BAB 2
LANDASAN TEORI

2.1 Tokopedia

Tokopedia merupakan sebuah platform jual beli berbasis online yang memungkinkan tiap individu dan para pelaku bisnis di Indonesia untuk mengembangkan dan memasarkan produk yang ingin dijual dengan mudah dan aman. Tokopedia menjadi platform *e-commerce* dengan pengunjung terbesar pada awal 2022 sebanyak 157,23 juta pengunjung [1]. Semua barang atau produk yang dijual sangat mudah ditemukan di situs toko online Tokopedia. Hal ini tentu sangat memudahkan pembeli yang ingin membeli suatu barang atau produk dikala tidak mempunyai waktu luang untuk belanja ke toko.

2.2 Analisis Sentimen

Analisis Sentimen merupakan proses memahami, mengekstraksi, dan mengolah data tekstual untuk mendapatkan informasi sentimental yang terkandung dalam sebuah kalimat opini [8]. Analisis sentimen adalah suatu proses pengolahan kata yang bertujuan untuk menganalisa opini, sentimen, evaluasi, penilaian, dan sikap terhadap suatu produk, topik, dan pendapat [9].

2.3 Text Preprocessing

Text preprocessing merupakan sebuah metode *Natural Languange Processing* (NLP) pada dokumen yang berupa teks. *Text preprocessing* merupakan proses pemilihan data teks agar hasilnya menjadi lebih terstruktur. *Text preprocessing* terdiri dari beberapa tahapan, yaitu :

2.3.1 Case Folding

Case Folding adalah suatu proses perubahan huruf kapital menjadi huruf kecil.

2.3.2 Tokenizing

Tokenizing atau tokenisasi merupakan proses memecah sekumpulankarakter pada teks menjadi satuan kata. Proses tokenisasi menghilangkan tanda baca

seperti tanda titik (.), tanda koma (,), dan angka yang terdapat pada teks. [10].

2.3.3 Filtering

Filtering adalah proses pemisahan kata-kata dari hasil tokenisasi dengan menggunakan algoritma *stoplist* (menghilangkan kata yang tidak penting) dan *wordlist* (menyimpan kata-kata penting). *Stopword* merupakan kata yang umum yang sering muncul dan tidak memiliki arti. Contoh *stopword* adalah “yang”, “di”, “dan”, dan lainnya [11].

2.3.4 Stemming

Stemming adalah proses mereduksi bentuk kata menjadi kata dasar. Tujuan proses stemming adalah menghilangkan bubuhan yang berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata [10].

2.4 Term Frequency-Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF-IDF) merupakan metode pembobotan yang menggabungkan dua konsep, yaitu *Term Frequency* dan *Document Frequency*[12]. *Term Frequency* adalah konsep pembobotan dengan mencari seberapa sering munculnya sebuah term dalam satu dokumen. *Document Frequency* adalah banyaknya jumlah dokumen di mana sebuah term itu muncul. Semakin kecil frekuensi kemunculan, maka semakin kecil juga nilai bobotnya[12].

2.5 Confusion Matrix

Confusion matrix merupakan suatu metode yang umum digunakan untuk menghitung tingkat akurasi dalam data mining. *Confusion Matrix* merupakan representasi hasil klasifikasi biner pada suatu dataset [12].

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Keterangan:

- TP (*True Positive*) : Data positif yang diprediksi benar.
- FP (*False Positive*) : Data negative yang diprediksi sebagai data positif.
- FN (*False Negative*) : Data Positif yang diprediksi sebagai data negatif.
- TN (*True Negative*) : Data negative yang diprediksi benar.

Dalam *Confusion Matrix* terdapat beberapa rumus umum untuk menghitung performa klasifikasi, yaitu *accuration, precision, recall*, dan *F-Measure*.

1. *Accuracy*

   Accuracy merupakan jumlah proporsi prediksi yang benar. Perhitungan *Accuracy* dapat dilihat pada rumus dibawah ini [12].

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

2. *Precision*

   Precision merupakan jumlah bagian data teks dokumen yang relevan yang memeriksa semua teks dokumen yang dipilih. Perhitungan *Precision* dapat dilihat pada rumus dibawah ini [12].

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

3. *Recall*

   Recall merupakan jumlah proporsi teks dokumen yang relevan yang memeriksa diantara semua teks dokumen relevan yang ada pada koleksi. Perhitungan *recall* dapat dilihat pada rumus di bawah ini [12].

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

4. *F1-Score*

   F1-Score adalah ukuran keberhasilan pengambilan parameter tunggal yang menggabungkan *recall* dan *precision*. Hasil perhitungannya diperoleh dari hasil mengalikan *precision* dan *recall* kemudian dibagi hasil penjumlahan *precision* dan *recall*, lalu dikalikan dengan 2. Perhitungan *f1-score* dapat dilihat pada rumus di bawah ini [13].

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall} \tag{2.4}$$

2.6 Algoritma Naive Bayes

Naive Bayes adalah metode klasifikasi dengan menggunakan metode probabilitas yang dikemukakan oleh ilmuan Inggris Thomas Bayes. *Naive Bayes* adalah pengklasifikasian probabilitas sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data yang diberikan. Algoritma menggunakan teorema Bayes dan mengasumsikan bahwa semua atribut

yang tidak saling bergantung atau atribut independen yang diberikan oleh nilai pada variabel kelas [14]. Rumus perhitungan persamaan dari Na¨ıve Bayes adalah sebagai berikut.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{2.5}$$

Keterangan :

X : Data dengan kelas yang belum diketahui
H : Hipotesis data merupakan suatu kelas spesifik
P(H—X): Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X—H): Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X