BAB III

METODOLOGI PENELITIAN

3.1 Metode Penelitian

Penelitian ini menerapkan kerangka kerja *Knowledge Discovery in Databases* (KDD) sebagai dasar proses *multimodal sentiment analysis* berbasis *Generative AI*. Pemilihan KDD dilakukan setelah menimbang kelebihan dan keterbatasan sejumlah alternatif, antara lain *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dan SEMMA (*Sample, Explore, Modify, Model, Assess*), dengan menyesuaikan kebutuhan dan karakter permasalahan yang dikaji. Untuk memperjelas pertimbangan tersebut, disajikan perbandingan ringkas pada Tabel 3.1 [103].

Tabel 3.1 Perbandingan Model *Data Mining* 15

Aspek	KDD	SEMMA	CRISP-DM
Fokus	Penemuan pengetahuan secara umum dalam basis data	Proses *Data Mining* yang disesuaikan dengan alat SAS	Standar lintas industri untuk *Data Mining*
Tahapan/Proses	*Selection*, *Pre-processing*, *Transformation*, *Data Mining*, Interpretation/*Evaluation*	Sample, Explore, Modify, Model, Assess	Business Understanding, Data Understanding, Data Preparation, Modeling, *Evaluation*, Deployment
Pengetahuan Awal	Memahami *domain* aplikasi dan tujuan	Tersirat dalam tahap Sample	Tersurat dalam tahap Business Understanding
Deployment	Integrasi setelah KDD	Tersirat dalam penilaian temuan	Jelas sebagai fase terpisah
Alat/Penggunaan	Umum, tidak terikat pada perangkat lunak tertentu	Terikat pada alat SAS *Enterprise Miner*	Dapat disesuaikan dengan berbagai industri dan alat
Dokumentasi	Terbatas	Dokumentasi dan panduan terfokus pada SAS	Terperinci dan komprehensif
Fleksibilitas	Tinggi, tetapi memerlukan keputusan ahli	Sedang, bergantung pada kemampuan alat SAS	Tinggi, dapat disesuaikan dengan berbagai industri dan proyek
Kekuatan	Dapat diterapkan secara luas, fokus pada ekstraksi pengetahuan	Terstruktur dan mudah diikuti	Sangat terorganisir, terdokumentasi dengan baik, mendukung kolaborasi
Kelemahan	Kurang terstruktur, mungkin kurang jelas dalam tahapan	Terbatas pada ekosistem SAS	Membutuhkan kustomisasi untuk industri tertentu

Tabel 3.1 merangkum tiga kerangka kerja *data mining* yang sering dipakai dengan penekanan, alur, dan tingkat keluwesan yang berbeda. KDD berorientasi

pada penemuan pengetahuan di basis data melalui tahapan Selection, *Pre-processing*, *Tr*ansformation, *Data Mining*, serta *Evaluation* [104]. Pendekatan ini relatif luwes namun strukturnya lebih longgar. SEMMA, yang dipopulerkan SAS Institute, menyajikan lima tahap utama yaitu *Sample*, *Explore*, *Modify*, *Model*, dan *Assess* [105]. Alurnya mudah diikuti tetapi cenderung terikat pada ekosistem SAS. CRISP-DM menjadi standar lintas industri yang paling sistematis dan terdokumentasi, dengan enam tahap *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [106]. Penerapannya sering memerlukan penyesuaian agar selaras dengan kebutuhan spesifik proyek.

Kerangka KDD dipilih karena menyediakan alur kerja yang terstruktur untuk mengubah data mentah menjadi pengetahuan yang bermakna sekaligus memberi ruang iterasi pada setiap tahap. Lima tahap utamanya, yakni *Selection*, *Pre-processing*, *Transformation*, *Data Mining*, dan *Evaluation*, saling terkait dalam satu siklus sehingga memudahkan integrasi beragam metode analitik serta penyesuaian ulang bila hasil evaluasi menunjukkan kebutuhan perbaikan. Dengan karakter tersebut, KDD mendukung proses analisis *end-to-end* yang sistematis dan tetap fleksibel terhadap pilihan teknik maupun sumber data yang beragam.

3.2 Tahapan Penelitian

Alur penelitian ini mengikuti kerangka *Knowledge Discovery in Databases* (KDD) yang mencakup tahap *Selection, Pre-processing, Transformation, Data Mining,* dan *Evaluation*. Setiap tahap dijalankan secara terstruktur dan dapat diulang bila diperlukan agar proses analisis *multimodal* berbasis *Generative AI* menghasilkan model yang kuat, akurat, dan mampu melakukan *sentiment analysis* pada ulasan konsumen secara andal.

Gambar 3.2 menggambarkan urutan proses dari pengumpulan data hingga penilaian kinerja model. Diagram tersebut menunjukkan bagaimana data teks dan gambar dihimpun, dibersihkan, dan diubah menjadi *dataset multimodal*, lalu dimanfaatkan dalam pelatihan model sebelum dievaluasi dengan metrik yang relevan. Uraian rinci untuk setiap langkah pada diagram sebagai berikut:

1. *Selection*

   Data ulasan produk *fashion* dihimpun dari Tokopedia dan Shopee pada periode Januari 2024-Agustus 2025 melalui otomasi *web scraping*. Di Tokopedia, ekstraksi dilakukan dengan Selenium WebDriver (meniru interaksi pengguna membuka tab ulasan, mengatur "Terbaru", *pagination*) lalu HTML terender diparsing menggunakan BeautifulSoup untuk memperoleh teks ulasan, *rating*, nama pengguna, tanggal, variasi/nama produk, dan tautan gambar, disimpan ke CSV [107][108]. Di Shopee, karena pembatasan Selenium, ekstraksi menggunakan JavaScript dari konsol *browser* yang menelusuri DOM dan mengekspor ke CSV [109]. Kurasi awal diberlakukan seragam, hanya ulasan be*rating* 1-2 (negatif) dan 4-5 (positif) dengan urutan "Terbaru" agar sinyal sentimen jelas.

2. *Pre-processing*

   Pada tahap *Pre-processing*, data dipersiapkan melalui pembersihan teks, verifikasi gambar, serta penataan ulang struktur *dataset* agar setiap modalitas (teks dan gambar) siap digunakan pada tahap pemodelan. Pembersihan awal mencakup normalisasi teks, *lowercasing*, *tokenization*, penghapusan *stopword*, dan *stemming* menggunakan Sastrawi, sekaligus penghapusan entri tidak valid seperti ulasan kosong, satu huruf, deret tanda baca, atau emoji tunggal. Secara paralel, tautan gambar diverifikasi dan diunduh, kemudian gambar disaring menggunakan deteksi objek YOLOv8 dengan ambang kepercayaan tertentu untuk memastikan bahwa gambar benar-benar menampilkan produk *fashion* yang relevan dengan ulasan. Setelah data teks dan gambar tersusun, dilakukan balancing awal berdasarkan empat komponen utama, yaitu ulasan positif bergambar, ulasan positif tanpa gambar, ulasan negatif bergambar, dan ulasan negatif tanpa gambar, dengan menerapkan teknik *undersampling* pada kelas mayoritas. Tahap ini dilanjutkan dengan kurasi visual lanjutan untuk mengidentifikasi gambar yang secara eksplisit menunjukkan kerusakan produk dan gambar yang merepresentasikan produk dalam kondisi baik, serta penyelarasan kembali dengan *review_id* agar hubungan antara teks dan gambar tetap konsisten.

3. *Transformation*

Tahap *Transformation* berfokus pada pengubahan data mentah yang telah dibersihkan menjadi representasi yang siap digunakan dalam pelatihan model. Untuk modalitas teks, *label* sentimen diturunkan dari *rating* ulasan dengan pemetaan 1-2 sebagai negatif dan 4-5 sebagai positif, kemudian sampel ulasan diperiksa ulang oleh ahli bahasa untuk mengurangi risiko ketidaksesuaian antara isi teks dan *label* sentimen. Untuk modalitas gambar, ulasan yang menyertakan foto menggunakan gambar asli hasil kurasi visual, sedangkan ulasan tanpa foto atau dengan jumlah gambar negatif yang terbatas dilengkapi dengan gambar sintetis yang dihasilkan menggunakan Stable Diffusion XL (SDXL) yang telah diadaptasi pada *domain fashion* melalui teknik LoRA. *Prompt* generatif dibangun berdasarkan terjemahan teks ulasan ke bahasa Inggris dan diarahkan pada beberapa kategori kerusakan utama (misalnya bahan tipis, jahitan tidak rapi, pakaian bolong) agar gambar sintetis selaras dengan narasi keluhan. Hasil tahap *Transformation* adalah *dataset multimodal* yang secara konsep seimbang, di mana pada sisi teks jumlah ulasan positif dan negatif dibuat sebanding, sedangkan pada sisi gambar terdapat himpunan gambar positif yang mencerminkan produk dalam kondisi baik dan gambar negatif (gabungan nyata dan sintetis) yang menggambarkan kerusakan produk; *dataset* ini kemudian dibagi ke dalam bagian latih, validasi, dan uji dengan skema proporsi tertentu (misalnya 80/10/10) [110].

4. *Data Mining*

Pada tahap *Data Mining*, klasifikasi sentimen dilakukan dengan menilai kinerja *un*imodal dan *multimodal* secara terstruktur. Untuk teks, digunakan BERT, RoBERTa, dan IndoBERT sebagai *backbone* bahasa yang merepresentasikan konteks secara dua arah, dengan RoBERTa mengandalkan optimasi skema pra-latih dan IndoBERT membawa keunggulan korpus Indonesia. Untuk gambar, dievaluasi tiga jalur visi, yakni CNN konvensional sebagai *baseline* ringan, ResNet-18 yang memanfaatkan residual *connection* untuk menjaga aliran gradien dan meningkatkan kedalaman efektif, serta DeiT (*Data-efficient Image Transformer*) yang memproyeksikan *patch*

gambar ke ruang *embedding* dan belajar representasi melalui mekanisme perhatian. Setelah memperoleh vektor fitur dari masing-masing modalitas, penelitian ini menggabungkan representasi pada tingkat fitur melalui *concatenation* ke dalam ruang bersama, sehingga informasi tekstual dan visual saling melengkapi. Skema fusi tersebut digunakan untuk membandingkan peningkatan akurasi dan kemampuan generalisasi dibandingkan model *unimodal*, sekaligus memeriksa konsistensi performa lintas kombinasi teks-gambar (BERT/RoBERTa/IndoBERT × CNN/ResNet-18/DeiT) dalam setting *multimodal*.

5. *Evaluation*

Tahap akhir adalah Evaluation, yakni menilai kinerja model klasifikasi sentimen yang telah dibangun. Penilaian berbasis *Confusion matrix* dengan empat komponen *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) yang kemudian digunakan untuk menghitung metrik utama, yaitu Akurasi (Rumus 2.1), Precision (Rumus 2.2), *recall* (Rumus 2.3), serta F1-*Score* (Rumus 2.4) sebagai ringkasan keseimbangan presisi dan cakupan prediksi [66].

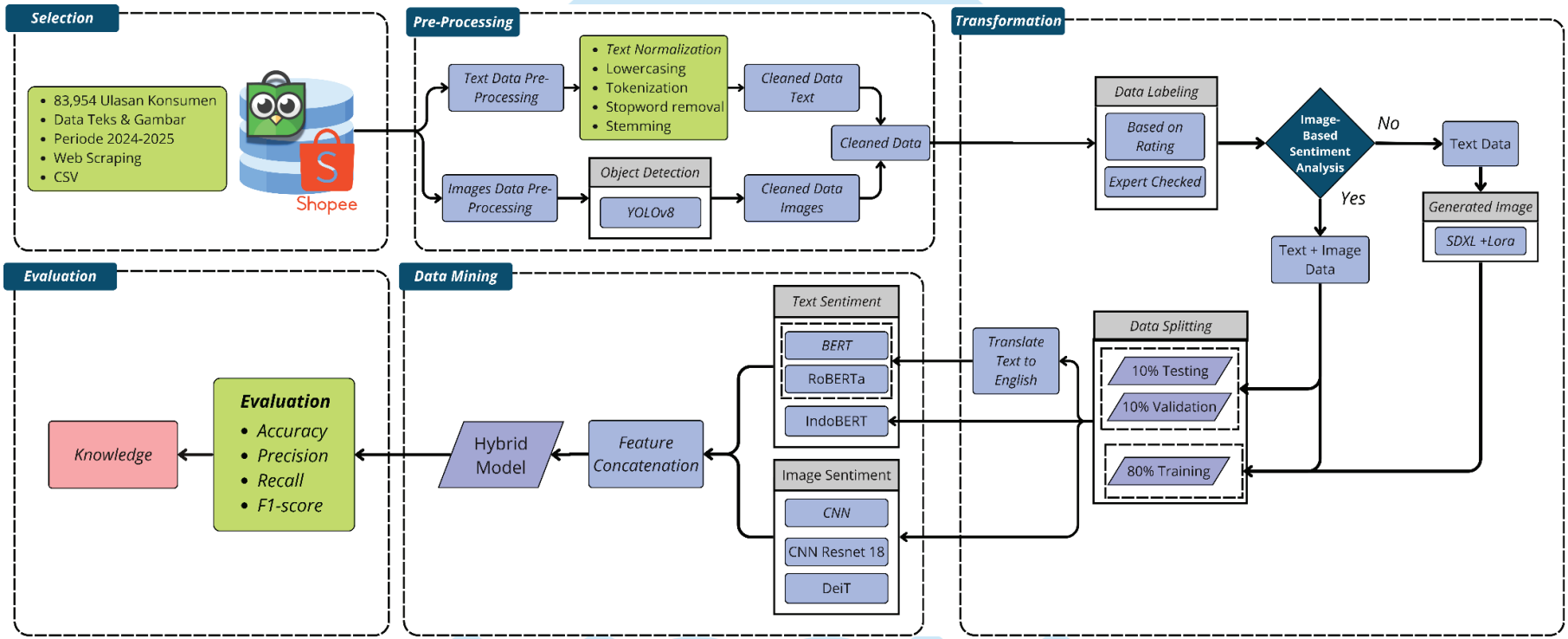
$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.1)$$

$$presisi = \frac{TP}{TP+FP} \quad (2.2)$$

$$Recall = \frac{TP}{TP+FN} \quad (2.3)$$

$$F1-Score = 2 \times \frac{presisi \times Recall}{presisi+Recall} \quad (2.4)$$

Akurasi merefleksikan porsi prediksi yang tepat, sedangkan presisi menunjukkan ketelitian saat model menyatakan kelas positif. *Recall* menggambarkan kemampuan model menangkap seluruh contoh positif yang benar, sementara F1-*Score* merangkum keseimbangan antara presisi dan *recall*. Di samping itu, waktu eksekusi dievaluasi untuk menilai efisiensi komputasi, sehingga himpunan metrik tersebut memberikan gambaran menyeluruh tentang kinerja dan efektivitas pendekatan *multimodal*.



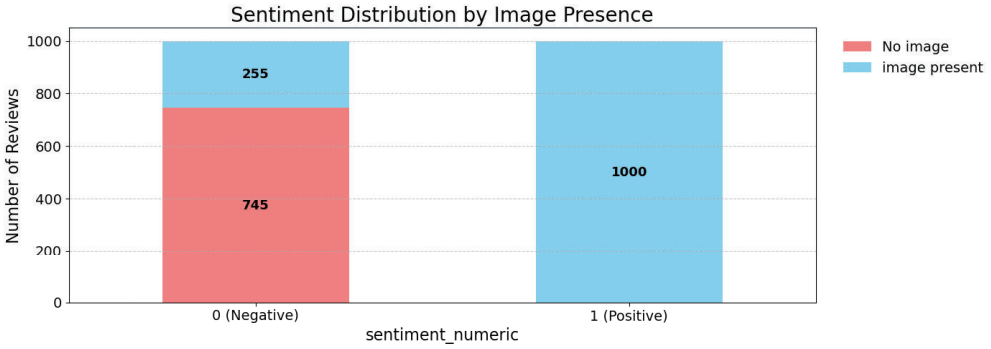
Gambar 3.1 Diagram Alur Penelitian

3.3 Teknik Pengumpulan Data

Pengumpulan data dilakukan secara otomatis melalui *web scraping* pada kategori *fa*shion di Tokopedia dan Shopee. Untuk Tokopedia, Python digunakan dengan Selenium dan BeautifulSoup guna mengekstraksi teks ulasan, *rating*, serta tautan atau berkas gambar dari elemen HTML yang relevan [42]. Pada Shopee, proses pengambilan dilakukan dengan JavaScript yang dijalankan di konsol *browser* untuk membaca DOM, memfilter rentang *rating*, dan mengekspor hasil dalam format terstruktur. Data teks disimpan sebagai CSV, sedangkan berkas visual dipertahankan dalam JPG/PNG sesuai resolusi asli. Ulasan yang tidak menyediakan gambar dilengkapi melalui generasi gambar sintetis menggunakan Stable Diffusion XL (SDXL) yang dipadukan dengan LoRA agar adaptif terhadap *domain fashion*. Seluruh korpus kemudian melalui pembersihan untuk menghapus duplikasi, entri kosong, serta file gambar yang rusak sehingga siap digunakan pada tahap analisis.

3.3.1 Populasi dan *Sample*

Populasi dalam penelitian ini mencakup seluruh ulasan konsumen produk *fashion* di Tokopedia dan Shopee yang tersedia dalam bentuk teks dan/atau gambar, dengan total 83.954 ulasan hasil *web scraping*. Setelah dilakukan pembersihan teks, penghapusan entri tidak valid, dan verifikasi gambar menggunakan YOLOv8, diperoleh 57.771 ulasan yang layak olah, terdiri atas 15.847 ulasan dengan gambar valid dan 41.864 tanpa gambar. Dari populasi terjangkau ini, dipilih sampel secara *purposive* sehingga diperoleh korpus akhir yang seimbang secara sentimen, yaitu 2.000 ulasan yang terdiri atas 1.000 ulasan bernuansa negatif dan 1.000 ulasan bernuansa positif. Sampel tersebut mencakup 1.255 entri dengan pasangan gambar dan 745 entri teks saja, dengan modalitas gambar terdiri atas 1.000 gambar positif nyata dan 1.000 gambar negatif yang dibentuk dari kombinasi 255 gambar negatif nyata dan 745 gambar negatif sintetis hasil generasi SDXL-LoRA yang disesuaikan dengan *domain fashion*.



Gambar 3.2 Distribusi *Dataset* 6

3.3.2 Periode Pengambilan Data

Periode pengambilan data dilakukan sejak Januari 2024 hingga Agustus 2025 dengan tujuan menangkap perilaku konsumen kategori *fashion* di Tokopedia secara terkini dan representatif. Rentang waktu ini dipilih karena mencakup beberapa fase aktivitas pengguna yang berbeda, seperti awal tahun, pertengahan tahun, hingga periode menjelang akhir tahun, ketika intensitas promosi, kampanye diskon, dan pergantian tren *fashion* cenderung meningkat dan lebih beragam. Komposisi periode tersebut memungkinkan data yang dikumpulkan tidak hanya merefleksikan kondisi pasar pada satu momen tertentu, tetapi juga mengakomodasi dinamika perubahan preferensi konsumen dari waktu ke waktu, sehingga analisis sentimen yang dihasilkan menjadi lebih akurat dan relevan terhadap konteks *e-commerce* modern.

3.4 Variabel Penelitian

Penelitian ini menggunakan *dataset* ulasan konsumen produk *fashion* dari *platform e-commerce* Tokopedia dan Shopee yang diperoleh melalui proses *web scraping*. Pada Tokopedia, pengambilan data dilakukan menggunakan Selenium dan BeautifulSoup untuk meniru interaksi pengguna serta mengekstraksi teks ulasan, informasi produk, *rating*, dan tautan gambar. Pada Shopee, data diekstraksi melalui penelusuran DOM berbasis JavaScript langsung dari konsol browser. Dalam penelitian ini digunakan dua kategori variabel, yaitu variabel independen dan variabel dependen.

3.4.1 Variabel Independen

Variabel independen merupakan variabel bebas yang dianggap memengaruhi variabel lain [111]. Dalam konteks penelitian ini, variabel independen berperan sebagai masukan utama pada proses klasifikasi sentimen *multimodal* dan terdiri atas informasi teks ulasan serta gambar produk (baik gambar asli maupun hasil generatif dari *Generative* AI). Secara rinci, variabel yang digunakan adalah:

1. *review_Text*

   Berisi teks lengkap ulasan yang ditulis oleh konsumen dan menjadi *input* utama bagi model analisis sentimen berbasis teks.

1. *Image_*link

   Memuat tautan atau berkas gambar yang menyertai ulasan konsumen, apabila ulasan tidak disertai gambar, gambar sintetis akan dibuat dari isi ulasan menggunakan Stable Diffusion v2-*Base*.

2. *product_name*

   Bagian ini berisi nama produk yang diulas dan dimanfaatkan sebagai informasi tambahan atau metadata dalam analisis.

3. *Variation*

   Berisi keterangan varian produk yang dibeli, seperti ukuran atau warna, yang berpotensi memengaruhi pengalaman dan penilaian konsumen.

4. *Rating*

   Memuat nilai penilaian konsumen terhadap produk dan dimanfaatkan sebagai dasar untuk membentuk label sentimen positif maupun negatif.

3.4.2 Variabel Dependen

Variabel dependen merupakan variabel terikat yang nilainya dipengaruhi oleh variabel independent [111]. Pada penelitian ini, variabel dependen berupa sentimen ulasan konsumen yang dibagi ke dalam dua kategori, yakni sentimen positif untuk ulasan dengan *rating* 4-5 dan sentimen negatif untuk ulasan dengan *rating* 1-2. Kategori sentimen tersebut dijadikan *ground truth* dalam proses pelatihan serta pengujian model *multimodal*. Kinerja model kemudian dievaluasi menggunakan metrik akurasi, presisi, *recall*, dan F1-*score* yang diperoleh dari kombinasi nilai

True Positive (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) pada hasil klasifikasi.

3.5 Teknik Analisis Data

Teknik analisis data pada penelitian ini disusun untuk mengukur kinerja pendekatan *unimodal*, baik berbasis teks maupun gambar, serta pendekatan *multimodal* yang menggabungkan keduanya dalam klasifikasi sentimen ulasan produk *fashion*. Analisis dilakukan dengan melatih dan membandingkan beberapa model *deep learning*, yaitu BERT, RoBERTa, dan IndoBERT untuk teks, serta CNN, ResNet-18, dan DeiT untuk gambar, termasuk gambar negatif hasil generasi SDXL-LoRA yang telah *domain adapted*.

3.5.1 Analisis Sentimen Berbasis Teks (*Text-Based Sentiment Analysis*)

Analisis sentimen berbasis teks dilakukan menggunakan tiga model *transformer*, yaitu BERT, IndoBERT, dan RoBERTa yang dipilih karena kemampuannya menangkap konteks semantik secara mendalam melalui *bidirectional attention mechanism*. Ketiga model tersebut menggunakan dimensi *embedding* dan ukuran *hidden layer* sebesar 768, sehingga representasi teks dapat diproyeksikan secara konsisten ke ruang vektor yang sama. Seluruh teks hasil *pre-processing* ditokenisasi dengan panjang maksimum 128 token dan diproses dengan *full fine-tuning* pada seluruh lapisan model. Proses *training* dijalankan dengan *batch size* 16 selama maksimal 10 *epoch* dengan *early stopping* untuk mencegah *overfitting*, menggunakan AdamW sebagai *optimizer* dengan *learning rate* $1{,}8 \times 10^{-5}$, *weight decay* 0,05, *warmup* 100 langkah, serta *dropout* 0,1 pada *hidden layer* dan *attention mechanism*.

3.5.2 Analisis Sentimen Berbasis Gambar (*Image-Based Sentiment Analysis*)

Analisis sentimen berbasis gambar dilakukan menggunakan tiga arsitektur visi, yaitu CNN konvensional sebagai *baseline*, ResNet-18 dengan *residual connection*, serta DeiT sebagai *vision transformer* yang mempelajari representasi visual berbasis *patch* dan *attention mechanism*. Seluruh model dilatih menggunakan gambar hasil kurasi, terdiri atas gambar ulasan asli dan gambar negatif sintetis yang dihasilkan oleh SDXL yang telah *domain adapted* dengan

LoRA. Ukuran *input* distandarkan pada resolusi 224×224 piksel, dengan *batch size* 32 dan 20 *training epochs*. *Optimization* menggunakan AdamW dengan *learning rate* 1×10^{-4} pada lapisan utama dan *weight decay* 2×10^{-3}, sedangkan *CrossEntropyLoss* diterapkan dengan *class weighting* serta *label smoothing* untuk menjaga kestabilan pelatihan pada distribusi yang tidak seimbang. Peningkatan generalisasi dicapai melalui rangkaian *augmentation* pada *train data*, termasuk *random resized crop*, *horizontal* dan *vertical flip*, rotasi hingga ±15°, *color jitter*, dan *random grayscale*, sementara *validation* dan *test data* hanya melalui *resize*, *center crop*, dan normalisasi.

3.5.3 Analisis Sentimen *Multimodal* (*Multimodal Sentiment Analysis*)

Analisis sentimen *multimodal* dilakukan dengan menggabungkan informasi dari dua sumber data, yaitu teks ulasan dan gambar produk, sehingga model dapat memanfaatkan konteks semantik sekaligus bukti visual secara bersamaan. Representasi teks diperoleh dari *transformer encoder* (BERT, IndoBERT, atau RoBERTa) melalui vektor [CLS] berdimensi 768, dengan skema *full fine-tuning* pada seluruh *layer* kecuali *embedding* dan empat *encoder layers* awal yang dibekukan untuk menjaga stabilitas *training*. Representasi visual diekstraksi dari *image backbone* (CNN, ResNet-18, atau DeiT) yang menerima *input image* berukuran 224×224 piksel dan menghasilkan vektor fitur berdimensi 512 untuk CNN/ResNet-18 atau 768 untuk DeiT, setelah melalui serangkaian *image transforms* berupa *resize*, *crop*, *flip*, rotasi, *color jitter*, dan normalisasi.

Kedua vektor fitur kemudian digabung pada level representasi (*feature-level fusion*), lalu diproyeksikan ke ruang berdimensi 256 melalui *fully connected layer* dengan aktivasi ReLU dan *dropout* 0,6, sebelum akhirnya diteruskan ke *output layer* dengan dua kelas. Proses *multimodal training* dijalankan dengan *batch size* 32 selama maksimal 20 *epochs* menggunakan AdamW sebagai *optimizer* dengan skema *differential learning rate* 2×10^{-6} untuk parameter *transformer* yang masih dilatih dan 2×10^{-5} untuk *image backbone* serta *classification head*. Regularisasi tambahan diterapkan melalui *weight decay* sebesar 2×10^{-3}, *label smoothing* 0,1 pada fungsi *loss CrossEntropyLoss*, serta *class weighting* untuk mengimbangi distribusi kelas, sementara proses *training* dipercepat dengan pemanfaatan *mixed*

presisi (AMP) dan pengaturan *adaptive learning rate* menggunakan ReduceLROnPlateau.

3.5.4 Analisis Perbandingan Model

Tahap ini membandingkan performa tiga kelompok model, yaitu model berbasis teks, berbasis gambar, dan model *multimodal*. Pada analisis teks, dibandingkan kinerja tiga *text encoder* (BERT, IndoBERT, dan RoBERTa). Pada analisis gambar, dievaluasi tiga jalur visi, yaitu CNN *baseline*, ResNet-18, dan DeiT sebagai *vision transformer*. Untuk pendekatan *multimodal*, dikaji berbagai kombinasi *text encoder* dan *image encoder* (BERT/IndoBERT/RoBERTa yang dipasangkan dengan CNN, ResNet-18, maupun DeiT) untuk melihat sejauh mana fusi teks-gambar memberikan peningkatan dibandingkan model *unimodal*. Perbandingan performa dilakukan menggunakan *Confusion matrix* yang diturunkan menjadi metrik akurasi, presisi, *recall*, dan F1-*score*, serta mempertimbangkan waktu eksekusi *training* dan *inference* guna menilai efektivitas sekaligus efisiensi masing-masing pendekatan.

3.5.5 Analisis Topik Ulasan

Analisis topik ulasan dilakukan dengan menggunakan file CSV hasil pra-pemrosesan dan translasi yang telah disiapkan sebelumnya, sehingga tahap ini tidak lagi bekerja pada teks mentah, melainkan pada *dataset* yang sudah dibersihkan dan diterjemahkan ke Bahasa Inggris. Seluruh 2.000 ulasan pada CSV dipisahkan menjadi dua subset berdasarkan label *sentiment*_numeric, yaitu df_0 untuk ulasan negatif dan df_1 untuk ulasan positif. Analisis teks berfokus pada kolom translated_*text* sebagai representasi teks yang seragam untuk pemodelan. Sebagai langkah eksplorasi awal, dibangun *Wordcloud* terpisah untuk masing-masing subset dengan cara menggabungkan seluruh teks translated_*text* dalam kelompok negatif dan positif, sehingga diperoleh gambaran visual mengenai kata-kata yang paling sering muncul pada tiap sentimen dan indikasi awal mengenai isu yang dominan dibahas konsumen.

Tahap berikutnya menggunakan kombinasi *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Latent Dirichlet Allocation* (LDA).

Representasi TF-IDF dibangun secara terpisah untuk ulasan negatif dan positif menggunakan *TfidfVectorizer* (*max_df* = 0,95, *min_df* = 2, *max_features* = 1000), menghasilkan dua matriks fitur, *tfidf_matrix*_0 untuk ulasan negatif dan *tfidf_matrix*_1 untuk ulasan positif. Kedua matriks ini kemudian dimodelkan dengan LDA (*LatentDirichletAllocation* pada scikit-learn) dengan jumlah topik awal *n_components* = 5 dan *random_state* = 42 untuk masing-masing kelompok sentimen. Dari model LDA diekstraksi kata-kata dengan bobot tertinggi pada setiap topik untuk keperluan interpretasi dan penamaan topik, sementara distribusi probabilitas topik per ulasan dihitung melalui lda.transform. Topik dominan tiap ulasan diperoleh dengan memilih topik berprobabilitas terbesar dan disimpan sebagai kolom *topic*, kemudian dipetakan ke label deskriptif *topic*_label berdasarkan analisis kata kunci. Selanjutnya dihitung distribusi jumlah ulasan per topik beserta persentasenya terhadap keseluruhan 2.000 ulasan dan divisualisasikan dengan *Wordcloud* per topik, sehingga dapat diringkas aspek-aspek apa yang paling sering dibahas dan seberapa besar kontribusinya dalam korpus ulasan secara keseluruhan.