positive* atau *false negative*, yang dapat membantu peneliti memahami

kelemahan model secara spesifik(Riyanto et al., 2023). Heydarian et al. (2022) menambahkan bahwa visualisasi *confusion matrix* memudahkan interpretasi performa model, terutama dalam kasus multikelas(Heydarian & Doyle, 2022).



Gambar 2.5 Visualisasi Confusion Matrix

Sumber**:** spotintelligence.com

Gambar ini menunjukkan empat komponen utama evaluasi model klasifikasi menggunakan *confusion matrix*: *true positive*, *true negative*, *false positive*, dan *false negative*. Visualisasi ini membantu peneliti memahami pola kesalahan dan meningkatkan akurasi model melalui penyesuaian strategi pelatihan

Interpretasi hasil evaluasi harus mencakup analisis mendalam mengenai bagaimana metrik-metrik tersebut menggambarkan performa model terhadap opini publik yang bergantung pada konteks. Pemilihan metrik evaluasi harus disesuaikan dengan tujuan analisis dan karakteristik data, sehingga kesimpulan yang dihasilkan

dapat memberikan gambaran akurat tentang efektivitas model dalam memahami sentimen publik(Aboulola & Umer, 2024).

2.5 Alat Bantu Analisis Teks Berbasis *Machine Learning*

Python, Scikit-learn, NLTK, dan Pandas dipilih karena mampu mengolah opini publik secara efisien mulai dari tahap prapemrosesan teks hingga evaluasi model. Kombinasi pustaka tersebut sangat relevan dalam menganalisis kebijakan Program Makan Bergizi Gratis, karena mampu menangani data komentar yang beragam di media sosial X, membersihkannya dari *noise*, mengekstraksi makna sentimen secara akurat, serta menguji performa algoritma secara terukur. Dengan dukungan ekosistem ini, analisis sentimen dapat dilakukan secara komprehensif sehingga hasilnya benar-benar mencerminkan persepsi masyarakat terhadap efektivitas dan penerimaan kebijakan tersebut.. Dalam konteks ini, tiga pustaka utama Scikit-learn, NLTK, dan Pandas memainkan peran fundamental dalam pipeline analisis teks, mulai dari praproses hingga tahap evaluasi model ((Darji & Goswami, 2024) (Raschka et al., 2020b) (Talati, 2021) .

2.5.1 Python dan Machine Learning

Python menawarkan lingkungan yang ideal untuk penelitian berbasis data karena sintaksisnya sederhana dan mendekati bahasa manusia. Hal ini mempermudah peneliti dalam menulis dan memahami kode tanpa kehilangan efisiensi komputasi. Menurut Talati (2021), Python dapat dianggap sebagai katalis bagi evolusi kecerdasan buatan karena kemampuannya menggabungkan fleksibilitas pemrograman dengan ekosistem pustaka yang sangat kaya (Talati,

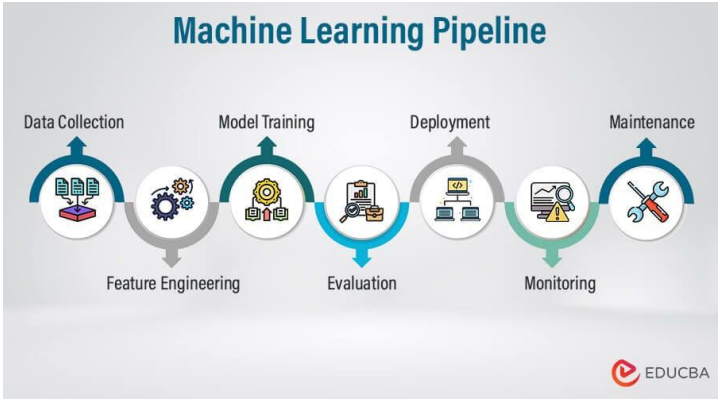
2021). Python mendukung berbagai pustaka seperti TensorFlow, PyTorch, dan Scikit-learn yang mempercepat implementasi algoritma pembelajaran mesin(Raschka et al., 2020b).

Python juga memiliki keunggulan lintas platform yang mempermudah replikasi penelitian. Dengan dokumentasi lengkap dan komunitas riset yang kuat, Python memungkinkan para peneliti untuk berbagi kode sumber dan hasil eksperimen secara terbuka, mendukung semangat *open science* dan kolaborasi global (Darji & Goswami, 2024). Oleh karena itu, Python tidak hanya sekadar alat bantu teknis, tetapi juga medium epistemologis yang memfasilitasi pertumbuhan pengetahuan ilmiah dalam bidang analisis teks berbasis *machine learning*.

Kemampuan Python dalam mengotomatisasi ekstraksi, pembersihan, serta klasifikasi data memungkinkan analisis dilakukan secara sistematis dan transparan, sehingga hasil penelitian dapat memberikan gambaran faktual tentang bagaimana masyarakat merespons kebijakan pemerintah terkait Program Makan Bergizi Gratis. Dengan demikian, Python menjadi fondasi penting yang memastikan analisis sentimen terhadap Program Makan Bergizi Gratis dilakukan secara reliabel, dapat direplikasi, dan mendukung proses evaluasi kebijakan berbasis data.

Proses *machine learning* menggunakan Python dalam analisis sentimen terhadap kebijakan Program Makan Bergizi Gratis dilakukan melalui tahapan sistematis. Pertama, data opini publik dikumpulkan dari media sosial X/Twitter menggunakan pustaka seperti *Pandas*, *Requests*, dan *Tweepy*. Data kemudian dipraproses menggunakan *NLTK* untuk menghapus *noise*, melakukan *tokenization*, *stopword removal*, serta *stemming* agar teks lebih bersih dan siap dianalisis.

Selanjutnya, fitur diekstraksi menggunakan metode TF-IDF atau *Bag of Words* dari *Scikit-learn* untuk mengubah teks menjadi representasi numerik. Dataset kemudian dibagi menjadi data latih dan data uji menggunakan *train_test_split* atau *stratified k-fold*. Model dibangun menggunakan algoritma *Naïve Bayes* atau *Support Vector Machine*, lalu dilatih dan diuji untuk memprediksi sentimen masyarakat. Evaluasi dilakukan menggunakan *accuracy*, *precision*, *recall*, dan *confusion matrix*. Hasilnya diinterpretasikan dan divisualisasikan melalui *Matplotlib*, disimpan menggunakan *joblib*, serta dapat dideploy untuk pemantauan kebijakan berbasis data.



Gambar 2.6 Alur Machine Learning

Sumber : Medium.com

Gambar di atas memperlihatkan pipeline (alur) machine learning: mulai dari pengumpulan data, pemrosesan (preprocessing), rekayasa fitur (feature engineering), pelatihan model, evaluasi, hingga deployment.

2.5.2 Peran Scikit-learn dalam Klasifikasi dan Evaluasi Model Teks

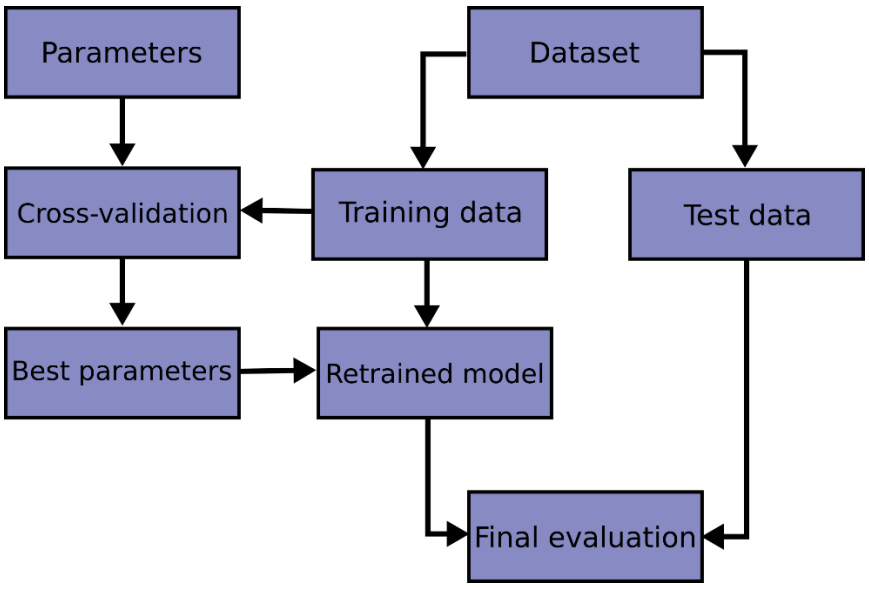
Scikit-learn merupakan pustaka yang sangat penting dalam implementasi *machine learning* untuk teks. Pustaka ini menawarkan API yang konsisten dan

mudah digunakan, mendukung berbagai algoritma klasik seperti *Support Vector Machine* (SVM), *Naive Bayes*, dan *Random Forest*. Menurut Raschka et al. (2020), Scikit-learn dirancang untuk mempermudah integrasi pipeline data, mulai dari tahap pelatihan hingga validasi model, dengan menyediakan fungsi-fungsi seperti *cross-validation*, *feature extraction*, dan *model evaluation*(Raschka et al., 2020a).

Scikit-learn sangat efektif untuk klasifikasi berbasis teks, seperti pada model *Naive Bayes* yang digunakan untuk menganalisis hubungan antara kondisi sebelum dan sesudah dalam spesifikasi kebutuhan perangkat lunak (Requirements et al., 2020). Iqbal et al. (2025) memperkuat temuan tersebut dengan penerapan Scikit-learn dalam analisis sentimen media sosial, yang mampu mengungkap pola opini publik (Iqbal et al., 2025). Selain itu, Yang et al. (2021) mengembangkan *Transformers-sklearn*, toolkit yang memadukan kekuatan *transformer models* dengan kemudahan Scikit-learn, khususnya untuk pemahaman bahasa medis (Yang et al., 2021). Kemampuan Scikit-learn dalam mengolah teks dalam skala besar memungkinkan peneliti mengekstraksi pola-pola opini publik dari media sosial secara efisien. Fitur *cross-validation* membantu memastikan bahwa model tidak hanya akurat, tetapi juga stabil dalam membaca variasi persepsi masyarakat terhadap kebijakan tersebut baik komentar positif terkait manfaat program, kekhawatiran mengenai pendanaan, maupun kritik terkait pelaksanaan di lapangan. Dengan dukungan metode evaluasi yang objektif seperti *precision*, *recall*, dan *F1-score*, Scikit-learn memungkinkan analisis yang dapat dipertanggungjawabkan secara teknis dan akademik, sehingga hasil sentimen publik dapat dijadikan bahan

pertimbangan yang valid bagi perumusan, perbaikan, atau sosialisasi kebijakan Program Makan Bergizi Gratis.

Tahapan proses Scikit-learn dalam klasifikasi dan evaluasi teks dimulai dari pengolahan komentar masyarakat tentang Program Makan Bergizi Gratis melalui prapemrosesan seperti pembersihan *noise*, normalisasi, dan tokenisasi agar opini informal menjadi data yang siap dianalisis. Scikit-learn kemudian mengekstraksi fitur menggunakan *CountVectorizer* atau *TF-IDF* untuk mengubah kata-kata seperti "bermanfaat" atau "tidak merata" menjadi bentuk numerik. Dataset dibagi menggunakan *train test split* atau *cross-validation* untuk menjaga keseimbangan sentimen sebelum model seperti *Naive Bayes* atau *SVM* dilatih membaca pola dukungan maupun kritik. Setelah diuji, performa model dievaluasi menggunakan akurasi, presisi, recall, dan F1-score. Hasil evaluasi ini membantu menafsirkan persepsi publik secara lebih objektif sehingga dapat menjadi dasar perbaikan dan penyesuaian kebijakan Program Makan Bergizi Gratis.



Gambar 2.7 Tahapan Proses Scikit-Learn

Sumber : scikit-learn.org

Gambar tersebut menunjukkan alur proses machine learning menggunakan Scikit-learn, mulai dari pembagian dataset menjadi data latih dan data uji, kemudian dilanjutkan dengan proses *cross-validation* untuk menguji berbagai parameter dan menentukan *best parameters* yang paling optimal. Parameter terbaik tersebut digunakan untuk melatih ulang model agar performanya meningkat sebelum akhirnya dilakukan *final evaluation* pada data uji. Secara ringkas, diagram ini menggambarkan bagaimana Scikit-learn melakukan *model selection*, *hyperparameter tuning*, dan evaluasi akhir secara terstruktur untuk menghasilkan model prediksi yang lebih akurat dan andal.

2.5.3 NLTK dan Praproses Data Teks dalam Analisis Sentimen

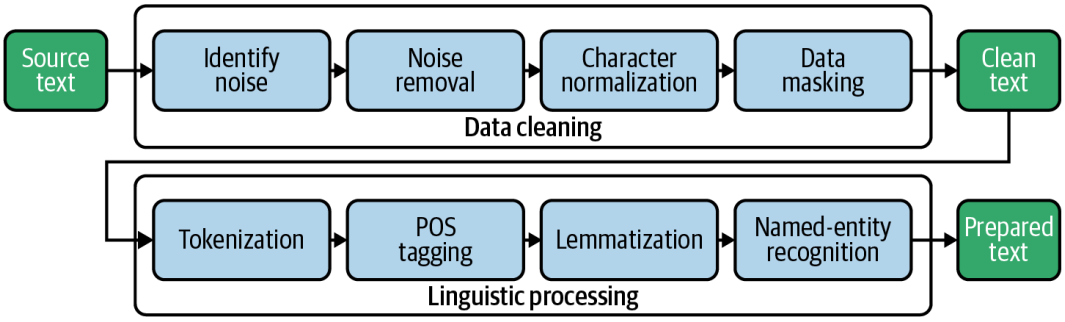
Penggunaan NLTK dalam praproses teks memiliki relevansi kuat bagi analisis kebijakan Program Makan Bergizi Gratis, terutama karena kebijakan ini memicu beragam opini publik di platform media sosial seperti X. Opini tersebut bersifat tidak terstruktur, mengandung *noise*, singkatan, emotikon, hingga variasi bahasa informal yang membutuhkan proses pembersihan sebelum dianalisis.

Sebelum analisis dilakukan, data teks perlu diproses agar bersih dan siap digunakan oleh model *machine learning*. Di sinilah NLTK (Natural Language Toolkit) berperan. NLTK menyediakan modul lengkap untuk tokenisasi, stemming, lemmatization, *stopword removal*, dan *POS tagging*. NLTK memungkinkan pembuatan pipeline praproses teks yang sistematis dan reproducible (Ogunrinde, 2025). Patel dan Passi (2020) juga menyoroti peran NLTK dalam analisis sentimen pada data Twitter selama Piala Dunia, di mana pustaka ini membantu

mengidentifikasi polaritas emosional dari tweet pengguna (Patel & Passi, 2020). Hal ini sangat relevan karena sentimen positif, negatif, dan netral terhadap Program Makan Bergizi Gratis dapat membantu pemerintah menilai tingkat penerimaan kebijakan, mengidentifikasi kritik dominan, serta memetakan aspek mana yang dipersepsikan publik sebagai berhasil atau bermasalah.

Penelitian terbaru oleh Nakib et al. (2024) menunjukkan bahwa NLTK dapat digunakan untuk menganalisis kepribadian berdasarkan data media sosial, sementara NLTK dalam analisis persepsi publik terhadap penyakit menular (Nakib et al., 2024) (Iparraguirre-villanueva et al., 2023). Hal ini menunjukkan bahwa pustaka ini tidak hanya berguna untuk analisis linguistik, tetapi juga dapat diterapkan dalam konteks sosial dan psikologis. Dengan demikian, penggunaannya pada Program Makan Bergizi Gratis bukan hanya untuk sekadar menganalisis teks, tetapi juga untuk memahami konteks emosi, kekhawatiran, harapan, dan motivasi masyarakat terkait keberlanjutan dan efektivitas program.

Tahapan proses NLTK dimulai dari *tokenization* untuk memecah komentar publik tentang Program Makan Bergizi Gratis menjadi unit kata yang mudah dianalisis. Dilanjutkan *stopword removal* untuk menghilangkan kata tidak penting sehingga fokus berada pada opini kunci. *Stemming* dan *lemmatization* digunakan menyederhanakan variasi kata, memperjelas makna sentimen. Selanjutnya, *feature extraction* seperti TF-IDF mengubah teks menjadi representasi numerik bagi model klasifikasi. Setelah model dilatih, proses evaluasi dilakukan menggunakan akurasi, presisi, dan *recall* guna menilai seberapa tepat model membaca persepsi masyarakat terhadap Program Makan Bergizi Gratis.



Gambar 2.8 Prapemrosesan Teks Menggunakan NLTK

Sumber : datacamp.com

Gambar tersebut menunjukkan alur prapemrosesan teks menggunakan NLTK yang terbagi dalam dua tahap utama. Data cleaning dimulai dari identifikasi *noise*, penghapusan *noise*, normalisasi karakter, hingga *data masking* untuk menghasilkan teks yang bersih. Setelah itu, tahap *linguistic processing* dilakukan melalui *tokenization*, *POS tagging*, *lemmatization*, dan *named-entity recognition* untuk menghasilkan teks yang telah siap digunakan dalam analisis atau pemodelan machine learning. Diagram ini menggambarkan bagaimana teks mentah diolah menjadi representasi linguistik yang terstruktur dan siap dianalisis.

2.5.4 Pandas sebagai Alat Pengelolaan dan Manipulasi Data

Pandas memiliki peran penting dalam manajemen data, khususnya pada tahap pengolahan data tabular sebelum masuk ke proses analisis model. Pustaka ini memungkinkan peneliti melakukan *data cleaning*, *aggregation*, dan *merging* dengan efisien. Dalam konteks NLP, Pandas sering digunakan untuk mengelola kumpulan data teks yang besar, seperti kumpulan tweet atau ulasan pengguna. Dengan integrasi erat bersama NumPy dan Matplotlib, Pandas menjadi jembatan antara analisis statistik dan visualisasi data.

Pandas memiliki peran krusial dalam pengelolaan data teks terkait opini publik terhadap Program Makan Bergizi Gratis. Dengan kemampuannya melakukan *data cleaning*, *aggregation*, dan *merging*, Pandas memungkinkan peneliti mengolah ribuan komentar masyarakat dari platform seperti X (Twitter) secara efisien sebelum masuk ke tahap analisis sentimen. Hal ini penting karena data opini publik sering berisi teks tidak terstruktur, duplikasi, atau informasi yang tidak relevan. Pandas juga membantu menggabungkan data teks dengan variabel lain, seperti waktu unggahan atau kategori sentimen, untuk menghasilkan analisis kebijakan yang lebih komprehensif.

Namun demikian, keterbatasan utama alat bantu ini terletak pada performa untuk data skala besar. Pustaka open-source seperti Pandas dan NLTK mengalami *bottleneck* dalam penggunaan memori saat menangani data dalam volume besar . Oleh karena itu, integrasi dengan kerangka kerja seperti Apache Spark diperlukan untuk mendukung pemrosesan paralel (Schröder et al., 2023).

Proses analisis sentimen Program Makan Bergizi Gratis menggunakan Pandas dimulai dari membaca dan mengimpor data komentar masyarakat ke dalam *dataframe*. Pandas kemudian digunakan untuk membersihkan data, seperti menghapus duplikasi, menangani nilai kosong, dan memfilter teks yang tidak relevan. Selanjutnya, data diolah melalui transformasi kolom, termasuk pelabelan sentimen dan pemetaan kategori. Pandas juga memfasilitasi eksplorasi awal, seperti menghitung distribusi sentimen dan tren komentar. Setelah fitur teks diekstraksi menggunakan pustaka lain, Pandas menggabungkan hasil prediksi model dan

evaluasi, sehingga memudahkan analisis akhir terhadap persepsi publik mengenai efektivitas program.

Gambar 2.9 Siklus Pembersihan Data Menggunakan Pandas

Sumber : analyticsvidhya.com

Gambar tersebut menggambarkan siklus pembersihan data menggunakan Pandas, yang dimulai dari proses mengimpor data mentah ke dalam *dataframe*. Setelah data masuk, tahap berikutnya adalah menggabungkan beberapa dataset melalui proses *merge* untuk menyatukan informasi yang relevan. Pandas kemudian digunakan untuk menangani data yang hilang dengan cara mengisi atau menghapus nilai kosong agar tidak memengaruhi analisis. Selanjutnya, dilakukan normalisasi untuk memastikan format data konsisten, diikuti dengan penghapusan duplikasi guna menjaga keakuratan informasi. Setelah seluruh tahapan selesai, data yang telah bersih kemudian disimpan atau diekspor kembali sebagai dasar untuk analisis lebih lanjut.

2.6 Penelitian Terdahulu

Table 2.1 Penelitian Terdahulu

No	Peneliti	Judul Penelitian	Tahun	Metode dan Data	Hasil
1.	Dey et al.	A Comparative Study of Support Vector Machine and Naïve Bayes Classifier for Sentiment Analysis on Amazon Product Reviews	2020	Data ulasan Amazon; metode SVM & NB menggunakan representasi TF–IDF	SVM menunjukkan akurasi lebih tinggi dalam menangani fitur teks kompleks; NB lebih cepat tetapi kurang akurat pada data ber-*noise*.
2.	Simarmata & Chrisinta	Studi Perbandingan Support Vector Machine dan Naive Bayes untuk Analisis Sentimen pada Kinerja Dosen	2023	Data kuesioner evaluasi dosen; penggunaan NB & SVM dengan preprocessing dasar	SVM lebih stabil dan akurat, sementara NB unggul pada kecepatan pemrosesan namun kalah dalam akurasi.
3.	Rahmadzani et al.	Perbandingan Metode Naive Bayes	2025	Data ulasan media sosial skincare;	NB kompetitif pada data

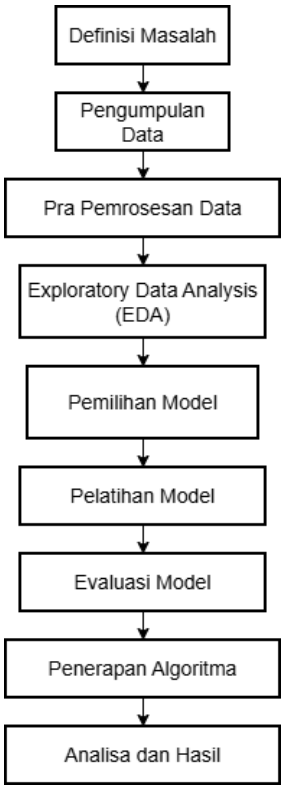
		dan Support Vector Machine (SVM) dalam Analisis Sentimen Skincare Female Daily		metode NB & SVM	bersih, tetapi SVM unggul dalam konsistensi akurasi, terutama untuk teks panjang yang kompleks.
4.	Rizki et al.	Comparison of SVM and NB in Sentiment Analysis of Tiktokshop App Reviews	2025	Data ulasan aplikasi Tiktokshop; metode SVM & NB	SVM memperoleh akurasi terbaik dengan TF–IDF, sedangkan NB unggul dalam efisiensi komputasi.
5.	Yogi et al.	Optimizing Sentiment Analysis in Social Media with Naive Bayes and Support Vector Machines	2025	Data sentimen media sosial dengan preprocessing NLP; model NB & SVM	SVM memberikan hasil akurasi lebih tinggi; NB tetap efektif untuk skenario ringan dan real-time.

6.	Penelitian Ini	Analisis Perbandingan Kinerja Algoritma Support Vector Machine dan Naïve Bayes Terhadap Sentimen Publik Pada Program Makan Bergizi Gratis di Media Sosial X	2025	Data sentimen publik dari media sosial X; preprocessing lengkap (tokenizing, stopword removal, stemming), representasi TF–IDF, penggunaan SVM & NB	Masih dalam proses penelitian.

Penelitian ini memiliki kebaruan yang kuat karena secara tegas memposisikan analisis sentimen publik sebagai alat evaluasi kebijakan, khususnya untuk Program Makan Bergizi Gratis yang merupakan inisiatif pemerintah yang masih baru dan perlu dikaji secara mendalam. Tidak seperti penelitian terdahulu yang cenderung berfokus pada ulasan produk, layanan digital, atau isu umum di media sosial, penelitian ini menempatkan analisis opini digital dalam konteks kebijakan publik sehingga menghasilkan pemahaman yang lebih strategis bagi pemerintah dalam membaca respons masyarakat. Dari sisi metodologis, penelitian ini memperkuat kontribusinya melalui rangkaian *preprocessing* yang lebih menyeluruh serta optimalisasi fitur TF–IDF, yang memungkinkan evaluasi performa *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) dilakukan dengan lebih akurat. Dengan demikian, penelitian ini tidak hanya menawarkan perbandingan algoritma, tetapi juga memberikan temuan praktis yang relevan untuk mendukung strategi komunikasi dan pengambilan keputusan berbasis data.

2.7 Kerangka Penelitian

Kerangka kerja penelitian ini disajikan dalam bentuk diagram alir penelitian sebagaimana ditampilkan pada gambar 2.7.



Gambar 2.10 Kerangka Kerja Penelitian

Langkah – langkah dalam kerangka kerja penelitian sebagai berikut:

1. Definisi Masalah

   Pada tahap ini, penelitian akan mengidentifikasi masalah yang perlu diselesaikan, yaitu bagaimana mengukur dan membandingkan efektivitas SVM dan Naive Bayes dalam analisis sentimen. Fokus utama adalah untuk mengetahui bagaimana masyarakat merespons Program Makan Bergizi Gratis yang diluncurkan oleh pemerintah. Hal ini akan dilihat melalui analisis sentimen dari data teks yang bersumber dari platform media sosial X.

2. Pengumpulan Data

   Data dikumpulkan melalui scraping menggunakan API yang disediakan dari media sosial X tersebut, dengan menggunakan kata kunci spesifik yang berkaitan dengan Program Makan Bergizi Gratis. Proses ini menekankan pengumpulan data autentik dan representatif agar model dapat dilatih dengan kualitas yang optimal. Seleksi relevansi, pembersihan duplikasi, dan verifikasi metadata dilakukan untuk memastikan kedua algoritma SVM dan NB memperoleh dataset yang konsisten sebelum diuji performanya. Hal ini penting agar perbandingan akurasi kedua model tidak bias oleh ketidakteraturan data.

3. Pra – Pemrosesan data

   Data teks yang telah dikumpulkan akan diproses melalui beberapa tahapan. Tahapan ini meliputi pembersihan teks (*cleaning*), *case folding*, *tokenizing*, *stopword removal*, *stemming*, serta normalisasi karakter. Tujuan utama proses ini adalah memastikan teks berada dalam format yang seragam sebelum dimodelkan. Representasi fitur akan menggunakan *TF–IDF* yang mengubah teks menjadi nilai numerik berbobot. Proses ini dirancang agar kedua algoritma diuji pada kondisi input yang adil dan optimal, sehingga perbedaan performa yang muncul benar-benar mencerminkan kemampuan masing-masing model, bukan karena perbedaan kualitas data input.

4. Exploratory Data Analysis (EDA)

   Sebelum pelatihan model, dilakukan analisis eksplorasi data (EDA). EDA dilakukan untuk memahami pola distribusi sentimen, frekuensi kata, sebaran

kelas, dan karakteristik konten publik. Visualisasi seperti *word cloud*, grafik distribusi polaritas, dan analisis panjang teks digunakan untuk membantu mengidentifikasi potensi ketidakseimbangan kelas yang dapat memengaruhi performa algoritma. Melalui EDA, peneliti dapat mengantisipasi faktor-faktor yang mungkin memberikan keunggulan atau kelemahan pada SVM maupun NB, sehingga analisis performa keduanya dapat dilakukan secara lebih komprehensif.

5. Pemilihan Model

   Model yang dipilih adalah dua algoritma klasik namun unggul dalam analisis sentimen: *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB). Pemilihan algoritma SVM dan Naïve Bayes bukan hanya karena keduanya populer dalam analisis sentimen, tetapi juga karena perbedaan karakteristik kinerja yang memungkinkan perbandingan yang bermakna. SVM dikenal unggul menangani data berdimensi tinggi seperti TF–IDF, sedangkan NB terkenal efisien dan sederhana dengan asumsi probabilistik. Perbedaan pendekatan inilah yang menjadi dasar studi komparatif untuk menentukan model yang paling sesuai untuk analisis kebijakan publik.

6. Pelatihan Model

   Model dilatih dengan menggunakan dataset berlabel yang telah diproses sebelumnya. Pelatihan dilakukan dengan pendekatan *supervised learning*, di mana data dibagi ke dalam *train set* dan *test set*. Proses pelatihan dilakukan menggunakan *pipeline* Scikit-learn memastikan alur kerja konsisten bagi kedua algoritma, sehingga perbandingan performa dapat dilakukan secara

objektif. Tahap ini menekankan kesetaraan kondisi pelatihan agar hasil evaluasi mencerminkan kemampuan nyata dari SVM dan NB dalam memahami pola sentimen.

7. Evaluasi Model

   Setelah model dilatih model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, serta validasi *stratified k-fold*. Evaluasi bertujuan mengukur performa dan reliabilitas model dalam mengklasifikasikan polaritas sentimen. Tahapan evaluasi ini bertujuan menegaskan model mana yang paling konsisten dan dapat diandalkan dalam mengklasifikasikan sentimen publik terhadap kebijakan, sehingga layak direkomendasikan sebagai model utama dalam analisis kebijakan digital.

8. Penerapan Algoritma

   Setelah pelatihan dan evaluasi model selesai, algoritma SVM dan Naive Bayes akan diterapkan pada dataset yang lebih luas atau data baru untuk menguji kemampuan generalisasi kedua model. Hasil dari penerapan ini akan menunjukkan seberapa efektif kedua algoritma dalam membaca dan memahami sentimen publik terhadap Program Makan Bergizi Gratis.

9. Analisa dan Hasil

   Tahap ini menguraikan hasil perbandingan performa SVM dan NB berdasarkan metrik evaluasi. Analisis ini juga menafsirkan sentimen publik terhadap Program Makan Bergizi Gratis dan mengaitkannya dengan kemampuan model dalam menangkap pola opini masyarakat.