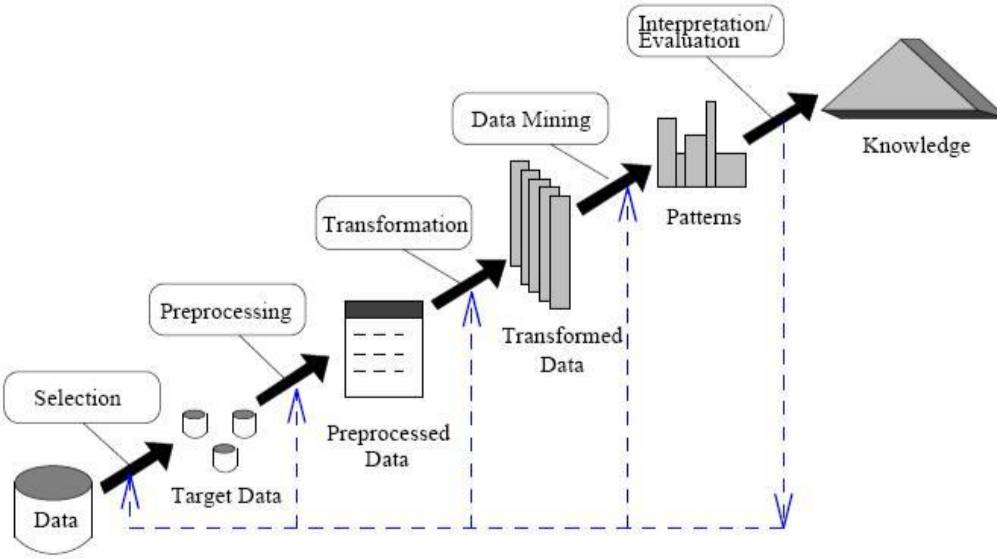
BAB II
LANDASAN TEORI

2.1. *Knowledge Discovery in Databases* (KDD)

Knowledge Discovery in Databases (KDD) adalah proses kompleks yang melibatkan serangkaian langkah sistematis untuk mengekstrak pengetahuan yang berguna dari data yang besar dan kompleks. Dalam dunia yang semakin bergantung pada data, KDD memainkan peran penting dalam membantu organisasi mengubah data mentah menjadi informasi yang dapat digunakan untuk pengambilan keputusan yang lebih baik

Tujuan utama dari KDD adalah untuk mengubah data menjadi pengetahuan yang berguna. Data yang kita miliki, terutama dalam jumlah besar, sering kali tampak acak atau tidak terstruktur. Dengan menggunakan teknik-teknik yang terorganisir, KDD bertujuan untuk menggali informasi yang relevan dan membuatnya dapat dipahami serta diterapkan dalam pengambilan keputusan. Dalam era digital saat ini, data dihasilkan dalam jumlah yang sangat besar dan sering kali sulit untuk dipahami secara langsung[1].

KDD bukan hanya sekadar satu langkah kerja, tetapi merupakan serangkaian tahapan yang terintegrasi. Proses ini sering digunakan di berbagai bidang seperti bisnis, kesehatan, pendidikan, dan teknologi, karena dapat memberikan wawasan yang membantu meningkatkan kinerja dan efisiensi sistem [2].



Gambar 2. 1. Tahap-tahap dalam *Knowledge Discovery in Databases* (KDD)

Sumber : Hermawati [3]

Proses KDD (*Knowledge Discovery in Databases*) adalah serangkaian tahapan yang digunakan untuk mengekstrak pengetahuan dari data besar dan kompleks. Proses ini terdiri dari beberapa langkah utama, dan setiap langkah memiliki tujuan serta tugas yang spesifik. Berikut adalah penjelasan pada setiap tahapan dalam proses KDD:

1. Pemilihan Data (Data Selection) Pada tahap ini, data yang relevan dari kumpulan data besar. Data yang dipilih harus memiliki nilai atau potensi untuk memberikan informasi yang berguna. Misalnya, jika analisis berfokus pada penjualan produk, maka data yang relevan mencakup informasi tentang transaksi penjualan, pelanggan, dan produk yang terlibat[4].
2. Pra-pemrosesan Data (Data Preprocessing) Data yang digunakan dalam KDD sering kali mengandung noise, data hilang, atau inkonsistensi yang

dapat memengaruhi kualitas hasil analisis. Oleh karena itu, tahap pra-pemrosesan bertujuan untuk membersihkan data agar siap digunakan. Ini termasuk menangani data yang hilang dengan pengisian atau penghapusan, mengidentifikasi dan menangani outlier, serta merubah format data untuk keseragaman. Tahap ini juga mencakup transformasi data, seperti normalisasi (agar data berada dalam skala yang sama) atau konversi data dalam format yang lebih mudah dianalisis. Selain itu, pada tahap ini, penting untuk memilih atribut yang relevan dan menghapus atribut yang tidak memberi informasi signifikan terhadap model yang akan dibangun.

3. Transformasi Data (Data Transformation) Setelah data diproses, transformasi dilakukan untuk memperbaiki struktur dan bentuk data yang lebih sesuai dengan analisis. Beberapa teknik yang digunakan dalam transformasi data termasuk penggabungan atribut (misalnya, menggabungkan beberapa kolom menjadi satu fitur yang lebih representatif), reduksi dimensi (seperti PCA, yang mengurangi jumlah fitur tetapi tetap mempertahankan informasi utama), dan encoding data untuk mengubah kategori menjadi format numeric [5].
4. Penambangan Data (*Data Mining*) Penambangan data adalah inti dari proses KDD, di mana teknik-teknik analitis diterapkan untuk menemukan pola atau pengetahuan yang tersembunyi dalam data. Di sinilah *Algoritma* pembelajaran mesin dan statistik digunakan untuk menganalisis data secara mendalam.

5. Evaluasi Pengetahuan (Evaluation of Knowledge**)** Setelah menemukan pola atau model yang potensial, penting untuk mengevaluasi sejauh mana hasil penambangan data tersebut relevan dan berguna. Evaluasi dilakukan dengan menguji kinerja model yang dihasilkan, menggunakan teknik evaluasi statistik untuk menilai keakuratan, presisi, recall, dan lainnya. Selain itu, perlu ada validasi terhadap model untuk memastikan bahwa pengetahuan yang diperoleh dapat diandalkan dan berlaku untuk data yang lebih luas.
6. Presentasi Pengetahuan (Knowledge Presentation) Tahap terakhir dalam KDD adalah menyajikan hasil pengetahuan yang diperoleh dalam bentuk yang mudah dipahami oleh pengambil keputusan. Ini melibatkan penggunaan visualisasi data, grafik, dan dashboard untuk menggambarkan pola atau tren yang ditemukan selama proses penambangan data. Hasil dari KDD harus disajikan dalam format yang jelas, dan interpretasi pengetahuan yang diperoleh harus relevan dengan kebutuhan dan tujuan pengguna.

2.2. *Data Mining*

Data Mining adalah proses eksplorasi data dalam jumlah besar untuk menemukan pola, tren, atau hubungan yang tersembunyi yang dapat digunakan untuk membuat keputusan yang lebih baik. Dalam dunia yang semakin terdigitalisasi dan data-driven, *Data Mining* menjadi salah satu disiplin ilmu yang sangat penting dalam berbagai sektor, mulai dari bisnis hingga penelitian ilmiah. *Data Mining* menggabungkan teknik-teknik dari bidang ilmu komputer, statistik, dan kecerdasan buatan untuk menganalisis data dan mengekstrak pengetahuan yang berharga. *Data Mining* memungkinkan untuk menemukan tren, pola, dan prediksi

masa depan, yang dapat diaplikasikan di berbagai bidang, termasuk bisnis, kesehatan, dan pendidikan. Dalam konteks ini, *Data Mining* membantu dalam mengambil keputusan strategis yang lebih efektif dan efisien, serta memberikan wawasan yang mendalam mengenai perilaku pengguna atau konsumen[6].

2.3. Analisis Sentimen

Analisis sentimen merupakan suatu proses yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini atau perasaan yang terkandung dalam data tekstual. Teknik ini memiliki aplikasi yang luas, terutama dalam evaluasi kepuasan pelanggan terhadap produk atau layanan. Dalam konteks ini, analisis sentimen digunakan untuk menggali perasaan pengguna terhadap pengalaman mereka dengan suatu produk, baik itu positif, negatif, maupun netral [7]. Sebagai contoh, platform seperti ulasan aplikasi atau media sosial sering kali menjadi sumber utama data yang digunakan dalam analisis sentimen. Dengan mengkategorikan sentimen yang diekstrak, perusahaan dapat memperoleh wawasan yang berharga mengenai persepsi pelanggan mereka serta mengidentifikasi area yang memerlukan perbaikan[8]

2.3.1. *Algoritma* dalam analisis sentimen

Dalam analisis sentimen, berbagai *Algoritma* telah dikembangkan untuk meningkatkan akurasi dan efisiensi dalam proses klasifikasi sentimen. Beberapa metode yang sering digunakan antara lain adalah Support Vector Machine (SVM) dan *Naïve Bayes* (NB), yang masing-masing memiliki kelebihan dan kekurangan dalam konteks analisis data teks[9]

SVM merupakan *Algoritma* yang kuat dalam klasifikasi data, terutama dalam memisahkan kelas dengan margin yang maksimal. Dengan memanfaatkan konsep hyperplane, SVM dapat bekerja dengan sangat baik dalam situasi di mana data bersifat non-linear dan memerlukan pemisahan yang jelas antara kelas [10]. Sementara itu, *Naïve Bayes* adalah metode statistik yang didasarkan pada teori probabilitas, yang mengklasifikasikan teks berdasarkan distribusi probabilistik kata-kata dalam dokumen. Meskipun sering dianggap lebih sederhana, metode ini dapat memberikan hasil yang memadai dengan kecepatan pemrosesan yang tinggi dan memerlukan sedikit data pelatihan [11]. Berdasarkan penelitian, kedua *Algoritma* ini memiliki kekuatan tersendiri, dengan *Naïve Bayes* lebih efisien dalam hal waktu komputasi, sementara SVM cenderung menghasilkan akurasi yang lebih tinggi pada dataset yang lebih besar dan kompleks.

Selain kedua metode tersebut, terdapat pula *Algoritma-Algoritma* lain yang dapat digunakan untuk analisis sentimen, seperti Bidirectional Long Short-Term Memory (BiLSTM). *Algoritma* ini menggabungkan kemampuan Recurrent Neural Networks (RNN) dengan arsitektur dua arah untuk meningkatkan kemampuan analisis konteks dalam kalimat [12]. Penggunaan BiLSTM dapat lebih efektif dalam menangani teks yang memiliki struktur lebih kompleks, misalnya dalam bentuk kalimat yang panjang atau ambigu.

Tabel 2. 1. Metode *Algoritma* untuk Analisis Sentimen

Algoritma	**Kelebihan**	**Kekurangan**
Naïve Bayes	Cepat dan efisien, mudah diimplementasikan	Akurasi rendah pada data besar atau data dengan fitur yang kompleks.

Support Vector Machine (SVM)	Akurasi tinggi, efektif pada data non-linear dan besar.	Membutuhkan waktu komputasi yang lebih lama pada dataset besar.
BiLSTM	Mampu menangani konteks kalimat yang lebih kompleks dan panjang	Memerlukan sumber daya komputasi yang lebih besar dan pelatihan yang lebih lama.

2.4. *Algoritma Naïve Bayes*

Naïve Bayes adalah *Algoritma* dalam pembelajaran mesin (machine learning) yang digunakan untuk klasifikasi berdasarkan teori probabilitas Bayes. *Algoritma* ini bekerja dengan mengasumsikan bahwa setiap fitur atau atribut input (dalam hal ini kata atau frasa dalam teks) bersifat independen satu sama lain, yang dikenal dengan sebutan naïve assumption (asumsi sederhana). Meskipun asumsi independensi ini seringkali tidak realistis dalam data dunia nyata, *Naïve Bayes* tetap memberikan hasil yang baik dalam banyak kasus, terutama untuk klasifikasi teks, seperti analisis sentimen.

Metode ini sangat populer dalam analisis sentimen karena kemampuannya yang baik dalam mengklasifikasikan teks menjadi kategori tertentu, seperti positif, negatif, atau netral, dengan menggunakan probabilitas dari fitur yang ada.

2.4.1. Prinsip Kerja *Algoritma Naïve Bayes*

Prinsip dasar dari *Algoritma Naïve Bayes* adalah berdasarkan teorema Bayes, yang menyatakan bahwa probabilitas suatu hipotesis H yang diberikan oleh data D dapat dihitung dengan rumus berikut:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Dimana:

1. P(H|D) adalah probabilitas posterior, yaitu probabilitas bahwa hipotesis H (misalnya, sentimen positif) benar diberikan data D (ulasan atau teks).
2. P(D|H) adalah likelihood, yaitu probabilitas data D muncul, diberikan hipotesis H.
3. P(H) adalah prior probability, yaitu probabilitas awal hipotesis H tanpa mempertimbangkan data.
4. P(D) adalah probabilitas data D, yang sering kali dianggap konstan dalam proses klasifikasi.

Naïve Bayes mengasumsikan bahwa setiap fitur dalam data (misalnya, kata-kata dalam teks) adalah independen satu sama lain. Dalam konteks analisis sentimen, ini berarti bahwa *Naïve Bayes* menghitung probabilitas untuk setiap kata dalam teks dan menggabungkannya untuk menentukan apakah ulasan tersebut positif, negatif, atau netral. Langkah-langkah kerja *Naïve Bayes* dalam analisis sentimen adalah sebagai berikut:

1. Pelatihan model *Naïve Bayes* dilatih menggunakan dataset yang sudah diberi label (misalnya, ulasan pelanggan yang diberi label positif atau negatif). Setiap kata atau frasa yang muncul dalam teks dihitung frekuensinya, dan dihitung probabilitas kondisional kata tersebut muncul dalam masing-masing kelas (positif atau negatif).
2. Penghitungan Probabilitas untuk masing-masing kelas (misalnya, positif atau negatif) dihitung berdasarkan jumlah kata-kata yang ada di dalam teks.

Probabilitas ini dihitung dengan rumus teorema Bayes untuk setiap kelas yang ada.

3. Klasifikasi Untuk teks baru, *Naïve Bayes* menghitung probabilitas masing-masing kelas (positif, negatif, atau netral) berdasarkan kata-kata yang ada dalam teks tersebut. Kelas yang memiliki probabilitas tertinggi akan dipilih sebagai prediksi untuk teks tersebut.

2.4.2. Evaluasi Performa Model

Evaluasi performa model. Evaluasi ini menggunakan beberapa metrik untuk mengukur keakuratan model dalam mengklasifikasikan sentimen, seperti akurasi, presisi, recall, dan F1-Score. Metrik-metrik ini memberikan gambaran yang lebih jelas tentang seberapa baik model dalam melakukan prediksi yang benar dan seimbang.

Tabel 2. 2. Metrik Evaluasi Model

Metrik	Deskripsi
Akurasi	Mengukur seberapa banyak prediksi yang benar dibandingkan dengan total prediksi yang dibuat oleh model. Akurasi dihitung dengan rumus: (TP + TN) / (TP + TN + FP + FN)
Presisi	Menunjukkan seberapa banyak prediksi positif yang benar dibandingkan dengan total prediksi positif yang dibuat oleh model. Dihitung dengan rumus: TP / (TP + FP)
Recall	Mengukur seberapa banyak kasus positif yang benar berhasil diprediksi sebagai positif oleh model. Dihitung dengan rumus: TP / (TP + FN)

F1-Score	Kombinasi dari presisi dan recall, memberikan gambaran yang lebih baik mengenai keseimbangan antara keduanya. Dihitung dengan rumus: 2 * (Presisi * Recall) / (Presisi + Recall)

Keterangan:

1. TP (True Positive): Kasus positif yang benar-benar teridentifikasi sebagai positif.
2. TN (True Negative): Kasus negatif yang benar-benar teridentifikasi sebagai negatif.
3. FP (False Positive): Kasus negatif yang salah teridentifikasi sebagai positif.
4. FN (False Negative): Kasus positif yang salah teridentifikasi sebagai negatif.
5. Tabel ini merangkum metrik evaluasi yang digunakan dalam penelitian ini untuk menilai kinerja model analisis sentimen berbasis *Naïve Bayes*.

2.5. Alat Bantu Program/Tools Pendukung

Alat bantu program atau tools pendukung adalah perangkat lunak atau komponen tambahan yang digunakan untuk memperkuat atau mempermudah penggunaan suatu aplikasi atau sistem. Dalam konteks pengolahan data dan analisis, alat bantu program ini sangat penting karena membantu pengguna dalam menyelesaikan tugas tertentu dengan cara yang lebih efisien dan efektif. Alat bantu ini sering kali meliputi fitur tambahan yang berfokus pada berbagai aspek seperti pemrosesan data, visualisasi, analisis statistik, serta integrasi dengan sistem lain.

2.5.1. *RapidMiner* Sebagai Alat Bantu Analisis

RapidMiner adalah salah satu platform perangkat lunak yang sangat populer dalam bidang data science, terutama untuk analisis data dan pemodelan prediktif. *RapidMiner* memungkinkan pengguna untuk membangun, melatih, dan mengevaluasi model analisis sentimen tanpa perlu menulis kode secara manual, yang menjadikannya alat yang sangat berguna dalam analisis teks, seperti analisis sentimen terhadap kepuasan pelanggan.

Gambar 2. 2. Tampilan Aplikasi *RapidMiner*

Sumber: https://mindmajix.com/*RapidMiner*-training

2.5.2. Persiapan dan Pengaturan Data dalam *RapidMiner*

Sebelum memulai analisis, data yang akan digunakan dalam *RapidMiner* perlu dipersiapkan terlebih dahulu. Data ulasan pelanggan yang dikumpulkan (baik dari e-commerce, media sosial, atau survei pelanggan) perlu diproses agar dapat digunakan dalam model analisis sentimen.

1. Import Data

RapidMiner memungkinkan penggunanya untuk mengimpor data dalam berbagai format, seperti CSV, Excel, atau bahkan langsung dari basis data. Dalam penelitian ini, file yang berisi ulasan pelanggan yang telah dilabeli sentimen

(positif, negatif, netral) perlu diimpor ke dalam *RapidMiner*. Pada penelitian ini penulis mengimpor data dalam format Excel

Gambar 2. 3. Tampilan Aplikasi Excel

Sumber: https://tech.hitekno.com/read/2023/02/19/143522/20

2. *Preprocessing Data*

Setelah data diimpor, langkah pertama adalah membersihkan data menggunakan modul preprocessing yang tersedia di *RapidMiner*, yang meliputi:

1. Tokenization: Memecah kalimat menjadi kata-kata.
2. Penghapusan Stopwords: Menghapus kata-kata umum yang tidak memberikan informasi penting bagi analisis sentimen.
3. Stemming: Mengubah kata menjadi bentuk dasarnya untuk mengurangi variasi kata.
4. Normalisasi Teks: Mengubah teks menjadi bentuk standar, misalnya dengan mengubah huruf besar menjadi huruf kecil atau menghapus karakter yang tidak relevan.

RapidMiner juga menyediakan alat untuk mengubah data teks ke dalam format yang dapat digunakan dalam *Algoritma* klasifikasi, seperti TF-IDF (Term Frequency-Inverse Document Frequency), yang memberi bobot lebih besar pada

kata-kata yang jarang muncul di banyak dokumen tetapi sering muncul dalam dokumen tertentu.

2.6. Metodologi Penelitian

Metodologi penelitian adalah pendekatan sistematis yang digunakan untuk melakukan penelitian. Berikut adalah langkah-langkah yang umumnya dilakukan dalam penelitian ini:

1. Pengumpulan Data

Data yang digunakan berasal dari komentar atau ulasan pelanggan tentang layanan servis. Ulasan tersebut bisa didapatkan melalui survei langsung kepada pelanggan atau melalui platform online tempat pelanggan memberikan feedback.

2. Pengolahan Data

Sebelum data dapat dianalisis, data mentah yang dikumpulkan perlu diproses terlebih dahulu. Proses ini meliputi pembersihan data, pemisahan teks menjadi kata-kata yang lebih kecil, penghilangan kata-kata yang tidak relevan, serta penyeragaman bentuk kata. Langkah-langkah ini penting untuk mempersiapkan data agar siap untuk dianalisis lebih lanjut.

3. Modeling

Setelah data diproses, tahap berikutnya adalah membangun model klasifikasi untuk menentukan jenis sentimen dalam setiap ulasan. Teknik yang digunakan adalah *Naïve Bayes*, yang berfungsi untuk mengklasifikasikan teks ulasan ke dalam kategori sentimen tertentu, seperti positif atau negatif.

4. Evaluasi Kinerja Model

Untuk memastikan bahwa model yang dibangun dapat melakukan prediksi yang akurat, dilakukan evaluasi dengan menggunakan sebagian data yang tidak digunakan saat pelatihan. Hasil dari evaluasi ini diukur dengan beberapa metrik seperti akurasi, precision, dan recall.

5. Analisis Hasil

Setelah model diuji, hasil klasifikasi sentimen dianalisis. Dengan melihat distribusi sentimen, kita dapat mengetahui seberapa banyak pelanggan yang merasa puas, tidak puas terhadap layanan yang diberikan. Selain itu, dapat dilihat apakah ada hubungan antara jenis layanan dengan tingkat kepuasan pelanggan.

2.6.1. Penelitian Terdahulu

Tabel 2. 3. Penelitian Terdahulu

Referensi Penelitian	**1**
Peneliti	Hidayati & Purwanto
Tahun	2023
Judul Penelitian	Studi Sentimen terhadap Kepuasan Pelanggan Menggunakan *Algoritma Naïve Bayes*
Metode yang digunakan	*Naïve Bayes*
Tempat Studi / Objek Penelitian	E-commerce Produk Elektronik
Hasil Penelitian	Penelitian ini menunjukkan bahwa *Naïve Bayes* dapat mengklasifikasikan sentimen pelanggan dengan akurasi 86% pada produk elektronik, termasuk perangkat mobile.
Referensi Penelitian	**2**

Peneliti	Nugroho & Wijaya
Tahun	2020
Judul Penelitian	Klasifikasi Sentimen pada Ulasan Pengguna Menggunakan *Naïve Bayes*
Metode yang digunakan	*Naïve Bayes*
Tempat Studi / Objek Penelitian	Ulasan Produk E-commerce
Hasil Penelitian	Model *Naïve Bayes* digunakan untuk mengklasifikasikan ulasan produk. Hasil menunjukkan bahwa akurasi mencapai 90% dalam mengidentifikasi sentimen pelanggan terhadap produk.
Referensi Penelitian	**3**
Peneliti	Pratama & Dewi
Tahun	2021
Judul Penelitian	Penggunaan *Algoritma Naïve Bayes* dalam Menganalisis Sentimen pada Media Sosial
Metode yang digunakan	*Naïve Bayes*
Tempat Studi / Objek Penelitian	Media Sosial (Twitter)
Hasil Penelitian	Analisis sentimen terhadap tweet terkait suatu produk dengan *Naïve Bayes* menghasilkan akurasi 92%, menunjukkan metode ini cocok untuk analisis media sosial.
Referensi Penelitian	**4**
Peneliti	Arifin & Suryani
Tahun	2022

Judul Penelitian	Analisis Sentimen pada Ulasan Layanan Pelanggan dengan Metode *Naïve Bayes*
Metode yang digunakan	*Naïve Bayes*
Tempat Studi / Objek Penelitian	Layanan Pelanggan (call center)
Hasil Penelitian	Hasil penelitian menunjukkan bahwa *Naïve Bayes* efektif untuk analisis sentimen terhadap kepuasan pelanggan, dengan akurasi mencapai 87%.
Referensi Penelitian	**5**
Peneliti	Setyawan & Amalia
Tahun	2023
Judul Penelitian	Analisis Kepuasan Pelanggan dalam Layanan Servis HP Menggunakan *Naïve Bayes*
Metode yang digunakan	*Naive Bayes*
Tempat Studi / Objek Penelitian	Layanan Servis HP (Samsung Service Center)
Hasil Penelitian	Penelitian ini menunjukkan bahwa *Naïve Bayes* dapat digunakan untuk menganalisis kepuasan pelanggan terhadap layanan servis HP dengan akurasi model 88%.