BAB 2
LANDASAN TEORI

Penelitian ini menerapkan beragam literatur sebagai acuan dan pedoman dalam menyusun laporan. *Review* literatur yang dilakukan menghasilkan identifikasi berbagai teori dan formula logika algoritma yang relevan, yang digunakan untuk pengembangan perangkat lunak yang dipaparkan dalam skripsi ini. Pendekatan tersebut memfasilitasi sintesis antara konsep teoretis dan implementasi praktis, memastikan bahwa pengembangan model didasarkan pada basis ilmiah yang kuat dan terkini.

2.1 Analisis Sentimen

Analisis sentimen merupakan suatu proses yang berfokus pada penentuan opini yang dinyatakan dalam bentuk teks dan dapat dikategorikan sebagai sentimen positif atau negatif [18]. Sebagai cabang dari *text mining*, analisis sentimen atau yang dikenal sebagai *opinion mining* menjadi subjek penelitian yang berfokus dalam menentukan persepsi atau subjektivitas masyarakat terhadap suatu topik, kejadian, atau permasalahan tertentu [19].

2.2 Support Vector Machine

Support Vector Machine (SVM) adalah sistem pembelajaran berbasis optimisasi yang menggunakan ruang fiktif dalam bentuk fungsi linier dalam fitur berdimensi tinggi. Dibandingkan dengan metode lainnya, SVM dianggap sebagai teknologi yang relatif baru [20].

2.3 TF-IDF

TF-IDF merupakan teknik pembobotan kata, di mana *Term Frequency* (TF) digunakan untuk menyimpan frekuensi kemunculan kata dalam suatu dokumen. *Document Frequency* (DF) mencatat frekuensi kemunculan dokumen, sementara *Inverse Document Frequency* (IDF) menampung nilai kebalikan dari DF [21].

2.4 Klasifikasi

Klasifikasi merupakan suatu proses pengelompokkan dengan menggunakan data pelatihan. Proses ini dapat melibatkan data kategori atau data berkelanjutan, di mana dilakukan pelabelan terhadap atribut yang menjadi keluaran atau kelas dari suatu rekaman [22].

2.5 Text-Preprocessing

Text-preprocessing adalah tahapan yang melibatkan serangkaian operasi yang dilakukan dalam mengolah data sebelum data final digunakan dalam algoritma data mining. Operasi yang dilakukan mencakup pembersihan data, seperti penanangan *noice* dan *inconsistent-data*, transformasi data menjadi bentuk yang sesuai dan reduksi data yang melibatkan seleksi dan ekstraksi fitur [23]. Beberapa operasi yang dapat dilakukan pada tahapan *text-preprocessing* adalah sebagai berikut:

1. Cleaning

   Cleaning merupakan operasi yang bertujuan untuk menghilangkan semua karakter di dalam text yang tidak termasuk alfabet. Maksud dari dilakukannya *cleaning* adalah untuk mengurangi karakter atau simbol yang tidak dikehendaki ataupun tidak memiliki makna di dalam melakukan analisis sentimen. [24]

2. Case Folding

   Case folding adalah operasi sederhana yang bertujuan untuk penyeragaman seluruh kata menjadi huruf kecil. Dalam *case folding* huruf yang diterima hanyalah a sampai z. Pada operasi *case folding* karakter selain huruf akan dihilangkan [24].

3. Tokenizing

   Tokenizing adalah operasi yang dilakukan untuk memisahkan rangkaian kalimat menjadi kata-kata yang tunggal [25]. Pemisahan dilakukan dengan tujuan untuk membentuk token berdasarkan pemisahan rangkaian kalimat tersebut [26].

4. Stemming

   Stemming adalah operasi untuk menghilangkan imbuhan yang menempel pada suatu kata dengan tujuan menemukan kata dasar pada kata tersebut [27].

5. Normalization

   Normalization adalah operasi untuk merubah kata-kata tidak standar seperti singkatan menjadi kata baku [23].

6. *Stopword Removal*

   Stopword Removal adalah operasi untuk menghilangkan kata sambung yang dinilai tidak bermakna dalam suatu analisis sentimen [25].

7. Labeling

   Labeling merupakan operasi pada dataset yang telah melewati sejumlah tahap dalam *text-preprocessing* dengan tujuan mendefinisikan kalimat yang mengandung nilai positif atau nilai negatif. Masing - masing kalimat akan diberikan label sebagai kalimat yang positif atau kalimat negatif [24].

2.6 Confussion Matrix

Confusion matrix adalah salah satu metode yang digunakan untuk mengevaluasi kinerja suatu sistem klasifikasi [28]. Pada dasarnya, *confusion matrix* menyajikan informasi perbandingan antara hasil klasifikasi yang diberikan oleh sistem dengan hasil klasifikasi yang seharusnya. Dalam pengukuran kinerja menggunakan confusion matrix, terdapat empat komponen yang menjadi hasil perbandingan dari proses klasifikasi. Keempat komponen tersebut melibatkan *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) yang masing-masing dirincikan sebagai berikut.

- TP merupakan data positif yang diprediksi dengan benar.
- FP merupakan data positif dengan prediksi salah.
- TN merupakan data negatif yang diprediksi dengan benar
- FN merupakan data negatif dengan prediksi salah.

Keempat komponen ini merupakan elemen penting dalam menghitung *Accuracy*, *Recall*, *Precision*, dan *F1-Score* dalam evaluasi sistem klasifikasi. Dengan memahami peran masing-masing komponen, kita dapat memperoleh pemahaman yang lebih baik tentang bagaimana setiap metrik memberikan gambaran tentang kinerja keseluruhan dari model klasifikasi yang dievaluasi. Definisi dan persamaan untuk menghitung metrik-metrik evaluasi dalam *Confusion Matrix* dijabarkan sebagai berikut.

1. *Accuracy*

   Accuracy adalah metrik yang digunakan untuk mengukur sejauh mana prediksi model sesuai dengan nilai sebenarnya. Dengan kata lain, *accuracy* menunjukkan proporsi total prediksi yang benar, baik itu prediksi positif maupun negatif, dibandingkan dengan seluruh *instance* yang ada. Rumus untuk menentukan *accuracy* dapat dilihat pada Persamaan 2.1.

$$Accuracy = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{Total\ Number\ of\ Instances} \tag{2.1}$$

2. *Recall*

   Recall adalah metrik yang digunakan untuk mengukur kemampuan model dalam mendeteksi semua *instance* positif dalam dataset. Recall menunjukkan proporsi *instance* positif yang benar-benar diidentifikasi dengan benar oleh model. Dengan kata lain, *recall* menjawab pernyataan: dari semua *instance* yang sebenarnya positif, berapa persen yang berhasil diprediksi positif oleh model. Rumus untuk menentukan *recall* dapat dilihat pada Persamaan 2.2.

$$Recall = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{2.2}$$

3. *Precision*

   Precision dalam *Confusion Matrix* adalah metrik yang digunakan untuk mengukur akurasi prediksi model pada kelas positif. *Precision* menunjukkan proporsi prediksi positif yang benar-benar positif. Dengan kata lain, *precision* menjawab pernyataan: dari semua *instance* yang diprediksi sebagai positif,

berapa persen yang benar-benar positif. Rumus untuk menentukan *precision* dapat dilihat pada Persamaan 2.3.

$$Precision = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}} \tag{2.3}$$

4. *F1-score*

   F1-score adalah metrik yang digunakan untuk mengukur keseimbangan antara *precision* dan *recall*. *F1-score* adalah rata-rata harmonis dari *precision* dan *recall* dan memberikan gambaran yang lebih baik tentang performa model ketika ada ketidakseimbangan antara kelas positif dan negatif. Rumus untuk menentukan *f1-score* dapat dilihat pada Persamaan 2.4.

$$F1\ Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

2.7 Word Cloud

Salah satu jenis alat visualisasi yang digunakan untuk menampilkan kata-kata yang paling umum dalam suatu kumpulan data teks adalah *word cloud*. Ukuran *font* masing-masing kata ditampilkan berkaitann dengan frekuensi kecumunculan dalam data; semakin besar *font* yang ditampilkan, semakin sering muncul kata tersebut di dalam sebuah data teks. *Word cloud* sangat membantu dalam menganalisis sentimen karena dapat memperlihatkan emosi atau opini yang paling sering diungkapkan dalam suatu teks.