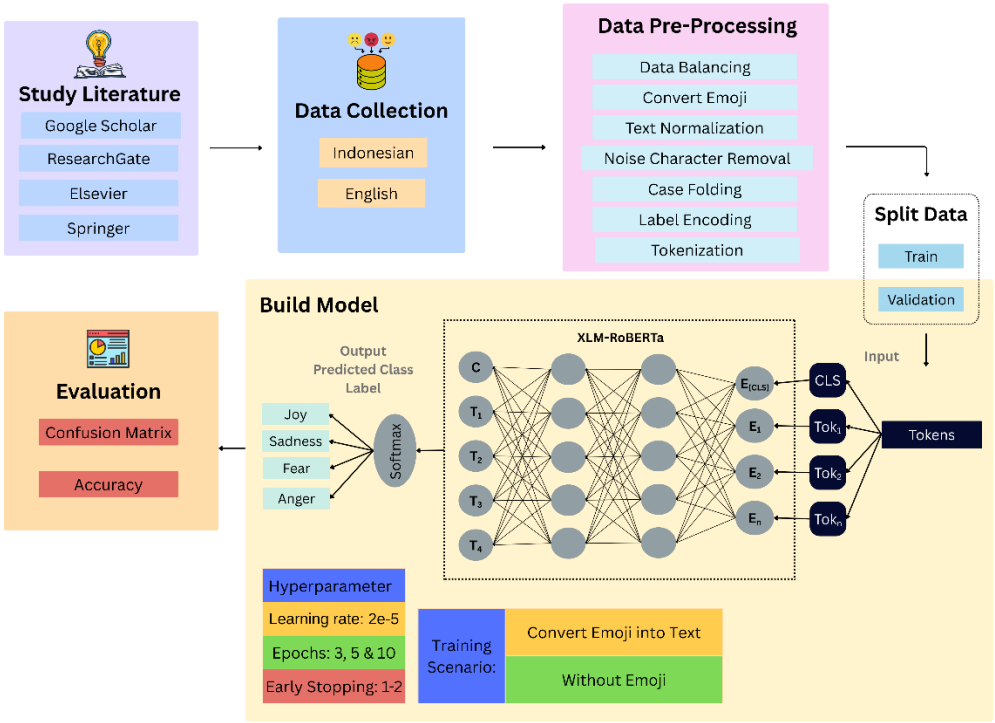
BAB III

METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Metodologi penelitian ini menjelaskan tahapan dari penelitian yang akan dilakukan. Terdapat lima tahapan dalam penelitian ini, diantaranya: *Study Literature, Data Collection, Data Pre-Processing, Build Model, Evaluation*. Tahapan penelitian disajikan pada Gambar 3.1.



Gambar 3.1 Tahapan penelitian

3.2 *Study Literature*

Tahap ini dilakukan dengan mengkaji berbagai referensi terkait klasifikasi emosi berbasis teks untuk memahami metode yang telah digunakan dalam penelitian sebelumnya. Pencarian referensi dilakukan melalui berbagai sumber akademik seperti *Google Scholar, Springer, Elsevier, dan ResearchGate*, dan yang

lainnya dengan fokus pada penelitian lima tahun terakhir yang membahas berbagai pendekatan dalam klasifikasi emosi teks. Kajian ini mencakup model berbasis *machine learning*, *deep learning*, dan *transformer*. Tahap ini juga mencakup kajian mengenai teknik pra-pemrosesan data. Studi-studi tersebut dapat menjadi acuan penelitian ini untuk menentukan strategi terbaik dalam membangun model klasifikasi emosi teks yang optimal.

3.3 *Data Collection*

Tahap ini dilakukan pengumpulan data. Dataset yang digunakan merupakan kumpulan dari postingan pengguna media sosial. Data yang dikumpulkan ini menggunakan teknik *crawling*. *Data crawling* adalah proses pengambilan data dari berbagai sumber di media sosial yang dapat dilakukan menggunakan *API* resmi atau *tools* pihak ketiga untuk mengekstrak data yang relevan dengan penelitian (Rifaldi et al., 2023).

Dataset Bahasa Inggris yang akan digunakan dalam penelitian ini diambil dari platform Kaggle dengan judul "*Emotion Classification NLP*". Dataset ini juga dikenal sebagai dataset WASSA-2017, yang terdiri sekitar 7.102 entri dan mencakup empat label emosi, yaitu *fear*, *anger*, *joy*, dan *sadness*. Dataset ini pada awalnya tersedia dalam format .txt melalui situs resmi WASSA, namun oleh pengguna Kaggle dataset tersebut telah dikonversi ke dalam format .csv tanpa mengubah isi atau struktur data aslinya. Dengan demikian, isi dataset tetap konsisten dengan versi aslinya yang digunakan dalam kompetisi WASSA-2017.

Dataset Bahasa Indonesia yang digunakan dalam penelitian ini diambil dari platform GitHub (https://github.com/ines179/Emotion-Detection_Indo*BERT*-

CNN-BiLSTM). Dataset tersebut terdiri 14.026 entri dengan enam label emosi, yaitu: *fear*, *disgust*, *surprise*, *anger*, *joy*, dan *sadness*. Hanya empat label emosi yang digunakan, yaitu *anger*, *fear*, *sadness*, dan *joy*. Pemilihan jumlah label ini disesuaikan dengan keterbatasan data yang tersedia pada dataset Bahasa Inggris, sehingga untuk menjaga konsistensi antara kedua dataset, hanya empat kelas emosi yang digunakan.

3.4 *Data Pre-Processing*

Data yang telah dikumpulkan selanjutnya diproses untuk memastikan kualitas data yang bersih sebelum digunakan dalam pelatihan model. Langkah ini mencakup evaluasi kualitas data, serta eksplorasi awal untuk memahami karakteristik dataset (Holstein et al., 2024). Data yang diambil dari media sosial mengandung banyak *noise* yang tidak diperlukan dalam penelitian, seperti username yang memiliki simbol (@) untuk penyebutan pengguna. Tahap ini mencakup berbagai langkah, seperti menghilangkan tanda baca, angka, mengganti kata informal, serta mengonversi emoji ke dalam teks.

3.4.1 *Data Balancing*

Proses data balancing dilakukan untuk mengatasi ketidakseimbangan data pada dataset agar model tidak bias terhadap kelas mayoritas. Dua pendekatan berbeda diterapkan sesuai kondisi masing-masing dataset. Dataset bahasa Inggris memiliki distribusi label cenderung tidak seimbang dengan jumlah data yang rendah pada kelas minoritas. Oleh karena itu, digunakan teknik *random oversampling*. Sebaliknya, pada dataset Bahasa Indonesia, distribusi label sudah seimbang, namun jumlah total datanya jauh lebih besar dibanding dataset Bahasa

Inggris. Teknik *random undersampling* diterapkan untuk menyelaraskan ukuran dataset. Pendekatan ini dipilih karena sederhana, efektif, dan tidak mengubah konten data asli, sehingga tetap menjaga konteks semantik yang penting dalam pemodelan berbasis teks (Tyas et al., 2023).

3.4.2 *Convert Emoji*

Emoji yang terdapat dalam teks dikonversi menjadi deskripsi teks. Pada dataset bahasa Inggris, konversi emoji dilakukan secara otomatis menggunakan *library* yang dapat mendeteksi dan menggantikan emoji dengan teks deskripsi. Pada dataset bahasa Indonesia, digunakan kamus khusus yang memetakan setiap emoji ke dalam deskripsi teks bahasa Indonesia. Langkah ini dilakukan untuk memastikan emoji dapat diproses dengan cara yang konsisten dan memberikan informasi emosi yang lebih jelas pada saat proses analisis (Yoo & Rayz, 2021). Contoh penerapan *convert* emoji ini disajikan dalam Tabel 3.1

Tabel 3.1 Contoh Penerapan *Convert Emoji*

Before	*After*
Parah sih gua belanja di @beststore dikasih cashback karna first order, mantap2! 👍	Parah sih gua belanja di @beststore dikasih cashback karna first order, mantap2! jempol

3.4.3 *Text Normalization*

Penggunaan kata tidak baku sangat umum dalam percakapan di media sosial. Tahap selanjutnya dilakukan normalisasi teks dengan mengganti kata informal dengan bentuk standar teks yang lebih formal atau baku. Pergantian ini dilakukan menggunakan sebuah kamus konversi yang berisi daftar kata tidak baku beserta kata dalam bahasa bakunya. Saat teks diproses, setiap kata yang terdeteksi sebagai

tidak baku akan dicocokkan dengan kamus ini dan diganti dengan versi yang lebih sesuai untuk analisis. Proses ini bertujuan untuk memastikan bahwa model dapat memahami makna sebenarnya dari teks tanpa terpengaruh oleh variasi bahasa informal yang sering digunakan di media sosial (Siino et al., 2024). Ilustrasi penerapan normalisasi teks disajikan dalam Tabel 3.2.

Tabel 3.2 Contoh Penerapan Normalisasi Teks

Before	*After*
Parah sih gua belanja di @beststore dikasih cashback karna first order, mantap2! jempol	Parah sih saya belanja di @beststore dikasih cashback karena first order, mantap2! jempol

3.4.4 *Noise Character Removal*

Tahap ini, dilakukan proses pembersihan teks dengan menghilangkan elemen-elemen yang tidak relevan untuk analisis emosi. Proses ini mencakup penghapusan URL, angka, sebagian tanda baca serta elemen lain yang tidak memberikan kontribusi langsung terhadap pemahaman emosi dalam teks. *Username* atau *mention* yang diawali dengan simbol '@' juga dihapus karena umumnya hanya digunakan untuk menandai pengguna atau menyertakan informasi tambahan yang tidak diperlukan dalam analisis (Siino et al., 2024). Terdapat beberapa tanda baca yang tetap dipertahankan, yaitu tanda seru (!), tanda tanya (?), dan elipsis (...). Ketiga tanda baca ini terbukti memiliki peran penting dalam mengekspresikan dan memperkuat makna emosional suatu teks. Tanda seru (!) sering diasosiasikan dengan emosi berintensitas tinggi seperti *anger* (marah) dan *joy* (senang), sementara tanda tanya (?) mencerminkan ketidakpastian atau kecemasan yang dapat ditemukan dalam emosi *fear* (takut) maupun *sadness* (sedih).

Adapun elipsis (…) digunakan untuk menyampaikan jeda dramatis yang umum dalam ekspresi kesedihan dan ketakutan (Li & Zou, 2024). Model diharapkan mampu lebih fokus dalam mengenali pola emosional secara akurat melalui seleksi elemen teks yang cermat. Contoh penghapusan simbol dan angka dapat dilihat pada Tabel 3.3.

Tabel 3.3 Contoh Penerapan *Noise Character Removal*

Before	*After*
Parah sih saya belanja di @beststore dikasih cashback karena first order, mantap2! jempol	Parah sih saya belanja di dikasih cashback karena first order mantap! jempol

3.4.5 *Case Folding*

Proses *case folding* dilakukan untuk membuat teks lebih seragam, yaitu dengan mengubah semua huruf menjadi huruf kecil (*lowercase*). Langkah ini dilakukan untuk menghilangkan perbedaan antara huruf besar dan kecil yang tidak memiliki makna khusus dalam analisis emosi (Aripin et al., 2023). Tabel 3.4 menyajikan contoh penerapan dari *case folding.*

Tabel 3.4 Contoh Penerapan *Case Folding*

Before	*After*
Parah sih saya belanja di dikasih cashback karena first order mantap! jempol	parah sih saya belanja di dikasih cashback karena first order mantap! jempol

3.4.6 *Label Encoding*

Proses pra-pemrosesan dilanjutkan dengan *label encoding*. *Label encoding* merupakan teknik yang digunakan untuk mengonversi label emosi yang semula berbentuk teks menjadi representasi numerik. Proses ini bertujuan agar label dapat dikenali oleh model saat proses pelatihan. Setiap label emosi diberi nilai numerik

yang unik, misalnya *anger*: 0, *joy*: 1, *fear*: 2, *sadness*: 3. Proses ini dilakukan menggunakan fungsi *LabelEncoder* dari pustaka *scikit-learn* (Herdian et al., 2024).

3.4.7 *Tokenization*

Teks yang telah melalui tahap pembersihan selanjutnya diproses dengan *tokenizing*. Tahap ini penting untuk mempersiapkan teks agar dapat dikenali oleh model. Teks diubah menjadi unit-unit kecil yang disebut token menggunakan teknik subword berbasis *Sentence Piece*. Pada model bahasa *XLM-RoBERTa*, token khusus seperti [CLS] digunakan di awal input untuk keperluan klasifikasi. Hasil dari proses ini berupa deretan token yang siap diproses lebih lanjut oleh model (Aripin et al., 2023).

3.5 *Build Model*

Model klasifikasi yang digunakan merupakan *XLM-RoBERTa* (*Cross Lingual Language Model*) untuk melakukan tugas klasifikasi teks ke dalam empat kategori emosi, yaitu *joy, sadness, anger, fear*. Model ini dipilih karena kemampuannya dalam memahami berbagai bahasa, termasuk bahasa Indonesia dan bahasa Inggris (Conneau et al., 2020). Dataset yang digunakan akan dibagi dengan rasio 80:20, di mana 80% data digunakan untuk pelatihan, sedangkan 20% digunakan untuk pengujian atau validasi. Rasio ini merupakan salah satu yang paling umum digunakan dalam *machine learning* (Joseph, 2022).

Konfigurasi *hyperparameter* dilakukan untuk meningkatkan performa model, dengan titik awal pada *learning rate* 2e-5 dan jumlah *epoch* 10, yang telah menunjukkan hasil efektif dalam penelitian terkait (Marthen & Novita, 2024). Sebagai bagian dari upaya optimasi, penelitian ini juga akan mengeksplorasi variasi

epoch terdekat, seperti 3 dan 5, untuk memahami pengaruh perubahan kecil pada performa model. Pemilihan jumlah *epoch* ini menjadi aspek penting, karena jika *epoch* terlalu sedikit maka model akan mengalami *underfitting*. Jika *epoch* terlalu banyak, model berisiko mengalami *overfitting* (Afaq & Rao, 2020). Rentang *epoch* dengan variasi nilai rendah, sedang, dan tinggi dievaluasi dalam penelitian ini untuk mengetahui pengaruhnya terhadap performa model.

Teknik *early stopping* diterapkan dalam penelitian ini untuk menyeimbangkan proses pelatihan. Terdapat keterkaitan antara nilai *patience* dan jumlah *epoch* tetapi, hubungan keduanya tidak bersifat linear dan tidak dapat dijadikan acuan baku karena sangat bergantung pada karakteristik model dan data yang digunakan (Hussein & Shareef, 2024). Nilai *patience* dalam penelitian ini ditetapkan dalam rentang proporsional terhadap jumlah *epoch*, yaitu antara 1 hingga 2, untuk menjaga efisiensi pelatihan sekaligus memastikan model tetap adaptif terhadap dinamika proses pembelajaran.

Diterapkan dua skenario pelatihan untuk mengevaluasi pengaruh penggunaan emoji dalam teks pada masing-masing dataset, yaitu: (1) melatih model menggunakan dataset yang mengonversi emoji kedalam teks deskriptif, (2) melatih model menggunakan dataset tanpa emoji. Perbedaan utama antara kedua skenario terletak pada tahap *pre-processing*: pada skenario pertama, emoji dikonversi menjadi deskripsi teks, sedangkan pada skenario kedua, seluruh emoji dihapus melalui proses *noise character removal*. Melalui skenario ini diharapkan dapat dianalisis bagaimana pengaruh keberadaan emoji terhadap performa model dalam klasifikasi emosi.

3.6 *Evaluation*

Tahap terakhir, untuk menilai kualitas performa model dalam melakukan klasifikasi emosi dilakukan proses evaluasi. Matrik yang akan digunakan yaitu *confusion matrix* dan nilai akurasi. *Confusion matrix* digunakan untuk memberikan gambaran yang rinci mengenai jumlah prediksi yang benar dan salah untuk setiap kelas emosi yang ada, sedangkan akurasi sebagai persentase prediksi yang benar dibandingkan dengan total prediksi yang dilakukan oleh model (Maghfiroh et al., 2023). Evaluasi juga mencakup analisis terhadap kelebihan dan kekurangan penelitian untuk meninjau efektivitas pendekatan yang digunakan dalam klasifikasi emosi pada teks bilingual. Proses evaluasi ini akan memberikan gambaran yang lebih baik tentang kemampuan model dalam membedakan berbagai emosi pada teks.