# BAB III DESAIN DAN IMPLEMENTASI SISTEM

* 1. **Metode Penelitian**

Penelitian ini menggunakan pendekatan eksperimen untuk mengembangkan model analisis sentimen berbasis emosi menggunakan algoritma hybrid CNN-LSTM. Metode penelitian terdiri dari beberapa tahap, yaitu pengumpulan data, preprocessing data, pengembangan arsitektur model, pelatihan, evaluasi, dan analisis hasil.

1. **Pengumpulan Data**

Dataset yang digunakan adalah data dari Twitter yang mencakup sekitar 4.400 entri. Dataset ini berisi teks pendek dengan label emosi yang telah dikategorikan menjadi lima kelas emosi utama, yaitu:

* + **Happy**
  + **Love**
  + **Sadness**
  + **Anger**
  + **Fear**

Data akan dibagi menjadi tiga subset:

* + **Data Training:** 70% dari total dataset untuk melatih model.
  + **Data Validasi:** 15% dari dataset untuk memvalidasi performa selama pelatihan.
  + **Data Uji:** 15% dari dataset untuk mengevaluasi performa akhir model.

1. **Preprocessing Data**

Proses preprocessing dilakukan untuk memastikan data dalam format yang dapat diproses oleh model. Langkah-langkahnya meliputi:

* + **Tokenisasi:** Memecah teks menjadi kata-kata individu atau token.
  + **Stopword Removal:** Menghapus kata-kata yang tidak relevan seperti "di", "ke", "dan".
  + **Stemming:** Mengembalikan kata ke bentuk dasar.
  + **Padding:** Menyelaraskan panjang teks agar konsisten.
  + **Encoding:** Mengonversi teks menjadi representasi numerik berbasis indeks.

1. **Pengembangan Model**

Model yang dirancang adalah arsitektur hybrid CNN-LSTM. CNN digunakan untuk menangkap pola lokal dalam teks, sedangkan LSTM digunakan untuk memahami hubungan temporal antar kata. Desain model mencakup tahap berikut:

* + Penambahan lapisan embedding untuk representasi vektor kata.
  + Lapisan CNN untuk mengekstraksi fitur lokal.
  + Lapisan LSTM bertingkat untuk menangkap konteks temporal dalam teks.

1. **Pelatihan dan Evaluasi**

Model dilatih menggunakan algoritma optimasi Adam dengan learning rate 0.005. Evaluasi dilakukan menggunakan metrik seperti akurasi, precision, recall, dan F1-score. Data validasi digunakan untuk memastikan bahwa model tidak overfitting.

* 1. **Desain Sistem**

Desain sistem ini memanfaatkan kombinasi Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM) untuk menghasilkan model analisis sentimen berbasis emosi yang mampu menangkap pola lokal (lokalitas) dan hubungan temporal dalam teks. Berikut adalah deskripsi terperinci setiap komponen dalam desain sistem.

**3.2.1 Input Data**

Input data untuk sistem ini berasal dari dataset Twitter yang mencakup 4.400 entri teks dengan label emosi. Setiap entri telah dikategorikan menjadi salah satu dari lima kelas emosi: **happy**, **love**, **sadness**, **anger**, dan **fear**. Teks mentah diproses melalui tahap-tahap preprocessing, termasuk tokenisasi, penghapusan stopword, dan stemming, untuk memastikan teks siap diolah oleh model. Panjang teks diseragamkan menggunakan padding agar semua data memiliki dimensi input yang sama.

**3.2.2 Lapisan Embedding**

Lapisan pertama dalam model ini adalah embedding. Lapisan ini digunakan untuk mengonversi kata-kata dalam teks menjadi representasi vektor berdimensi tetap. Representasi ini didasarkan pada embedding pretrained (seperti GloVe atau FastText) untuk memanfaatkan pengetahuan semantik dan sintaksis yang telah dilatih sebelumnya.

Setiap kata dalam teks direpresentasikan sebagai vektor berdimensi 200. Embedding matrix yang digunakan tidak dapat di-train ulang (non-trainable) agar dapat memanfaatkan keunggulan representasi pretrained secara optimal. Output dari lapisan ini adalah tensor dengan dimensi (batch size, sequence length, embedding dimension)\text{(batch size, sequence length, embedding dimension)}(batch size, sequence length, embedding dimension).

**3.2.3 Lapisan Convolutional Neural Network (CNN)**

Setelah lapisan embedding, data diteruskan ke lapisan CNN untuk mengekstraksi fitur lokal dari teks. CNN dirancang untuk mendeteksi pola-pola penting dalam teks, seperti n-gram (kombinasi kata yang sering muncul bersama). CNN menggunakan beberapa ukuran filter untuk menangkap berbagai tingkat granularitas informasi.

* **Konfigurasi CNN:**
  + Jumlah filter: 128 untuk setiap ukuran filter.
  + Ukuran filter: 3, 4, dan 5 (menggambarkan panjang kombinasi kata yang berbeda).
  + Fungsi aktivasi: ReLU (Rectified Linear Unit) untuk menghasilkan non-linearitas.
  + Global Max Pooling: Setelah fitur dihasilkan oleh CNN, pooling dilakukan untuk mereduksi dimensi fitur menjadi ukuran tetap sambil mempertahankan informasi penting dari setiap filter.

Output dari lapisan ini adalah representasi fitur lokal teks yang akan diteruskan ke lapisan berikutnya.

**3.2.4 Lapisan Long Short-Term Memory (LSTM)**

Setelah fitur lokal diekstraksi oleh CNN, data diteruskan ke lapisan LSTM untuk memahami hubungan temporal antar kata dalam teks. LSTM digunakan karena kemampuannya untuk mempertahankan konteks kata dalam urutan panjang.

Model ini menggunakan arsitektur bidirectional LSTM untuk menangkap konteks dari kedua arah teks (maju dan mundur). Dengan demikian, model dapat lebih akurat memahami konteks emosi yang terkandung dalam teks.

* **Konfigurasi LSTM:**
  + **Layer 1:** 256 unit dengan dropout 0.2 dan recurrent dropout 0.2. Lapisan ini dirancang untuk menangkap pola temporal tingkat tinggi.
  + **Layer 2:** 128 unit dengan dropout 0.1 dan recurrent dropout 0.2. Lapisan ini mendeteksi pola yang lebih spesifik dari data.
  + **Layer 3:** 128 unit dengan dropout 0.2 dan recurrent dropout 0.2. Lapisan ini menyelesaikan proses ekstraksi pola temporal dan menghasilkan output akhir untuk lapisan dense.

Output dari lapisan LSTM adalah tensor berdimensi tetap yang merepresentasikan hubungan temporal antar kata dalam teks.

**3.2.5 Lapisan Dense dan Softmax**

Output dari lapisan LSTM diteruskan ke lapisan dense (fully connected layer) dengan aktivasi softmax. Lapisan ini bertugas menghasilkan probabilitas untuk setiap kelas emosi, sesuai dengan jumlah kategori dalam dataset (5 emosi: happy, love, sadness, anger, fear).

* **Konfigurasi Lapisan Dense:**
  + Jumlah neuron: 5 (sesuai dengan jumlah kelas emosi).
  + Fungsi aktivasi: Softmax, untuk memastikan output adalah probabilitas yang dijumlahkan hingga 1.

Probabilitas yang dihasilkan akan digunakan untuk menentukan kelas emosi akhir dari teks input.

**3.2.6 Optimasi dan Kompilasi**

Model dikompilasi menggunakan fungsi loss categorical crossentropy karena ini adalah tugas klasifikasi multikelas. Algoritma optimasi Adam digunakan untuk mempercepat konvergensi, dengan learning rate 0.005 yang dipilih berdasarkan eksperimen awal.

**3.2.7 Evaluasi Model**

Evaluasi model dilakukan dengan menggunakan subset data validasi dan uji. Metrik yang digunakan mencakup:

* **Akurasi:** Untuk mengukur persentase prediksi benar.
* **Presisi:** Untuk mengevaluasi keakuratan prediksi positif.
* **Recall:** Untuk mengukur sensitivitas terhadap kelas yang benar.
* **F1-Score:** Kombinasi harmonis antara presisi dan recall.

Evaluasi dilakukan pada data uji untuk memastikan bahwa model tidak overfitting pada data pelatihan.

**3.2.8 Arsitektur Sistem**

Berikut adalah deskripsi arsitektur sistem secara keseluruhan:

1. **Input Data:** Dataset Twitter yang telah diproses (preprocessed).
2. **Embedding Layer:** Mengonversi kata menjadi vektor.
3. **CNN Layer:** Mengekstraksi pola lokal dengan beberapa ukuran filter.
4. **Global Max Pooling:** Mengurangi dimensi fitur tanpa kehilangan informasi penting.
5. **Bidirectional LSTM:** Menganalisis hubungan temporal dalam teks.
6. **Dense Layer dengan Softmax:** Menghasilkan prediksi probabilitas untuk lima kelas emosi.