BAB II
TINJAUAN PUSTAKA

2.1 Landasan Teori

1. Analisis Sentimen

Analisis sentimen, atau yang juga dikenal sebagai opinion mining, adalah proses komputasional untuk mengidentifikasi dan mengekstraksi informasi subjektif dari teks. Menurut penelitian terbaru dalam bidang NLP, analisis sentimen telah berkembang dari pendekatan machine learning tradisional hingga model deep learning berbasis transformer yang lebih canggih (Talaat, 2023). Analisis sentimen umumnya mengklasifikasikan teks ke dalam kategori positif, negatif, atau netral berdasarkan opini atau emosi yang terkandung di dalamnya.

Dalam konteks e-commerce, analisis sentimen memainkan peran penting dalam memahami kepuasan pelanggan dan mengidentifikasi area yang memerlukan perbaikan. Penelitian menunjukkan bahwa analisis sentimen pada review produk dapat membantu meningkatkan sistem rekomendasi dan personalisasi layanan (Liu, 2022).

2. IndoBERT (Indonesian BERT)

IndoBERT adalah model pre-trained berbasis arsitektur BERT (Bidirectional Encoder Representations from Transformers) yang dilatih khusus untuk bahasa Indonesia. Model ini dikembangkan oleh (Wilie, et al., 2020) sebagai bagian dari benchmark IndoNLU (Indonesian Natural Language Understanding) yang mencakup 12 tugas pemrosesan bahasa

Indonesia. IndoBERT memanfaatkan mekanisme attention bidirectional yang memungkinkan model untuk memahami konteks kata dari kedua arah (kiri dan kanan), menghasilkan representasi yang lebih kaya dibandingkan model unidirectional.

Penelitian terbaru menunjukkan bahwa IndoBERT telah mencapai performa state-of-the-art pada berbagai tugas NLP bahasa Indonesia, termasuk analisis sentimen dengan akurasi mencapai 92% (Ahmadian, Abidin, Riza, & Muchtar, 2024). Model ini telah digunakan secara luas dalam berbagai aplikasi seperti deteksi cyberbullying, klasifikasi berita palsu, dan analisis sentimen media sosial (Setiawan, Iswavigra, & Anggiratih, 2025)

Keunggulan utama IndoBERT dibandingkan BERT multilingual adalah kemampuannya dalam menangani karakteristik unik bahasa Indonesia, termasuk kompleksitas morfologi, sintaksis yang fleksibel, dan ambiguitas semantik. Model ini dilatih pada korpus besar teks bahasa Indonesia yang mencakup berbagai domain, sehingga dapat menangkap nuansa dan konteks yang spesifik untuk bahasa Indonesia (Wang, 2025) (Perwira, Permadani, Purnamasari, & Agusdian, 2025).

3. XGBoost (Extreme Gradient Boosting)

XGBoost adalah algoritma ensemble learning berbasis gradient boosting yang dikembangkan oleh (Chen & Guestrin, 2016). Algoritma ini merupakan implementasi yang dioptimalkan dari gradient boosted decision trees yang dirancang untuk kecepatan dan performa tinggi. XGBoost

bekerja dengan membangun ensemble dari weak learners (pohon keputusan dangkal) secara sekuensial, di mana setiap pohon baru memperbaiki kesalahan dari pohon sebelumnya.

Dalam konteks analisis sentimen, XGBoost telah terbukti efektif dalam mengklasifikasikan teks dengan akurasi tinggi. Penelitian oleh Nsaif & Abd (2022) menunjukkan bahwa XGBoost mencapai akurasi 95.16% pada klasifikasi sentimen postingan politik. Keunggulan XGBoost meliputi kemampuan menangani data tidak seimbang, regularisasi untuk mencegah overfitting, dan kemampuan paralel processing yang meningkatkan efisiensi komputasi (Wang, 2025).

XGBoost juga memiliki beberapa fitur advanced seperti handling missing values secara otomatis, built-in cross-validation, dan berbagai parameter tuning yang memungkinkan optimalisasi model. Penelitian terbaru menunjukkan bahwa XGBoost dapat meningkatkan performa klasifikasi ketika dikombinasikan dengan feature engineering yang tepat dan hyperparameter tuning (Chandrasekaran, Dhanasekaran, Moorthy, & Oli, 2024)

4. Model Hybrid dalam Analisis Sentimen

Model hybrid menggabungkan kekuatan dari berbagai algoritma atau teknik untuk mencapai performa yang lebih baik dibandingkan pendekatan tunggal. Dalam analisis sentimen, pendekatan hybrid umumnya menggabungkan kemampuan deep learning dalam ekstraksi fitur dengan algoritma machine learning untuk klasifikasi. Penelitian terbaru

menunjukkan bahwa model hybrid BERT-XGBoost dapat mencapai akurasi superior dengan menggabungkan pemahaman kontekstual BERT dengan kemampuan optimalisasi XGBoost (Mushtaq Gardazi, et al., 2025)

Ahmadian, Abidin, Riza, & Muchtar (2024) dalam penelitiannya mendemonstrasikan bahwa model hybrid IndoBERT dengan BiLSTM dan attention mechanism mencapai F1-score yang lebih tinggi dibandingkan IndoBERT tunggal pada tugas klasifikasi emosi dan sentimen. Pendekatan hybrid memungkinkan model untuk memanfaatkan representasi semantik yang kaya dari transformer model sambil memanfaatkan kemampuan klasifikasi yang robust dari algoritma ensemble learning.

5. Preprocessing Teks Bahasa Indonesia

Preprocessing adalah tahap krusial dalam analisis sentimen yang meliputi serangkaian teknik untuk membersihkan dan mentransformasi teks mentah menjadi format yang sesuai untuk pemrosesan lebih lanjut. Untuk teks bahasa Indonesia, preprocessing umumnya mencakup: (1) Case folding - mengubah semua teks menjadi huruf kecil, (2) Cleaning - menghapus karakter khusus, URL, dan angka, (3) Tokenization - memecah teks menjadi token individual, (4) Stopword removal - menghapus kata-kata umum yang tidak bermakna, (5) Normalization - mengubah kata tidak baku menjadi baku, dan (6) Stemming - mengubah kata menjadi bentuk dasarnya (Idris, Rifai, & Tania, 2025). Perlu dicatat bahwa dari keenam tahapan umum tersebut, stemming tidak diterapkan dalam penelitian ini karena IndoBERT menggunakan tokenisasi berbasis WordPiece yang telah memecah kata

berimbuhan menjadi sub-kata, sehingga variasi morfologi bahasa Indonesia sudah tertangani secara implisit tanpa memerlukan stemming eksplisit.

Penelitian menunjukkan bahwa kualitas preprocessing sangat mempengaruhi performa model analisis sentimen. Ulhaq & Suprayogi (2025)menemukan bahwa penggunaan normalisasi kata dan stemming dapat meningkatkan akurasi klasifikasi hingga 5% pada dataset review bahasa Indonesia.

2.2 Kerangka Teori

Penelitian ini mengintegrasikan tiga pilar teori utama guna menciptakan fondasi yang kokoh bagi implementasi model hybrid IndoBERT-XGBoost dalam mengklasifikasikan ulasan pengguna Tokopedia.

1. NLP Berbasis Transformer dan Transfer Learning

   Landasan pertama berfokus pada pemanfaatan IndoBERT sebagai ekstraktor fitur. Berdasarkan teori *Natural Language Processing* (NLP) berbasis *Transformer*, IndoBERT mampu memahami konteks semantik bahasa Indonesia secara dua arah (*bidirectional*) melalui mekanisme *self-attention*. Didukung oleh teori Transfer Learning, model ini memungkinkan pengetahuan linguistik dari korpus besar bahasa Indonesia untuk diadaptasi (melalui *fine-tuning*) guna menyelesaikan tugas klasifikasi sentimen secara lebih spesifik dan efisien.

2. Ensemble Learning dan Regularisasi (XGBoost)

   Pilar kedua melibatkan penggunaan XGBoost sebagai algoritma klasifikasi. Berdasarkan teori *Ensemble Learning*, khususnya *Gradient*

Boosting, XGBoost bekerja dengan membangun pohon keputusan secara sekuensial untuk mengoreksi galat dari tahapan sebelumnya. Keunggulan XGBoost terletak pada teori Regularisasi yang mencakup teknik *shrinkage* dan *subsampling*, yang secara efektif meminimalkan risiko *overfitting* saat memproses representasi fitur semantik yang dihasilkan oleh IndoBERT.

3. Sinergi Model Hybrid

   Pendekatan ini didasarkan pada teori Model Hybrid yang menggabungkan kekuatan *Deep Learning* dengan *Machine Learning* konvensional. Melalui prinsip *Complementary Strengths*, IndoBERT berperan dalam menangkap nuansa bahasa yang kompleks, sementara XGBoost mengoptimalkan pola klasifikasi. Integrasi ini membentuk sebuah *pipeline* sistematis: IndoBERT mengonversi teks menjadi vektor semantik, yang kemudian dieksekusi oleh XGBoost untuk memprediksi kategori sentimen secara akurat.

Pendukung Kontekstual dan Metodologi

a. Analisis Sentimen E-commerce: Mengacu pada teori *Expectation-Confirmation*, sentimen pengguna dipahami sebagai refleksi dari perbandingan antara ekspektasi dan realitas layanan aplikasi. Model hybrid di sini bertindak sebagai instrumen otomatisasi untuk memetakan pola feedback dari data ulasan berskala besar.
b. Preprocessing Teks: Mengingat sifat ulasan aplikasi yang seringkali informal, teori pra-pemrosesan bahasa Indonesia diterapkan melalui

normalisasi dan pembersihan *noise* untuk memastikan kualitas data sebelum masuk ke tahap pemodelan.

c. Evaluasi Model: Keandalan model diukur menggunakan metrik performa standar (*Accuracy, Precision, Recall,* dan *F1-score*) serta divalidasi melalui *Cross-Validation* untuk menjamin stabilitas dan ketahanan (*robustness*) hasil prediksi.

Kesimpulan: Integrasi teoretis ini memperkuat hipotesis bahwa penggabungan pemahaman bahasa mendalam dari IndoBERT dengan efisiensi klasifikasi XGBoost akan menghasilkan performa yang lebih unggul dibandingkan penggunaan model secara mandiri dalam konteks e-commerce di Indonesia.

2.3 Tinjauan Umum: E-commerce dan Perilaku Pengguna Digital

E-commerce (electronic commerce) merupakan kegiatan jual beli barang dan jasa yang dilakukan secara elektronik melalui jaringan internet. Dalam dua dekade terakhir, pertumbuhan e-commerce di seluruh dunia, khususnya di Asia Tenggara, telah mengalami akselerasi yang luar biasa didorong oleh penetrasi internet yang semakin tinggi, meningkatnya kepemilikan smartphone, dan perubahan perilaku belanja masyarakat. Indonesia sebagai negara dengan populasi terbesar keempat di dunia menjadi salah satu pasar e-commerce yang paling dinamis, dengan total nilai transaksi digital yang terus meningkat setiap tahunnya.

Salah satu karakteristik unik ekosistem e-commerce adalah tingginya intensitas interaksi pengguna dalam bentuk ulasan (review) dan penilaian (rating). Setiap transaksi yang diselesaikan memberikan kesempatan kepada pembeli untuk meninggalkan umpan balik yang kemudian menjadi panduan bagi calon pembeli

berikutnya. Fenomena ini menciptakan apa yang disebut sebagai User-Generated Content (UGC) yang bersifat masif, tidak terstruktur, dan sangat berharga. Tokopedia, sebagai salah satu platform marketplace terkemuka di Indonesia dengan lebih dari 100 juta pengguna aktif, menghasilkan jutaan data ulasan setiap bulannya yang mencerminkan pengalaman nyata konsumen terhadap produk, layanan pengiriman, maupun antarmuka aplikasi.

Namun demikian, volume data ulasan yang sangat besar tersebut tidak dapat dianalisis secara manual dengan efisien. Dibutuhkan pendekatan komputasional yang mampu mengolah ribuan hingga jutaan teks ulasan secara otomatis untuk mengekstraksi informasi yang bermakna, khususnya terkait sentimen atau opini pengguna. Kebutuhan inilah yang mendorong perkembangan bidang analisis sentimen berbasis kecerdasan buatan.

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah cabang ilmu kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia (bahasa alami). NLP mencakup serangkaian teknik dan algoritma yang memungkinkan komputer untuk memahami, menginterpretasikan, dan menghasilkan teks atau ucapan manusia secara bermakna. Secara umum, pipeline NLP meliputi beberapa tahapan: tokenisasi (pemecahan teks menjadi unit-unit dasar), analisis morfologi, penandaan kelas kata (POS tagging), parsing sintaktik, pengenalan entitas bernama (Named Entity Recognition), hingga pemahaman semantik dan pragmatik.

Perkembangan NLP telah melalui beberapa era yang signifikan. Era pertama ditandai dengan pendekatan berbasis aturan (rule-based), di mana linguist

secara manual mendefinisikan aturan-aturan gramatikal dan leksikal. Era kedua hadir dengan metode statistik dan machine learning klasik seperti Naive Bayes, SVM, dan model n-gram yang memanfaatkan frekuensi kemunculan kata. Era ketiga, yang saat ini tengah berjalan, didominasi oleh model deep learning berbasis jaringan saraf tiruan (neural network) yang mampu mempelajari representasi fitur secara otomatis dari data besar tanpa rekayasa fitur manual. Puncak dari evolusi ini ditandai dengan kemunculan arsitektur Transformer yang merevolusi hampir seluruh tugas NLP.

Tantangan khusus NLP untuk bahasa Indonesia meliputi: morfologi yang kaya (afiksasi yang kompleks), tidak adanya pembatasan kata yang jelas seperti dalam bahasa Jepang atau Cina, penggunaan bahasa informal dan slang yang tinggi terutama di media sosial dan ulasan aplikasi, serta keterbatasan sumber daya linguistik (corpus, leksikon) dibandingkan bahasa Inggris. Hal ini mendorong pengembangan model NLP yang dilatih secara khusus pada korpus berbahasa Indonesia.

2.5 Analisis Sentimen dalam Konteks E-Commerce

Analisis sentimen (sentiment analysis) atau opinion mining adalah bidang dalam NLP yang bertujuan mengidentifikasi dan mengekstraksi opini subjektif dari teks. Tujuan utamanya adalah menentukan sikap penulis terhadap suatu topik, yang umumnya diklasifikasikan ke dalam kategori positif, negatif, atau netral. Analisis sentimen dapat dilakukan pada beberapa tingkat granularitas: tingkat dokumen (satu label sentimen untuk keseluruhan teks), tingkat kalimat (sentimen per

kalimat), hingga tingkat aspek (Aspect-Based Sentiment Analysis/ABSA) yang mengidentifikasi sentimen terhadap fitur-fitur spesifik.

Dalam konteks e-commerce, analisis sentimen memainkan peran strategis yang penting. Bagi penyedia platform, hasil analisis sentimen dapat digunakan untuk: (1) memantau kepuasan pengguna secara real-time dalam skala besar; (2) mengidentifikasi masalah produk atau layanan sebelum berkembang menjadi krisis; (3) memperbaiki sistem rekomendasi berbasis opini pengguna; dan (4) membandingkan persepsi pengguna terhadap kompetitor. Pendekatan analisis sentimen pada data ulasan e-commerce berkembang dari metode leksikon berbasis kamus kata sentimen, metode machine learning dengan fitur TF-IDF dan Bag-of-Words, hingga model deep learning modern yang mampu menangkap konteks semantik yang kompleks.

2.6 Arsitektur Transformer dan Mekanisme Self-Attention

Arsitektur Transformer diperkenalkan oleh (Vaswani, et al., 2017)dalam makalah "Attention Is All You Need" sebagai alternatif dari arsitektur sekuensial seperti RNN dan LSTM yang memiliki keterbatasan dalam menangani dependensi jarak jauh dan paralelisasi komputasi. Transformer sepenuhnya dibangun di atas mekanisme attention, khususnya self-attention (juga disebut intra-attention), yang memungkinkan model untuk secara langsung mengukur hubungan dan relevansi antara setiap pasang token dalam sebuah kalimat tanpa harus memproses secara berurutan.

Mekanisme self-attention bekerja dengan mentransformasi setiap token input menjadi tiga vektor: Query (Q), Key (K), dan Value (V). Skor attention

dihitung dengan mengalikan matriks Query dengan transpose matriks Key, kemudian dinormalisasi dengan faktor skala akar dari dimensi vektor (sqrt(d_k)) dan dilewatkan melalui fungsi Softmax. Representasi akhir tiap token diperoleh sebagai kombinasi tertimbang dari seluruh vektor Value dengan bobot dari skor attention tersebut. Proses ini memungkinkan model untuk secara adaptif memfokuskan perhatian pada bagian teks yang paling relevan untuk memahami setiap token.

Multi-head attention memperluas konsep ini dengan menjalankan beberapa mekanisme attention secara paralel (setiap "head" berfokus pada aspek representasi yang berbeda), kemudian menggabungkan hasilnya. Dikombinasikan dengan mekanisme positional encoding untuk menyandikan informasi posisi token, residual connection, dan layer normalization, arsitektur Transformer menghasilkan representasi kontekstual yang sangat kaya untuk setiap token dalam kalimat.

2.7 BERT dan Transfer Learning untuk NLP

BERT (Bidirectional Encoder Representations from Transformers) dikembangkan oleh (Devlin, Chang, Lee, & Toutanova, 2019) dari Google sebagai model bahasa pre-trained berbasis Transformer encoder. Keunggulan utama BERT dibandingkan model sebelumnya (seperti GPT dan ELMo) terletak pada sifat bidirectionalnya yang sejati: BERT memproses konteks dari kedua arah secara bersamaan (kiri dan kanan), bukan secara unidirectional atau shallow bidirectional. Hal ini memungkinkan BERT menghasilkan representasi kontekstual yang jauh lebih kaya dan akurat.

BERT dilatih dengan dua tugas pre-training secara bersamaan: (1) Masked Language Modeling (MLM), di mana sebagian token input ditutupi secara acak dan model harus memprediksi token yang ditutupi berdasarkan konteks sekitarnya; dan (2) Next Sentence Prediction (NSP), di mana model belajar memahami hubungan antar kalimat dengan memprediksi apakah dua kalimat berurutan atau tidak. Pre-training dilakukan pada korpus teks yang sangat besar (Wikipedia dan BookCorpus) sehingga BERT mempelajari representasi bahasa yang kaya dan general.

Transfer learning dalam konteks BERT berarti mengambil model yang sudah dilatih pada tugas umum (pre-training) dan mengadaptasinya untuk tugas spesifik (fine-tuning) dengan data yang lebih sedikit. Proses fine-tuning cukup menambahkan lapisan klasifikasi (classification head) di atas model BERT dan melatih seluruh jaringan pada dataset tugas spesifik dengan learning rate kecil. Pendekatan ini sangat efisien karena sebagian besar pengetahuan linguistik umum sudah tertanam dalam bobot BERT selama pre-training, sehingga fine-tuning hanya perlu menyesuaikan representasi untuk domain tertentu.

2.8 IndoBERT: BERT untuk Bahasa Indonesia

IndoBERT adalah model pre-trained berbasis BERT yang dikembangkan secara khusus untuk bahasa Indonesia oleh (Wilie, et al., 2020) sebagai bagian dari benchmark IndoNLU (Indonesian Natural Language Understanding). Berbeda dengan model BERT multilingual (mBERT) yang dilatih pada 104 bahasa sekaligus sehingga kapasitas untuk bahasa Indonesia menjadi terbatas, IndoBERT dilatih secara eksklusif pada korpus teks berbahasa Indonesia yang besar dan beragam,

mencakup data dari Wikipedia bahasa Indonesia, berita daring, dan berbagai sumber web lainnya.

Keunggulan IndoBERT dibandingkan mBERT untuk tugas-tugas NLP bahasa Indonesia antara lain: (1) pemahaman yang lebih dalam terhadap morfologi bahasa Indonesia yang kompleks seperti sistem afiksasi (prefiks, sufiks, infiks, dan konfiks); (2) representasi yang lebih akurat untuk kosakata khusus bahasa Indonesia termasuk kata serapan, istilah daerah, dan bahasa informal; (3) performa lebih tinggi pada benchmark IndoNLU yang mencakup 12 tugas NLP bahasa Indonesia. IndoBERT-base-p2 yang digunakan dalam penelitian ini merupakan varian phase 2 yang dilatih pada data yang lebih besar dan berkualitas lebih tinggi dibandingkan versi sebelumnya.

2.9 XGBoost dan Pendekatan Ensemble Learning

Ensemble learning adalah paradigma machine learning yang menggabungkan prediksi dari beberapa model (disebut weak learners atau base learners) untuk menghasilkan prediksi akhir yang lebih akurat dan robust dibandingkan model tunggal manapun. Terdapat tiga strategi utama dalam ensemble learning: (1) Bagging (Bootstrap Aggregating), yang melatih model secara paralel pada subset data berbeda (contoh: Random Forest); (2) Boosting, yang melatih model secara sekuensial di mana setiap model baru berfokus pada kesalahan model sebelumnya; dan (3) Stacking, yang melatih meta-model untuk menggabungkan prediksi dari beberapa model base.

XGBoost (eXtreme Gradient Boosting) dikembangkan oleh Chen dan Guestrin (2016) sebagai implementasi yang sangat dioptimalkan dari Gradient

Boosted Decision Trees. XGBoost bekerja dengan membangun pohon keputusan (decision tree) secara sekuensial, di mana setiap pohon baru dilatih untuk memperbaiki residual (sisa kesalahan) dari ensemble pohon-pohon sebelumnya menggunakan pendekatan gradient descent pada ruang fungsi. Keunggulan kritis XGBoost dibandingkan implementasi gradient boosting standar meliputi: efisiensi komputasi tinggi melalui pemrosesan paralel dan teknik histogram-based splitting; regularisasi bawaan (L1 dan L2) untuk mencegah overfitting; penanganan missing values secara otomatis; serta built-in cross-validation. Kombinasi keunggulan ini menjadikan XGBoost sebagai salah satu algoritma machine learning paling populer dan efektif dalam berbagai kompetisi data science.

2.10 Model Hybrid Deep Learning dan Machine Learning

Model hybrid dalam konteks klasifikasi teks adalah pendekatan yang menggabungkan kekuatan dua paradigma berbeda: Deep Learning (DL) untuk ekstraksi fitur semantik yang kaya, dan Machine Learning (ML) tradisional untuk proses klasifikasi yang efisien dan terkontrol. Motivasi utama pendekatan ini adalah untuk memanfaatkan keunggulan masing-masing paradigma sekaligus menutupi kelemahannya. Model deep learning seperti BERT sangat unggul dalam memahami konteks semantik yang kompleks, ambiguitas, dan nuansa bahasa, namun cenderung bersifat black-box dan membutuhkan sumber daya komputasi yang besar. Sebaliknya, algoritma ML seperti XGBoost lebih interpretatif, efisien secara komputasi, dan sangat baik dalam mengoptimalkan pola klasifikasi dari representasi fitur yang sudah diekstraksi.

Pola umum arsitektur hybrid DL-ML adalah sebagai berikut: model deep learning (dalam hal ini IndoBERT) digunakan sebagai feature extractor, menghasilkan representasi vektor berdimensi tinggi yang kaya akan informasi semantik dan kontekstual dari teks input. Representasi vektor ini kemudian diserahkan kepada model ML (XGBoost) sebagai input fitur untuk melakukan klasifikasi akhir. Dengan cara ini, IndoBERT tidak perlu melatih lapisan klasifikasi sendiri, dan XGBoost mendapatkan fitur input yang jauh lebih informatif dibandingkan fitur tradisional seperti TF-IDF. Berbagai penelitian telah membuktikan bahwa kombinasi BERT dengan XGBoost secara konsisten menghasilkan performa yang lebih baik dibandingkan penggunaan salah satu model secara mandiri pada berbagai tugas klasifikasi teks.