BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian Terdahulu		
1	Jurnal	Journal of Agricultural Engineering, Vol. 14, No. 5, 2025
	Judul	Comparison of Machine Learning Models for Classifying Consumer Sentiment of Coffee Shops on Social Media X
	Tahun	2025
	Penulis	Agung Putra Pamungkas, Adam Mahendra, Ibnu Wahid Fakhrudin Aziz
	Masalah	Bagaimana membandingkan performa beberapa algoritma machine learning klasik dalam mengklasifikasikan sentimen pelanggan berdasarkan data tweet.
	Framework	Tahapan umum: Data Collection – Preprocessing – TF-IDF – Modeling – Evaluation
	Metode	Naïve Bayes, SVM, Logistic Regression
	Hasil	LR 79%, SVM 78%, NB 75%
	Adopsi	Dijadikan acuan utama: metodologi ML klasik dan struktur eksperimen diadopsi, dengan tambahan deep learning IndoBERTweet dan framework CRISP-DM.
2	Jurnal	Jurnal Ilmiah Teknik Informatika, Vol. 19 No.1 (Mei 2025)
	Judul	Model Analisis Sentimen pada Kendaraan Listrik Menggunakan Algoritma IndoBERTweet dan IndoBERT
	Tahun	2025
	Penulis	Belinda Eka Sarah Dewi
	Masalah	Meningkatnya opini publik terhadap kendaraan listrik di Indonesia menimbulkan kebutuhan untuk memahami persepsi masyarakat melalui analisis sentimen di media sosial X (Twitter).
	Framework	NLP Sentiment Analysis Pipeline (Data Collection – Preprocessing – Tokenization – Fine-tuning – Evaluation)
	Metode	IndoBERTweet, IndoBERT (Transformer Models)

	Hasil	IndoBERTweet akurasi 82.40%, F1-score 82.38%, IndoBERT Akurasi 75.98%
	Adopsi	implementasi fine-tuning IndoBERTweet dan evaluasi multi-metrik (accuracy, precision, recall, F1-score). Hasil riset ini memperkuat argumen pemilihan IndoBERTweet karena terbukti unggul 6,4% dibanding IndoBERT dalam konteks bahasa informal.
3	Jurnal	PIKSEL, Vol. 12(2), Sept 2024
	Judul	A Comparative Analysis of MultinomialNB, SVM, and BERT on Garuda Indonesia Twitter Sentiment
	Tahun	2024
	Penulis	Budi Prasetyo, Ahmad Yusuf Al-Majid, Suharjito
	Masalah	Garuda Indonesia menghadapi krisis reputasi dan perlu mengetahui sentimen pelanggan di Twitter untuk strategi perbaikan.
	Framework	Tahapan analisis berbasis NLP (Data → Preprocessing → Modeling → Evaluation)
	Metode	MultinomialNB, SVM, BERT
	Hasil	BERT unggul dengan akurasi 75.6%, SVM akurasi 71.6%, NB akurasi 64.8%
	Adopsi	Digunakan untuk menunjukkan relevansi perbandingan machine learning vs deep learning pada domain bahasa Indonesia.
4	Jurnal	IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 14 No. 3, June 2025
	Judul	Modeling Sentiment Analysis of Indonesian Biodiversity Policy Tweets Using IndoBERTweet
	Tahun	2025
	Penulis	Mohammad Teduh Uliniansyah, Asril Jarin, Agung Santosa, Gunarso
	Masalah	Mengukur efektivitas IndoBERTweet dalam Analisis sentimen kebijakan biodiversitas di Twitter sulit karena bahasa informal dan kompleks.
	Framework	Deep Learning Pipeline (BERT Embeddings, Cross-Validation, Statistical Validation)
	Metode	IndoBERTweet, Logistic Regression, SVM
	Hasil	IndoBERTweet unggul dengan mean F1 dan akurasi tertinggi ($p < 0.05$), akurasi 78.99% dan F1 score 0.7633, lalu untuk nilai tertinggi

		algoritma klasik akurasi 71.44% dan F1 score 0.6688
	Adopsi	Jadi dasar penggunaan IndoBERTweet untuk teks berbahasa Indonesia di Twitter, serta pembanding performa antara model deep learning dan algoritma klasik (Naïve Bayes, SVM) dalam konteks analisis sentimen pelanggan Fore Coffee.
5	Jurnal	Journal of Intelligent Computing and Health Informatics (JICHI), Vol. 2 No. 1, Mar 2021
	Judul	Sentiment Analysis on Coffee Consumer Perceptions on Social Media Twitter Using Multinomial Naïve Bayes
	Tahun	2021
	Penulis	Nurul Qomariah
	Masalah	Persepsi masyarakat terhadap kopi di Twitter belum banyak dianalisis secara otomatis.
	Framework	Text Mining Pipeline
	Metode	Multinomial Naïve Bayes + TF-IDF
	Hasil	Akurasi 94%, Precision 99%, Recall 88%
	Adopsi	Mengadopsi konsep domain coffee-related sentiment dan pemanfaatan model NB berbasis TF-IDF.
6	Jurnal	TECHNOVATE Journal, Vol. 2 No. 1, Jan 2025
	Judul	Sentiment Analysis of Twitter (X) Comments on the Cyanide Coffee Case Using Comparison of Naïve Bayes and K-NN Method Results
	Tahun	2025
	Penulis	I Gusti Agung Ayu Manik Ulandari, I Komang Arya Ganda Wiguna, I Made Dedy Setiawan
	Masalah	Kasus kopi sianida memicu diskusi besar di Twitter dan perlu diketahui pola sentimen public.
	Framework	Tahapan Eksperimen (Preprocessing – Splitting – Modeling – Evaluation)
	Metode	Naïve Bayes, K-NN
	Hasil	NB akurasi 87.76%, K-NN 80.60%
	Adopsi	Mengadopsi struktur uji perbandingan dua algoritma klasik dan skenario split rasio data.
7	Jurnal	Proceedings of the 4th International Conference on ICT in Business, Industry & Government (ICTBIG 2024)
	Judul	Twitter Sentiment Analysis Based Classification Model incorporating Multinominal Naive Bayes Classifiers
	Tahun	2024

	Penulis	Disha Purohit, Ajay Kumar Sharma, Narendra Singh Rathore, Mayank Patel
	Masalah	Klasifikasi sentimen tweet (positif/negatif) pada data Twitter skala besar yang penuh noise (stopwords, punctuation, user handles)
	Framework	Data collection → Pre-processing → Feature extraction → Classification → Performance evaluation (F1, precision, recall, confusion matrix)
	Metode	Preprocessing (hapus stopwords/punctuation/handles), Count Vectorizer, Naïve Bayes
	Hasil	Peningkatan akurasi dari baseline 57% menjadi 87% pada tweet
	Adopsi	Dijadikan dasar penerapan algoritma Multinomial Naïve Bayes sebagai model klasik pembanding, serta evaluasi performa multi-metrik (accuracy, precision, recall, dan F1-score).
8	Jurnal	International Journal of Engineering Trends and Technology (IJETT), Vol. 70, Issue 12, Desember 2022
	Judul	Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact Using BERT Method
	Tahun	2022
	Penulis	Fransiscus, Abba Suganda Girsang
	Masalah	Kebijakan PPKM menimbulkan pro-kontra, seperti efektif dari sisi kesehatan, tetapi berdampak negatif pada ekonomi masyarakat. Diperlukan pemetaan opini publik untuk mengevaluasi kebijakan
	Framework	NLP Transformer Pipeline
	Metode	IndoBERT, SVM, Naïve Bayes
	Hasil	F1: BERT 84%, SVM 70%, NB 83%
	Adopsi	Diadopsi untuk validasi superioritas model transformer dan perbandingan dengan algoritma klasik.
9	Jurnal	2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)
	Judul	Development of the Recommended Coffee Shops Application Based Twitter Sentiment Analysis
	Tahun	2022

	Penulis	Petipol Nilpao, Nopparuj Suetrong, Nitjaree Nanta, Natthanan Promsuk
	Masalah	Informasi ulasan di media sosial seperti Twitter sangat banyak, tetapi tidak terstruktur, sehingga sulit dimanfaatkan langsung oleh konsumen.
	Framework	Data Mining & System Deployment
	Metode	Naïve Bayes + TF-IDF + React.js deployment
	Hasil	Akurasi 86%
	Adopsi	Diadopsi domain coffee shop + pendekatan NB-TF IDF untuk studi pelanggan Fore Coffee.
10	Jurnal	Journal of Internet and Software Engineering (JISE), Vol. 5, No. 1, Mei 2024
	Judul	Analisis Sentimen Berbasis Aspek pada Ulasan Pengguna Aplikasi Starbucks Menggunakan Algoritma Support Vector Machine
	Tahun	2024
	Penulis	Muhammad Adin Palimbani, Rochana Prih Hasuti, Rian Adam Rajagede
	Masalah	Banyaknya ulasan negatif aplikasi Starbucks namun belum ada analisis aspek usability.
	Framework	ABSA (Aspect-Based Sentiment Analysis) Framework
	Metode	SVM (Linear, Polynomial, RBF) + GridSearchCV
	Hasil	Akurasi 88.96%
	Adopsi	Mengadopsi penggunaan SVM dan tuning hyperparameter sebagai referensi optimasi.
11	Jurnal	JOIV : Int. J. Inform. Visualization, 9(2) March 2025 796-807
	Judul	Evaluation of Sentiment Analysis Methods for Social Media Application
	Tahun	2025
	Penulis	Jose Octavian Leandro, Melissa Indah Fianty
	Masalah	Banyaknya ulasan negatif dan positif di Google Play Store yang memengaruhi reputasi TikTok dan kebutuhan untuk menemukan algoritma paling akurat dalam mengklasifikasikan sentimen pengguna.
	Framework	Comparative ML Framework
	Metode	SVM, Naïve Bayes
	Hasil	Akurasi SVM : 84%, Akurasi Naïve Bayes : 84.27%
	Adopsi	Diadopsi multi-metric evaluation (accuracy, precision, recall, F1) untuk hasil komparatif

Table 2.1 Penelitian Terdahulu

Berdasarkan beberapa penelitian terdahulu yang ditampilkan pada Tabel 2.1, dapat disimpulkan bahwa metode *machine learning* dan *deep learning* telah banyak digunakan dalam menganalisis sentimen di media sosial, khususnya Twitter. Penelitian yang dilakukan oleh Agung Putra Pamungkas et al. (2025) menunjukkan bahwa algoritma *Naïve Bayes*, *Support Vector Machine*, dan *Logistic Regression* mampu melakukan klasifikasi sentimen dengan tingkat akurasi yang cukup baik, meskipun masih terdapat ruang untuk peningkatan performa[5]. Sementara itu, penelitian lain yang menggunakan pendekatan *deep learning* seperti *IndoBERTweet* membuktikan adanya peningkatan hasil karena kemampuannya memahami konteks bahasa secara lebih mendalam, terutama pada teks yang bersifat informal.

Dari berbagai studi yang telah dilakukan, terlihat bahwa *machine learning* klasik masih sering digunakan karena kemudahannya dalam penerapan dan interpretasi hasil. Namun, model berbasis *deep learning* kini mulai banyak diterapkan karena mampu menangkap hubungan antar kata dalam kalimat secara kontekstual[12]. Hasil penelitian Belinda Eka Sarah Dewi (2025) dan Mohammad Teduh Uliniansyah et al. (2025) memperkuat keunggulan *IndoBERTweet* dibandingkan model klasik, terutama dalam menangani bahasa tidak baku dan penggunaan slang di Twitter[6].

Selain itu, beberapa penelitian terdahulu juga menyoroti pentingnya tahapan preprocessing seperti *case folding*, *stopword removal*, dan *stemming* untuk meningkatkan akurasi model. Penggunaan *TF-IDF* juga terbukti efektif dalam membentuk representasi fitur pada pendekatan klasik, sebagaimana ditunjukkan dalam studi oleh Disha Purohit et al. (2024) dan Nurul Qomariah (2021). Dengan demikian, proses *preprocessing* menjadi faktor penting dalam menentukan performa akhir model[13].

Penelitian ini mengadopsi beberapa konsep dari studi-studi sebelumnya, khususnya dalam hal metodologi dan penggunaan algoritma, tetapi dengan penyesuaian terhadap konteks data pelanggan Fore Coffee di Indonesia. Perbedaan utama penelitian ini terletak pada penerapan pendekatan komparatif antara tiga algoritma *machine learning* (*Naïve Bayes, SVM*, dan *Logistic Regression*) dan satu algoritma *deep learning* (*IndoBERTweet*)[14]. Dengan

demikian, penelitian ini diharapkan dapat memperkaya literatur mengenai analisis sentimen berbahasa Indonesia, sekaligus memberikan pemahaman baru mengenai perbandingan performa algoritma klasik dan modern dalam konteks ulasan pelanggan terhadap produk kopi lokal di media sosial Twitter.

2.2 Analisis Gap Penelitian Terdahulu

Berdasarkan sebelas penelitian terdahulu, sebagian besar studi analisis sentimen berbahasa Indonesia menggunakan model *machine learning* klasik seperti *Naïve Bayes*, *Logistic Regression*, atau *SVM*, dengan tingkat akurasi yang bervariasi pada rentang 70–88%. Namun, masih sangat sedikit penelitian yang mengintegrasikan model *transformer* modern seperti *IndoBERTweet* dan secara eksplisit membandingkan performanya dengan algoritma tradisional dalam konteks data media sosial berbahasa Indonesia.

Selain itu, penelitian sebelumnya belum secara khusus menerapkan pendekatan ini pada *domain* ulasan pelanggan Fore Coffee, meskipun brand ini memiliki tingkat interaksi publik yang tinggi di media sosial. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan membandingkan performa algoritma *machine learning* dan *IndoBERTweet* pada dataset ulasan Fore Coffee sekaligus mengevaluasi apakah penggunaan model *transformer* mampu meningkatkan akurasi dibandingkan metode yang digunakan dalam penelitian sebelumnya.

Dengan demikian, penelitian ini tidak hanya memberikan konteks objek yang berbeda, namun juga berpotensi menghasilkan peningkatan performa model melalui pemanfaatan pendekatan *deep learning* berbasis *transformer* yang lebih terbaru.

2.3 Teori yang berkaitan

2.3.1 Analisis Sentimen

Analisis sentimen merupakan metode untuk mengenali, mengambil, dan menentukan kategori emosi atau opini yang terdapat dalam sebuah teks. Teknik ini digunakan untuk mengidentifikasi kecenderungan sikap suatu pernyataan, apakah bersifat netral, negatif, atau positif. Bidang kajian ini termasuk dalam ranah *Natural Language*

Processing (*NLP*) dan telah diterapkan secara luas pada sektor bisnis, pemerintahan, politik, hingga layanan publik[15].

Pada penelitian ini, analisis sentimen diimplementasikan pada kumpulan tweet yang berkaitan dengan Fore Coffee. Tahapan yang dilakukan meliputi preprocessing teks seperti normalisasi, pemisahan kata (*tokenization*), penghapusan *stopword*, hingga proses pelabelan sentimen sesuai konteks tweet. Setelah melalui tahapan tersebut, data kemudian diolah dan diklasifikasikan menggunakan algoritma machine learning untuk mengetahui bagaimana persebaran opini pengguna Twitter terhadap Fore Coffee.

2.3.2 Twitter sebagai Sumber Data

Twitter (atau X) merupakan salah satu platform media sosial dengan format teks pendek yang sangat populer digunakan dalam penelitian analisis sentimen. Karakteristik utama Twitter, seperti batas panjang karakter, penggunaan bahasa informal, singkatan, emoji, dan tagar, menjadikannya sumber data yang dinamis dan representatif untuk menangkap opini publik[16].

Selain itu, Twitter menyediakan Application Programming Interface (API) yang memungkinkan peneliti untuk melakukan pengambilan data secara otomatis (scraping) berdasarkan kata kunci tertentu. Data yang diperoleh umumnya mencakup teks tweet, tanggal unggahan, *user metadata*, serta jumlah interaksi. Karena tweet bersifat publik dan real-time, data ini sering digunakan untuk menganalisis tren opini terhadap topik, merek, atau isu sosial tertentu.

2.4 Framework dan Algoritma yang digunakan

Penelitian ini menggunakan tiga algoritma *machine learning* klasik, yaitu *Naïve Bayes*, *Logistic Regression*, dan *Support Vector Machine* (*SVM*), serta satu algoritma *deep learning*, yaitu *IndoBERTweet*, dengan tujuan untuk melakukan perbandingan kinerja dalam analisis sentimen berbasis data tweet berbahasa Indonesia. Pemilihan ketiga algoritma *machine learning* klasik tersebut didasarkan pada jurnal-jurnal acuan yang relevan dan banyak

digunakan dalam penelitian analisis sentimen sebelumnya, khususnya pada domain media sosial. Algoritma *Naïve Bayes*, *Logistic Regression*, dan *SVM* sering dijadikan baseline model karena memiliki karakteristik yang berbeda namun saling melengkapi dalam menangani data teks berdimensi tinggi, sehingga cocok digunakan sebagai pembanding dalam penelitian ini.

Naïve Bayes dipilih karena merupakan algoritma probabilistik yang sederhana, efisien, dan memiliki performa yang cukup baik pada tugas klasifikasi teks, terutama ketika dikombinasikan dengan representasi fitur seperti *Term Frequency–Inverse Document Frequency* (*TF-IDF*). *Logistic Regression* digunakan karena mampu menghasilkan model klasifikasi yang stabil dan mudah diinterpretasikan, serta telah terbukti efektif dalam berbagai penelitian analisis sentimen berbasis teks. Sementara itu, *Support Vector Machine* (*SVM*) dipilih karena kemampuannya dalam menangani data berdimensi tinggi dan menemukan batas keputusan yang optimal, sehingga sering memberikan performa yang unggul pada permasalahan klasifikasi teks pendek seperti tweet. Ketiga algoritma ini digunakan dengan tujuan untuk meningkatkan hasil akurasi dibandingkan penelitian-penelitian terdahulu, serta untuk mengetahui sejauh mana peningkatan performa dapat dicapai melalui penyesuaian preprocessing dan strategi penanganan data yang diterapkan.

Metode klasik seperti *Naïve Bayes*, *Logistic Regression*, dan *SVM* telah terbukti efektif dalam pemrosesan teks pendek, sedangkan *IndoBERTweet* dipilih karena merupakan model berbasis *transformer* yang telah dilatih menggunakan lebih dari 500 juta tweet berbahasa Indonesia sehingga diharapkan memiliki pemahaman konteks yang lebih baik dibandingkan pendekatan tradisional.

2.4.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Penelitian ini menggunakan pendekatan *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) sebagai kerangka kerja dalam melakukan analisis sentimen. *CRISP-DM* dipilih karena memiliki struktur yang sistematis dan fleksibel untuk mengolah data besar, termasuk data teks tidak terstruktur seperti tweet. Tahapan-tahapan utama dalam metode ini meliputi :

1. **Business Understanding**

   Tahap ini berfokus pada pemahaman konteks bisnis dan tujuan penelitian. Peneliti mengidentifikasi permasalahan utama, yaitu bagaimana membandingkan performa algoritma *machine learning* (*Naïve Bayes, Logistic Regression,* dan *SVM*) dengan algoritma *deep learning* (*IndoBERTweet*) dalam menganalisis sentimen pelanggan terhadap Fore Coffee di Twitter. Tujuan dari tahap ini adalah menentukan arah penelitian dan hasil yang ingin dicapai.

2. **Data Understanding**

   Tahapan ini mencakup proses pengumpulan data tweet yang relevan dengan kata kunci terkait Fore Coffee, serta eksplorasi awal terhadap karakteristik data. Analisis dilakukan untuk memahami jumlah data, variasi bahasa, serta potensi adanya *noise* seperti spam, duplikasi, dan kata tidak relevan yang dapat memengaruhi hasil klasifikasi.

3. **Data Preparation**

   Pada tahap ini dilakukan serangkaian proses preprocessing dengan pendekatan *Natural Language Processing (NLP)*, meliputi *case folding*, *tokenization*, *stopword removal*, *stemming* menggunakan library *Sastrawi***,** serta labeling sentimen menjadi positif, negatif, dan netral. Tujuannya adalah menyiapkan data yang bersih dan seragam sebelum dilakukan proses pelatihan model.

4. **Modeling**

   Tahap ini menerapkan empat algoritma klasifikasi, yaitu:

   1. **Naïve Bayes** sebagai model probabilistik yang sederhana dan cepat.
   2. **Support Vector Machine** (***SVM***) sebagai model klasifikasi yang bertujuan menemukan *hyperplane* yang paling optimal untuk memisahkan data berdasarkan kelas sentimennya.
   3. **Logistic Regression** sebagai model statistik yang mengklasifikasikan data dengan mengestimasi probabilitas berbasis fungsi logistic.

4. **IndoBERTweet** sebagai model *transformer* berbahasa Indonesia yang memahami konteks kata secara dua arah. Setiap algoritma diuji menggunakan data yang sama agar hasil perbandingan performanya adil dan objektif.

5. **Evaluation**

   Evaluasi dilakukan dengan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai performa tiap algoritma. Selain itu, digunakan *confusion matrix* untuk melihat pola klasifikasi dan kesalahan prediksi dari masing-masing model.

6. **Deployment**

   Tahap terakhir berupa penyajian hasil analisis dan visualisasi perbandingan kinerja model dalam bentuk grafik dan tabel agar mudah dipahami. Hasil dari penelitian ini diharapkan menjadi acuan untuk penelitian lanjutan dalam pengembangan sistem analisis opini publik secara otomatis berbasis NLP.

2.4.2 Naïve Bayes

Naïve Bayes merupakan algoritma berbasis probabilitas yang menggunakan Teorema Bayes untuk menentukan kemungkinan suatu teks termasuk dalam kelas sentimen tertentu[17]. Algoritma ini mengasumsikan bahwa antar kata bersifat independen. Dalam penelitian ini, *Naïve Bayes* digunakan untuk mengklasifikasikan tweet ke dalam kategori positif, negatif, dan netral berdasarkan nilai peluang kemunculan kata.

Secara umum, rumus dari *Teorema Bayes* dituliskan sebagai berikut:

$$P(C \mid X) = \frac{P(X \mid C) \times P(C)}{P(X)}$$

Keterangan:

- $P(C \mid X)$: Probabilitas sebuah dokumen X termasuk dalam kelas sentimen C (*posterior probability*).

- $P(X \mid C)$: Probabilitas munculnya fitur (kata) X jika diketahui kelas C (*likelihood*).
- $P(C)$: Probabilitas awal dari kelas C (*prior probability*).
- $P(X)$: Probabilitas total dari fitur X yang muncul di semua kelas (*evidence*).

Dalam praktiknya, algoritma ini mengasumsikan bahwa setiap fitur atau kata bersifat independen satu sama lain (independensi bersyarat). Asumsi inilah yang membuat algoritma ini disebut *naïve* (naif). Meskipun demikian, pendekatan ini terbukti sangat efisien terutama untuk teks pendek seperti tweet.

Kelebihan utama *Naïve Bayes* adalah kecepatan dalam proses pelatihan dan prediksi karena hanya memerlukan perhitungan probabilitas antar kata. Namun, kekurangannya terletak pada asumsi independensi antar fitur yang tidak selalu sesuai dengan kondisi nyata, serta sensitivitas terhadap data yang tidak seimbang.

Pada penelitian ini, *Naïve Bayes* digunakan untuk mengklasifikasikan tweet pelanggan menjadi tiga kategori sentimen, yaitu positif, negatif, dan netral. Hasil probabilitas tertinggi dari masing-masing kelas akan menentukan kategori akhir dari sebuah tweet.

2.4.3 Support Vector Machine (SVM)

SVM bekerja dengan mencari *hyperplane* yang paling optimal dalam melakukan pemisahan data dari kelas yang berbeda dengan margin maksimal. Dalam konteks analisis sentimen, setiap tweet direpresentasikan sebagai vektor hasil transformasi *TF-IDF*. *SVM* dipilih karena mampu menangani data berdimensi tinggi, menghasilkan performa yang stabil, dan memberikan akurasi yang tinggi dalam pengklasifikasian teks pendek[18].

Secara matematis, bentuk umum dari *hyperplane* dalam SVM dapat ditulis sebagai:

$$w \cdot x + b = 0$$

Keterangan:

- w : Vektor bobot yang menentukan arah dan orientasi *hyperplane*.
- x : Vektor fitur (misalnya hasil *TF-IDF* dari teks).
- b : Bias atau konstanta pergeseran posisi *hyperplane*.

Tujuan dari SVM adalah menemukan nilai w dan b yang memaksimalkan margin antara dua kelas yang berbeda. Margin ini diukur dari jarak antara *hyperplane* dengan data terdekat dari masing-masing kelas, yang disebut *support vectors*.

Jika data tidak dapat dipisahkan secara linier, SVM menggunakan pendekatan *kernel trick* untuk memetakan data ke dalam ruang berdimensi lebih tinggi sehingga pemisahan linier menjadi mungkin dilakukan. Jenis *kernel* yang umum digunakan adalah *linear*, *polynomial*, dan *radial basis function (RBF)*[19].

Kelebihan utama *SVM* adalah kemampuannya menangani data berdimensi tinggi seperti hasil *TF-IDF vectorization*, serta menghasilkan performa yang stabil pada dataset teks yang kompleks. Namun, kelemahannya adalah waktu pelatihan yang relatif lama terutama untuk data besar.

Dalam penelitian ini, *SVM* digunakan untuk mengklasifikasikan tweet pelanggan berdasarkan representasi fitur hasil *TF-IDF*. Model ini diharapkan memberikan performa tinggi karena sifatnya yang kuat terhadap *overfitting* dan kemampuannya menangani data yang memiliki batas keputusan kompleks.

2.4.4 Logistic Regression

Logistic Regression adalah salah satu metode *machine learning* yang digunakan untuk melakukan tugas klasifikasi, baik dua kelas maupun lebih, dengan pendekatan berbasis statistik. Tidak seperti *regresi linear*

yang menghasilkan nilai numerik kontinu, *Logistic Regression* memanfaatkan fungsi logistik atau sigmoid untuk mengonversi output menjadi nilai probabilitas dalam rentang 0 hingga 1.

Secara matematis, fungsi logistik dituliskan sebagai berikut:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Keterangan:

1. $P(y = 1 \mid x)$ = probabilitas suatu sampel termasuk dalam kelas positif
2. β_0 = intercept (bias)
3. β_i = koefisien dari setiap fitur x_i
4. e = bilangan eksponensial

Dalam konteks analisis sentimen, *Logistic Regression* digunakan untuk memetakan setiap tweet ke dalam kategori sentimen (positif, negatif, atau netral) berdasarkan nilai probabilitas yang dihasilkan. Jika probabilitas melebihi ambang batas tertentu (misalnya 0.5), maka tweet diklasifikasikan ke dalam kelas positif, sedangkan jika di bawah ambang batas, diklasifikasikan ke dalam kelas lain[20].

Logistic Regression sering kali memberikan hasil yang stabil dan akurat untuk data teks yang telah diubah ke dalam representasi numerik seperti *TF-IDF*. Kelebihan algoritma ini terletak pada kesederhanaan, efisiensi komputasi, serta interpretabilitas yang tinggi karena setiap fitur memiliki bobot yang jelas dalam memengaruhi hasil klasifikasi. Namun, kekurangannya adalah performa *Logistic Regression* dapat menurun apabila data bersifat non-linear atau memiliki hubungan antar fitur yang kompleks.

Dalam penelitian ini, *Logistic Regression* digunakan sebagai salah satu pembanding dalam kelompok algoritma *machine learning* klasik bersama dengan *Naïve Bayes* dan *SVM*. Ketiga model tersebut akan dibandingkan performanya dengan algoritma *deep learning* *IndoBERTweet* untuk mengetahui pendekatan mana yang paling efektif dalam menganalisis sentimen pelanggan Fore Coffee di Twitter.

2.4.5 IndoBERTweet

IndoBERTweet merupakan model *deep learning* berbasis arsitektur *transformer* yang dikembangkan khusus untuk bahasa Indonesia[21]. Model ini dilatih menggunakan lebih dari 500 juta tweet berbahasa Indonesia sehingga sangat relevan untuk penelitian berbasis media sosial. Berbeda dengan algoritma *machine learning* klasik, *IndoBERTweet* mampu memahami konteks antar kata dalam dua arah, sehingga hasil klasifikasi sentimen menjadi lebih akurat. Dalam penelitian ini, *IndoBERTweet* digunakan untuk mengukur sejauh mana pendekatan *deep learning* dapat meningkatkan performa dibandingkan metode klasik[6].

Proses kerja utama *IndoBERTweet* meliputi tiga tahap:

1. **Tokenization** : Setiap teks diubah menjadi token menggunakan *IndoBERTweet tokenizer* agar sesuai dengan format input model.

2. **Embedding Representation** : Token diubah menjadi representasi vektor numerik menggunakan lapisan *embedding*.

3. **Fine-tuning** : Model *pre-trained* disesuaikan kembali (*fine-tuned*) terhadap dataset penelitian agar dapat mengenali pola sentimen pada tweet pelanggan Fore Coffee.

Secara konseptual, rumus dasar fungsi aktivasi dalam model *transformer* dapat digambarkan melalui fungsi perhatian (*self-attention mechanism*) berikut:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Keterangan:

- Q : *Query vector*, representasi dari kata yang sedang diproses.
- K : *Key vector*, representasi dari semua kata dalam urutan teks.
- V : *Value vector*, representasi konteks yang akan dikalikan dengan bobot perhatian.

- d_k : Dimensi dari vektor kunci untuk menormalkan hasil perkalian.

Fungsi *attention* inilah yang membuat *IndoBERTweet* mampu memahami hubungan antar kata dalam satu kalimat maupun antar kalimat. Model ini sangat cocok digunakan untuk bahasa informal seperti tweet, karena telah dilatih menggunakan korpus media sosial yang kaya akan variasi bahasa, singkatan, dan gaya penulisan tidak baku.

Dalam penelitian ini, *IndoBERTweet* digunakan untuk membandingkan kinerja algoritma *deep learning* terhadap model *machine learning* klasik. Harapannya, pendekatan *transformer* ini dapat memberikan hasil dengan akurasi dan *F1-score* yang lebih tinggi dibandingkan dengan *Naïve Bayes* dan *SVM*.

2.5 Tools dan Software yang digunakan

Penelitian ini menggunakan berbagai perangkat dan *library* untuk mendukung seluruh tahapan proses, mulai dari pengumpulan data hingga analisis hasil. Seluruh proses dijalankan di lingkungan *Google Colab* yang berbasis *cloud*, sehingga memudahkan peneliti dalam menulis dan menjalankan kode tanpa perlu instalasi tambahan di perangkat lokal. Selain itu, *Colab* menyediakan dukungan GPU yang mempercepat proses pelatihan model, khususnya untuk algoritma berbasis *deep learning* seperti *IndoBERTweet*.

1. **X Developer (Twitter API)**

   X Developer digunakan untuk memperoleh data tweet secara langsung dari platform Twitter. Melalui akses *Application Programming Interface (API)*, peneliti dapat mengumpulkan data berupa teks, waktu unggahan, dan informasi lain yang relevan berdasarkan kata kunci tertentu. Penggunaan *X Developer* memungkinkan proses pengambilan data dilakukan secara otomatis, efisien, dan terarah sesuai kebutuhan penelitian. Data yang diambil

kemudian disimpan dalam format *comma-separated values (CSV)* untuk diolah pada tahap berikutnya[22].

2. Google Colab

Google Colab merupakan platform berbasis *cloud* yang digunakan untuk menjalankan kode Python secara interaktif. Peneliti menggunakan *Colab* untuk melakukan seluruh tahapan penelitian, mulai dari *preprocessing* data, pelatihan model, hingga evaluasi hasil. Kelebihan utama *Colab* adalah tersedianya dukungan GPU yang membantu mempercepat proses komputasi model *deep learning*, serta integrasi dengan *Google Drive* yang mempermudah penyimpanan dataset dan hasil eksperimen[23].

3. Python

Python digunakan sebagai bahasa pemrograman utama karena memiliki ekosistem *library* yang luas untuk analisis data, *machine learning*, dan *natural language processing (NLP)*. Bahasa ini digunakan dalam setiap tahapan *CRISP-DM*, mulai dari pengolahan teks, pembentukan model, hingga visualisasi hasil. Struktur sintaks yang sederhana dan dukungan komunitas yang luas membuat *Python* menjadi pilihan ideal untuk penelitian berbasis data[24].

4. Pandas dan NumPy

Pandas digunakan untuk memproses dan menganalisis data dalam bentuk tabel, seperti membaca file *CSV* hasil *scraping*, membersihkan kolom yang tidak relevan, serta mengelola dataset yang telah diberi label sentimen. *NumPy* digunakan untuk melakukan perhitungan matematis, seperti operasi matriks dan vektor yang menjadi dasar dalam pembentukan model *machine learning*. Kedua *library* ini saling melengkapi dan menjadi pondasi penting dalam tahap *data preprocessing[25]*.

5. Sastrawi

Sastrawi merupakan *library* pengolahan teks berbahasa Indonesia yang digunakan untuk melakukan *stemming*, yaitu mengubah kata berimbuhan menjadi bentuk dasarnya. Proses ini bertujuan agar sistem

dapat mengenali kata yang memiliki makna sama namun bentuknya berbeda, sehingga hasil analisis menjadi lebih konsisten. Penggunaan *Sastrawi* penting dalam tahap *preprocessing* agar model dapat memahami struktur bahasa Indonesia dengan lebih baik[26].

6. **NLTK (Natural Language Toolkit)**

   NLTK digunakan untuk melakukan *tokenization*, penghapusan tanda baca, serta penghapusan *stopword*. *Tokenization* membantu memisahkan teks menjadi kata-kata terpisah agar mudah diolah lebih lanjut, sedangkan *stopword removal* bertujuan untuk menghapus kata umum seperti "dan", "yang", atau "atau" yang tidak memiliki makna penting dalam analisis sentimen. *Library* ini digunakan pada tahap awal *preprocessing* untuk memastikan teks bersih sebelum diubah ke dalam representasi numerik[27].

7. **Scikit-learn**

   Scikit-learn digunakan untuk mengimplementasikan algoritma *machine learning* seperti *Naïve Bayes* dan *Support Vector Machine (SVM)*. Selain itu, *library* ini juga digunakan untuk membagi data menjadi bagian pelatihan dan pengujian (*train-test split*), melakukan transformasi teks menggunakan *Term Frequency–Inverse Document Frequency (TF-IDF)*, serta menghitung metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*. *Scikit-learn* dipilih karena bersifat modular, mudah digunakan, dan mendukung berbagai fungsi analisis klasifikasi teks[28].

8. **Hugging Face Transformers**

   Hugging Face Transformers digunakan untuk memanggil dan melakukan *fine-tuning* terhadap model *IndoBERTweet*. *Library* ini menyediakan *tokenizer* dan model *transformer* yang telah dilatih sebelumnya dengan korpus berbahasa Indonesia, sehingga dapat digunakan untuk tugas klasifikasi teks dengan hasil yang lebih baik. Dalam penelitian ini, *library* tersebut mempermudah proses penerapan model *deep learning* tanpa perlu membangun arsitektur model dari awal[29].

9. Matplotlib dan Seaborn

Matplotlib dan *Seaborn* digunakan untuk menampilkan hasil analisis dalam bentuk visual seperti grafik dan diagram. Kedua *library* ini digunakan untuk membuat *confusion matrix*, grafik perbandingan akurasi antar algoritma, serta visualisasi distribusi sentimen pada dataset. Visualisasi membantu peneliti memahami performa model dengan lebih jelas dan memudahkan penyajian hasil analisis dalam laporan penelitian[30].