BAB 2
LANDASAN TEORI

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 [9]. 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 [10].

2.2 Klasifikasi

Klasifikasi merupakan suatu proses pengelompokan 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 [11].

2.3 *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 [12]. Konsep dasar SVM dapat dijelaskan sebagai upaya menemukan hyperplane optimal yang bertindak sebagai pemisah antara dua kelas di ruang input [13]. SVM dirancang khusus untuk data klasifikasi dengan cara mencari margin terbesar antara hyperplane pemisah dengan kedua kelas data. Margin sendiri adalah jarak antara hyperplane dengan titik terdekat dari masing-masing kelas. Dalam implementasinya, SVM memiliki beberapa tahapan utama [14]:

2.3.1 Pemilihan Data *Training*

Tahap awal dimulai dengan pemilihan data *training* yang representatif. Data ini akan digunakan untuk pembelajaran dan pengembangan model klasifikasi.

Kualitas dan representativitas data *training* sangat mempengaruhi performa akhir model SVM.

2.3.2 Penentuan Hyperplane

Setelah data *training* dipilih, SVM akan mencari hyperplane optimal dengan rumus:

$$W \cdot X + b = 0 \tag{2.1}$$

dimana W adalah vektor bobot, X adalah vektor input, dan b adalah bias.

2.3.3 Penerapan Kernel

Untuk kasus data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk mentransformasi data ke dimensi yang lebih tinggi. Beberapa kernel yang umum digunakan meliputi:

- Kernel Linear
- Kernel Polynomial
- Kernel RBF (Radial Basis Function)

2.3.4 Evaluasi Model

Evaluasi performa SVM biasanya menggunakan metrics seperti:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.4}$$

dimana TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative) adalah komponen confusion matrix. SVM telah terbukti efektif dalam berbagai aplikasi klasifikasi, termasuk pengenalan pola, klasifikasi teks, dan analisis sentimen. Keunggulan utama SVM adalah kemampuannya dalam menangani data berdimensi tinggi dan memberikan solusi global optimal [13].

2.4 Seleksi Fitur *Chi Square*

Chi-Square dalam konteks analisis sentimen pemindahan ibu kota dapat digunakan untuk mengukur korelasi antara fitur (kata-kata dalam komentar) dengan label sentimen (positif/negatif) [15]. Perhitungannya menggunakan rumus:

$$K = \sum \frac{(O-E)^2}{E} \tag{2.5}$$

dimana:

- O adalah frekuensi observasi kemunculan kata dalam suatu kelas sentimen
- E adalah frekuensi yang diharapkan

Semakin tinggi nilai *Chi-Square* suatu kata, semakin kuat korelasinya dengan label sentimen, sehingga kata tersebut layak dijadikan fitur untuk klasifikasi.

2.4.1 Implementasi

Dalam implementasinya, langkah-langkah yang dilakukan adalah:

1. Membuat matriks observasi frekuensi kata (O) yang muncul di tweet positif dan negatif:

$$O = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \tag{2.6}$$

dimana f_{ij} adalah frekuensi kata i pada kelas sentimen j

2. Menghitung nilai harapan (E) untuk setiap kata menggunakan total marginal:

$$E_{ij} = \frac{R_i \times C_j}{N} \tag{2.7}$$

dimana:

- R_i = jumlah total baris ke-i
- C_j = jumlah total kolom ke-j
- N = total keseluruhan data

3. Menghitung nilai *Chi-Square* untuk setiap kata:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.8}$$

4. Melakukan seleksi fitur berdasarkan nilai threshold:
   - Jika $\chi^2 > threshold$: kata dijadikan fitur
   - Jika $\chi^2 \leq threshold$: kata dihapus

Fitur-fitur yang terpilih dari hasil seleksi menggunakan *Chi-Square* ini kemudian akan digunakan sebagai input untuk klasifikasi menggunakan algoritma SVM. Metode ini dapat membantu mengurangi dimensi fitur dan meningkatkan akurasi klasifikasi sentimen. Berikut pada Tabel 2.1 merupakan contoh tabel kontingensi dari hasil seleksi *Chi-Square*.

Tabel 2.1. Tabel contoh kontingensi *Chi-Square* untuk setiap kata

Kata	Sentimen		Total
	Positif	Negatif	
White collar	O_{11}	O_{12}	R_1
Blue collar	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

2.5 Text-Preprocessing

Text-preprocessing adalah tahap penting dalam pengolahan data teks sebelum data tersebut digunakan dalam algoritma data mining. Tahapan ini mencakup berbagai operasi untuk membersihkan dan menyiapkan data, serta mengubahnya menjadi format yang sesuai. Beberapa operasi utama dalam text-preprocessing adalah sebagai berikut:

1. ***Cleaning***: Operasi ini bertujuan untuk menghilangkan karakter non-alfabet dari teks. Tujuannya adalah untuk mengurangi noise dan simbol yang tidak relevan dalam analisis sentimen [16].

2. ***Tokenizing***: Operasi ini memisahkan rangkaian kalimat menjadi kata-kata individu. Tujuannya adalah untuk membentuk token dari teks yang diberikan [17].

3. ***Stemming***: Stemming bertujuan untuk menghilangkan imbuhan dari kata sehingga hanya menyisakan kata dasar. Hal ini membantu dalam menemukan bentuk dasar dari kata tersebut [18].

4. ***Normalization***: Operasi ini merubah kata-kata tidak standar, seperti singkatan, menjadi bentuk kata baku [19].

5. ***Labeling***: Labeling adalah proses mendefinisikan kalimat yang telah diproses sebagai kalimat yang memiliki nilai positif atau negatif. Setiap kalimat diberi label sesuai dengan sentimen yang terkandung [16].

6. ***Stopword Removal***: Operasi ini menghilangkan kata sambung atau kata yang dianggap tidak bermakna dalam analisis sentimen [20].