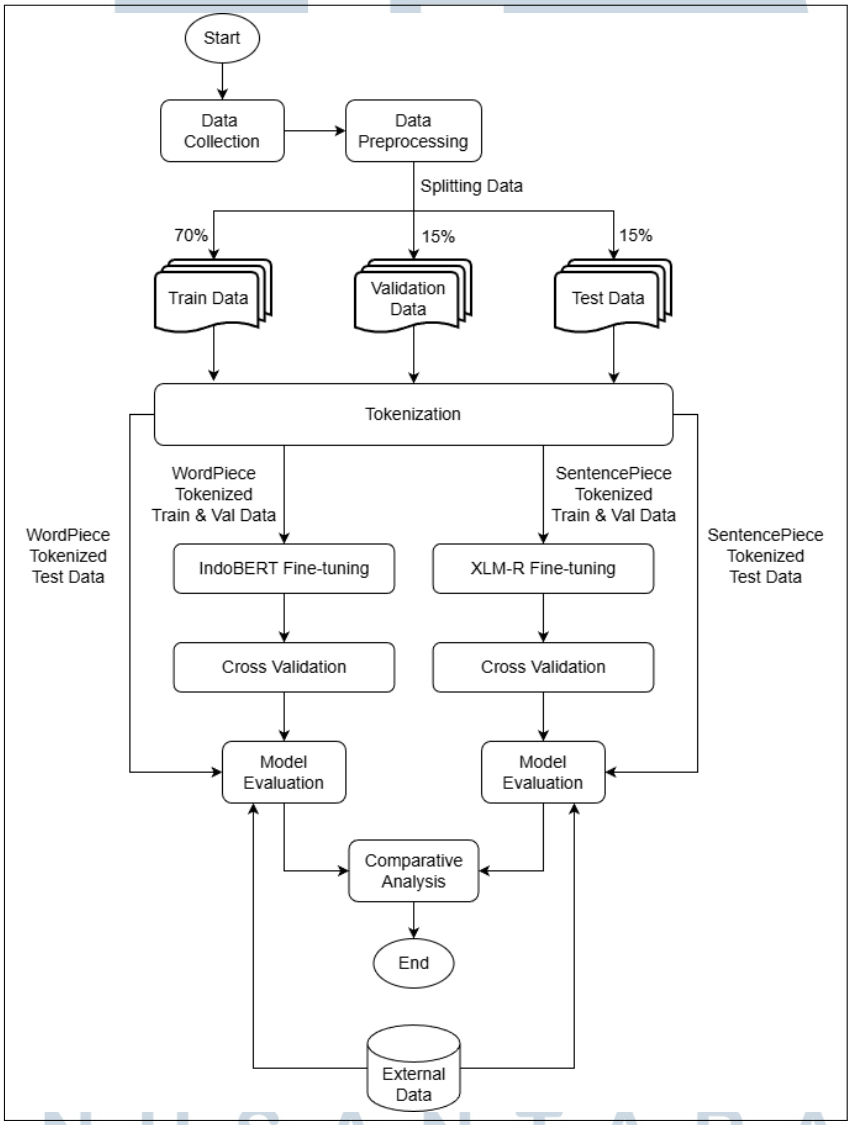
BAB 3
METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk melakukan komparasi kinerja dua model, yaitu *IndoBERT* dan *XLM-RoBERTa*, dalam mengklasifikasikan emosi pada teks berbahasa Indonesia. Bab ini menjelaskan tahapan metodologi penelitian yang disajikan pada Gambar 3.1.



Gambar 3.1. Diagram alur metodologi penelitian

3.1 Spesifikasi dan Environment Eksperimen

Penelitian ini dilaksanakan menggunakan perangkat dengan spesifikasi dan konfigurasi lingkungan sebagaimana ditunjukkan pada Tabel 3.1.

Tabel 3.1. Spesifikasi perangkat dan lingkungan eksperimen

Komponen	**Spesifikasi**
GPU	NVIDIA GeForce RTX 5070 (12 GB VRAM)
CPU	AMD Ryzen 7 9700X
RAM	32 GB DDR5
Sistem Operasi	Windows 11 + WSL 2 (Ubuntu 22.04 LTS)
Virtual *Environment*	Anaconda Environment

Konfigurasi tersebut digunakan secara konsisten selama seluruh eksperimen, baik pada tahap pelatihan model *IndoBERT* maupun *XLM-RoBERTa*.

3.2 Pengumpulan Data

Penelitian ini menggunakan tiga sumber dataset utama yang berasal dari studi terdahulu dan repositori publik, yaitu dataset Riccosan et al. [13], dataset Saputri et al. [12], serta satu dataset eksternal yang diperoleh dari Kaggle [28]. Dua dataset pertama digunakan sebagai *internal dataset* untuk proses pelatihan, validasi, dan pengujian internal, sedangkan dataset Kaggle digunakan secara terpisah sebagai *external test set* untuk mengevaluasi kemampuan generalisasi model.

Dataset Riccosan et al. dikumpulkan melalui Twitter API dengan proses penyaringan berbasis kata kunci emosi berbahasa Indonesia. Proses anotasi dilakukan oleh beberapa annotator independen, kemudian dievaluasi menggunakan metrik *inter-annotator agreement* berupa Cohen's Kappa dan Fleiss's Kappa. Nilai kesepakatan yang dilaporkan masing-masing sebesar 0,5679 dan 0,5657, yang termasuk dalam kategori *moderate agreement* [13]. Hal ini menunjukkan bahwa label emosi yang dihasilkan cukup konsisten antar-annotator dan layak digunakan untuk penelitian lanjutan.

Dataset Saputri et al. dikembangkan melalui pengumpulan tweet berbahasa Indonesia yang mengandung ekspresi emosi eksplisit. Proses pelabelan dilakukan secara manual oleh beberapa annotator yang memiliki pengalaman dalam penandaan data linguistik. Validasi kualitas anotasi dilakukan melalui pemeriksaan

konsistensi label dan diskusi resolusi konflik ketika terjadi perbedaan penilaian [12]. Meskipun tidak secara eksplisit melibatkan pakar psikologi klinis, prosedur anotasi manual dan validasi antar-annotator yang diterapkan telah mengikuti praktik umum dalam penelitian NLP berbasis emosi.

Pelabelan kelas emosi pada kedua dataset internal mengacu pada kerangka teori emosi dasar yang diperkenalkan oleh Ekman [14]. Lima kelas emosi yang digunakan, yaitu *gembira*, *marah*, *sedih*, *takut*, dan *cinta*, merupakan subset emosi dasar yang paling dominan muncul dalam teks media sosial berbahasa Indonesia serta paling konsisten dilaporkan pada penelitian terdahulu. Dengan demikian, pemilihan kelas tidak asal, melainkan berlandaskan teori emosi dan karakteristik data.

Distribusi kelas pada masing-masing dataset internal ditunjukkan pada Tabel 3.2.

Tabel 3.2. Distribusi kelas emosi pada dataset internal

Kelas Emosi	**Riccosan et al.**	**Saputri et al.**	**Total**
Marah	1.130	916	2.046
Takut	911	1.042	1.953
Gembira	1.275	1.127	2.402
Cinta	760	907	1.667
Sedih	1.003	664	2.007
Total	**5.079**	**4.656**	10.075

Kedua dataset internal tersebut kemudian digabungkan sehingga menghasilkan total 10.075 sampel teks. Meskipun distribusi kelas pada dataset tidak sepenuhnya seimbang secara numerik, perbedaan jumlah antar kelas berada dalam rentang yang relatif kecil dengan rasio 1,22 : 1 untuk kelas terbanyak dengan kelas paling sedikit. Oleh karena itu, dataset ini dikategorikan sebagai *relatively balanced dataset* [2]. Untuk mengantisipasi potensi bias akibat perbedaan distribusi kelas, evaluasi model tidak hanya mengandalkan metrik akurasi, tetapi juga menggunakan metrik berbasis *macro-average* serta analisis performa per kelas.

Sebagai tambahan, penelitian ini menggunakan satu dataset eksternal yang diperoleh dari Kaggle [28] dengan total 1.932 sampel teks. Dataset ini digunakan secara eksklusif sebagai data uji generalisasi dan tidak dilibatkan dalam proses pelatihan. Alasan penggunaan dataset Kaggle sebagai *external test set* adalah untuk

mengevaluasi ketahanan (*robustness*) model terhadap data dengan karakteristik berbeda. Berbeda dari dataset internal yang berasal dari teks asli pengguna media sosial, dataset Kaggle merupakan data berbahasa Inggris yang telah diterjemahkan ke bahasa Indonesia menggunakan sistem penerjemahan mesin. Perbedaan gaya bahasa, struktur kalimat, dan morfologi hasil terjemahan ini memberikan skenario uji yang lebih menantang bagi model, sehingga mampu mengungkap kemampuan generalisasi model di luar distribusi data pelatihan.

Distribusi kelas pada dataset eksternal ditunjukkan pada Tabel 3.3.

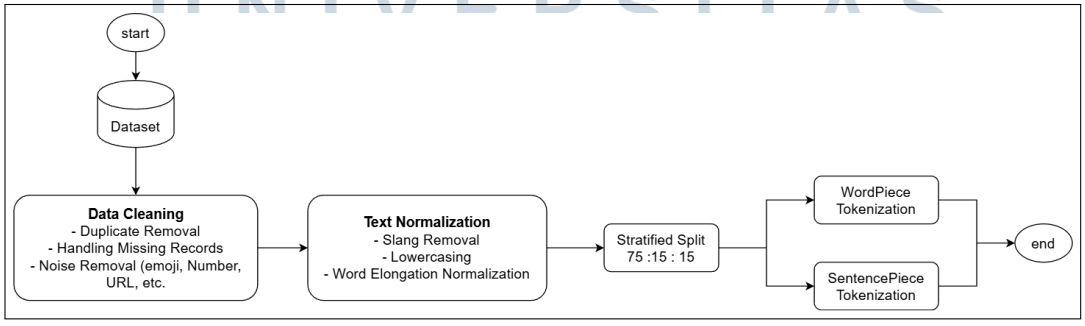
Tabel 3.3. Distribusi kelas emosi pada dataset eksternal (Kaggle)

Kelas Emosi	**Jumlah Sampel**
Cinta	159
Gembira	695
Marah	275
Takut	223
Sedih	580
Total	**1.932**

Penggunaan dataset eksternal ini memungkinkan evaluasi generalisasi model secara lebih objektif dan menghindari kesimpulan yang hanya berlaku pada data dengan distribusi serupa dengan data pelatihan.

3.3 Preprocessing Data

Tahap *preprocessing* dilakukan untuk menyiapkan teks agar sesuai dengan karakteristik model berbasis *Transformer* tanpa menghilangkan informasi semantik yang relevan terhadap ekspresi emosi. Alur preprocessing yang diterapkan pada penelitian ini ditunjukkan pada Gambar 3.2.



Gambar 3.2. Diagram alur preprocessing data

3.3.1 Data Cleaning

Proses pertama dalam *preprcessing* adalah *data cleaning* yang dilakukan untuk menghilangkan komponen yang tidak relevan serta memastikan integritas data. Tahap ini mencakup penghapusan data duplikat, penanganan nilai kosong, serta pembuangan elemen yang tidak memberikan kontribusi semantik seperti emoji, angka, URL, dan simbol non-alfabet. Langkah-langkah tersebut mengikuti praktik umum dalam pemrosesan data media sosial, yang umumnya mengandung variasi teks informal dan derau linguistik yang tinggi [17, 9].

3.3.2 Normalisasi Teks

Normalisasi teks dilakukan untuk mereduksi variasi penulisan yang bersifat ortografis tanpa mengubah atau menghilangkan informasi emosional utama pada teks. Normalisasi mencakup konversi seluruh teks menjadi huruf kecil, penyelarasan kata gaul dan singkatan menggunakan kamus slang, serta penanganan perpanjangan huruf (*word elongation*) seperti *"senenggg"* menjadi *"seneng"*. Langkah ini dilakukan secara selektif dengan tujuan meningkatkan kesesuaian teks masukan terhadap kosakata tokenizer berbasis *subword*, khususnya WordPiece yang digunakan oleh IndoBERT. Variasi penulisan ekstrem dapat menyebabkan token tidak dikenali oleh kosakata tetap (*fixed vocabulary*) dan direpresentasikan sebagai token `[UNK]`, yang berpotensi menghilangkan informasi leksikal penting. Oleh karena itu, normalisasi pada penelitian ini tidak bertujuan menghapus kata bermuatan emosi, melainkan menyelaraskan bentuk penulisan agar tetap dapat dipetakan secara optimal ke dalam representasi token model [10, 9].

3.3.3 Pembagian Data 70:15:15

Dataset dibagi menggunakan proporsi 70:15:15 untuk data pelatihan, validasi, dan pengujian dengan metode *stratified split*. Pendekatan stratifikasi dipilih untuk memastikan distribusi kelas emosi pada setiap subset tetap konsisten dengan distribusi dataset keseluruhan. Alokasi 70% data untuk pelatihan bertujuan menyediakan jumlah sampel yang memadai bagi proses fine-tuning model Transformer yang memiliki jumlah parameter besar. Sementara itu, 15% data validasi digunakan untuk memantau performa model selama pelatihan dan mendeteksi indikasi overfitting tanpa memengaruhi hasil evaluasi akhir. Sisanya, 15% data uji disimpan sepenuhnya terpisah dan hanya digunakan untuk evaluasi

kinerja model setelah proses pelatihan selesai. Proporsi ini sejalan dengan praktik umum dalam penelitian NLP berbasis Transformer dan telah digunakan secara efektif pada studi deteksi emosi berbahasa Indonesia sebelumnya [9, 2].

3.3.4 Tokenisasi: WordPiece dan SentencePiece

Tokenisasi dilakukan setelah proses pembagian data untuk mencegah kebocoran informasi antar subset. Dua teknik tokenisasi digunakan sesuai dengan model pralatih, yaitu WordPiece untuk IndoBERT dan SentencePiece untuk XLM-R. Pemilihan tokenisasi ini mengikuti skema tokenisasi bawaan masing-masing model dan bertujuan menjaga konsistensi representasi input dengan parameter hasil pra-pelatihan.

3.4 Modelling

Tahap *modelling* bertujuan untuk mendefinisikan arsitektur model yang digunakan dalam penelitian serta memastikan kesetaraan struktur antar model yang dibandingkan. Pada penelitian ini, dua model bahasa pralatih, yaitu *IndoBERT* dan *XLM-RoBERTa*, digunakan sebagai *backbone* untuk tugas klasifikasi emosi teks berbahasa Indonesia. IndoBERT dan XLM-R digunakan dalam bentuk model pralatih tanpa modifikasi pada lapisan *encoder*. Seluruh lapisan *encoder* beserta bobot pra-latihnya dipertahankan, sehingga proses pelatihan selanjutnya berfokus pada penyesuaian representasi terhadap tugas klasifikasi emosi.

Untuk keperluan klasifikasi, kedua model dilengkapi dengan *classification head* yang identik, berupa satu lapisan *dense* dengan jumlah neuron sesuai jumlah kelas emosi. Lapisan ini menerima representasi vektor dari token khusus `[CLS]` dan menghasilkan skor *logits* untuk setiap kelas. Selanjutnya, fungsi aktivasi *softmax* digunakan untuk mengonversi *logits* menjadi probabilitas kelas. Penyamaan struktur *classification head* dilakukan untuk memastikan bahwa perbandingan kinerja tidak dipengaruhi oleh perbedaan arsitektur, melainkan mencerminkan perbedaan kemampuan representasi model pralatih.

3.5 Cross Validation

Dalam penelitian ini digunakan skema *5-fold cross validation* dengan pendekatan *stratified split*. Pemilihan 5-fold cross validation didasarkan pada pertimbangan keseimbangan antara stabilitas estimasi performa dan efisiensi

komputasi. Penelitian terdahulu menyebutkan bahwa 5-fold merupakan konfigurasi yang umum digunakan pada model dengan kompleksitas tinggi seperti Transformer karena mampu memberikan estimasi performa yang stabil tanpa meningkatkan biaya komputasi secara signifikan [29]. Selain itu *cross validation* dilakukan hanya pada data pelatihan dan validasi, sementara data uji internal serta dataset eksternal tidak dilibatkan dalam proses ini. Dengan demikian, hasil evaluasi akhir sepenuhnya merefleksikan kemampuan generalisasi model terhadap data yang tidak pernah dilihat selama proses pelatihan.

Hasil *cross validation* tidak digunakan sebagai hasil akhir penelitian, melainkan sebagai dasar untuk memastikan bahwa konfigurasi pelatihan dan proses *fine-tuning* menghasilkan performa yang konsisten dan tidak sensitif terhadap variasi pembagian data. Evaluasi akhir dan perbandingan kinerja antar model tetap dilakukan menggunakan data uji internal dan dataset eksternal yang sepenuhnya terpisah dari proses pelatihan.

3.6 Fine-Tuning Model

Fine-tuning merupakan proses penyesuaian bobot model pralatih terhadap tugas spesifik dengan menggunakan dataset berlabel. Berbeda dengan tahap pemodelan yang bersifat struktural, fine-tuning berfokus pada proses pembelajaran parameter agar representasi yang telah dipelajari selama pretraining dapat diadaptasi secara optimal untuk tugas klasifikasi emosi. Pada penelitian ini, fine-tuning dilakukan dengan memperbarui seluruh parameter model (*full fine-tuning*), baik pada lapisan *encoder* maupun *classification head*. Pendekatan ini dipilih karena telah terbukti memberikan performa yang lebih baik dibandingkan pembekuan sebagian lapisan pada tugas klasifikasi teks berbasis Transformer [5, 24].

3.6.1 Konfigurasi Hyperparameter

Untuk memastikan perbandingan yang adil, konfigurasi hyperparameter disamakan pada kedua model. Pemilihan hyperparameter didasarkan pada rekomendasi literatur fine-tuning Transformer serta hasil penelitian deteksi emosi berbahasa Indonesia sebelumnya.

- **Jumlah Epoch: 3**

  Jumlah epoch ditetapkan sebanyak 3 berdasarkan temuan bahwa model

berbasis BERT umumnya mencapai konvergensi optimal dalam rentang 2 hingga 4 epoch [5, 24]. Penggunaan epoch yang relatif kecil bertujuan untuk menghindari overfitting, khususnya pada dataset berukuran menengah. Konfigurasi serupa juga digunakan pada penelitian emosi berbahasa Indonesia oleh Setiawan et al. dan Nugroho et al. [9, 10].

- **Batch Size: 32**
  Batch size 32 dipilih karena memberikan keseimbangan antara stabilitas gradien dan efisiensi komputasi pada fine-tuning Transformer [5]. Studi Setiawan et al. dan Nugroho et al. juga melaporkan penggunaan batch size 32 memiliki performa yang baik untuk pemodelan emosi berbasis Transformer [9, 10].

- **Learning Rate:** 2×10^{-5}
  Learning rate ditetapkan sebesar 2×10^{-5}, yang termasuk dalam rentang learning rate yang direkomendasikan untuk fine-tuning model berbasis BERT dan RoBERTa dengan kisaran 5×10^{-5} hingga 2×10^{-5}. Nilai ini dikenal mampu memberikan keseimbangan yang baik antara kecepatan konvergensi dan stabilitas pelatihan [5, 30]. Konfigurasi learning rate yang sama juga digunakan pada penelitian deteksi emosi teks menggunakan model berbasis *Transformers* [9, 10].

3.6.2 Analisis Dinamika Fine-Tuning Berdasarkan Jumlah Epoch

Selain pelatihan utama dengan 3 epoch, penelitian ini juga mengevaluasi perilaku model selama fine-tuning pada rentang epoch 1 hingga 10. Analisis ini bertujuan untuk mengamati dinamika pembelajaran model, termasuk pola konvergensi, stabilitas performa, serta indikasi overfitting. Evaluasi dilakukan dengan memantau perubahan nilai loss dan metrik evaluasi pada data validasi dan melihat bagaimana hasil performanya pada data uji di setiap epochnya. Pendekatan ini memungkinkan identifikasi titik optimal pelatihan, yaitu ketika peningkatan performa mulai melambat atau terjadi degradasi pada data validasi. Temuan ini didasarkan dengan laporan Iskoko et al. [31], yang menyatakan bahwa fine-tuning IndoBERT dengan jumlah epoch berlebih tidak selalu menghasilkan peningkatan performa pada data baru.

3.7 Evaluasi Model

Evaluasi model dilakukan untuk menilai kemampuan klasifikasi emosi dari model yang telah melalui proses fine-tuning, baik pada data dengan distribusi serupa dengan data pelatihan maupun pada data baru dengan karakteristik berbeda. Evaluasi dirancang secara berlapis agar hasil yang diperoleh tidak hanya merepresentasikan performa numerik, tetapi juga mencerminkan kemampuan generalisasi serta stabilitas model. Secara umum, evaluasi dalam penelitian ini terdiri atas tiga tahap utama, yaitu evaluasi pada data uji internal, uji generalisasi menggunakan dataset eksternal, serta analisis performa per kelas emosi. Pendekatan ini dipilih untuk menghindari kesimpulan yang bias apabila evaluasi hanya didasarkan pada satu metrik agregat seperti akurasi.

3.7.1 Evaluasi pada Data Uji Internal

Evaluasi utama dilakukan menggunakan data uji internal yang diperoleh dari pembagian dataset dengan skema stratified split. Data uji ini tidak terlibat dalam proses pelatihan, validasi, maupun cross validation, sehingga hasil evaluasi mencerminkan kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya namun masih berasal dari distribusi yang sama dengan data pelatihan.

Kinerja model diukur menggunakan beberapa metrik evaluasi standar pada tugas klasifikasi multikelas, yaitu akurasi, presisi, recall, dan F1-score. Selain akurasi, penelitian ini secara khusus menekankan penggunaan metrik berbasis *macro-average*. Pendekatan ini dipilih untuk memastikan bahwa setiap kelas emosi memberikan kontribusi yang setara terhadap penilaian performa, mengingat distribusi kelas yang meskipun relatif seimbang tetap menunjukkan variasi jumlah sampel antar kelas.

3.7.2 Analisis Performa per Kelas Emosi

Selain evaluasi agregat, penelitian ini juga melakukan analisis performa pada tingkat kelas emosi secara individual. Analisis per kelas dilakukan dengan mengamati nilai presisi, recall, dan F1-score untuk masing-masing kategori emosi, serta melalui visualisasi *confusion matrix*. Pendekatan ini digunakan untuk menjawab kebutuhan harmonisasi performa antar kelas, bukan untuk menilai model secara parsial. Analisis per kelas memungkinkan identifikasi pola kesalahan

klasifikasi yang tidak terlihat apabila evaluasi hanya berfokus pada metrik agregat. Misalnya, model dapat memiliki nilai akurasi yang tinggi, tetapi tetap gagal mengenali emosi tertentu dengan baik akibat tumpang tindih semantik antar kelas.

Dengan demikian, evaluasi per kelas tidak dimaksudkan sebagai pengganti evaluasi agregat, melainkan sebagai pelengkap yang memberikan konteks interpretatif terhadap nilai akurasi dan F1-score secara keseluruhan. Pendekatan ini sejalan dengan praktik evaluasi pada penelitian klasifikasi emosi dan sentimen, di mana variasi performa antar kelas sering kali menjadi indikator penting dalam menilai kualitas model [27].

3.7.3 Uji Generalisasi Model pada Dataset Eksternal

Untuk menilai kemampuan generalisasi model secara lebih ketat, penelitian ini melakukan evaluasi tambahan menggunakan dataset eksternal yang sepenuhnya terpisah dari data internal. Dataset ini tidak digunakan pada tahap pelatihan, validasi, cross validation, maupun pemilihan hyperparameter. Uji generalisasi bertujuan untuk mengevaluasi apakah model mengalami overfitting terhadap distribusi data pelatihan. Apabila performa model menurun secara signifikan pada dataset eksternal, hal tersebut mengindikasikan bahwa model terlalu menyesuaikan diri dengan karakteristik data pelatihan. Sebaliknya, performa yang relatif konsisten antara data uji internal dan data eksternal menunjukkan bahwa model memiliki kemampuan generalisasi yang baik.