BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

Penelitian Terdahulu 1	
Judul	Analisis Sentimen Ulasan Pengguna Aplikasi Google Play Menggunakan *Naïve Bayes*
Nama Penulis	Andriani Nurian, Betha Nurina Sari [31]
Jurnal	JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Tahun	2023
Permasalahan	Di era digital dengan meningkatnya penggunaan aplikasi mobile, penting untuk memahami kepuasan dan ketidakpuasan pengguna. Khususnya, aplikasi Dana yang menyediakan layanan keuangan, memiliki ulasan bervariasi yang perlu dianalisis untuk meningkatkan kualitas layanan.
Algoritma	*Naïve Bayes*
Temuan	Mayoritas ulasan pengguna terhadap aplikasi Dana positif (60%), dengan ulasan negatif sekitar 12% dan netral 28%. Akurasi tertinggi yang diperoleh adalah 85% dengan presisi 79%, *Recall* 85%, dan *F1-Score* 80%.
Pembahasan	Temuan menunjukkan respons positif dominan dari pengguna, namun adanya ulasan negatif menandakan area untuk perbaikan. Analisis sentimen mengungkapkan pentingnya feedback pengguna dalam pengembangan aplikasi. Permasalahan teknis dan layanan pelanggan adalah aspek yang sering dikeluhkan.
Relevansi	Studi ini menunjukkan penggunaan efektif *Naïve Bayes* dalam analisis sentimen ulasan aplikasi, mirip dengan pendekatan peneliti menggunakan SVM dan *Naïve Bayes* untuk ulasan TikTok. Ini menggarisbawahi pentingnya analisis sentimen dalam mengevaluasi feedback pengguna dan pengaruhnya pada pengembangan aplikasi.
Penelitian Terdahulu 2	
Judul	Analisis Sentimen Ulasan Pengguna Tiktok pada Google Play Store Berbasis TF-IDF dan *Support Vector Machine*
Nama Penulis	Sukirman, Sajiah, Nursuci Putri Husain, Anastasya Febriana Syam, Ragil Mustikosari [32]
Jurnal	Journal of System and Computer Engineering (JSCE)
Tahun	2024
Permasalahan	Kesulitan dalam mengelola dan memahami ulasan pengguna TikTok dalam jumlah besar secara manual
Algoritma	TF-IDF, SVM
Temuan	Menurut hasil evaluasi penelitian berikut, pendekatan yang diajukan berhasil mencapai tingkat akurasi yang signifikan, mencapai 84% pada skenario pembagian data latih 70% dan data uji 30%.
Pembahasan	Klasifikasi sentimen ulasan TikTok menggunakan TF-IDF dan SVM menunjukkan pemahaman mendalam terhadap respons pengguna.
Relevansi	Penelitian ini relevan karena menggunakan TF-IDF dan SVM untuk analisis sentimen, teknik yang sama yang dapat diaplikasikan untuk mengklasifikasikan ulasan TikTok. Pendekatan ini menawarkan metode efektif untuk memproses dan menganalisis data teks besar, yang penting untuk penelitian.
Penelitian Terdahulu 3	

Judul	Analisis Sentimen Aplikasi TikTok Menggunakan Algoritma *Naïve Bayes* dan *Support Vector Machine*
Nama Penulis	Friska Aditia Indriyani, Ahmad Fauzi, Sutan Faisal [33]
Jurnal	TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika
Tahun	2023
Permasalahan	Pengaruh aplikasi TikTok pada anak di bawah umur dan penggunaan aplikasi ini dalam bisnis, serta pendapat masyarakat yang disalurkan melalui ulasan di Google Play Store.
Algoritma	*Naïve Bayes*, SVM
Temuan	Klasifikasi ulasan positif dan negatif aplikasi TikTok di Google Play Store menunjukkan 76.7% ulasan berlabel positif dan 23.3% negatif. Metode *Naïve Bayes* mencapai akurasi 79%, sedangkan metode SVM lebih tinggi dengan akurasi 84%.
Pembahasan	Klasifikasi menggunakan metode *Naïve Bayes* dan SVM untuk menilai sentimen positif dan negatif di dalam ulasan aplikasi TikTok di Google Play Store. Metode SVM menunjukkan hasil yang lebih baik dibanding *Naïve Bayes* berdasarkan pengujian yang dilakukan.
Relevansi	Penelitian ini relevan untuk memahami sentimen masyarakat terhadap aplikasi TikTok yang dapat memiliki dampak negatif pada anak di bawah umur, serta untuk memahami pendapat masyarakat yang berpengaruh dalam bidang bisnis melalui platform media sosial.

Penelitian Terdahulu 4	
Judul	Analisis Sentimen Opini Publik Terhadap Program Televisi Indonesian Lawyers Club
Nama Penulis	Nico Nathanael Wilim, Raymond Sunardi Oetama [34]
Jurnal	IJNMT (International Journal of New Media Technology)
Tahun	2021
Permasalahan	Penurunan jumlah sentimen positif terhadap program ILC di Twitter, berbanding terbalik dengan kemenangan Panasonic Gobel Award.
Algoritma	K-Nearest Neighbor, *Naïve Bayes* Classifier, dan Decision Tree.
Temuan	Varian sentimen tahunan menunjukkan bahwa ILC menang dalam sentimen positif pada 2018 dibandingkan dengan Mata Najwa, tetapi kalah pada 2019.
Pembahasan	Fluktuasi sentimen publik terhadap program televisi dapat dikuantifikasi menggunakan analisis sentimen dan dapat mempengaruhi persepsi kualitas dan penerimaan program.
Relevansi	Penelitian ini menunjukkan keefektifan kombinasi metode dalam analisis sentimen, yang dapat diterapkan pada penelitian untuk membandingkan SVM dan *Naïve Bayes* dalam analisis ulasan pengguna TikTok. Insight ini memperkaya pemahaman tentang penerapan dan perbandingan metode analisis sentimen dalam konteks digital yang berbeda.
Penelitian Terdahulu 5	
Judul	*Sentiment Analysis for Assessment of Hotel Services Review using Feature Selection Approach based-on Decision Tree*
Nama Penulis	Dyah Apriliani, Taufiq Abidin, Edhy Sutanta, Amir Hamzah, Oman Somantri [35]
Jurnal	International Journal of Advanced Computer Science and Applications
Tahun	2020
Permasalahan	Kesulitan dalam interpretasi sentimen ulasan hotel di media sosial dan penentuan parameter optimal untuk meningkatkan akurasi analisis sentimen.
Algoritma	*Feature Selection, Decision Tree, Naïve Bayes*, SVM
Temuan	Penggunaan metode *Feature Selection* (FS) berbasis DT meningkatkan akurasi analisis sentimen ulasan hotel hingga 88.54%.
Pembahasan	Studi ini menekankan pentingnya pemilihan parameter dan penggunaan algoritma optimasi lain untuk meningkatkan akurasi model analisis sentimen,

	menghasilkan sistem rekomendasi yang lebih akurat untuk penilaian layanan hotel.
Relevansi	Penelitian ini relevan dengan penulis yang menganalisis sentimen ulasan pengguna TikTok karena menunjukkan pentingnya teknik seleksi fitur dan optimasi parameter dalam meningkatkan akurasi analisis sentimen, yang dapat diaplikasikan pada analisis sentimen ulasan aplikasi TikTok.
Penelitian Terdahulu 6	
Judul	Komparasi Algoritma *Support Vector Machine* Dan Random Forest Untuk Analisis Sentimen Metaverse
Nama Penulis	Putri Kumala Sari, Ryan Randy Suryono [36]
Jurnal	Jurnal MNEMONIC
Tahun	2024
Permasalahan	Analisis sentimen publik terhadap metaverse menggunakan data dari media sosial X, dengan tantangan mengenali sentimen secara akurat.
Algoritma	SVM, *Random Forest*
Temuan	*Random Forest* mencapai akurasi 91%, sedangkan SVM mencapai 90%. Penerapan SMOTE meningkatkan keseimbangan dataset dan kemampuan mengenali sentimen positif.
Pembahasan	Random Forest lebih unggul daripada SVM dalam analisis sentimen metaverse, dengan trade-off antara *Recall* dan precision yang lebih seimbang.
Relevansi	Penelitian ini relevan karena menunjukkan efektivitas penggunaan algoritma Random Forest dan SVM dalam analisis sentimen, yang bisa diaplikasikan pada penulis yang menganalisis sentimen ulasan pengguna TikTok.
Penelitian Terdahulu 7	
Judul	*The Right Sentiment Analysis Method of Indonesian Tourism in Social Media Twitter*
Nama Penulis	Cristian Steven, Wella [37]
Jurnal	IJNMT (International Journal of New Media Technology)
Tahun	2020
Permasalahan	Meningkatnya penggunaan media sosial untuk ekspresi opini memerlukan analisis sentimen yang akurat untuk industri pariwisata, termasuk untuk memahami sentimen publik terhadap pariwisata Bali.
Algoritma	*Naive Bayes*, *Neural Network*, *K-Nearest Neighbor*, *Support Vector Machines*, dan *Decision Tree*
Temuan	SVM dianggap sebagai algoritma paling tepat dengan AUC 0.805, menunjukkan klasifikasi yang baik. Sentimen publik terhadap Bali lebih banyak positif.
Pembahasan	Analisis menunjukkan SVM efektif dalam industri pariwisata untuk analisis sentimen. Penelitian mendukung penggunaan analisis sentimen dalam strategi pemasaran pariwisata.
Relevansi	Penelitian ini relevan karena menunjukkan penerapan dan perbandingan beberapa metode analisis sentimen dalam konteks pariwisata, yang dapat memberi insight untuk penelitian penulis terkait ulasan TikTok.
Penelitian Terdahulu 8	
Judul	Klasifikasi Ulasan Aplikasi TikTok Menggunakan Algoritma K-Nearest Neighbor dan Chi Square
Nama Penulis	Sandrina Ferani Aisyah Putri, I Wayan Supriana [38]
Jurnal	Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA)
Tahun	2024
Permasalahan	Ulasan di Google Play Store tentang aplikasi TikTok yang memiliki dampak baik positif maupun negatif terhadap pengalaman pengguna dan performa aplikasi.
Algoritma	K-Nearest Neighbor (KNN) dan Chi Square

Temuan	Penggunaan seleksi fitur Chi Square bersama algoritma KNN (dengan k = 9) meningkatkan akurasi klasifikasi hingga 86.22%, berbanding dengan KNN tanpa Chi Square (hanya 77.04% dengan k = 11).
Pembahasan	Studi ini menunjukkan bahwa penggabungan metode KNN dengan seleksi fitur Chi Square secara signifikan dapat meningkatkan performa dalam mengklasifikasikan ulasan pengguna TikTok menjadi positif atau negatif.
Relevansi	Penelitian ini relevan karena menunjukkan penerapan dan perbandingan beberapa metode analisis sentimen pada aplikasi Tiktok

Seluruh penelitian ini menggunakan area penelitian sama yaitu dengan menggunakan area penelitian TikTok. Berdasarkan hasil jurnal dalam Tabel 2.1, dapat disimpulkan bahwa jurnal 1 dan 3 menjadi acuan utama dalam penelitian ini karena topik yang diangkat berkaitan dengan analisis sentimen aplikasi di Google Play Store. Jurnal 1 mengaplikasikan *Naïve Bayes* untuk analisis sentimen aplikasi Dana, memperlihatkan keefektivitasan *Naïve Bayes* dalam mengidentifikasi kepuasan pengguna. Jurnal 3 fokus pada aplikasi TikTok menggunakan *Naïve Bayes* dan SVM, dengan SVM menunjukkan hasil yang lebih baik.

Penelitian terdahulu lainnya juga memberikan wawasan penting. Jurnal 2 menggunakan TF-IDF dan SVM untuk ulasan TikTok, yang relevan dengan metode yang digunakan dalam penelitian ini. Jurnal 4 hingga 6 menunjukkan pentingnya pemilihan fitur dan optimasi parameter dalam analisis sentimen, menggunakan metode seperti *K-Nearest Neighbor*, *Decision Tree*, dan *Random Forest*. Jurnal 7 menyoroti efektivitas SVM dalam analisis sentimen pariwisata di media sosial. Jurnal 8 menggunakan *K-Nearest Neighbor* dan *Chi Square* untuk analisis sentimen aplikasi TikTok, menunjukkan peningkatan akurasi yang signifikan.

Dalam penelitian ini, metode *Support Vector Machine* (SVM) dan *Naïve Bayes* diadopsi untuk menganalisis sentimen ulasan pengguna TikTok di Google Play Store. Kedua metode ini sebelumnya telah menunjukkan hasil yang baik dalam berbagai konteks analisis sentimen. Penelitian ini bertujuan untuk membandingkan tingkat akurasi yang dicapai melalui implementasi SVM dan Naïve Bayes, serta mengintegrasikan hasilnya menggunakan confusion matrix untuk menampilkan nilai precision, recall, dan F1-Score. Evaluasi performa model juga akan mencakup penggunaan Area Under Curve (AUC) sebagai metrik tambahan, yang sebelumnya tidak termasuk dalam evaluasi di jurnal 1 dan 3.

2.2 Teori tentang Topik

2.2.1 Analisis Sentimen

Analisis sentimen, sering juga disebut penggalian opini, adalah teknik yang digunakan untuk memahami dan menilai perasaan yang tersirat dalam teks. Proses ini bergantung pada teknologi pemrosesan bahasa alami dan linguistik komputasi untuk mengidentifikasi dan mengklasifikasikan pandangan, perasaan, dan sikap yang diungkapkan dalam berbagai jenis teks [21]. Tujuan utama dari analisis sentimen adalah untuk menyelidiki berbagai aspek dari pendapat individu, seperti emosi, penilaian, dan reaksi mereka terhadap subjek tertentu. Dalam praktiknya, analisis sentimen melibatkan pemilahan teks berdasarkan polaritasnya positif, negatif, atau netral sehingga memungkinkan peneliti dan praktisi untuk menilai dan membandingkan reaksi atau pendapat terhadap produk, layanan, atau topik [21]. Dengan demikian, analisis sentimen tidak hanya memungkinkan pemahaman lebih dalam tentang bagaimana teks mengkomunikasikan emosi tetapi juga memberikan dasar untuk mengukur sejauh mana pendapat dan sikap dapat mempengaruhi persepsi publik terhadap suatu isu.

Selain itu, analisis sentimen juga mempertimbangkan aspek-aspek lain dari teks, seperti intensitas emosi dan konteks [31]. Beberapa sistem analisis sentimen lanjutan mampu mendeteksi sarkasme dan ironi, yang dapat mengubah sentimen yang sebenarnya. Ini adalah tantangan khusus dalam analisis sentimen karena sarkasme sering kali memerlukan pemahaman mendalam tentang konteks dan nuansa bahasa. Dengan demikian, analisis sentimen tidak hanya memungkinkan pemahaman lebih dalam tentang bagaimana teks mengkomunikasikan emosi tetapi juga memberikan dasar untuk mengukur sejauh mana pendapat dan sikap dapat mempengaruhi persepsi publik terhadap suatu isu [21]. Alat ini sangat penting dalam dunia yang semakin terhubung, di mana opini publik dapat berdampak besar pada reputasi dan kesuksesan organisasi.

2.2.2 TikTok

Indonesia berada di posisi kedua dengan jumlah pengguna TikTok sebanyak 106 juta, di mana sebagian besar dari mereka adalah anak-anak di bawah umur, termasuk pelajar sekolah [5]. TikTok, sebagai platform media sosial berbasis video, telah berkembang menjadi fenomena global yang memengaruhi berbagai aspek kehidupan, khususnya di kalangan generasi muda. Aplikasi Tiktok memungkinkan pengguna untuk membuat, berbagi, dan menemukan konten kreatif dalam format video singkat. TikTok tidak hanya menjadi media ekspresi diri dan hiburan, tetapi juga telah memainkan peran penting dalam pemasaran digital dan pengaruh budaya. Algoritma canggih aplikasi TikTok mendorong penemuan konten yang personal dan relevan, menjadikannya sarana yang efektif untuk penyebaran tren dan ide-ide baru. Dengan demografi pengguna yang didominasi oleh remaja dan anak muda, TikTok telah menjadi bagian integral dari landskap media sosial kontemporer, memberikan wawasan unik tentang perilaku dan preferensi generasi baru [41].

2.2.3 Google Play Store

Google Playstore adalah platform distribusi digital yang dioperasikan oleh Google. Google Playstore merupakan platform distribusi konten digital yang memungkinkan pengguna ponsel pintar berbasis Android untuk mendownload berbagai aplikasi serta konten online lain yang tersedia, termasuk buku elektronik, film, permainan, dan banyak lagi [23]. Google Playstore menyediakan sarana bagi pengembang untuk menerbitkan dan mendistribusikan aplikasi mereka kepada pengguna global. Google Playstore juga memungkinkan pengguna untuk menjelajahi, mengunduh, dan membeli aplikasi serta konten digital lainnya.

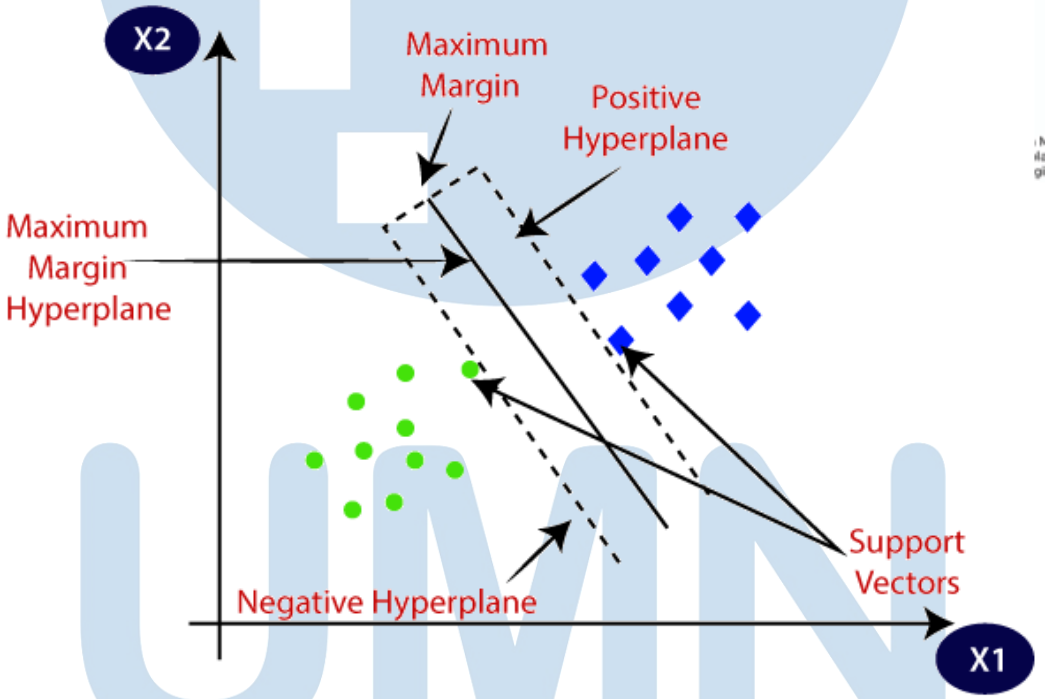
2.3 Framework, Algoritma, dan Teknik dalam Penelitian

2.3.1 Text Mining

Text mining merupakan metode untuk menemukan pola dalam teks dan mengubahnya menjadi data atau laporan yang sistematis dari kumpulan data

yang awalnya tidak terstruktur [24]. Tujuan utama dari text mining adalah mengenali kata-kata kunci yang mendefinisikan isi informasi, yang kemudian memfasilitasi analisis keterkaitan antar informasi tersebut. Proses ini seringkali diibaratkan sebagai siklus informasi yang intensif, dimana pengguna berinteraksi dengan berbagai dokumen dalam jangka waktu tertentu dengan menggunakan perangkat analisis. Berbeda dengan data mining yang lebih umum diidentifikasikan dengan ekstraksi pola dari basis data yang terstruktur, text mining menggali pola dari sumber data teks yang tidak terstruktur, meskipun kedua proses ini saling terkait dalam kerangka penelitian data mining [43].

2.3.2 Support Vector Machine



Gambar 2. 1 Hyperlane pada SVM
Sumber: [25]

Support Vector Machine (SVM) adalah algoritma *machine learning* yang digunakan untuk klasifikasi biner, yang artinya algoritma ini mengategorikan data ke dalam dua kelas, seperti positif dan negatif. Cara kerja SVM adalah dengan pertama-tama mempelajari data latih, mengamati distribusi datanya, dan mengidentifikasi titik-titik data yang ekstrem, atau yang dikenal sebagai outlier. Titik-titik ekstrem ini lantas dijadikan acuan sebagai vektor

pendukung yang akan menentukan pembentukan garis pemisah atau hyperplane untuk kelas positif dan negatif, guna mendapatkan hyperplane dengan margin terbesar yang mungkin, sebagaimana ditunjukkan dalam Gambar 2.1[25]. Oleh karena itu, SVM akan bekerja lebih efektif ketika distribusi data antar kelas dapat dikenali dengan jelas.

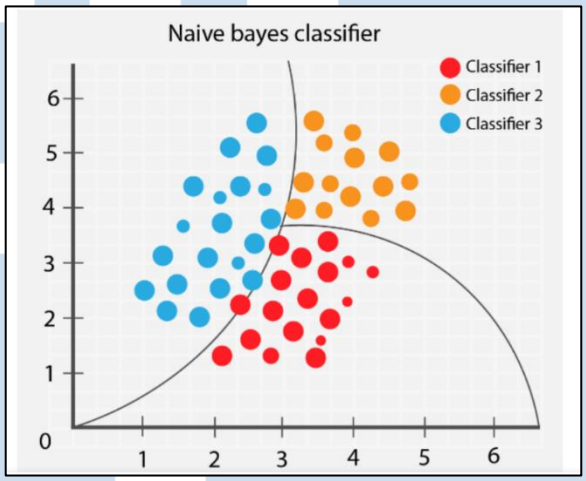
Dalam Support Vector Machine (SVM), kernel berfungsi sebagai alat transformasi yang mengizinkan pemisahan data yang tidak linearly separable di ruang aslinya untuk dipecahkan dalam ruang dimensi yang lebih tinggi di mana mereka bisa linearly separable. Ini dilakukan melalui pemetaan data ke ruang fitur yang lebih kompleks tanpa perlu menghitung secara eksplisit koordinat data dalam ruang tersebut, yang sering kali menghemat banyak waktu komputasi dan sumber daya. Berikut adalah penjelasan dari setiap kernel [44]:

1. **Linear Kernel:** Ini adalah kasus paling sederhana dari SVM, di mana pemisah dibuat langsung dalam ruang fitur tanpa memerlukan pemetaan ke dimensi yang lebih tinggi. Ini efektif untuk set data yang sudah bisa dibedakan secara jelas dengan garis atau pemisah lurus.
2. **Polynomial Kernel**: Kernel ini mengubah data ke ruang fitur dimana batas keputusan menjadi fungsi polinomial dari input asli. Tingkat polinomial dapat ditentukan sebagai parameter.
3. **Radial Basis Function (RBF) Kernel:** IRBF dapat memetakan sampel ke dalam ruang dimensi tak hingga, sehingga sangat efektif untuk kasus di mana hubungan antara kelas tidak linear.
4. **Sigmoid Kernel:** Mirip dengan fungsi aktivasi dalam jaringan saraf, kernel ini menghasilkan output yang menyerupai fungsi sigmoid, yang memungkinkan transformasi non-linear pada data untuk klasifikasi.

Pemilihan kernel yang tepat tergantung pada data dan jenis masalah yang dihadapi. Kernel memungkinkan SVM untuk memperoleh solusi yang fleksibel dan kuat bahkan untuk masalah klasifikasi yang kompleks. Dengan memanfaatkan kernel, SVM dapat menemukan hyperplane optimal di ruang

fitur yang ditransformasi, memastikan pemisahan kelas yang efektif dan meningkatkan akurasi prediksi.

2.3.3 Naïve Bayes



Gambar 2. 2 Gambar *Naive Bayes* Classifier
Sumber: [26]

Pada Gambar 2.2 adalah *Naïve Bayes* yang merupakan metode klasifikasi statistik yang berbasis pada *Teorema Bayes*. Metode ini sangat populer dalam machine learning karena sifatnya yang sederhana namun efektif, terutama dalam tugas-tugas klasifikasi teks. Algoritma NB adalah suatu metode prediksi yang sederhana dan berbasis probabilitas, yang didasarkan pada penggunaan *Teorema Bayes* dengan asumsi independensi yang sangat kuat [26]. Inti dari *Naïve Bayes* adalah asumsi independensi bersyarat, yang berarti bahwa keberadaan fitur tertentu dalam suatu kelas tidak bergantung pada keberadaan fitur lainnya. Model ini menghitung probabilitas suatu event dengan mengasumsikan independensi antara prediktor. Meskipun asumsi independensi sering kali dianggap naif, yang menjadi asal usul nama metode ini, *Naïve Bayes* terbukti efektif dalam banyak kasus praktis. Biasanya, algoritma ini cepat dan mudah untuk diimplementasikan, membuatnya cocok untuk dataset besar. *Naïve Bayes* sering digunakan dalam klasifikasi email spam, analisis sentimen, dan pengenalan pola. Meskipun metode *Naïve Bayes* memiliki keterbatasan dalam

hal asumsi independensi dan kesulitan dengan data input nol, *Naïve Bayes* tetap menjadi alat yang penting dalam *toolbox machine learning*.

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \quad (2.1)$$

Keterangan:

a. A adalah kelas atau kategori yang ingin diprediksi
b. B adalah fitur atau bukti yang diobservasi
c. (P(A|B)): Probabilitas *posterior*, yaitu probabilitas dari hipotesis (A) diberikan bukti (B). Dalam konteks *Naïve Bayes*, ini merupakan probabilitas dari kelas tertentu (misalnya, sentimen positif atau negatif) diberikan sebuah fitur/atribut tertentu dari data.
d. (P(B|A)): *Likelihood*, yaitu probabilitas dari bukti (B) diberikan hipotesis (A) benar. Dalam analisis sentimen, ini bisa diartikan sebagai probabilitas munculnya kata-kata tertentu dalam ulasan, asumsikan ulasan tersebut memiliki sentimen tertentu.
e. (P(A)): Probabilitas *prior*, yaitu probabilitas awal dari hipotesis (A) sebelum diperbarui dengan bukti baru. Ini merupakan probabilitas awal dari kelas tertentu sebelum melihat data.
f. (P(B)): Probabilitas bukti, yaitu probabilitas dari bukti (B) terjadi. Dalam banyak kasus, ini dihitung sebagai normalisasi untuk memastikan bahwa probabilitas *posterior* bersifat probabilitas valid.

2.3.4 *Term Frequency Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah teknik penting dalam text mining dan pemrosesan bahasa alami yang digunakan untuk mengukur pentingnya sebuah kata dalam dokumen dalam kumpulan dokumen atau corpus. TF-IDF mengukur pentingnya kata dalam dokumen dengan memperhitungkan seberapa sering kata tersebut muncul, memberi bobot lebih pada kata-kata yang dianggap informatif. Pengukuran ini kemudian dapat diaplikasikan untuk menentukan sentimen dalam teks [27]. Dengan menggabungkan TF dan IDF, TF-IDF memungkinkan analisis untuk menilai pentingnya kata secara lebih akurat dalam konteks corpus yang lebih luas. Dalam

algoritma TF-IDF, perhitungan bobot (W) untuk setiap dokumen dilaksanakan menggunakan formula khusus, yang ditunjukkan oleh Rumus 2.2.

$$\boldsymbol{WDT = TFDT * IDFT} \tag{2.2}$$

a. WDT: Bobot dari term T dalam dokumen D. Ini mengindikasikan seberapa penting atau relevan term tersebut dalam konteks dokumen tersebut.
b. TFDT: *Term Frequency* (TF) dari term T dalam dokumen D. TF mengukur frekuensi kemunculan term dalam dokumen, dengan asumsi bahwa semakin sering suatu term muncul dalam dokumen, semakin penting atau relevan term tersebut dalam dokumen.
c. IDFT: *Inverse Document Frequency* (IDF) dari term T. IDF mengukur kebalikan dari frekuensi dokumen yang mengandung term T dalam keseluruhan corpus atau kumpulan dokumen. Tujuan dari IDF adalah untuk mengurangi bobot dari term-term yang muncul di banyak dokumen dan dianggap kurang informatif atau kurang spesifik. Dengan kata lain, IDF memberikan bobot lebih kepada term-term yang jarang muncul karena dianggap lebih unik atau spesifik terhadap topik tertentu.

2.3.5 Confusion Matrix

Confusion Matrix merupakan alat evaluasi yang digunakan untuk mengukur kinerja suatu model pembelajaran mesin dalam melakukan klasifikasi [28]. Matriks ini disajikan dalam bentuk tabel yang menyederhanakan jumlah prediksi yang tepat dan tidak tepat, terorganisir dalam sebuah tabel dua dimensi yang mencerminkan keakuratan model klasifikasi. Confusion Matrix adalah alat fundamental untuk mengevaluasi efektivitas model analisis sentimen dalam membedakan antara sentimen positif dan negatif dalam data teks. Matriks ini terdiri dari empat elemen utama yang menunjukkan keakuratan prediksi model: True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). Tabel 2.2 mengilustrasikan distribusi hasil klasifikasi ini.

Tabel 2.2 Tabel Confusion Matrix

Actual	Predicted	
	+	-
+	*True Positive*	*False Positive*
-	*False Negative*	*True Negative*

Tabel 2.2 merupakan representasi dari *Confusion Matrix*. Ini adalah tabel yang digunakan untuk mengevaluasi kinerja algoritma klasifikasi dalam machine learning. Tabel ini dibagi menjadi empat bagian untuk menggambarkan empat kemungkinan hasil yang dapat terjadi saat melakukan prediksi:

a. *True Positive* (TP): Ini adalah kasus-kasus dimana model dengan benar memprediksi kelas positif, yaitu model mengatakan positif dan sebenarnya itu benar.
b. *False Positive* (FP): Ini adalah kasus-kasus dimana model salah memprediksi kelas positif, yaitu model mengatakan positif tetapi sebenarnya itu salah.
c. *False Negative* (FN): Ini adalah kasus-kasus dimana model salah memprediksi kelas negatif, yaitu model mengatakan negatif tetapi sebenarnya itu salah.
d. *True Negative* (TN): Ini adalah kasus-kasus dimana model dengan benar memprediksi kelas negatif, yaitu model mengatakan 'negatif' dan sebenarnya itu benar.

Perhitungan berikut digunakan untuk mengukur kinerja model klasifikasi untuk prediksi:

a. *Accuracy*

Accuracy merupakan rasio prediksi yang tepat baik untuk kelas positif maupun negatif terhadap seluruh prediksi yang dibuat. Hal ini dihitung dengan menjumlahkan data yang secara akurat terprediksi sebagai positif (*True Positive*s - TP) dan negatif (*True Negative*s - TN) kemudian dibagi dengan jumlah total data. Secara matematis, rumus akurasi dinyatakan sebagai:

$$Accuracy = \frac{\text{TP} + \text{TN}}{TP + FP + FN + TN}\ x\ 100\% \quad (2.3)$$

b. Precision

Presisi mengukur proporsi prediksi positif yang benar-benar merupakan kasus positif sebenarnya. Rumusnya melibatkan pembagian jumlah *True Positive*s dengan total prediksi positif (jumlah *True Positive*s ditambah dengan *False Positive*s):

$$Precision\ Positive = \frac{\text{TP}}{TP + FP}\ x\ 100\% \quad (2.4)$$

$$Precision\ Negative = \frac{\text{TN}}{TN + FN}\ x\ 100\% \quad (2.5)$$

c. *Recall*

Recall, atau sensitivitas, mengukur kemampuan model dalam mengidentifikasi semua kasus positif yang sebenarnya dari keseluruhan data positif atau negatif yang ada. *recall* dihitung sebagai berikut:

$$Recall\ Positive = \frac{\text{TP}}{TP + FN}\ x\ 100\% \quad (2.6)$$

$$Recall\ Negative = \frac{\text{TN}}{TN + FP}\ x\ 100\% \quad (2.7)$$

d. *F1- Score*

F1-Score adalah rata-rata harmonis dari *precision* dan *recall*, memberikan ukuran tunggal kinerja pada kedua dimensi tersebut. *F1-Score* dihitung sebagai berikut:

$$F1 - Score\ = \frac{2\ \times\ \text{Precision}\ \times Recall}{\text{Precision} + Recall}\ x\ 100\% \quad (2.8)$$

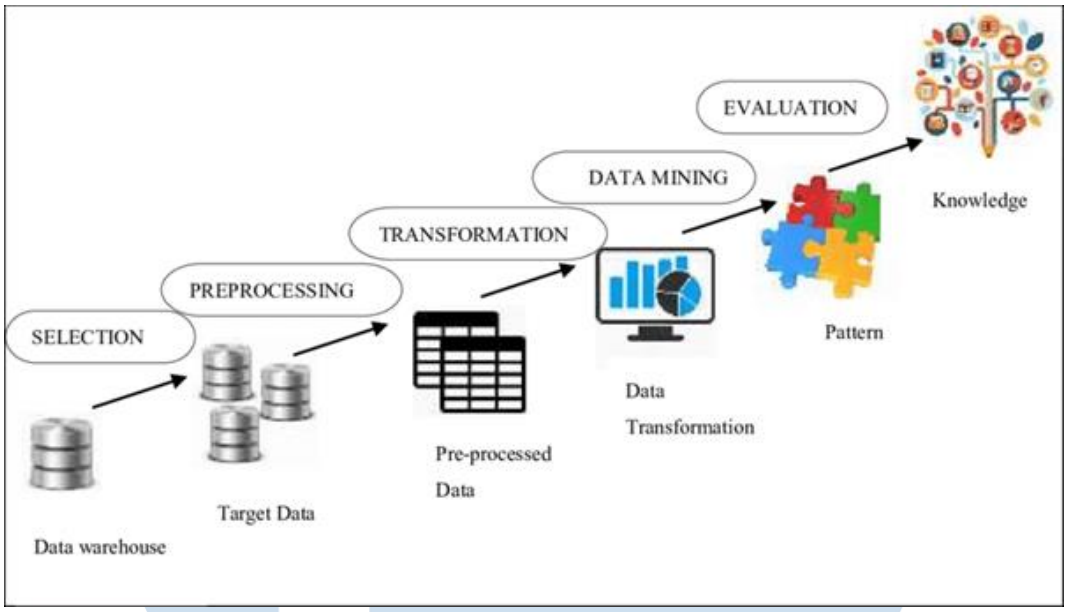
2.3.6 Area Under Curve

Area Under Curve (AUC) merupakan ukuran yang berasal dari kurva *Receiver Operating Characteristic* (ROC) yang digunakan untuk mengevaluasi model klasifikasi dalam menetapkan threshold atau ambang batas. Threshold ini penting untuk memisahkan antara kelas positif dan negatif dalam klasifikasi biner [29]. AUC mengukur sejauh mana kurva mampu membedakan antara kelas-kelas yang berbeda. Nilai AUC berkisar antara 0 dan 1. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sangat baik, di mana model dengan sempurna dapat membedakan antara positif dan negatif kelas dengan sedikit sampai tidak ada kesalahan. Sebaliknya, nilai AUC yang mendekati 0 menunjukkan kinerja yang sangat buruk, di mana model melakukan kesalahan hampir di semua prediksi. Semakin mendekati nilai AUC mendekati 1, semakin unggul kualitas model yang digunakan untuk melakukan prediksi. Pada Tabel 2.3 adalah kategori dari setiap hasil nilai AUC [40].

Tabel 2.3 Kategori Dari Hasil Nilai AUC

Nilai AUC	Kategori
0.90 – 1.00	*Excellent classification*
0.80 – 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.50	*Poor classification*
0.50 – 0.60	*Failure classification*

2.3.7 Knowledge Discovery In Database (KDD)



Gambar 2. 3 Proses KDD
Sumber: [39]

Penelitian ini mengadopsi metodologi *Knowledge Discovery in Databases* (KDD), yang umumnya dikenal sebagai proses KDD. Proses ini mengikuti alur kerja yang telah ditetapkan dan terdiri dari lima tahap utama [39]:

a. *Selection*

Tahap awal dari metode KDD dalam penelitian ini fokus pada pengumpulan data yang relevan dengan keperluan studi. Data ini diekstraksi dari aplikasi Tiktok dari platform Google Playstore melalui data collection dengan menggunakan alat *Jupyter Notebook*. Setelah pengumpulan data, dilakukan seleksi lanjutan untuk mengklasifikasikan sentimen menjadi dua kategori utama: sentimen positif dan negatif.

b. *Pre-processing*

Sebelum memulai analisis *Data Mining*, perlu dilakukan pembersihan datasebagai bagian penting dari proses KDD. Ini termasuk membersihkan data yang tidak terstruktur dan menghilangkan nilai-nilai yang hilang untuk memastikan integritas dan relevansi data yang akan digunakan dalam penelitian.

c. *Transformation*

Pada tahap transformasi, data yang telah diseleksi dan dibersihkan dipersiapkan dan ditransformasikan untuk memenuhi kebutuhan proses *Data Mining* berikutnya.

d. *Data Mining*

Menggunakan teknik *Data Mining*, model dibangun dari data yang telah ditransformasikan untuk mendapatkan hasil yang optimal. Penelitian ini akan menggunakan dua metode yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes*. Akan ada perbandingan kinerja antara kedua metode ini untuk menentukan yang paling efektif.

e. *Evaluation*

Tahap terakhir dalam proses KDD adalah evaluasi, yang dilakukan setelah model selesai dibangun. Tujuan dari evaluasi ini adalah untuk mengukur kinerja model yang telah dibuat. Dalam evaluasi, akan dihasilkan nilai-nilai seperti *Confusion Matrix* untuk mengukur *accuracy*, *F1-Score*, *recall*, dan *precision*. Selain itu, juga akan ditinjau nilai *Area Under Curve* (AUC) yang berfungsi untuk menampilkan performa komprehensif dari setiap model yang diuji.

2.4 Teori tentang Tools yang digunakan

2.4.1 Python

Python, yang dikenal sebagai bahasa pemrograman tingkat tinggi dengan orientasi objek dan bersifat sumber terbuka, memiliki aplikasi yang sangat beragam, mulai dari pengembangan situs web, pengelolaan data, hingga kreasi permainan [30]. Menyediakan perpustakaan sumber terbuka yang kaya dan detil, *Python* diakui memiliki kemampuan untuk mengatasi berbagai tantangan seperti *Big Data*, *Data Mining*, *Data Science*, dan *Deep Learning*, serta bidang yang sedang naik daun, yaitu pembelajaran mesin [30]. Dengan demikian, *Python* dipandang sebagai pilihan yang efisien untuk pengembangan kecerdasan buatan karena kesederhanaannya.

2.4.2 Jupyter Notebook

Jupyter Notebook adalah aplikasi *web open-source* yang memungkinkan pengguna untuk membuat dan berbagi dokumen yang mengandung kode live, persamaan, visualisasi, dan teks naratif. Ini digunakan untuk *Data Cleaning* dan *Transformation*, numerical simulation, statistical modeling, machine learning, dan banyak lagi. *Jupyter Notebook* mendukung lebih dari 40 bahasa pemrograman, termasuk *Python*, R, Julia, dan Scala. *Python* telah menjadi alat yang sangat populer di kalangan data scientists karena memungkinkan analisis data yang interaktif dan kolaboratif serta pendidikan yang efektif dalam bidang *Data Science* dan pemrograman.