BAB I
PENDAHULUAN

1.1 Latar Belakang

Penjualan *smartphone* di Indonesia mengalami peningkatan yang signifikan, di mana pada kuartal III tahun 2025 meningkat sebesar 12% menurut laporan terbaru dari *International Data Corporation* [1]. Tokopedia menjadi salah satu *e-commerce* terbesar di Indonesia yang mampu menghimpun jutaan penjual dan pembeli dalam satu ekosistem digital [2]. Ulasan pembeli dalam sistem *e-commerce* secara umum dianggap sebagai sumber informasi yang penting karena mencerminkan pengalaman, perasaan, dan keinginan pembeli dalam melakukan pembelian suatu produk [3]. Analisis sentimen dapat diimplementasikan untuk menganalisis ulasan pembeli dan mengklasifikasikan polaritas sentimennya. Namun, volume ulasan yang terus bertambah serta pembeli yang cenderung menggunakan bahasa Indonesia ataupun bahasa asing informal penuh dengan singkatan, *slang*, dan ketidakbakuan membuat pengklasifikasian ulasan pembeli menjadi tidak efisien sehingga penjual kehilangan data dan informasi penting yang seharusnya dapat dimanfaatkan [4, 5].

Untuk mengatasi masalah di atas, terdapat beberapa penelitian mengenai analisis sentimen sejenis yang sudah pernah dilakukan. Penelitian yang dilakukan oleh [6] menggunakan algoritma SVM (*Support Vector Machine*) dengan TF-IDF (*Term Frequency-Inverse Document Frequency*) sebagai *feature extraction* untuk melakukan analisis sentimen pada *official store smartphone* yang ada di *platform* Shopee. Penelitian ini mendapatkan hasil *accuracy* sebesar 77% dengan nilai *F1-score* terbaik sebesar 86%. Penelitian yang dilakukan oleh [7] memanfaatkan algoritma Naïve Bayes untuk melakukan analisis sentimen terhadap dua *smartphone flagship* pada tahun 2023 yaitu Samsung S22 Ultra dan Xiaomi 12 Pro di *platform* media sosial seperti Twitter, Youtube, dan GSMArena dengan total 3.206 ulasan. Penelitian ini memperoleh hasil akurasi terbaik sebesar 66% dan *F1-score* sebesar 63%. Pelabelan sentimen pada penelitian tersebut dilakukan secara otomatis menggunakan *library* Python yaitu TextBlob yang masih terbatas pada bahasa, sarkasme, dan pemahaman struktur kalimat. Penelitian lain yang dilakukan oleh [8] melakukan analisis sentimen *review* hotel Indonesia dari *website* Traveloka berjumlah 2.500 baris data menggunakan algoritma LSTM. Penelitian ini mendapatkan hasil akurasi sebesar 85,96%. Tantangan yang dihadapi dari penelitian-penelitian sebelumnya sebagian besar disebabkan oleh jumlah *dataset* yang tidak melebihi 5.000 baris dan ketidakseimbangan antarkelas klasifikasi sehingga

mengakibatkan model cenderung mempelajari kelas mayoritas dan tidak dapat mendeteksi kelas minoritas yang ada [9].

Untuk mengatasi kekurangan pada penelitian sebelumnya, maka pada penelitian ini digunakan salah satu model *Transformer* yaitu IndoBERT yang merupakan turunan dari BERT (*Bidirectional Encoder Representations from Transformers*) yang secara spesifik dilatih menggunakan korpus bahasa Indonesia berjumlah 4 miliar kata (Indo4B) [10]. IndoBERT dirancang dengan arsitektur *encoder bidirectional* yang memungkinkannya memahami konteks suatu kata dengan mempertimbangkan informasi dari kedua arah, sehingga menjadikannya unggul dalam tugas-tugas seperti analisis sentimen dan ekstraksi informasi [11]. Penelitian yang dilakukan oleh [12] melakukan analisis sentimen dan membandingkan model *machine learning* seperti Naïve Bayes dan SVM dengan IndoBERT pada 1.500 ulasan produk kategori kecantikan, peralatan makan dan rumah tangga di Tokopedia. Algoritma Naïve Bayes, SVM dan IndoBERT mendapatkan akurasi berturut-turut sebesar 76.40%, 81.30% dan 88.33%. Penelitian lain yang dilakukan oleh [13] mengevaluasi IndoBERT dan IndoGPT pada 5.432 ulasan produk kecantikan dengan pembagian kelas positif, negatif dan netral sebanyak 3,288, 1,299, dan 845 baris yang menghasilkan akurasi sebesar 83% disusul oleh IndoGPT dengan akurasi sebesar 80%.

Berdasarkan uraian di atas maka pada penelitian ini dilakukan analisis sentimen penjualan *smartphone* pada Tokopedia menggunakan IndoBERT. Pada penelitian ini jumlah *dataset* yang digunakan berjumlah 8.203 baris yang kemudian dilakukan pelabelan data secara manual untuk memastikan kualitas pelabelan yang konsisten pada data yang beragam. Selanjutnya akan dilakukan tahapan *data preprocessing* yang meliputi *data cleaning, case folding*, *tokenizing*, *normalization, stopword removal* dan *lemmatization* yang menghasilkan data yang terstruktur dan lebih seragam. Berdasarkan uraian tersebut, maka akan dibuat penelitian yang dituangkan melalui Tugas Akhir dengan judul **"ANALISIS SENTIMEN PENJUALAN SMARTPHONE PADA TOKOPEDIA MENGGUNAKAN INDOBERT"**.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka permasalahan yang dihadapi pada penelitian ini yaitu:

1. Pengklasifikasian ulasan produk *smartphone* di Tokopedia yang tidak efisien dan efektif dikarenakan volume ulasan produk yang semakin bertambah setiap tahunnya.

2. Kurangnya jumlah *dataset* selama proses pelatihan yang menyebabkan model yang bias pada kelas tertentu.
3. Ulasan produk yang menggunakan bahasa yang sangat beragam, informal dan singkatan-singkatan baru sehingga proses klasifikasi ulasan menjadi tidak efisien.

1.3 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah membangun model IndoBERT yang dapat mengatasi ulasan yang mengandung bahasa informal dan *slang* untuk melakukan tugas analisis sentimen penjualan *smartphone* di Tokopedia agar lebih efisien dan efektif dalam memprediksi ulasan melalui peningkatan jumlah *dataset* yang digunakan.

1.4 Manfaat

Penelitian ini diharapkan dapat memberikan manfaat:

1. Model yang dihasilkan dapat membantu dalam pengklasifikasian ulasan pembeli *smartphone*.
2. Memberikan kontribusi dalam pengembangan ilmu pengetahuan yang berhubungan dengan analisis sentimen pada ulasan berbasis teks.

1.5 Ruang Lingkup

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data ulasan yang digunakan pada penelitian ini adalah hasil *scraping* pada rentang waktu Januari 2024 sampai dengan Mei 2026 dari *platform* Tokopedia dengan kata kunci *smartphone*, *handphone* dan hp berjumlah 8.203 ulasan.
2. Data yang diambil hanya berfokus pada konten teks yang berbahasa Indonesia.
3. *Dataset* ulasan yang digunakan akan dibagi menjadi 80:20 yang artinya 80% untuk *data training* dan 20% untuk *data testing* serta *data validation*.
4. Proses pengolahan data pada penelitian ini menggunakan bahasa pemrograman Python dengan bantuan *library* seperti *scikit-learn* dan *nltk*.
5. Metrik pengukuran kinerja model dengan *training* dan *validation loss*, *confusion matrix* dan *classification report* dengan mengukur nilai *accuracy*, *precision*, *recall* dan *F1-score*
6. Model terbaik yang dihasilkan dalam penelitian ini akan diaplikasikan ke dalam bentuk CLI (*Command Line Interface*) pada *Google Colab*.