

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Media sosial YouTube telah bertransformasi menjadi ruang publik digital yang krusial bagi masyarakat Indonesia untuk menyuarakan aspirasi politik, salah satunya terlihat pada fenomena Demo DPR RI 2025. Melalui fitur kolom komentar, masyarakat memberikan beragam reaksi, mulai dari dukungan, kritik tajam, hingga pernyataan netral terhadap situasi tersebut. Namun, volume data komentar yang sangat besar dan bersifat tidak terstruktur membuat pemantauan opini publik secara manual menjadi tidak efisien. Oleh karena itu, diperlukan sebuah pendekatan Natural Language Processing (NLP) melalui analisis sentimen untuk mengklasifikasikan pola opini publik secara otomatis ke dalam kategori positif, negatif, dan netral sebagai bahan evaluasi terhadap kebijakan pemerintah.

Pemanfaatan teknik machine learning dalam analisis sentimen menjadi solusi utama untuk menangani data skala besar tersebut. Keberhasilan klasifikasi sangat bergantung pada pemilihan algoritma yang tepat serta kualitas data yang telah melalui proses preprocessing dan ekstraksi fitur. Dalam konteks ini, penggunaan kamus leksikon seperti Indonesian Sentiment Lexicon (InSet) berperan penting untuk memberikan label awal secara objektif sebelum data diproses oleh model. Tantangan utama dalam analisis sentimen teks media sosial terletak pada keberagaman gaya bahasa, penggunaan kata tidak baku, serta ambiguitas kalimat yang memerlukan algoritma klasifikasi dengan tingkat akurasi dan stabilitas yang tinggi.

Penelitian terdahulu telah banyak mengeksplorasi penggunaan berbagai algoritma untuk klasifikasi teks, namun hasil akurasi seringkali bervariasi tergantung pada domain data dan teknik optimasi yang digunakan. Beberapa studi menunjukkan bahwa algoritma berbasis probabilitas sangat cepat secara komputasi, sementara algoritma berbasis geometri ruang lebih unggul dalam menangani kompleksitas fitur teks. Namun, masih terdapat celah penelitian dalam menemukan parameter yang paling optimal untuk menangani data sentimen politik yang bersifat

dinamis dan cenderung tidak seimbang (*imbalanced*). Hal inilah yang mendasari perlunya dilakukan studi komparasi untuk melihat sejauh mana perbedaan performa antara algoritma dasar dengan algoritma yang telah dioptimasi.

Penelitian ini mengangkat perbandingan antara algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) karena keduanya memiliki karakteristik matematis yang bertolak belakang. *Naïve Bayes* dipilih sebagai perwakilan model probabilistik yang bekerja dengan asumsi independensi fitur yang kuat, menjadikannya efisien namun cenderung kaku dalam menangani hubungan antar-kata yang kompleks. Di sisi lain, SVM dipilih sebagai metode yang bekerja berdasarkan prinsip *Structural Risk Minimization* (SRM) yang memetakan data ke dalam ruang dimensi tinggi untuk mencari *hyperplane* pemisah optimal. Perbandingan ini difokuskan pada kemampuan kedua metode dalam memproses ekstraksi fitur *Bag-of-Words* pada dataset Demo DPR RI 2025 yang memiliki tingkat derau (*noise*) tinggi.

Fokus utama dari studi komparasi ini lebih mengarah pada pengembangan skema *enhanced SVM* melalui variasi *parameter tuning* untuk mengungguli performa standar *Naïve Bayes*. Melalui penggunaan *GridSearchCV*, penelitian ini melakukan eksperimen terhadap berbagai kombinasi *hyperparameter*, seperti nilai *Cost* (C) dan jenis *Kernel*, guna meningkatkan sensitivitas model terhadap kelas-kelas sentimen yang sulit dibedakan. Dengan melakukan optimasi tersebut, diharapkan penelitian ini dapat membuktikan bahwa pendekatan SVM yang ditingkatkan mampu memberikan akurasi yang lebih stabil dan kredibel dibandingkan model klasifikasi dasar. Hasil akhir dari penelitian ini diharapkan dapat memberikan rekomendasi algoritma terbaik bagi pengembangan sistem pemantauan opini publik di masa depan.

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan secara rinci di atas, maka perumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana Bagaimana performa perbandingan antara algoritma Support Vector Machine (SVM) dan Naïve Bayes dalam mengklasifikasikan opini publik terkait Demo DPR RI 2025 di YouTube?
2. Sejauh mana penggunaan Hyperparameter Tuning melalui metode GridSearchCV dapat meningkatkan akurasi dan stabilitas algoritma SVM dibandingkan dengan performa standar algoritma Naïve Bayes?
3. Bagaimana tingkat efektivitas ekstraksi fitur Bag-of-Words (BoW) saat diimplementasikan pada kedua algoritma tersebut dalam menangani karakteristik data teks media sosial yang memiliki tingkat derau (noise) tinggi? (default)?

1.3 Batasan Masalah

Agar pembahasan dalam penelitian ini lebih terarah dan tidak meluas dari topik utama, penulis menetapkan batasan-batasan masalah sebagai berikut:

1. Sumber Data: Data yang digunakan adalah komentar publik yang diambil dari tiga video YouTube populer video tersebut dikategorikan sebagai video populer dengan justifikasi bahwa ketiga video tersebut menempati peringkat teratas dalam hasil pencarian, memiliki intensitas interaksi yang tinggi dengan jumlah komentar kumulatif mencapai lebih dari 19.000 data mentah, yang membahas topik "Demo DPR RI 2025". Pengambilan data dilakukan menggunakan *YouTube Data API v3*.
2. *Preprocessing*: Tahapan pra-pengolahan data mencakup pembersihan (*cleaning*), penyeragaman huruf (*case folding*), normalisasi kata tidak baku, dan penghapusan kata hubung (*stopword removal*).
3. Algoritma: Penelitian ini hanya membandingkan dua algoritma klasifikasi, yaitu *Support Vector Machine* (SVM) dan *Multinomial Naive Bayes*.
4. Pelabelan: Proses pelabelan sentimen dilakukan secara otomatis menggunakan metode berbasis kamus (*Lexicon-Based*) dengan kamus InSet (*Indonesian Sentiment Lexicon*).

5. Tools: Implementasi program dilakukan menggunakan bahasa pemrograman Python dengan bantuan *library Scikit-Learn* di lingkungan Google Colab.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan

Tujuan dilaksanakannya penelitian ini adalah:

1. Menganalisis hasil perbandingan performa antara algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes* dalam mengklasifikasikan opini publik mengenai Demo DPR RI 2025 di media sosial YouTube.
2. Mengevaluasi peningkatan akurasi dan stabilitas model *Support Vector Machine* (SVM) melalui implementasi *Hyperparameter Tuning* menggunakan metode *GridSearchCV* dibandingkan dengan model standar *Naïve Bayes*.
3. Menguji efektivitas penggunaan teknik ekstraksi fitur *Bag-of-Words* (BoW) pada kedua algoritma klasifikasi dalam mengolah data teks yang memiliki kompleksitas bahasa dan derau (noise) yang tinggi.

1.4.2 Manfaat

Hasil penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Teoritis

Hasil penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan ilmu pengetahuan di bidang Informatika, khususnya *Data Science* dan *Natural Language Processing* (NLP). Temuan mengenai perbandingan performa SVM dan *Naive Bayes* dengan optimasi parameter ini dapat menjadi referensi empiris bagi peneliti selanjutnya yang menangani dataset teks media sosial yang tidak terstruktur.

2. Manfaat Praktis

Bagi pemerintah maupun instansi terkait, hasil analisis sentimen ini dapat digunakan sebagai bahan evaluasi berbasis data untuk memahami aspirasi dan tingkat kepuasan masyarakat. Bagi masyarakat umum, penelitian ini memberikan gambaran objektif mengenai peta opini publik yang berkembang di ruang digital.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai analisis sentimen pada media sosial telah menjadi topik yang sangat berkembang dalam beberapa tahun terakhir. Berbagai algoritma pembelajaran mesin diterapkan untuk membedah opini publik yang tersebar di platform digital. Berikut adalah uraian mendalam mengenai 15 penelitian terdahulu yang relevan dengan topik ini, di mana setiap penelitian memberikan wawasan unik mengenai metode dan karakteristik data.

Media sosial YouTube menyediakan ruang bagi masyarakat untuk menyampaikan opini melalui kolom komentar. Melimpahnya data teks pada platform tersebut memungkinkan dilakukannya analisis sentimen untuk mengetahui respons dan persepsi publik terhadap suatu fenomena secara lebih mendalam (Arita, 2025).

Penelitian yang dilakukan oleh Huwaida dkk. (2024) berfokus pada pemetaan opini publik terkait kebijakan strategis nasional. Dalam studi ini, peneliti menggunakan algoritma *Naive Bayes* untuk mengklasifikasikan komentar masyarakat menjadi sentimen positif dan negatif. Hasil penelitian menunjukkan bahwa algoritma ini cukup efektif dalam menangani data teks YouTube yang memiliki volume besar, dengan temuan bahwa sentimen negatif cenderung mendominasi diskusi akibat kekhawatiran publik terhadap dampak lingkungan dan ekonomi dari pemindahan ibu kota (Huwaida dkk., 2024).

Selanjutnya, Misrun dkk. (2023) dalam penelitiannya menyoroti penggunaan analisis sentimen dalam konteks politik elektoral. Penelitian ini menegaskan bahwa komentar YouTube merupakan sumber data yang kaya untuk memotret polarisasi opini masyarakat terhadap figur politik. Hasil klasifikasi menunjukkan bahwa *Naive Bayes* mampu memproses data dengan cepat dan memberikan gambaran umum mengenai dukungan serta kritik yang diterima oleh tokoh politik tersebut menjelang tahun pemilu (Chely Aulia Misrun dkk., 2023).

Fauzan (2023) melakukan mengevaluasi respons mahasiswa dan masyarakat umum terhadap program pendidikan pemerintah. Dengan memanfaatkan *Multinomial Naive Bayes*, penelitian ini berhasil mengidentifikasi pola kepuasan dan keluhan pengguna program. Salah satu temuan pentingnya adalah bahwa *preprocessing* data yang baik sangat mempengaruhi akurasi model *Naive Bayes*, terutama dalam menangani istilah-istilah akademik yang spesifik (Fauzan, 2023).

Dalam domain teknologi, Ardiansyah dkk. (2025) membandingkan kinerja dua algoritma untuk melihat respons masyarakat terhadap adopsi teknologi baru. Meskipun fokus utamanya adalah komparasi dengan *Random Forest*, penelitian ini memberikan wawasan penting mengenai tahapan *preprocessing* data komentar YouTube yang seringkali penuh dengan *spam* dan karakter tidak baku yang perlu dibersihkan secara intensif agar tidak bias (Ardiansyah dkk., 2025).

Harpizon dkk. (2022) mengeksplorasi analisis sentimen pada konten keagamaan Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes". Penelitian ini menunjukkan fleksibilitas algoritma *Naive Bayes* yang tidak hanya cocok untuk isu politik atau produk, tetapi juga efektif untuk mengklasifikasikan sentimen pada konten dakwah yang memiliki karakteristik bahasa yang lebih emosional dan bercorak religius (Al Rasyid Harpizon dkk., 2022).

Muhayat, Fauzi, dan Indra (2022) dalam penelitiannya menemukan bahwa SVM memiliki kemampuan generalisasi yang sangat baik dalam menangani data komentar video YouTube yang memiliki dimensi fitur yang tinggi (*high-dimensional space*). Studi ini menjadi landasan penting yang menunjukkan bahwa SVM bisa menjadi alternatif yang lebih tangguh (*robust*) dibandingkan metode probabilistik sederhana (Muhayat dkk., 2022).

Mola dkk. (2024) melakukan studi komparatif langsung yang sangat relevan dengan penelitian ini hasil eksperimen mereka menunjukkan bahwa SVM secara signifikan mengungguli *Naive Bayes* dalam hal akurasi dan *recall*. Peneliti menyimpulkan bahwa kemampuan SVM dalam menemukan *hyperplane* pemisah yang optimal menjadikannya lebih andal untuk memetakan sentimen publik terhadap lembaga legislatif yang seringkali ambigu (Mola dkk., 2024).

Penelitian fundamental oleh Firdaus, Asror, dan Herdiani (2021) memberikan kontribusi pada metode pelabelan data. Studi ini memvalidasi penggunaan pendekatan berbasis leksikon (*Lexicon-Based*) untuk bahasa Indonesia. Mereka membuktikan bahwa penggunaan kamus sentimen yang komprehensif (seperti InSet) dapat menggantikan pelabelan manual yang memakan waktu, terutama untuk dataset bervolume besar, dengan tingkat akurasi yang dapat dipertanggungjawabkan (Firdaus dkk., 2021).

Permana dkk. (2025) yang menunjukkan selisih akurasi tipis antara SVM dan Naïve Bayes pada domain esports menjadi urgensi utama dilakukannya pengujian kembali pada domain politik. Berbeda dengan industri game, isu Demo DPR RI 2025 memiliki karakteristik bahasa yang lebih kompleks, penuh sarkasme, dan penggunaan slang yang masif. Penggunaan kedua algoritma ini bertujuan untuk membuktikan apakah ketangguhan SVM yang dioptimasi (*enhanced*) melalui parameter tuning tetap mampu mengungguli Naïve Bayes saat dihadapkan pada tingkat derau (*noise*) dan ambiguitas data politik yang tinggi. Dengan demikian, penelitian ini memvalidasi sejauh mana perbedaan domain data dapat mempengaruhi signifikansi performa antara model berbasis probabilitas dan model berbasis geometri ruang. (Permana dkk., 2025).

Firsttama, Arifiyanti, dan Kartika (2024) mengangkat isu internasional dalam "Analisis Sentimen Komentar Youtube Konferensi Tingkat Tinggi G20 Menggunakan Metode *Naive Bayes*". Penelitian ini menyimpulkan bahwa metode *Naive Bayes* sangat cocok untuk klasifikasi teks pendek dengan respons cepat. Namun, penelitian ini juga mencatat keterbatasan algoritma tersebut ketika dihadapkan pada kalimat sarkasme atau sindiran halus yang sering muncul dalam diskusi politik tingkat tinggi (Firsttama dkk., 2024).

Siswono dkk. (2024) melakukan penelitian pada pentingnya penanganan ketidakseimbangan kelas (*imbalanced dataset*) dalam data sentimen politik. Mereka menemukan bahwa *Naive Bayes* cenderung bias ke kelas mayoritas jika tidak dilakukan penyeimbangan data, sebuah wawasan yang penting untuk desain metodologi penelitian selanjutnya (Prayoga Siswono dkk., 2024).

Supian dkk. (2024) memberikan bukti kuat mengenai superioritas SVM. Dalam studi ini, SVM mencatatkan akurasi sebesar 94%, jauh melampaui *Naive Bayes* yang berhenti di angka 91%. Peneliti menyimpulkan bahwa SVM lebih stabil dalam menangani *noise* dan data *sparse* yang merupakan karakteristik umum dari data teks media sosial seperti Twitter dan YouTube (Supian dkk., 2024).

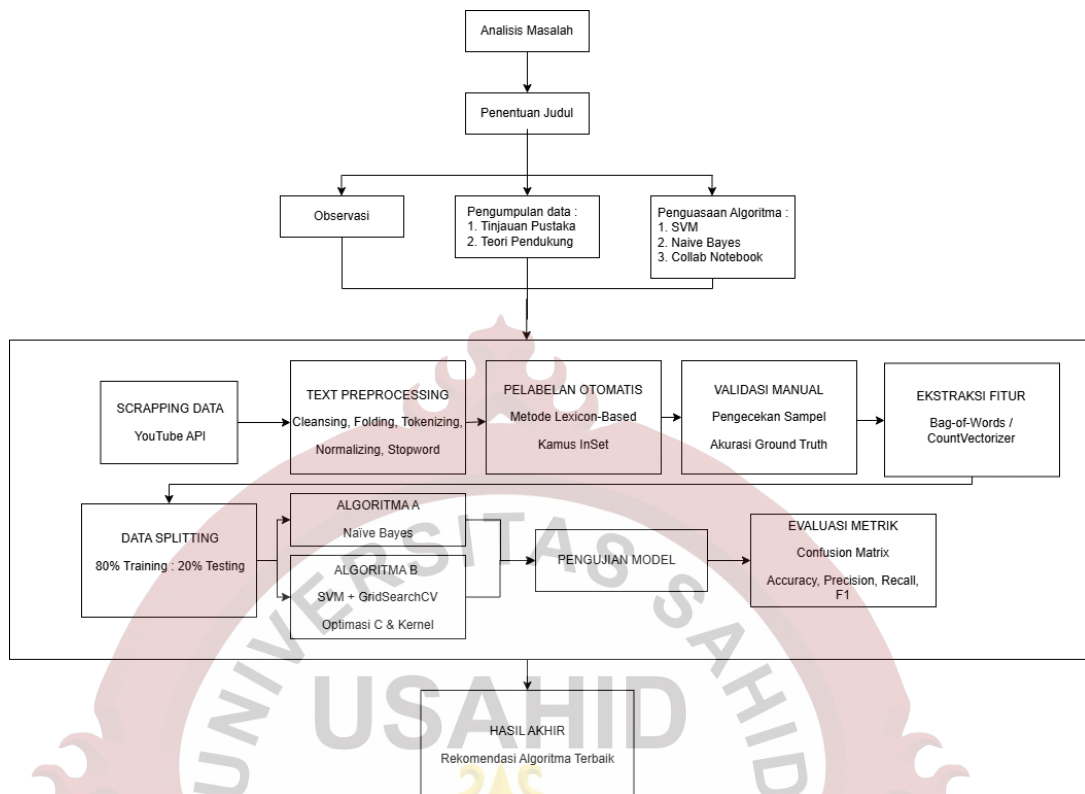
Nugroho dan Susanto (2025) menekankan aspek pemantauan *real-time*. Penelitian ini menunjukkan bahwa *Naive Bayes* adalah pilihan yang tepat untuk sistem yang membutuhkan kecepatan klasifikasi tinggi untuk memantau tren opini publik yang berubah cepat selama masa pemilihan umum (Wahyu Nugroho & Susanto, 2025).

Badri dkk. (2023) meneliti dampak kebijakan ekonomi hasil penelitian mereka menyoroti tantangan utama dalam analisis sentimen di Indonesia, yaitu penggunaan bahasa tidak baku. Mereka membuktikan bahwa teknik pra-pengolahan data yang tepat, khususnya normalisasi kata *slang*, adalah kunci utama untuk meningkatkan akurasi model *Naive Bayes* secara signifikan (Badri dkk., 2023).

SVM bekerja dengan mencari *hyperplane* optimal yang memisahkan antar kelas sentimen. Keunggulan utama dari metode ini adalah kemampuannya dalam menangani data berdimensi tinggi dan meminimalisir kesalahan klasifikasi pada dataset yang kompleks (Setiawati dkk., 2025).

2.2 Kerangka Pemikiran

Kerangka pemikiran penelitian ini menggambarkan alur kerja yang sistematis dan terstruktur dalam melaksanakan penelitian, mulai dari tahap awal hingga penarikan kesimpulan. Kerangka ini berfungsi sebagai panduan untuk memastikan bahwa setiap tahapan penelitian dilakukan secara logis dan runtut, sebagaimana ditunjukkan pada Gambar 2.1 berikut:



Gambar 2. 1 Kerangka Pemikiran

1. Analisis Masalah

Penelitian ini didasari oleh munculnya fenomena media sosial sebagai wadah penyaluran aspirasi politik, salah satunya pada isu "Demo DPR RI 2025". Analisis masalah dilakukan untuk mengidentifikasi kompleksitas data teks yang dihasilkan, seperti volume data yang sangat besar, penggunaan bahasa yang tidak baku, sarkasme, serta adanya *data imbalance* (ketidakseimbangan jumlah sentimen). Hal ini menciptakan tantangan bagi algoritma klasifikasi standar dalam memberikan prediksi yang akurat.

2. Penentuan Judul

Berdasarkan identifikasi masalah, ditentukanlah judul penelitian: "Perbandingan Kinerja Algoritma *Support Vector Machine* (SVM) dan *Naive Bayes* untuk Analisis Sentimen". Penentuan judul ini bertujuan untuk memberikan batasan yang jelas bahwa fokus utama penelitian adalah pada aspek komparasi metrik performa antara model berbasis probabilitas dan model berbasis geometri ruang pada domain data politik.

3. Observasi

Tahap observasi dilakukan dengan menelaah secara langsung interaksi pengguna pada kolom komentar video terkait di YouTube. Langkah ini bertujuan untuk memahami karakteristik linguistik dataset, seperti penggunaan singkatan, kata-kata slang, serta dialek tertentu yang dominan muncul. Observasi ini menjadi landasan dalam menentukan strategi *preprocessing* dan penyusunan kamus normalisasi kata.

4. Pengumpulan Data (Studi Literatur)

Selain data primer, peneliti melakukan pengumpulan referensi dari jurnal ilmiah, buku, dan penelitian terdahulu mengenai *Natural Language Processing* (NLP). Tahap ini berfungsi untuk membangun fondasi teoretis yang kuat mengenai algoritma SVM dan *Naive Bayes*, serta teknik evaluasi model yang diakui secara akademis.

5. Penguasaan Algoritma (SVM, Naive Bayes, & Tools)

Peneliti melakukan pendalaman teknis terhadap logika matematika di balik algoritma *Naive Bayes* dan *Support Vector Machine* (pencarian *hyperplane* optimal). Selain algoritma, tahap ini juga mencakup penguasaan perangkat lunak Google Colaboratory dan pustaka Python seperti *Scikit-Learn* yang digunakan sebagai mesin pengolah data.

6. *Scrapping* Dataset (YouTube Data API v3)

Proses pengambilan data dilakukan menggunakan metode *web scraping* melalui API resmi YouTube. Tahap ini memungkinkan pengambilan data secara legal dan terstruktur. Peneliti mengekstraksi metadata komentar dalam skala besar (>19.000 data) dari video-video populer untuk memastikan bahwa sampel data yang diolah memiliki tingkat representasi opini publik yang tinggi.

7. *Text Preprocessing*

Ini merupakan tahap krusial untuk mentransformasi teks mentah menjadi data yang bersih (*clean data*). Prosesnya meliputi:

- *Cleansing & Case Folding*: Menghapus elemen non-teks (URL, simbol, angka) dan menyamakan format huruf.
- *Tokenization*: Pemecahan kalimat menjadi unit kata.
- *Normalization*: Mengubah kata tidak baku menjadi kata formal sesuai KBBI.
- *Stopword Removal*: Membuang kata-kata yang tidak memiliki nilai informatif terhadap sentimen.
- *Stemming*: Mengembalikan kata ke bentuk dasar untuk mengurangi variasi kosakata.

8. Pelabelan Data (*Lexicon-Based*)

Setiap data komentar diberikan label kelas (Positif, Negatif, atau Netral) menggunakan metode *Lexicon-Based*. Proses ini melibatkan Kamus InSet (*Indonesian Sentiment Lexicon*) yang memberikan skor sentimen pada setiap kata. Hasil akumulasi skor tersebut menentukan kategori akhir dari sebuah komentar secara otomatis.

9. Validasi Manual (*Quality Control*)

Untuk memastikan validitas data, dilakukan pengecekan manual pada sebagian sampel dataset yang telah dilabeli secara otomatis. Validasi ini bertujuan untuk mengoreksi kesalahan pelabelan otomatis yang mungkin terjadi akibat penggunaan sarkasme atau konteks kalimat yang kompleks, sehingga *ground truth* yang digunakan untuk melatih model benar-benar akurat.

10. Ekstraksi Fitur (*Bag-of-Words*)

Data teks yang telah bersih diubah menjadi format numerik menggunakan teknik *Bag-of-Words* (*CountVectorizer*). Proses ini memetakan frekuensi kemunculan kata ke dalam matriks angka, sehingga algoritma *machine learning* dapat melakukan komputasi matematika terhadap pola-pola sentimen tersebut.

11. Data *Splitting*

Dataset dibagi secara acak namun proporsional menjadi dua bagian. 80% Data Latih digunakan oleh algoritma untuk mempelajari karakteristik dan pola dari setiap label sentimen, sedangkan 20% Data Uji digunakan sebagai simulasi data baru untuk menguji kecerdasan model setelah proses pelatihan selesai.

12. Tahap Pembangunan Model (SVM & *Naive Bayes*)

Pembangunan model dilakukan secara paralel. Untuk *Naive Bayes*, model dibangun menggunakan asumsi independensi antar fitur. Sementara untuk Enhanced SVM, model dibangun dengan menerapkan GridSearchCV. Ini adalah proses pencarian parameter optimal (C , γ , dan $kernel$) secara otomatis guna menghasilkan garis pemisah (*hyperplane*) paling efektif untuk membedakan kategori sentimen.

13. Tahap Pengujian Model

Model yang telah dilatih kemudian diuji dengan memberikan input berupa 20% data uji. Pada tahap ini, model diminta untuk memprediksi label sentimen tanpa melihat "kunci jawaban" aslinya. Kemampuan model dalam memberikan prediksi yang benar pada tahap inilah yang akan dinilai sebagai performa sistem.

14. Evaluasi Metrik (*Confusion Matrix*)

Evaluasi dilakukan menggunakan Confusion Matrix untuk membandingkan label aktual dengan label prediksi. Dari matriks ini dihasilkan nilai:

- *Accuracy*: Persentase ketepatan prediksi secara umum.
- *Precision*: Tingkat ketepatan antara data yang diminta dengan hasil prediksi.
- *Recall*: Tingkat keberhasilan model dalam menemukan kembali informasi.
- *F1-Score*: Rata-rata harmonis antara *precision* dan *recall*.

15. Hasil Akhir

Tahap terakhir adalah melakukan analisis terhadap hasil perbandingan metrik dari kedua algoritma. Peneliti menyimpulkan algoritma mana yang paling unggul untuk dataset Demo DPR RI 2025 Hasil ini menjadi rekomendasi ilmiah bagi pengembangan sistem analisis sentimen di masa mendatang.

2.3 Landasan Teori

2.3.1 Analisis Sentimen (*Sentiment Analysis*)

Analisis sentimen, atau sering disebut sebagai *opinion mining*, adalah cabang dari *Natural Language Processing* (NLP) yang bertujuan untuk mengekstrak, mengidentifikasi, dan mengategorikan informasi subjektif dari teks sumber. Proses ini melibatkan komputasi linguistik untuk menentukan apakah opini yang terkandung dalam dokumen, kalimat, atau aspek tertentu bernilai positif, negatif, atau netral. Sebagaimana dijelaskan oleh Nugroho dkk.(2025), analisis sentimen menjadi alat krusial bagi pemerintah maupun instansi terkait untuk memantau aspirasi publik secara *real-time* di media sosial tanpa harus melakukan survei manual yang memakan waktu lama.

2.3.2 Pra-pengolahan Data Text (*Text Preprocessing*)

Data teks yang berasal dari komentar media sosial bersifat sangat kotor (noisy) dan tidak terstruktur. Tahapan pra-pengolahan data teks adalah serangkaian teknik yang bertujuan untuk membersihkan dan menstandarisasi data mentah agar algoritma pembelajaran mesin dapat mengekstraksi fitur secara optimal (Buntoro dkk., 2023). Tanpa pra-pengolahan yang memadai, model akan mengalami kesulitan dalam mengenali pola kata akibat adanya ambiguitas dan redundansi data. Langkah-langkah kritis dalam tahap ini meliputi:

1. *Cleaning* Menghapus elemen non-informatif seperti URL, tanda baca, dan angka, serta menyeragamkan seluruh teks menjadi huruf kecil (*lowercase*) (Rizkiyani dkk., 2024).
2. *Case Folding* : merupakan tahapan awal yang fundamental dalam pra-pengolahan data teks (*text preprocessing*) yang bertujuan untuk menyeragamkan seluruh karakter huruf dalam dataset menjadi huruf kecil

(*lowercase*). Proses ini sangat krusial dalam analisis data tekstual karena algoritma komputer bersifat *case-sensitive*, yang artinya komputer memperlakukan karakter huruf kapital dan huruf kecil sebagai entitas digital yang berbeda secara kode ASCII/Unicode (Buntoro dkk., 2023).

3. **Normalisasi** : Proses krusial dalam bahasa Indonesia untuk mengonversi bahasa slang, singkatan, dan *typo* menjadi bentuk baku sesuai kaidah tata bahasa agar dapat dikenali oleh kamus sentimen maupun model klasifikasi.
4. **Stopword Removal** : Memecah kalimat menjadi satuan kata (*token*) dan mengeliminasi kata-kata fungsional yang memiliki frekuensi kemunculan tinggi namun tidak bermakna sentimen (seperti "dan", "yang", "di") (Hidayatullah dkk., 2022).

2.3.3 Ekstraksi Fitur (*Bag-of-Word*)

Metode *Bag-of-Words* (BoW) dipilih dalam penelitian ini karena kemampuannya dalam menyederhanakan representasi data teks yang kompleks menjadi format numerik yang efisien untuk diolah oleh algoritma klasifikasi (Kurniawan dkk., 2019). BoW bekerja dengan cara mengekstraksi kosakata unik dari seluruh korpus untuk membentuk sebuah kamus fitur, kemudian menghitung frekuensi kemunculan setiap kata tersebut di dalam setiap baris komentar. Pemilihan metode ini didasarkan pada karakteristik komentar YouTube yang cenderung singkat dan padat opini, di mana keberadaan kata kunci tertentu jauh lebih krusial untuk menentukan polaritas sentimen dibandingkan dengan mempertimbangkan urutan kata atau struktur tata bahasa yang kompleks (Rizkiyani dkk., 2024). Selain itu, representasi data berupa angka bulat (*integer*) yang dihasilkan oleh teknik *CountVectorizer* pada BoW sangat selaras dengan prinsip probabilitas pada algoritma *Multinomial Naive Bayes* dan ketangguhan dimensi pada SVM, sehingga memungkinkan proses pelatihan model berjalan lebih stabil dan cepat meskipun melibatkan dataset berskala besar (Sari dkk., 2019).

2.3.3 Perlabelan Berbasis Leksikon (*Lexicon-Based*)

Pendekatan leksikon menentukan nilai polaritas sebuah dokumen berdasarkan akumulasi bobot kata yang ditemukan dalam kamus sentimen. Metode ini bersifat transparan karena setiap prediksi didasarkan pada daftar kata yang telah diverifikasi (Taboada dkk., 2021). Dalam penelitian ini, digunakan kamus InSet yang dikembangkan khusus untuk bahasa Indonesia. Proses perhitungan skor polaritas secara matematis dirumuskan pada Persamaan 2.1 sebagai berikut:

$$Score = \sum_{i=1}^n W_{pos} - \sum_{j=1}^m W_{neg} \quad (2.1)$$

Dimana :

- $Score$: Nilai total polaritas sentimen kalimat.
- W_{pos} : Bobot kata positif yang ditemukan dalam kalimat.
- W_{neg} : Bobot kata negatif yang ditemukan dalam kalimat.
- $\sum$: Simbol penjumlahan dari seluruh kata yang cocok dengan kamus.

Jika hasil $Score > 0$, maka kalimat diklasifikasikan sebagai Positif. Jika $Score < 0$, diklasifikasikan sebagai Negatif, dan jika $Score = 0$, diklasifikasikan sebagai Netral (Birjali dkk., 2021).

2.3.4 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin berbasis *supervised learning* yang bekerja dengan cara membangun sebuah *hyperplane* atau bidang pemisah di dalam ruang dimensi tinggi. Tujuan utama SVM adalah menemukan garis pemisah yang memiliki jarak terjauh (*maximum margin*) antara dua kelas data yang berbeda (Cortes & Vapnik, dalam tinjauan modern oleh Wankhade dkk., 2022). SVM dikenal sangat unggul dalam menangani data teks karena ketahanannya terhadap *overfitting* dan efisiensinya pada data dengan dimensi besar. Persamaan keputusan klasifikasi linear SVM didefinisikan pada Persamaan 2.2 sebagai berikut:

$$f(x) = \text{sign}(w \cdot x + b) \quad (2.2.)$$

Dimana :

- $f(x)$: Hasil keputusan klasifikasi (output kelas sentimen, misal +1 atau -1).
- sign : Fungsi untuk menentukan tanda positif atau negatif dari hasil

perhitungan.

- w : Vektor bobot (*weight vector*) yang menentukan orientasi *hyperplane*.
- x : Vektor input data (nilai fitur kata dari Bag-of-Words).
- b : Nilai bias (*intercept*) untuk menggeser posisi *hyperplane* agar optimal.

Prinsip utama SVM adalah menemukan nilai w dan b yang memaksimalkan jarak (margin) antara *hyperplane* dengan titik data terdekat (*support vectors*) dari masing-masing kelas.

2.3.5 Algoritma Naïve Bayes

Naïve Bayes adalah algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi antar fitur. Meskipun asumsi ini seringkali dianggap terlalu sederhana (*naïve*), algoritma ini terbukti sangat efisien dan cepat dalam menangani klasifikasi teks berskala besar (Zhang dkk., 2021). Varian Multinomial Naïve Bayes sering dipilih untuk analisis sentimen karena kemampuannya memodelkan frekuensi kata secara efektif. Rumus probabilitas posteriornya ditunjukkan pada Persamaan 2.3 sebagai berikut:

$$P(c|d) = \frac{(P(c) \prod_{i=1}^n P(w_i|c))}{P(d)} P(d) \quad (2.3)$$

Dimana :

- $P(c|d)$: Peluang dokumen d masuk ke dalam kelas c (Posterior Probability).
- $P(c)$: Peluang prior kemunculan kelas c dalam dataset (Prior Probability).
- $P(w_i|c)$: Peluang kemunculan kata w_i di dalam kelas c (Likelihood).
- $P(d)$: Probabilitas dokumen d (Evidence), yang bernilai konstan untuk semua kelas.

2.3.6 Optimasi Model dengan Grid Search Cross Validation

Grid Search Cross Validation (GridSearchCV) merupakan teknik optimasi hyperparameter yang bekerja dengan cara mencoba seluruh kombinasi parameter yang telah ditentukan secara sistematis untuk menemukan performa terbaik pada sebuah model (Wankhade dkk., 2022). Dalam algoritma *Support Vector Machine* (SVM), pemilihan parameter seperti nilai C (Cost) dan jenis kernel sangat menentukan akurasi akhir.

GridSearchCV melakukan evaluasi terhadap setiap kombinasi parameter tersebut menggunakan metode cross-validation. Menurut Buntoro dkk. (2023), penggunaan GridSearchCV dalam penelitian analisis sentimen bertujuan untuk menghilangkan unsur subjektivitas peneliti dalam memilih parameter, sehingga model SVM yang dihasilkan merupakan model yang paling optimal secara matematis bagi dataset yang digunakan.

2.3.7 K-Fold Cross Validation

K-Fold Cross Validation adalah metode evaluasi statistik yang digunakan untuk menilai kemampuan generalisasi sebuah model pembelajaran mesin. Teknik ini bekerja dengan membagi dataset menjadi k bagian (atau fold) yang berukuran sama. Dalam penelitian ini, digunakan 5-Fold Cross Validation, di mana dataset dibagi menjadi 5 bagian: 4 bagian digunakan sebagai data latih dan 1 bagian digunakan sebagai data uji (Buntoro dkk., 2023).

Proses ini dilakukan sebanyak 5 kali putaran (iterasi), sehingga setiap *fold* pernah merasakan menjadi data uji satu kali. Keuntungan utama dari metode ini adalah mengurangi variansi pada hasil evaluasi dan memastikan bahwa performa model (seperti akurasi 90,17%) tidak hanya terjadi secara kebetulan pada satu set pembagian data saja, melainkan konsisten pada seluruh distribusi data (Zhang dkk., 2021).

2.3.8 Evaluasi Model

Pengukuran kinerja model dilakukan untuk menilai sejauh mana algoritma mampu melakukan klasifikasi secara akurat pada data yang belum pernah dilihat sebelumnya. Metrik evaluasi dihitung menggunakan *Confusion Matrix* yang memetakan hasil prediksi model terhadap data aktual (Buntoro dkk., 2023).

Pengukuran kinerja model dalam penelitian ini menggunakan empat metrik utama. Pertama adalah Akurasi untuk mengukur rasio prediksi benar secara keseluruhan sebagaimana terdapat pada Persamaan 2.4 Kedua, Presisi yang mengukur ketepatan hasil prediksi sistem sesuai Persamaan 2.5 Ketiga, Recall untuk mengukur keberhasilan penemuan kembali informasi sesuai Persamaan 2.6 Terakhir, untuk melihat keseimbangan antara presisi dan recall, digunakan F1-Score yang ditunjukkan pada Persamaan 2.7.

- 1) Akurasi (*Accuracy*) Mengukur rasio prediksi benar (positif dan negatif) dari keseluruhan data.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.4)$$

- 2) Presisi (*Precision*) Mengukur tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$Presisi = \frac{TP}{TP+FP} \quad (2.5)$$

- 3) Recall Mengukur tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

$$Recall = \frac{TP}{TP+FN} \quad (2.6)$$

- 4) F1-Score Merupakan rata-rata harmonik (harmonic mean) dari presisi dan recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.7)$$

Dimana :

- *TP* : True Positive
- *TN* : True Negative
- *FP* : False Positive
- *FN* : False Negative

Selain akurasi, penelitian ini juga menggunakan metrik Presisi, Recall, dan F1-Score untuk mendapatkan gambaran performa yang lebih komprehensif, terutama jika terdapat ketidakseimbangan kelas pada dataset.