BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Berbagai peneliti telah melakukan penelitian terdahulu tentang analisis sentimen pada platform e-commerce dengan menggunakan berbagai metode dan model. Penelitian tertentu lebih berkonsentrasi pada tingkat akurasi dari model yang digunakan dalam proses klasifikasi sentimen, sementara penelitian lainnya lebih berkonsentrasi pada mengidentifikasi sentimen pengguna terhadap layanan atau fitur tertentu yang tersedia pada platform e-commerce. Selain itu, penelitian menggunakan data ulasan pengguna untuk mengetahui bagaimana pelanggan melihat layanan yang ditawarkan oleh platform digital. Untuk meningkatkan kemampuan klasifikasi teks dalam analisis sentimen, berbagai teknik machine learning dan deep learning telah digunakan. Namun, penelitian yang secara khusus membandingkan kinerja model Support Vector Machine (SVM) dengan model IndoBERT dalam pendekatan Aspect Based Sentiment Analysis terhadap ulasan pengguna e-commerce berbahasa Indonesia masih sangat sedikit. Dengan melakukan analisis perbandingan kedua model tersebut pada ulasan pengguna aplikasi Tokopedia & Shopee—yang dijelaskan dalam rangkuman penelitian sebelumnya pada tabel berikut—kesenjangan tersebut dipenuhi oleh penelitian ini.

Tabel 2. 1 Penelitian Terdahulu

Judul	Penulis & Tahun	Algoritma	Hasil
Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding & CNN	Syaiful, Ester Irawati, Joan Santoso & 2023	IndoBERT & CNN	Untuk itu dapat diambil kesimpulan bahwa metode CNN lebih baik dari pada metode IndoBERT untuk deteksi aspek
Analisis Sentimen Produk Hijab Pada E-Commerce Tokopedia Menggunakan	Ghina Tri Fadilah, Lailil Muflikhah & 2025	SVM & IndoBERT	Hasil pengujian menunjukkan bahwa kombinasi IndoBERT dan SVM memberikan

Algoritma Support Vector Machine Dan Indobert Embedding			performa yang baik dalam analisis sentimen
Komparasi Model Indobert Dan Indogpt Untuk Analisis Sentimen Pada Produk E-Commerce	Fajar Moch, Rangga Gelar & 2025	IndoBERT & IndoGPT	Hasilevaluasi menunjukkan bahwa IndoBERT secara konsisten unggul pada seluruh metrik dan di kedua skenario pengujian, dengan accuracymencapai 83persendan F1-scorekelas positif hingga 96persen
Eksplorasi Sentimen Pengguna pada Aplikasi E-Commerce dengan Deep Learning	Ahmad Kamal, Renita Asri & 2025	IndoBert & BERT	Hasil eksperimen menunjukkan bahwa model IndoBERT secara konsisten unggul dalam seluruh metrik evaluasi utama
Metode Aspect - Based Sentiment Analysis (Absa) Pada Ulasan Songket Di Google Maps Menggunakan Indobert	Fitri Salamah, Siti Fatimah & 2026	IndoBert	Hasil ini menunjukkan bahwa model tersebut dapat secara akurat menangkap opini pelanggan tentang kualitas produk, harga, dan layanan, memberikan wawasan berharga untuk membantu UMKM
Hybrid Fine-Tuning Indobert Dan Ensemble Tf-Idf Logistic Regression Untuk Analisis Sentimen Ulasan Aplikasi Zalora	Al Ikhsan & 2026	Hybrid Fine Tuning Indobert & Ensemble TF-IDF Logistic Regresi	model memperoleh akurasi 86,77%, macro-F181,71%, dan weighted-F186,76%

Comparison of IndoBERT and SVM Performance in Sentiment Analysis of Digital Education Platforms	Aldina Bonaria Siva1) , Robet2) , Leony Hok	IndoBERT & SVM	**SVM unggul dari sisi accuracy**, yaitu mencapai **0,888**, karena model ini sangat kuat dalam mengenali kelas positif yang jumlah datanya dominan. IndoBERT lebih seimbang dalam mengenali sentimen positif, negatif, dan netral, meskipun accuracy-nya sedikit lebih rendah dibandingkan SVM.
Comparison of IndoBERT and SVM Algorithm to Perform Aspect Based Sentiment Analysis using Hierarchical Dirichlet Process	Sheila Prima Octarini1 , Alfi Yusrotis Zakiyyah2 , Kartika Purwandari3	IndoBERT & SVM	SVM lebih baik daripada IndoBERT untuk ABSA review fashion Tokopedia pada dataset yang digunakan. SVM menghasilkan akurasi dan precision lebih tinggi, sehingga dipilih untuk implementasi website.
Sentiment Analysis of Cyber Attacks in Bank Syariah Indonesia Using SVM and Indobert Method	Chandra Apriyadi*1 , Styawati	IndoBERT & SVM	IndoBERT lebih baik daripada SVM karena mampu menghasilkan akurasi dan F1-score yang lebih tinggi. Selain itu, hasil LDA menunjukkan bahwa sentimen negatif publik banyak membahas masalah mobile banking, transfer, ATM, tarik tunai,

			saldo, dan keamanan dana nasabah.
Sentiment Analysis Comparison of Two E-Commerce Platforms Using Random Forest, Support Vector Machine, Logistic Regression, and IndoBERT	Theresia Vania Davita Suyana	Random Forest, Support Vector Machine, Logistic Regression, and IndoBERT	Model IndoBERT sangat efektif ketika dataset besar, seperti pada Shopee, karena mampu memahami konteks bahasa Indonesia dengan baik. Namun, untuk dataset yang lebih kecil seperti Tokopedia, Random Forest lebih stabil dan menghasilkan akurasi lebih tinggi dibanding IndoBERT.

Berdasarkan Tabel 2.1, penelitian terdahulu menunjukkan bahwa analisis sentimen pada ulasan pengguna e-commerce telah banyak dilakukan dengan pendekatan machine learning dan deep learning, seperti yang ditunjukkan dalam Tabel 2.1. Model berbasis transformer seperti IndoBERT banyak digunakan karena mampu memahami konteks bahasa Indonesia lebih baik daripada metode konvensional. Beberapa penelitian juga telah membandingkan algoritma Support Vector Machine (SVM) dengan IndoBERT untuk menentukan model mana yang melakukan proses klasifikasi sentimen dengan paling efisien. Namun, karena hasil penelitian masih beragam, belum ada kesimpulan yang konsisten tentang model terbaik.

Perbedaan hasil penelitian terlihat pada beberapa studi yang menyatakan bahwa IndoBERT menghasilkan performa yang lebih tinggi dibandingkan SVM karena mampu menangkap hubungan kontekstual antar kata dengan lebih baik. Sebaliknya, terdapat penelitian lain yang menunjukkan bahwa SVM mampu menghasilkan akurasi yang lebih tinggi dibandingkan IndoBERT, khususnya pada dataset dengan representasi fitur TF-IDF atau jumlah data yang tidak terlalu besar. Variasi hasil tersebut menunjukkan

bahwa performa model sangat dipengaruhi oleh karakteristik dataset, proses preprocessing, metode pelabelan, serta domain penelitian yang digunakan. Dengan demikian, masih diperlukan penelitian yang membandingkan kedua model menggunakan dataset dan proses pengolahan data yang sama sehingga hasil evaluasi dapat dibandingkan secara lebih objektif.

Selain itu, sebagian besar penelitian sebelumnya berkonsentrasi pada analisis sentimen secara umum, juga dikenal sebagai analisis sentimen pada tingkat dokumen. Analisis ini hanya menghasilkan klasifikasi positif, negatif, atau netral untuk ulasan secara keseluruhan. Metode ini tidak dapat menemukan faktor khusus yang menyebabkan perasaan pengguna. Aspect-Based Sentiment Analysis (ABSA) belum banyak digunakan untuk menilai aplikasi e-commerce Indonesia yang menggunakan beberapa aspek penting secara bersamaan, seperti harga, produk, pengiriman, dan sistem aplikasi. Kondisi ini menunjukkan bahwa penelitian masih dapat mengembangkan analisis sentimen untuk memberikan informasi lebih lanjut tentang elemen layanan yang memengaruhi kepuasan pengguna.

Hasil penelitian sebelumnya menunjukkan bahwa masih ada beberapa celah penelitian yang perlu diteliti lebih lanjut. Studi sebelumnya tentang perbandingan model Support Vector Machine (SVM) dan IndoBERT belum mencapai kesimpulan yang konsisten; beberapa penelitian menunjukkan bahwa IndoBERT lebih baik, sementara penelitian lain menunjukkan bahwa SVM dapat memberikan hasil klasifikasi yang lebih baik untuk fitur tertentu dari dataset. Perbedaan hasil menunjukkan bahwa karakteristik data, proses preprocessing, dan teknik pelabelan yang digunakan masih memengaruhi kinerja kedua model. Penerapan *Aspect Based Sentiment Analyst* (ABSA) masih terlihat minoritas dalam penelitian. Sebagian besar penelitian lebih berfokus pada analisis sentimen secara keseluruhan atau hanya menggunakan ABSA pada satu subjek penelitian. Oleh karena itu, belum banyak penelitian yang membandingkan SVM dan IndoBERT dalam konteks ABSA menggunakan ulasan dari dua platform e-commerce terbesar di Indonesia, Shopee dan Tokopedia.

Penelitian ini memberikan kontribusi dengan menerapakan *Aspect Based Sentiment Analysis* (ABSA) dengan menggunakan 4 aspek utama dengan aplikasi Tokopedia dan Shopee. Sehingga hasilnya lebih objektif, penelitian ini membandingkan kinerja model Support Vector Machine (SVM) dan IndoBERT dengan dataset, tahapan preprocessing, dan skenario evaluasi yang sama. Studi ini tidak hanya mengevaluasi model; itu juga menerapkannya ke dalam aplikasi berbasis Streamlit, yang dapat secara interaktif menampilkan hasil prediksi aspek, sentimen, dan tingkat kepercayaan model. Oleh karena itu, diharapkan bahwa penelitian ini akan membantu mengembangkan metode dan menerapkan analisis sentimen berbasis aspek untuk ulasan e-commerce berbahasa Indonesia.

2.2 Teori Terkait

2.2.1 E-Commerce

E-commerce adalah perdagangan yang dilakukan secara elektronik melalui jaringan internet, yang memungkinkan transaksi jual beli dilakukan di mana pun dan kapan pun [17]. Sistem ini memungkinkan penjual dan pembeli untuk bertransaksi secara digital melalui platform online, termasuk pemesanan produk, pembayaran, dan pengiriman barang. Perkembangan internet dan perangkat mobile telah mendorong pertumbuhan e-commerce di banyak negara, termasuk Indonesia [18]. Berbagai platform perdagangan online seperti Tokopedia, Shopee, dan Lazada telah mempermudah orang untuk berbelanja secara online. E-commerce juga memungkinkan bisnis untuk memperluas jangkauan pasar mereka tanpa harus memiliki toko fisik. Akibatnya, e-commerce menjadi salah satu bidang penting yang berkontribusi pada perkembangan ekonomi digital [19][20].

Seiring dengan peningkatan jumlah pengguna internet dan digitalisasi sistem pembayaran, tren e-commerce Indonesia terus berkembang. Menurut data Bank Indonesia, transaksi e-commerce di Indonesia mencapai sekitar Rp227,8 triliun pada semester pertama tahun 2022 dan diperkirakan meningkat hingga Rp489 triliun pada tahun 2022. Nilai ini terus meningkat menjadi sekitar Rp572 triliun pada tahun 2023 dan Rp689 triliun pada tahun

2024, masing-masing. Meningkatnya kepercayaan masyarakat terhadap transaksi online serta kemudahan sistem pembayaran digital juga mendorong pertumbuhan sektor e-commerce Indonesia. Meningkatnya aktivitas transaksi digital disebabkan oleh konsumen yang semakin terbiasa berbelanja secara online [21].

Jumlah pengguna e-commerce dan transaksi di Indonesia telah meningkat pesat, bersama dengan nilai transaksi yang terus meningkat. Jumlah transaksi e-commerce mencapai sekitar 1,44 miliar transaksi pada kuartal III tahun 2025, meningkat sekitar 20,5% setiap tahunnya, menurut data Bank Indonesia [22]. Pertumbuhan ini menunjukkan bahwa orang semakin sering menggunakan platform digital untuk memenuhi kebutuhan sehari-hari. Selain itu, seiring dengan kemajuan ekonomi digital dan inovasi teknologi, nilai transaksi e-commerce di Indonesia diperkirakan akan terus meningkat. Selain itu, menurut beberapa laporan, dalam beberapa tahun ke depan, pasar e-commerce Indonesia dapat mencapai puluhan miliar dolar. Ini menempatkan Indonesia di antara pasar e-commerce terbesar di Asia Tenggara [23].

2.2.2 Aspect-Based Sentiment Analyst (ABSA)

Analisis Sentimen Berbasis Aspek (ABSA) adalah variasi dari analisis sentimen yang bertujuan untuk mengidentifikasi sentimen berdasarkan aspek tertentu dari suatu produk atau layanan. Berbeda dengan analisis sentimen konvensional, yang hanya menghasilkan hasil yang berkaitan dengan semua sentimen, ABSA mampu memberikan analisis yang lebih khusus pada aspek tertentu dari suatu objek [24]. Metode ini memungkinkan sistem untuk mengidentifikasi perasaan pengguna yang positif dan negatif. Misalnya, dalam e-commerce, fitur seperti kualitas produk, layanan pengiriman, harga, dan fitur aplikasi dapat dievaluasi. Perusahaan dapat mendapatkan data yang lebih rinci tentang bagaimana pelanggan menggunakan layanan dengan menggunakan metode ini. Oleh karena itu, ABSA menjadi alat yang sangat penting untuk menilai pendapat pelanggan [25].

Untuk mendapatkan hasil analisis yang lebih mendalam, terdapat beberapa tahapan penting yang dilakukan selama proses penerapan ABSA. Langkah pertama adalah menemukan fitur atau elemen yang disebutkan dalam teks ulasan pengguna. Setelah berhasil menemukan fitur tersebut, langkah berikutnya adalah menentukan perasaan yang terkait dengan fitur tersebut [26]. Dalam proses ini, proses pemrosesan bahasa alami biasanya digunakan untuk memahami bagaimana kata-kata dalam suatu kalimat berhubungan satu sama lain. Selain itu, untuk mengklasifikasikan sentimen berdasarkan elemen yang telah diidentifikasi, machine learning atau deep learning juga dapat digunakan. Metode ini memungkinkan sistem untuk menghasilkan analisis yang lebih mendalam tentang persepsi pengguna terhadap berbagai aspek layanan [27].

Metode ABSA telah mengalami perkembangan besar dalam beberapa tahun terakhir seiring dengan kemajuan teknologi pengolahan bahasa alami. Untuk meningkatkan akurasi dalam proses identifikasi aspek dan klasifikasi sentimen, beberapa penelitian telah mengembangkan model berbasis machine learning dan deep learning. Dengan memahami hubungan kontekstual antar kata dalam suatu kalimat, model berbasis transformer seperti BERT dan IndoBERT banyak digunakan. Model-model ini dapat mengidentifikasi aspek dan sentimen dengan lebih akurat. Selain itu, model deep learning memungkinkan sistem untuk menangani dataset teks yang lebih besar dengan lebih baik. Oleh karena itu, ABSA telah berkembang menjadi salah satu metode yang paling efisien untuk menganalisis opini pengguna pada berbagai platform digital, termasuk platform e-commerce seperti Tokopedia & Shopee [28].

2.2.3 Analisa Sentimen

Analisis sentimen adalah bagian dari proses pengolahan bahasa alami (NLP) yang digunakan untuk menemukan, mengekstraksi, dan mengklasifikasikan pendapat atau emosi yang ada dalam teks. Teknik analisis sentimen bertujuan untuk mengetahui sikap atau persepsi seseorang terhadap suatu produk, layanan, atau masalah tertentu yang biasanya ditulis dalam teks

[29]. Analisis sentimen biasanya mengelompokkan opini ke dalam kategori seperti sentimen positif, negatif, atau netral. Metode ini banyak digunakan untuk mengumpulkan data dari forum diskusi, platform media sosial, dan ulasan pengguna pada platform digital. Analisis sentimen membantu bisnis memahami bagaimana pelanggan melihat barang atau jasa mereka. Oleh karena itu, analisis sentimen adalah salah satu teknik penting dalam pengolahan data teks yang membantu pengambilan keputusan berbasis data [30].

Seiring perkembangan penelitian, analisis sentimen tidak hanya dilakukan pada tingkat dokumen (*document-level sentiment analysis*), tetapi juga berkembang menjadi analisis pada tingkat kalimat (*sentence-level sentiment analysis*) dan tingkat aspek (*aspect-level sentiment analysis*). Pendekatan ini memungkinkan sistem memberikan hasil analisis yang lebih rinci karena mampu mengidentifikasi objek atau aspek tertentu yang menjadi penyebab munculnya suatu opini. Oleh karena itu, analisis sentimen banyak diterapkan dalam berbagai bidang, seperti analisis ulasan produk, pelayanan publik, media sosial, kesehatan, hingga sistem pendukung keputusan [31]. Oleh karena itu, analisis sentimen berkembang seiring dengan kemajuan dalam teknologi pengolahan bahasa alami.

Dalam implementasinya, analisis sentimen berkembang menjadi beberapa pendekatan utama, yaitu Lexicon-Based Sentiment Analysis, Machine Learning-Based Sentiment Analysis, Deep Learning-Based Sentiment Analysis. Masing-masing pendekatan memiliki karakteristik, kelebihan, dan keterbatasan yang berbeda sehingga pemilihannya disesuaikan dengan karakteristik data, kebutuhan penelitian, serta tujuan analisis. Penelitian ini menggunakan pendekatan Lexicon-Based untuk proses pelabelan sentimen, sedangkan proses klasifikasi dilakukan menggunakan pendekatan Machine Learning melalui SVM dan Deep Learning menggunakan IndoBERT.

2.2.3.1. Lexicon Based Labelling

Lexicon-Based Labeling adalah metode pelabelan sentimen yang bekerja dengan cara mencocokkan kata-kata dalam teks dengan sebuah kamus sentimen (lexicon) yang telah dibuat sebelumnya, di mana setiap kata dalam kamus tersebut sudah memiliki skor atau polaritas sentimen yang telah ditentukan secara manual oleh para ahli linguistik. Karena metode ini termasuk dalam kategori pendekatan yang berbasis aturan dan tidak memerlukan proses pelatihan model pelatihan komputer, keputusan tentang kategorisasi sentimen sepenuhnya bergantung pada keberadaan dan nilai kata dalam lexicon [32]. SentiWordNet, VADER, dan AFINN adalah beberapa lexicon yang banyak digunakan, dan untuk Bahasa Indonesia, InSet (Indonesian Sentiment Lexicon) memetakan kata-kata dalam Bahasa Indonesia dengan skor positif dan negatifnya [33].

Labeling berbasis *lexicon* biasanya dimulai dengan tokenisasi teks. Setelah itu, setiap token dicocokkan dengan lexicon untuk mendapatkan skor sentimen tertentu. Skor-skor ini kemudian digabungkan untuk menghasilkan skor sentimen total untuk kalimat atau dokumen. Berikut adalah rumus dasar:

$$S(d) = \sum_{i=1}^{n} \text{score}(w_i)$$

Rumus diatas merupakan skor sentimen dokumen dd d adalah S(d), wiw_i wi adalah kata ke-ii i dalam dokumen, dan skor(wi)\text{score}(w_i) adalah skor sentimen kata dalam lexicon. Label sentimen kemudian dibuat dengan menggunakan hasil akhir S(d)S(d) dan threshold yang telah ditentukan [34]. Meskipun metode ini sederhana dan mudah dipahami, labeling yang berbasis lexicon memiliki beberapa keterbatasan yang cukup signifikan. Metode ini tidak dapat memahami konteks setiap kalimat secara menyeluruh, jadi mungkin salah label kalimat yang mengandung kata negatif (seperti "tidak bagus"), sarkasme, atau kata ambigu yang artinya bergantung pada konteks. Selain itu, hasil pelabelan sangat bergantung pada kelengkapan dan kualitas kosa kata yang digunakan. Jika sebuah kata tidak ada dalam kosa kata yang digunakan,

maka kata tersebut dianggap nol dan diabaikan, yang dapat menyebabkan hasil pelabelan yang tidak adil. Oleh karena itu, dalam banyak penelitian, labeling yang berbasis lexicon digunakan sebagai tahap awal otomasi pelabelan data sebelum data dilatih untuk model deep learning seperti IndoBERT, yang digunakan dalam pipeline modeling ini [35].

2.2.3.2. Machine Learning Based Sentiment Analyst

Machine Learning-Based Sentiment Analysis merupakan pendekatan yang melakukan klasifikasi sentimen berdasarkan pola yang dipelajari dari data berlabel. Berbeda dengan pendekatan lexicon yang menggunakan aturan atau kamus sentimen, metode machine learning mempelajari hubungan antara fitur teks dan label sentimen melalui proses pelatihan (*training*). Algoritma yang umum digunakan antara lain **Naïve Bayes**, **Decision Tree**, **Random Forest**, **Logistic Regression**, dan **Support Vector Machine (SVM)**. Keunggulan pendekatan ini adalah kemampuannya dalam mengenali pola klasifikasi yang lebih kompleks dibandingkan metode berbasis aturan[36].

$$y = f(x)$$

Pendekatan machine learning memiliki beberapa kelebihan, seperti kemampuan menghasilkan akurasi yang tinggi pada data berlabel dan proses inferensi yang relatif cepat setelah model selesai dilatih. Namun demikian, performa model sangat dipengaruhi oleh kualitas preprocessing, representasi fitur, distribusi data, serta jumlah data pelatihan. Oleh karena itu, tahapan preprocessing dan ekstraksi fitur menjadi faktor penting dalam meningkatkan performa klasifikasi sentimen berbasis machine learning.

2.2.3.3. Deep Learning Based Sentiment Analysis

Deep Learning-Based Sentiment Analysis merupakan perkembangan dari pendekatan machine learning yang menggunakan jaringan saraf tiruan (*Artificial Neural Network*) untuk mempelajari representasi teks secara otomatis tanpa memerlukan proses rekayasa fitur (*feature engineering*) secara manual. Model deep learning seperti **CNN**, **LSTM**, dan terutama **Transformer** mampu memahami hubungan antar kata dalam suatu

kalimat sehingga menghasilkan representasi kontekstual yang lebih baik dibandingkan metode konvensional. Dalam penelitian bahasa Indonesia, salah satu model transformer yang banyak digunakan adalah **IndoBERT [37].**

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

Perkembangan terbaru menunjukkan bahwa arsitektur **Transformer** menjadi pendekatan yang paling dominan dalam analisis sentimen karena menggunakan mekanisme *self-attention* yang mampu mempelajari hubungan antar kata secara dua arah (*bidirectional*). Mekanisme ini memungkinkan model memahami konteks kalimat secara lebih menyeluruh dibandingkan arsitektur sebelumnya seperti CNN dan LSTM, terutama pada kalimat dengan ketergantungan kata yang panjang. Salah satu implementasi Transformer untuk bahasa Indonesia adalah **IndoBERT**, yaitu model *pre-trained language model* yang telah dilatih menggunakan korpus bahasa Indonesia dalam jumlah besar sehingga memiliki kemampuan memahami karakteristik bahasa Indonesia secara lebih baik. Berbagai penelitian terbaru menunjukkan bahwa model berbasis Transformer, termasuk IndoBERT, mampu menghasilkan performa yang lebih tinggi dibandingkan metode *machine learning* konvensional pada tugas klasifikasi teks dan analisis sentimen. Oleh karena itu, penelitian ini menggunakan IndoBERT sebagai representasi pendekatan *deep learning* untuk dibandingkan dengan algoritma Support Vector Machine (SVM) dalam melakukan klasifikasi sentimen berbasis aspek [38].

2.2.4 Natural Language Processing (NLP)

Natural Language Processing atau biasa disebut NLP, merupakan bagian dari kecerdasan buatan (AI) yang berfokus pada kemampuan computer untuk dapat memahami dan melakukan interpretasi serta menghasilkan Bahasa manusia secara alami [39]. NLP menggabungkan

ilmu linguistic komputasional dengan Teknik pembelajaran mesin (Machine Learning) dan deep learning. Dalam analisis sentiment, NLP memiliki peran sebagai fondasi dalam memungkinkan system untuk mengekstraksi makna dari teks ulasan pengguna maupun mengindetikasi opini. Proses ini melibatkan tahapan preprocessing yang bertujuan untuk mengubah data teks menjadi representasi numerik yang dapat diproses oleh algoritma [40].

Salah satu konsep fundamental dalam NLP adalah representasi teks ke dalam bentuk vektor numerik yang dikenal sebagai word embedding . Teknik klasik seperti TF-IDF (Term Frequency-Inverse Document Frequency) merepresentasikan bobot sebuah kata dalam dokumen menggunakan rumus:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{df(t)}\right)$$

Pada rumus diatas adalah rumus dari penerapan NLP dalam machine learning yang dimana $TF(t,d)$adalah frekuensi kemunculan term $t$dalam dokumen d, Nadalah total jumlah dokumen, dan $df(t)$adalah jumlah dokumen yang mengandung term t. Pendekatan ini kemudian berkembang menjadi teknik berbasis neural network seperti Word2Vec dan GloVe yang mampu menangkap hubungan semantik antar kata [41]. Dalam konteks penelitian ini, representasi teks yang dihasilkan melalui proses NLP menjadi input penting bagi model Support Vector Machine (SVM) maupun IndoBERT dalam melakukan klasifikasi sentimen berbasis aspek (Aspect-Based Sentiment Analysis/ABSA) pada ulasan aplikasi Tokopedia dan Shopee.

2.2.5 Large Language Model (LLM)

Large Language Model atau bisa disebut LLM merupakan model Bahasa berskala besar yang dilatih pada miliaran token teks sehingga memahami konteks, semantic, dan relasi antar kata. Indobert adalah implementasi arsitektur BERT yang dilatih khusus pada korpus Bahasa

Indonesia, sehingga memiliki nuansa Bahasa slang, struktur kalimat pasif dan konteks local. Dengan pendekatan **zero-shot**, model tidak dilatih ulang pada data berlabel sentimen namun langsung digunakan sebagaimana adanya. Zero-shot dilakukan karena IndoBERT sudah memahami hubungan semantic antar kalimat melalui tugas *Natural Language Inference* (NLI) yang mempunyai kemampuan menentukan apakah satu kalimat logis mengikuti kalimat lain [42].

Pada saat ero-shot classification, input yang dikirim ke IndoBERT bukan sekadar teks biasa, melainkan pasangan kalimat dalam format NLI: **premise** adalah teks asli yang ingin dianalisis, sedangkan **hypothesis** adalah label sentimen yang diubah menjadi kalimat natural, isalnya *"Teks ini mengungkapkan sentimen positif"*. IndoBERT memproses pasangan ini melalui 12 layer Transformer, menghasilkan representasi vektor pada token CLS yang merangkum makna keseluruhan pasangan kalimat tersebut. Vektor [CLS] ini kemudian diumpankan ke **NLI classification head** yang menghasilkan tiga skor mentah (logits) — yaitu skor untuk *entailment* (teks mendukung hypothesis), *neutral*, dan *contradiction* (teks menolak hypothesis). Proses ini diulang untuk setiap kandidat label sentimen (positif, negatif, netral), sehingga model menghasilkan skor entailment untuk masing-masing label secara terpisah [43].

Skor entailment mentah yang dihasilkan untuk setiap label belum bisa langsung dibandingkan karena nilainya bisa berupa bilangan negatif atau sangat besar. Di sinilah fungsi **Softmax** berperan. Rumus pada softmax adalah :

$$P(\text{label}_i \mid \text{teks}) = \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}}$$

Rumus diatas merupakan rumus dalam proses data ulasan sedang dilakukan labeling secara otomatis dengan cara mengambil semua skor entailment (s_1, s_2, ..., sC), mengeksponensiasikannya dengan fungsi *e^x* agar semua nilai menjadi positif, lalu membaginya dengan total seluruh

nilai eksponen sehingga hasil akhirnya adalah distribusi probabilitas yang jumlahnya tepat 1,0. Misalnya jika skor entailment untuk *positif* = 3.2, *negatif* = 0.8, *netral* = 1.1, maka setelah Softmax diperoleh probabilitas ~0.78, ~0.07, ~0.15 — dan model menyimpulkan sentimen **positif** karena memiliki probabilitas tertinggi. Inilah keunggulan zero-shot dengan IndoBERT: tanpa satu pun data training berlabel, model mampu mengklasifikasikan sentimen hanya bermodal pemahaman bahasa yang sudah dimilikinya [44].

2.2.6 Mini LM-12 Aspect Labeling

Dalam penelitian ini, model yang digunakan adalah para-multilingual-*MiniLM-L12-v2*. Model ini termasuk dalam keluarga sentence embedding transformers dan memiliki kemampuan untuk mengubah kalimat atau paragraf menjadi representasi vektor numerik. Keluarga ini juga memiliki kemampuan untuk melakukan tugas seperti mencari secara semantic, memcluster, dan mengukur kemiripan antar teks. Model ini dapat digunakan pada data ulasan berbahasa Indonesia karena kemampuan multilingualnya, terutama karena ulasan pengguna e-commerce sering berbentuk kalimat pendek, tidak baku, dengan variasi makna yang beragam [45].

Konsep Sentence-BERT, yang dikembangkan oleh Reimers dan Gurevych, diciptakan untuk menghasilkan representasi kalimat yang bermakna secara semantik dan dapat dibandingkan dengan cosine similarity. Konsep ini mendukung penggunaan MiniLM-L12 dalam penelitian ini [46]. Metode ini meningkatkan efisiensi proses pencarian kemiripan antar kalimat dibandingkan dengan penggunaan BERT secara langsung untuk membandingkan pasangan kalimat satu per satu. Karena proses labeling aspek membutuhkan perbandingan kemiripan makna antara teks ulasan pengguna dan kategori aspek yang telah ditentukan, konsep penerapan frasa sesuai digunakan dalam penelitian ini.

Secara arsitektural, MiniLM dikembangkan melalui pendekatan deep self-attention distillation, yang berarti proses kompresi model transformer

yang lebih besar menjadi model yang lebih kecil tetapi tetap mempertahankan kemampuan pemahaman bahasa yang baik. Wang et al. menjelaskan bahwa, karena model ini dapat meniru mekanisme self-attention dari model transformer yang lebih besar, model ini tetap efektif untuk berbagai tugas pemrosesan bahasa alami, terlepas dari ukuran dan kebutuhan kompu yang berbeda [47]. Untuk alasan akademis, MiniLM-L12 adalah pilihan yang tepat untuk penelitian ini karena proses aspek labeling membutuhkan model yang dapat menangkap hubungan semantik teks sambil tetap efisien dalam mengolah banyak ulasan pengguna [48].

2.2.7 Data Preprocessing

Data preprocessing merupakan tahapan awal dalam pengolahan data yang bertujuan untuk meningkatkan kualitas data sebelum digunakan pada proses analisis maupun pemodelan *machine learning* dan *deep learning*. Pada penelitian berbasis *Natural Language Processing* (NLP), data teks yang diperoleh dari media sosial atau ulasan pengguna umumnya masih mengandung berbagai bentuk *noise*, seperti penggunaan huruf kapital yang tidak konsisten, tanda baca, angka, simbol, *emoji*, *hyperlink*, maupun kata-kata yang tidak memiliki makna terhadap proses analisis. Kondisi tersebut dapat memengaruhi representasi fitur dan menurunkan kinerja model klasifikasi apabila tidak dilakukan proses pembersihan data terlebih dahulu. Oleh karena itu, data preprocessing menjadi salah satu tahapan yang sangat penting untuk menghasilkan data yang lebih bersih, terstruktur, dan siap digunakan pada proses pemodelan[49].

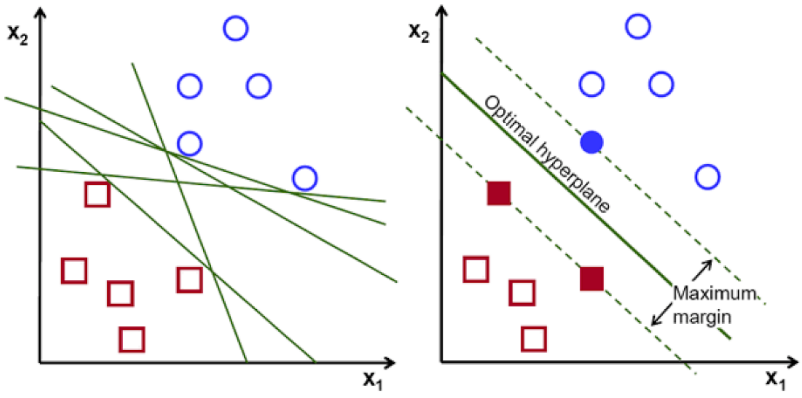
Data preprocessing biasanya terdiri dari sejumlah teknik yang dimaksudkan untuk menormalisasi dan menyederhanakan representasi teks sehingga algoritma lebih mudah memprosesnya. Case folding, cleansing, tokenization, stopword removal, dan stemming adalah beberapa teknik yang umum digunakan. Masing-masing metode melakukan hal-hal yang berbeda untuk mencapai tujuan tertentu, seperti menyeragamkan format penulisan, menghilangkan karakter yang tidak diperlukan, membagi kalimat menjadi simbol, menghilangkan kata yang tidak

memiliki informasi penting, dan mengembalikan kata ke bentuk aslinya. Tahapan preprocessing yang tepat dapat meningkatkan kualitas fitur yang dihasilkan, mengurangi dimensi data, dan membantu model mengenali pola sentimen secara lebih baik [50].

2.3 Framework/Algoritma yang digunakan

Penelitian ini memakai model Support Vector Machine (SVM) dan IndoBERT dipilih karena keduanya mewakili dua pendekatan yang berbeda untuk klasifikasi teks. Karena SVM sangat baik dalam mengolah data teks yang sangat besar, terutama ketika teks diwakili menggunakan pembobotan kata seperti TF-IDF. Dalam kasus ulasan pengguna aplikasi e-commerce, data teks memiliki banyak kata yang berbeda, sehingga SVM dapat digunakan sebagai model dasar yang efektif untuk mengklasifikasikan sentimen berdasarkan pola kata yang muncul di setiap aspek. Menurut beberapa penelitian, kombinasi TF-IDF dan SVM masih banyak digunakan dalam analisis sentimen. Ini karena keduanya dapat menghasilkan klasifikasi data teks yang sangat baik. Berikut penjelasan dari model SVM dan IndoBERT.

1.3.1. SVM (Support Vector Machine)



Gambar 2. 1 SVM

Gambar 2.4 (source: https://www.mathworks.com/discovery/support-vector-machine.html) adalah algoritma machine learning yang digunakan untuk melakukan regresi dan klasifikasi dataset adalah Support Vector Machine (SVM) [51]. Pengenalan pola, klasifikasi teks, dan analisis

sentimen adalah beberapa bidang di mana metode ini banyak digunakan, pertama kali diperkenalkan oleh Vladimir Vapnik pada tahun 1995. Prinsip utama SVM adalah menemukan hyperplane terbaik yang dapat memisahkan data ke dalam dua kelas dengan margin terbesar. Margin adalah jarak antara hyperplane dan titik data terdekat dari masing-masing kelas, yang disebut sebagai support vector [52]. Diharapkan bahwa model dapat melakukan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya dengan margin yang maksimal. Akibatnya, SVM dianggap sebagai algoritma yang memiliki kemampuan klasifikasi yang baik, terutama untuk data yang sangat besar [53].

SVM bekerja dalam proses klasifikasi teks dengan mengubah data teks menjadi representasi numerik dengan menggunakan teknik ekstraksi fitur seperti TF-IDF atau Bag of Words. Setelah data diubah menjadi representasi numerik, algoritma SVM mencari hyperplane yang memiliki kemampuan untuk memisahkan data berdasarkan label kelas yang telah diberikan. SVM dapat menggunakan fungsi kernel untuk memetakan data ke dalam ruang yang lebih besar untuk pemisahan yang lebih baik jika data tidak dapat dipisahkan secara linear. SVM memiliki beberapa jenis kernel yang dapat digunakan, seperti linear kernel, polynomial kernel, radial basis function kernel, dan sigmoid kernel. Pada penelitian ini, kernel yang digunakan adalah linear kernel. Pilihan kernel linear didasarkan pada karakteristik data teks yang diwakili menggunakan TF-IDF, yang menghasilkan fitur berdimensi tinggi. Kernel linear dinilai sesuai karena mampu memisahkan data secara efisien pada ruang fitur dan sering digunakan dalam tugas klasifikasi teks [54].

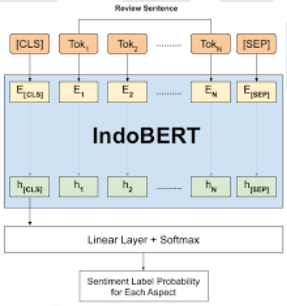
$$w \cdot x + b = 0$$

$$\min \quad \frac{1}{2}|w|^2$$

Rumus diatas merupakan rumus cara kerja *Hyperlane Optimal* pada algoritma SVM yang memiliki kemampuan untuk memisahkan dua kelas data dengan margin yang paling besar, dapat ditemukan dengan

menggunakan rumus dasar *Support Vector Machine* (SVM). Hyperplane tersebut dinyatakan dalam persamaan $w \cdot x + b = 0$, di mana $w$merupakan vektor bobot (weight vector), $x$adalah vektor fitur dari data, dan $b$adalah nilai bias [55]. Dalam proses klasifikasi, setiap data $x_i$dengan label $y_i$harus memenuhi kondisi $y_i(w \cdot x_i + b) \geq 1$, yang bertujuan agar data berada pada sisi hyperplane yang benar sesuai kelasnya. Untuk mendapatkan hyperplane terbaik, SVM meminimalkan fungsi objektif $\frac{1}{2} || w ||^2$, yang bertujuan untuk memaksimalkan jarak atau margin antara dua kelas data. Dengan memaksimalkan margin tersebut, model dapat meningkatkan kemampuan generalisasi terhadap data baru yang belum pernah dilatih sebelumnya . Oleh karena itu, pendekatan ini membuat SVM menjadi salah satu algoritma yang efektif dalam melakukan klasifikasi data, termasuk pada data teks seperti analisis sentiment [56]. Kernel linear dipilih karena data ulasan pengguna telah direpresentasikan ke dalam bentuk numerik menggunakan TF-IDF, yang menghasilkan ruang fitur yang besar. Kernel linear dinilai sesuai untuk karakteristik data teks seperti ini karena mereka memiliki proses komputasi yang lebih sederhana dan mampu melakukan pemisahan kelas secara efisien.

1.3.2. IndoBERT



Gambar 2. 2 IndoBERT

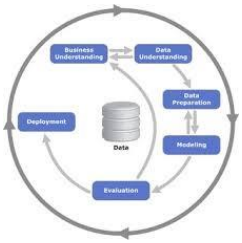
Gambar 2.5 (source: https://indolem.github.io/IndoBERT/) adalah algoritma IndoBERT yang merupakan model bahasa berbasis pembelajaran mendalam yang dirancang untuk meningkatkan pemahaman teks berbahasa Indonesia. IndoBERT adalah adaptasi dari arsitektur BERT (Bidirectional Encoder Representations from Transformers) yang dikembangkan oleh

Google. Dilatih menggunakan dataset bahasa Indonesia yang besar, model ini memiliki kemampuan pemahaman yang lebih baik tentang struktur bahasa, konteks kata, dan hubungan antar kalimat dibandingkan dengan metode pembelajaran mesin konvensional [57]. Arsitektur transformer model ini memungkinkan pemrosesan teks secara dua arah (bidirectional). Dengan kemampuan ini, IndoBERT banyak digunakan dalam berbagai tugas pemrosesan bahasa alami, seperti klasifikasi teks, analisis sentimen, pengenalan entitas bernama, dan menjawab pertanyaan. IndoBERT sering digunakan dalam penelitian analisis sentimen karena mampu menciptakan representasi teks yang lebih kontekstual.

$$[\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V]$$

Rumus utama yang digunakan dalam model IndoBERT berasal dari mekanisme **Scaled Dot-Product Attention** pada arsitektur transformer yang juga digunakan pada BERT[58]. Persamaan attention dituliskan sebagai $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, yang berfungsi untuk menghitung tingkat perhatian antara setiap kata dalam suatu kalimat. Pada persamaan tersebut, Q(Query), K(Key), dan V(Value) merupakan representasi vektor dari kata-kata dalam teks yang digunakan untuk menentukan hubungan antar kata. Perkalian matriks QK^T digunakan untuk menghitung tingkat kesamaan antara kata satu dengan kata lainnya, sedangkan pembagian dengan $\sqrt{d_k}$ bertujuan untuk menstabilkan nilai perhitungan agar tidak terlalu besar. Selanjutnya, fungsi **softmax** digunakan untuk mengubah nilai tersebut menjadi bobot probabilitas yang menunjukkan tingkat perhatian setiap kata terhadap kata lainnya. Dengan mekanisme ini, model IndoBERT mampu memahami konteks kata secara lebih mendalam sehingga menghasilkan representasi teks yang lebih akurat dalam berbagai tugas Natural Language Processing seperti analisis sentiment [59].

1.3.3. CRISP-DM



Gambar 2. 3 CRISP-DM

Gambar 2.6 (source: https://www.dicoding.com/blog/crisp-dm-tahapan-studi-kasus-kelebihan-dan-kekurangan/) adalah metode yang digunakan untuk topik ini adalah CRISP-DM. CRISP-DM (Cross Industry Standard Process for Data Mining) adalah salah satu metodologi standar yang banyak digunakan dalam proses pengolahan dan analisis data untuk menghasilkan informasi yang berharga. Metodologi ini dibuat pada tahun 1996 oleh beberapa konsultan dan perusahaan teknologi data mining sebagai kerangka kerja yang sistematis untuk menjalankan proyek data mining. CRISP-DM mencakup tahapan yang terorganisir yang dimulai dengan pemahaman masalah bisnis dan berakhir dengan evaluasi hasil model yang dihasilkan. Karena metode ini memberikan pedoman yang jelas untuk setiap tahap penelitian, ia banyak digunakan dalam penelitian tentang machine learning, data mining, dan analisis data. Metode ini memungkinkan proses penelitian dilakukan secara lebih terarah dan sistematis. Karena itu, CRISP-DM dipilih sebagai metodologi untuk mengolah dan menganalisis data ulasan dalam penelitian ini.

Metode CRISP-DM terdiri dari enam tahapan utama yang saling terkait: pemahaman bisnis, pemahaman data, persiapan data, modeling, evaluasi, dan penerapan. Setiap tahapan berfungsi untuk mendukung proses analisis data secara menyeluruh. Dalam penelitian ini, tahapan CRISP-DM digunakan untuk mengevaluasi sentimen ulasan pengguna e-commerce dengan membandingkan kinerja algoritma Support Vector Machine dan model IndoBERT. Metode ini memungkinkan proses pengolahan data teks dilakukan secara sistematis mulai dari pengumpulan data hingga evaluasi model. CRISP-DM juga membantu peneliti memastikan bahwa temuan

analisis sesuai dengan tujuan penelitian. Oleh karena itu, pendekatan ini sangat relevan digunakan dalam penelitian yang melibatkan analisis sentimen berbasis mesin dan pengajaran mendalam. Berikut tahapan CRISP-DM pada topik ini:

1. **Business Understanding**

   Tahap pertama CRISP-DM adalah memahami masalah yang ingin diselesaikan dan tujuan penelitian. Pada titik ini, peneliti menentukan tujuan dan kebutuhan analisis. Tujuan utama penelitian ini adalah untuk menganalisis sentimen ulasan pengguna pada aplikasi Tokopedia dan membandingkan kinerja dua model klasifikasi, Support Vector Machine dan IndoBERT. Selain itu, penelitian ini juga ingin mengetahui model mana yang dapat melakukan analisis sentimen berbasis aspek dengan paling baik. Proses analisis dapat dilakukan secara lebih terarah jika tujuan penelitian dipahami secara jelas.

2. **Data Understanding**

   Tahap kedua terdiri dari memperoleh pemahaman tentang data yang akan digunakan dalam penelitian. Pada titik ini, peneliti mengumpulkan data dan mengeksplorasi kumpulan data tersebut. Data yang digunakan dalam penelitian ini berasal dari ulasan pengguna untuk aplikasi Tokopedia, yang tersedia di Google Play Store. Data dikumpulkan melalui teknik scraping. Analisis awal dilakukan setelah data dikumpulkan untuk memahami variabel seperti jumlah data, distribusi teks, dan perbedaan pendapat pengguna.

3. **Data Preparation**

   Sebelum data digunakan dalam proses pemodelan, proses pembersihan dan persiapan data dikenal sebagai tahap persiapan data. Tahap ini sangat penting dalam penelitian analisis sentimen karena data teks sering mengandung banyak kata yang tidak relevan. Dalam tahap ini, teks dibersihkan, case folding, tokenisasi, penghapusan stopword, dan stemming dilakukan. Tahap ini bertujuan untuk menghasilkan dataset yang bersih dan siap digunakan dalam proses pelatihan model

pembelajaran mesin. Untuk model Support Vector Machine, teknik ekstraksi fitur seperti TF-IDF digunakan untuk mengubah data yang telah diproses menjadi representasi numerik

4. **Modeling**

   Proses pembuatan model machine learning atau deep learning untuk melakukan analisis terhadap dataset yang telah diproses dikenal sebagai tahap modeling. Dua model yang berbeda digunakan dalam penelitian ini: Support Vector Machine sebagai model pembelajaran mesin dan IndoBERT sebagai model pembelajaran mendalam berbasis transformer. Sementara itu, IndoBERT meningkatkan pemahaman konteks bahasa melalui proses fine-tuning terhadap dataset ulasan pengguna, model SVM digunakan untuk melakukan klasifikasi sentimen berdasarkan fitur teks yang telah diekstraksi. Tujuan dari langkah ini adalah untuk membuat model klasifikasi sentimen yang dapat digunakan untuk membandingkan kinerja kedua metode.

5. **Evaluation**

   Tujuan dari tahap evaluasi adalah untuk mengevaluasi kinerja model yang telah dibangun pada tahap sebelumnya. Pada tahap ini, kinerja model diukur dengan menggunakan metrik seperti ketepatan, ketepatan, recall, dan skor F1. Kemudian, hasil evaluasi dari kedua model dibandingkan untuk menentukan model mana yang paling efektif dalam melakukan analisis sentimen berbasis aspek. Tahap ini juga digunakan untuk memastikan bahwa model yang dibuat memenuhi tujuan penelitian.

6. **Deployment**

   Penyebaran hasil analisis adalah tahap terakhir CRISP-DM. Dalam penelitian ini, deployment dilakukan dengan menyajikan hasil analisis sentimen serta perbandingan performa model. Diharapkan hasil penelitian ini memberikan informasi tentang persepsi pengguna terhadap layanan aplikasi Tokopedia serta rekomendasi tentang model analisis sentimen berbasis aspek yang paling efektif. Dengan demikian,

hasil penelitian ini dapat digunakan sebagai referensi untuk pengembangan penelitian lanjutan.

2.4 Tools/software yang digunakan

2.4.1. Python

Gambar 2. 4 Python

Gambar 2.5 Python merupakan bahasa pemrograman yang berorientasi objek dan memiliki struktur data tingkat tinggi. Bahasa ini terkenal, serbaguna, dan banyak digunakan. Diciptakan oleh Guido van Rossum, Python pertama kali dirilis pada tahun 1991 dengan penekanan pada kemudahan penggunaan, dan juga kejelasan, kemudahan dalam pemahaman kode, mengintegrasikan kemampuan, kejelasan sintaks. Dengan tujuan menyederhanakan tugas programmer. Dan juga Python memiliki kelebihan yaitu ; efisiensi waktu dalam pembuatan program, kemudahan dalam pengembangan. Python memiliki fitur yang dimana masing masing fitur seperti ; Sintaks yang mudah dipelajari membuat mudah untuk dipelajari oleh pemula. Bahasa yang diinterpretasikan yang dimana kode dieksekusi baris per baris, mudah dalam pengujian. Library dukungan seperti ; Pandas, Numpy, Tensorflow. Python juga memiliki kegunaan yang lain yaitu seperti pada Data Science yaitu library seperti Pandas, Numpy, Matplotlib, Scipy. Pada Machine Learning library meliputi Tensorflow, Keras, Scikit learn.

2.4.2. Visual Studio Code

Gambar 2. 5 Visual Studio Code

Visual Studio Code (VS Code), yang banyak digunakan oleh pengembang perangkat lunak dan peneliti data science, adalah salah satu editor kode Microsoft yang paling ringan namun kuat. Untuk menulis, mengedit, dan menjalankan kode Python yang digunakan dalam pengumpulan data, preprocessing, pelatihan model, dan tahap deployment, VS Code digunakan sebagai lingkungan pengembangan utama dalam penelitian ini. Dengan mendukung berbagai ekstensi seperti Python dan Jupyter Notebook, VS Code memudahkan proses pengembangan secara menyeluruh dalam satu platform yang terintegrasi, sehingga alur kerja penelitian dapat dilakukan secara lebih efisien..