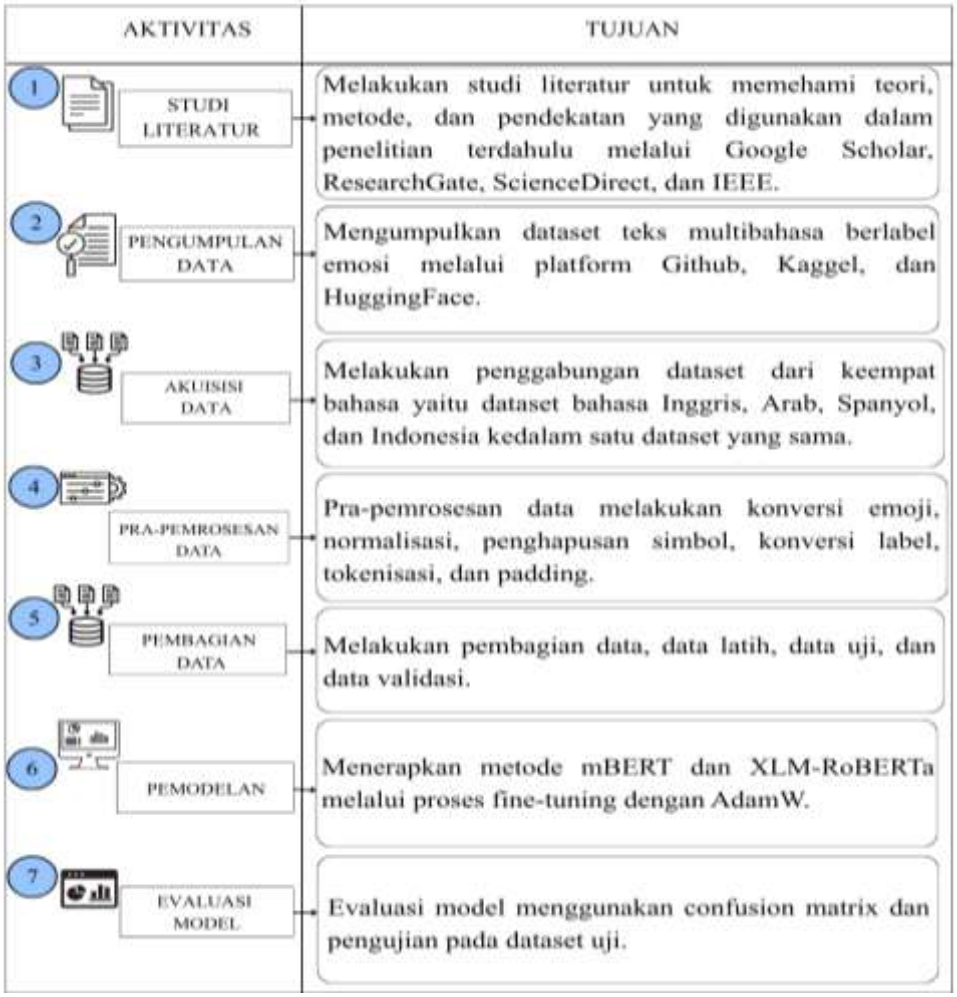
BAB III

METODOLOGI

3.1 Tahapan Penelitian

Penelitian ini menggunakan model klasifikasi *deep learning*, khususnya mBERT dan XLM-RoBERTa untuk klasifikasi emosi dalam teks *multilingual*. Penelitian ini dilakukan secara menyeluruh mulai dari pengumpulan data hingga evaluasi model, dengan tujuan mengoptimalkan akurasi di berbagai konteks teks dapat dilihat pada Gambar 3.1.



Gambar 3. 1 Tahapan Penelitian

3.1.1 Studi Literatur

Studi literatur pada penelitian ini, dilakukan tinjauan pustaka untuk memahami perkembangan, pendekatan, dan tantangan dalam klasifikasi emosi berbasis teks, khususnya dalam konteks multibahasa dan *cross-lingual*. Tujuan dari tahap ini adalah untuk mengidentifikasi metode yang digunakan dalam penelitian sebelumnya, jenis data dan model yang diterapkan, serta potensi kesenjangan penelitian. Pencarian literatur dilakukan di berbagai sumber akademis seperti *Google Scholar, Research Gate, ScienceDirect*, IEEE menggunakan kata kunci seperti *emotion classification, multilingual emotion classification, cross-lingual emotion detection, transformer-based models,* mBERT, dan XLM-RoBERTa. Literatur ditinjau meliputi artikel jurnal, makalah konferensi, serta disertasi dan tesis magister yang relevan dengan topik penelitian.

3.1.2 Pengumpulan Data

Data dikumpulkan dari empat sumber dataset publik dengan karakteristik bahasa yang berbeda, yaitu bahasa Indonesia, dataset SemEval Task 1 untuk bahasa Spanyol, dan dataset bahasa Inggris di dapatkan dari dataset SemEval Task 1 tetapi versi github, lalu bahasa Arab mengambil dari dataset publik di github. Penggunaan empat bahasa ini bertujuan untuk mendukung analisis klasifikasi emosi *multilingual*. Dataset pertama adalah dataset bahasa Indonesia yang bersumber dari repositori Github (https://github.com/ines179/Emotion-Detection_IndoBERT-CNN-BiLSTM), dengan total 14.048 data. Dataset bahasa Indonesia memiliki enam label emosi yang terdiri dari kegembiraan, terkejut, kesedihan, ketakutan, kemarahan, dan

rasa jijik. Dataset kedua bahasa Arab menggunakan dataset publik (Amr Alkhatib, 2018), dengan total 10.065 data, dataset ini memiliki 8 label emosi di antaranya none, anger, joy, sadness, love, fear, sympathy, dan surprise. Dataset Ketiga dataset Spanyol di dapatkan dari dataset SemEval Task 1 (Felipebravom, 2026) dengan total 7.094 data, memiliki 11 label emosi diantaranya *Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Trust*. Dataset keempat dataset bahasa Inggris dari SemEval Task 1 versi github (Hridoy, 2021) dengan total 20.000 data, dataset ini memiliki 6 label emosi yaitu *anger, love, surprise, fear, joy, sadness*. Dari keempat dataset yang berbeda tersebut, dilakukan penggabungan dataset. Sampel dari keempat dataset tersebut bisa dilihat dari tabel 3.1. tabel 3.2, tabel 3.3, dan tabel 3.4.

Tabel 3. 1 Sampel Dataset Bahasa Indonesia

No	***Message***	**Emosi**
1	Sinyal 3 sering banget mati pas lagi nelpon, panik banget nggak bisa lanjut ngobrol sama temen	1.0
2	Jaringan 3 sering gangguan pas lagi telepon penting, bikin panik takut obrolan putus dan nggak bisa lanjut, komunikasi jadi terganggu	1.0
3	Kok bisa begini XL? Kuota sudah terisi tapi masih belum sampai. Sungguh menakutkan, bagaimana bisa seperti ini?	1.0
4	Saya mencoba masuk ke akun layanan, tetapi kok tidak berhasil? Saya sudah coba reset password, tapi tetap tidak bisa login. Jadi saya khawatir banget, takutnya akun saya dihack. @Telkomsel mohon bantuannya, saya tidak bisa mengakses layanannya. 😨 #Tidak BisaLogin #TolongBantuan	1.0

5	Indosat sering putus tiba-tiba, jadi saya khawatir jika ada urusan penting dengan atasan	1.0

Tabel 3.1 menunjukkan sampel dataset bahasa Indonesia yang merepresentasikan keluhan pengguna layanan telekomunikasi. Memiliki label emosi yang beragam.

Tabel 3. 2 Sampel Dataset SemEval Bahasa Arab

No	**ID**	**Teks**	***Label***
1	1	.. الاوليمبياد الجايه هكون لسه ف الكليه	None
2	2	عجز الموازنه وصل ل93.7 % من الناتج المحلي يعني لسه اقل من 7 و نفلس و بهايم لسه يتابعوا الاوليمبياد %	Anger
3	3	xD كتنا نيله ف حظنا الهباب	Sadness
4	4	...جميعنا نريد تحقيق اهدافنا لكن تونس تالقت في حراسه المرمي	Joy
5	5	الاوليمبياد نظامها مختلف .. ومواعيد المونديال مكانتش مقرفه ولا حاجه كانت معقوله	None

Tabel 3.2 menunjukkan sampel data teks dan label emosi dari dataset SemEval Task 1 bahasa Arab yang terdiri dari kolom ID, Teks, dan label emosi.

Tabel 3. 3 Sampel Dataset SemEval Bahasa Inggris

No	**Teks**	***Label***
1	i didnt feel humiliated	Sadness
2	i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake	Sadness
3	im grabbing a minute to post i feel greedy wrong	Anger
4	i am ever feeling nostalgic about the fireplace i will know that it is still on the property	Love
5	i am feeling grouchy	Anger

Tabel 3.3 menampilkan sampel data teks dan label emosi dari dataset SemEval Task 1 bahasa Inggris versi github yang berisi teks dan label emosi yang beragam. Teks dalam dataset ini umumnya berupa kalimat motivasi, opini, atau percakapan media sosial yang mencerminkan berbagai ekspresi emosi.

Tabel 3. 4 Sampel Dataset SemEval Bahasa Spanyol

No	**Teks**	***Label***
1	@aliciaenp Ajajjaa somos del clan twitteras perdidas pa eventos "importantes"	Joy
2	@AwadaNai la mala suerte del gato fichame la cara de help me pls	Sadness
3	@audiomano A mí tampoco me agrado mucho eso. Especialmente por tratarse de él. No hay justificación.	Anger
4	Para llevar a los bebes de un lugar a otro debemos cantarles canciones... Quiero cantarles Gunaa' nibiina (La llorona, en Zapoteco)	Joy
5	@DalasReview me encanta la terrible hipocresia y doble moral que tiene esta gente, claro, cuando ella te lo quita ILEGALMENTE no importa...	Anger

Tabel 3.4 menunjukkan contoh sampel data teks dan label emosi dari dataset SemEval Task 1 bahasa Spanyol. Teks yang digunakan berasal dari media sosial dengan gaya bahasa informal, sehingga mencerminkan ekspresi emosi yang lebih natural dan kontekstual.

3.1.3 Akuisisi Data

Dataset SemEval Task 1 memiliki format dataset berbentuk teks (.txt) dan Setiap bahasa dipisahkan menjadi tiga subset file terpisah, yaitu train.txt, dev.txt, dan test.txt, yang berfungsi untuk pembagian data pelatihan, validasi, dan pengujian. Selain itu, peneliti ini juga mengintegrasikan dataset bahasa Indonesia yang diambil dari repositori GitHub public terkait klasifikasi emosi multibahasa. Dataset bahasa Inggris SemEval-2018 Task 1 versi Github memiliki satu dataset train. Dataset Arab dari Github memiliki satu dataset train.

Tabel 3. 5 Akuisisi Dataset

Dataset	***Train***	***Validation***	***Test***
SemEval (Spanyol)	3.561	679	2.854
Bahasa Arab	10.065	-	-
SemEval (Inggris) versi Github	20.000	-	-
Bahasa Indonesia	14.048	-	-
Total	**47.674**	**679**	**2.854**

Tabel 3.5 menyajikan statistik jumlah data pada setiap subset pelatihan, validasi, dan pengujian dari setiap sumber dataset berbagai bahasa Spanyol, Arab, Inggris, dan Indonesia.

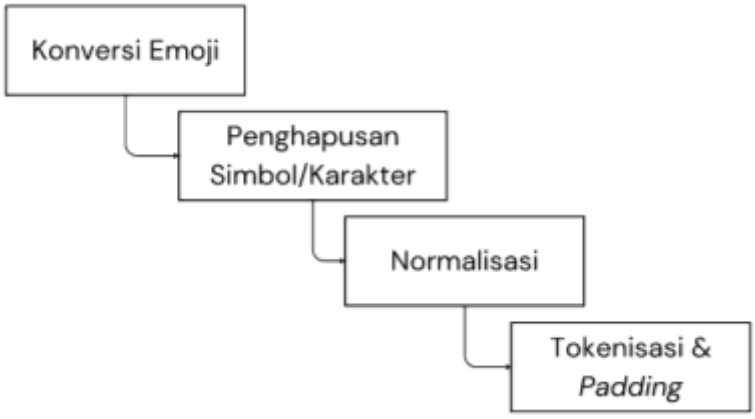
3.1.4 Pemetaan Label

Dataset yang digunakan pada penelitian ini memiliki perbedaan taksonomi label emosi antara SemEval Task 1 dan dataset bahasa Indonesia. Dataset SemEval menggunakan 11 kategori label emosi yang lebih spesifik, meliputi *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, dan trust*. Sementara itu, dataset Indonesia menggunakan 6

kategori label berdasarkan basic emotion yang terdiri dari *joy, surprise, sadness, fear, anger, dan disgust.* Dataset Inggris memiliki 6 kategori label, dan dataset bahasa Arab memiliki 8 label emosi. Proses pemetaan ini bertujuan untuk menyederhanakan label emosi pada dataset SemEval Task 1, Inggris, dan Arab menjadi 6 label *joy, surprise, sadness, fear, anger, dan disgust.*

3.1.5 Pra-pemrosesan Data

Tahap ini data yang sudah dikumpulkan akan diproses untuk diambil informasi yang terkandung di dalamnya karena masih banyak data yang kotor dan adanya simbol-simbol yang tidak diperlukan. Pra-pemrosesan data dilakukan dengan cara menghapus data yang tidak perlu atau tidak sesuai agar membuat data lebih mudah untuk diproses, sehingga menghasilkan data yang sesuai dengan kebutuhan penelitian. Pada penelitian ini, metode pemrosesan yang dilakukan adalah konversi emoji, penghapusan simbol-simbol atau karakter, normalisasi, tokenisasi dan padding. Proses ini penting dilakukan untuk menghasilkan data yang sama dan optimal. Tahapan pra-pemrosesan data yang dilakukan pada penelitian ini dapat dilihat pada Gambar 3.2.



Gambar 3. 2 Tahapan Pra-pemrosesan

Gambar 3.2 menunjukkan pra-pemrosesan data dilakukan untuk memastikan data berada pada format yang terstruktur, mengurangi potensi kesalahan pemrosesan, dan memaksimalkan kemampuan model dalam memahami konteks semantik pada teks multibahasa sebelum pada tahap pelatihan dan evaluasi model (Faishal Basbeth, 2024). Proses ini sangat penting dalam analisis data dan pembelajaran mesin karena membantu dalam menghasilkan data yang berkualitas dan dapat diandalkan untuk keperluan analisis lebih lanjut (Ishak, 2023).

3.1.6 Konversi Emoji

Konversi emoji dilakukan untuk mentransformasikan simbol menjadi teks agar makna emosional yang terkandung di dalamnya dapat diinterpretasikan oleh model. Proses konversi dilakukan secara otomatis menggunakan *library* emoji pada Python, yang mencocokkan setiap karakter emoji dengan deskripsi standarnya dalam bahasa Inggris. Tahap ini digunakan karena daripada menghapus emoji, karena emoji sering kali menjadi indikator yang kuat dalam menentukan label emosi pada teks multibahasa ini. Contoh penerapan dari konversi emoji ke teks misalnya emoji 😢 di konversikan menjadi *anguished_face*.

3.1.7 Penghapusan Simbol atau Karakter

Penghapusan simbol atau karakter dilakukan untuk membersihkan dataset dari elemen-elemen *noise* yang tidak memiliki kontribusi semantik bagi model, seperti tautan URL, mention, dan tagar, serta karakter angka. Proses ini diimplementasikan menggunakan teknik *Regular Expression* (Regex) yang

menghapus berbagai tanda baca namun tetap mempertahankan tanda seru (!), tanda tanya (?), dan ellipsis (…) karena dianggap masih membawa nilai ekspresi emosional yang penting dalam teks.

3.1.8 Normalisasi

Normalisasi teks bertujuan untuk menyelaraskan variasi penulisan teks agar menjadi format yang standar, sehingga mengurangi kompleksitas pada saat ekstraksi fitur. Proses ini meliputi perbaikan kata tidak baku (slang) atau singkatan menjadi kata baku sesuai pedoman bahasa masing-masing, serta pembersihan karakter berulang yang tidak perlu. Normalisasi sangat penting dalam penelitian *cross-lingual* ini untuk memastikan bahwa kata yang memiliki makna sama namun ditulis dengan gaya yang berbeda. Contoh penerapan normalisasi teks pengubahan kata non formal ke formal misalnya "Gak" menjadi "Tidak".

3.1.9 Tokenisasi dan *Padding*

Tahap tokenisasi dilakukan setelah teks melalui tahap pembersihan data. Tahap tokenisasi dan *padding* dilakukan menggunakan *tokenizer* bawaan dari model mBERT dan XLM-RoBERTa. Tokenisasi berfungsi untuk mengubah teks menjadi deretan token, sedangkan *padding* digunakan untuk menyamakan panjang setiap urutan token agar dapat diproses dalam bentuk batch. Proses ini dilakukan secara otomatis oleh *tokenizer* masing-masing model.

Metode ini memecah kata-kata yang jarang muncul menjadi unit-unit yang lebih kecil (subkata), sehingga model dapat menangani masalah *Out-of-*

Vocabulary (OOV) dan memahami struktur bahasa yang kompleks secara lebih efektif. Selain memecah teks, proses ini juga menghasilkan tiga komponen utama yang dibutuhkan model:

1. *Input* IDs

   Representasi numerik unik untuk setiap token berdasarkan kamus (*vocabulary*) model.

2. *Attention Mask*

   Serangkaian angka biner (0 dan 1) yang memberitahu model bagian mana yang merupakan teks asli dan bagian mana yang merupakan *padding* (pengisi kosong). Terlampir pada persamaan (2.3).

3. *Special Tokens*

   Penambahan token khusus seperti <s> dan </s> (pada XLM-RoBERTa) atau [CLS] dan [SEP] (pada mBERT) untuk menandai awal dan akhir kalimat.

3.1.10 *Data Balancing*

Proses menyeimbangkan data untuk mengatasi masalah ketidakseimbangan jumlah sampel pada setiap kelas emosi. Setelah penggabungan dataset multibahasa, distribusi label menunjukkan perbedaan yang cukup signifikan. Penelitian ini menerapkan metode *class weight loss* pada fungsi *cross-entropy loss*, yaitu bobot yang berbeda pada setiap kelas berdasarkan frekuensi kemunculannya. Kelas dengan jumlah data lebih sedikit akan diberikan bobot yang lebih besar, sehingga kesalahan prediksi pada kelas tersebut memiliki batasan yang lebih tinggi selama proses pelatihan model.

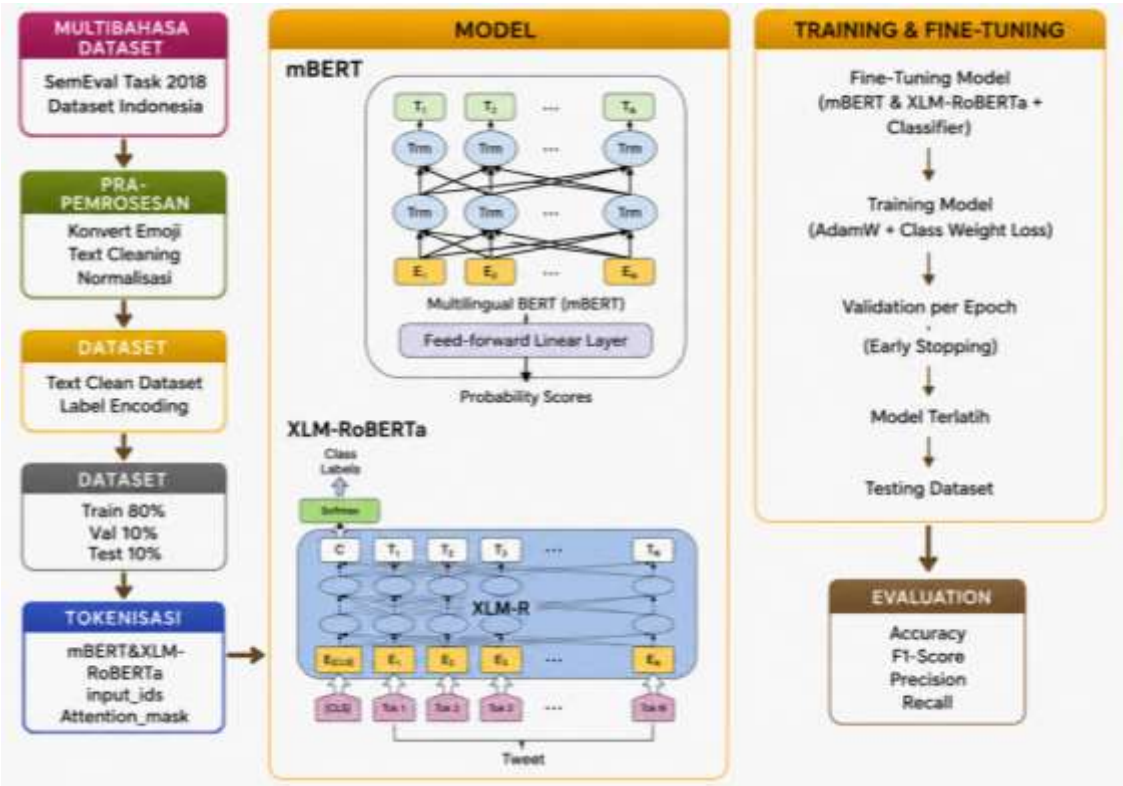
3.1.11 Pembagian Data

Penelitian ini membagi dataset menjadi tiga bagian yaitu data latih, validasi, dan uji. Pembagian dataset menggunakan rasio 80:10:10 untuk data latih 80%, 10% untuk data validasi, dan 10% untuk data uji (Rizky & Hidayat, 2025).

3.1.12 *Label Encoding*

Label Encoding adalah metode yang mengubah data kategorikal yang direpresentasikan sebagai label teks menjadi format numerik (Herdian dkk., 2024). Tahapan ini bertujuan untuk label agar dikenali oleh model saat proses pelatihan. Misalnya setiap label emosi diubah menjadi nilai numerik *anger*: 0, *disgust*: 1.

3.1.13 Implementasi Model



Gambar 3. 3 Tahapan Pemodelan

Gambar 3.3 menunjukkan tahapan pemodelan yang dilakukan penelitian ini menggunakan pendekatan *Deep Learning* yaitu *Bidirectional Encoder Representations from Transformer* (BERT). Model yang digunakan dalam klasifikasi emosi multibahasa adalah *Multilingual* BERT (mBERT) dan XLM-RoBERTa. Kedua model tersebut merupakan model berbasis *transformer* yang dilatih pada korpus multibahasa berskala besar, sehingga memiliki kemampuan dalam memahami struktur linguistik yang beragam pada berbagai bahasa (Hao & Liu, 2025). Dalam penelitian ini, kedua model diadaptasi untuk tugas klasifikasi emosi teks multibahasa melalui proses *fine-tuning.* Proses ini bertujuan untuk menyesuaikan parameter model dengan karakteristik linguistik dan distribusi data pada dataset yang digunakan yaitu bahasa Inggris, bahasa Arab, bahasa Spanyol, dan bahasa Indonesia. Model diharapkan mampu menangkap representasi semantik dan konteks emosional secara lebih optimal pada masing-masing bahasa. Skema pembagian data yang digunakan dalam penelitian ini adalah 80% untuk data pelatihan, 10% data validasi, dan 10% untuk data pengujian (Rizky & Hidayat, 2025).

Multilingual BERT (mBERT) yaitu versi terbaru dari BERT yang dilatih pada 104 bahasa berbeda (Manias dkk., 2023). Model mBERT telah dilatih dan disesuaikan menggunakan kumpulan data pelatihan dari dataset (Horbach dkk., 2024). *Multilingual* yang menjadi ciri model ini memudahkan pemahaman semantik hubungan antara kata-kata dalam semua 104 bahasa tersebut. Selain itu, versi huruf besar maupun huruf kecil dari mBERT digunakan dalam pendekatan yang diusulkan ini, karena versi huruf besar

mBERT berguna dalam bahasa yang dimana aksen menjadi peran penting. Model mBERT berbeda dengan yang lainnya, model mBERT dilatih dalam berbagai bahasa dan menggunakan *Masked Language Modeling* (MLM) (Manias dkk., 2023).

Cross-Lingual Language Model - RoBERTa (XLM-RoBERTa) model bahasa yang mempunyai multibahasa dilatih pada 2,5 TB data *CommonCrawl* yang baru dibuat dan bersih dalam 100 bahasa (Manias dkk., 2023). Pada penelitian ini model XLM-RoBERTa yang telah di *pre-trained* pada data multibahasa berukuran besar. *Fine-tuning* teknik adaptasi model dengan memanfaatkan bobot awal dari model pra-latih, kemudian menyesuaikan untuk menyelesaikan tugas yang spesifik.

Tabel 3. 6 Perbandingan Model dengan Peneliti Terdahulu

Penelitian	**Dataset**	**Model**	***Cross-lingual***	**Jumlah Label Emosi**
(Hassan dkk., 2022)	SemEval-2018	mBERT, AraBERT, BetoBERT	Ya	11
(Fatima et al. 2025)	SemEval-2025 Task 11	Afro-XLM-R	Tidak	6
(Elfaik & Nfaoui, 2021)	SemEval-2018 Task 1 Ec-Ar	AraBERT word embeddings,attention-based LSTM and BiLSTM	Tidak	
(Ahanin dkk., 2023)	GoEmotions dan SemEval-2018	Bi-LSTM, BERT	Tidak	11
(Anita Fatmawati dkk., 2025)	Inggris	SVM	Tidak	8
(Ameer, Bölücü, Siddiqui, dkk., 2023)	SemEval-2018 Task 1 dan Ren-CEps	Bi-LSTM, RoBERTa, XLNet	Tidak	11 dan 8

Penelitian	Dataset	Model	*Cross-lingual*	Jumlah Label Emosi
(Deborah S dkk., 2018)	SemEval-2018 Task 1	Multilayer Perception (MLP)	Tidak	11
(Mansy dkk., 2022)	SemEval-2018 Task 1	MARBERT, BiLSTM, Bi-GRU	Tidak	11
Penelitian ini	**SemEval-2018 Task 1**	**mBERT, XLM-R**	**Yes**	**6**

Penelitian ini diterapkan parameter yang sama untuk kedua model yaitu batch size 16, learning rate 2e-5, early stopping 3, dan epoch 10. Penelitian ini juga mengatasi dua penyelesaian, yaitu untuk mengatasi kompleksitas dataset *multilingual* yang memiliki variasi pola semantik antarbahasa, digunakan model mBERT dan XLM-RoBERTa. Dilakukan perbandingan performa model sebelum dan sesudah proses *fine-tuning* dengan menyesuaikan seluruh parameter model menggunakan dataset *Multilingual* penggabungan dataset SemEval-2018 Task 1 untuk bahasa Spanyol, dataset bahasa Inggris, dataset bahasa Arab, dan dataset bahasa Indonesia, sehingga model dapat dilakukan pra-latihannya terhadap karakteristik emosi dan ekspresi linguistik pada berbagai bahasa. Model mBERT dan XLM-RoBERTa menjadi solusi dengan menggunakan kosakata bersama dan dilatih sebelumnya pada dataset *multilingual* yang luas. Jadi sangat baik untuk mengidentifikasi toksisitas di berbagai bahasa (A.Akshaya dkk., 2025). Model tersebut dipilih karena kemampuannya yang unggul dalam menangani representasi *cross-lingual* serta performanya yang lebih stabil dibandingkan model *multilingual* lainnya pada berbagai tugas pemrosesan bahasa alami (Magistra Apta Paramarta dkk., 2025). Penelitian ini menerapkan teknik *early stopping,* yaitu mekanisme

penghentian pelatihan secara otomatis ketika nilai *validation loss* tidak mengalami penurunan dalam sejumlah epoch tertentu.

3.1.14 Evaluasi Model

Evaluasi model bertujuan untuk menilai kelayakan dan kinerja model klasifikasi emosi yang telah dilatih dalam mengenali dan memprediksi label emosi pada teks multibahasa. Tahap evaluasi penelitian ini dilakukan dengan mengukur akurasi, *precision, recall, f1-score* dan matriks evaluasi lainnya pada data uji, serta menganalisis hasil prediksi menggunakan *confusion matrix. Confusion matrix* digunakan untuk menggambarkan hubungan antara label emosi yang diprediksi oleh model dan label emosi aktual pada data uji, sehingga dapat diketahui jumlah prediksi yang benar maupun kesalahan klasifikasi pada setiap kelas emosi. Dengan *confusion matrix*, performa model dapat dianalisis secara lebih rinci, termasuk kemampuan model dalam membedakan antar kelas emosi yang memiliki karakteristik sama.