BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Text Mining

Text mining adalah proses untuk mengambil pola atau informasi penting dari dokumen teks. Metode ini melibatkan mengubah dokumen teks menjadi bentuk lain agar dapat mengidentifikasi pola atau pengetahuan yang terkandung di dalamnya (Hwang et al., 2025). Meskipun sering disamakan dengan *data mining*, *text mining* memiliki fokus yang berbeda. Salah satu perbedaan utama antara *data mining* dengan *text mining* yaitu jenis data yang diproses. *Data mining* memproses data terstruktur dari sistem seperti *database*, ERP (*Enterprise Resource Planning*) dan CRM (*Customer Relationship Management*). Sementara itu, *text mining* memproses data tidak terstruktur yang ditemukan dalam dokumen, *email*, media sosial, dan web (Hassani et al., 2020).

Proses *text mining* dimulai dengan menerapkan teknik *natural language processing* (NLP) untuk memproses teks tidak terstruktur melalui tahapan *pre-processing*, seperti tokenisasi, normalisasi, dan penghapusan *stop word*. Setelah teks distandarkan, informasi penting diekstraksi dan dianalisis menggunakan metode *data mining* seperti klasifikasi atau *clustering* untuk menemukan pola tersembunyi dalam data. Pendekatan ini memungkinkan transformasi teks mentah menjadi informasi yang berguna secara sistematis dan terstruktur (Bagheri et al., 2023).

2.1.2 Klasifikasi Teks

Klasifikasi teks adalah proses mengkategorikan dan mengelompokkan data teks ke dalam kategori tertentu untuk mempermudah proses analisis. Teknik ini memanfaatkan metode *natural language processing*, *machine learning*, dan *data mining* untuk mengekstraksi pola bermakna, mengelompokkan data teks, serta mengubah teks tidak terstruktur menjadi terstruktur sehingga dapat dianalisis lebih lanjut. Berbagai algoritma sering digunakan dalam klasifikasi teks, termasuk metode tradisional seperti *Naïve Bayes* dan *Support Vector Machines* (SVM) yang dikenal dengan efisiensinya pada data berskala kecil. Selain itu, algoritma berbasis *deep learning* seperti *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN), dan model *Transformer* seperti *BERT* telah memberikan hasil yang signifikan pada tugas klasifikasi dengan data berskala besar. Algoritma berbasis *transfer learning*, seperti *BERT*, sering digunakan untuk memanfaatkan representasi bahasa yang telah dilatih sebelumnya untuk berbagai tugas, termasuk analisis sentimen sehingga menghasilkan akurasi yang lebih tinggi dalam memahami konteks dan pola teks (Taha et al., 2024).

2.1.3 Emosi

Emosi adalah komponen mendasar dari setiap bahasa yang digunakan untuk mengungkapkan perasaan seseorang terhadap berbagai hal (Prasad et al., 2022). Beragam emosi muncul dalam berbagai kondisi dan situasi yang dialami, di mana kondisi tersebut memicu pola pikir dan tindakan, yang pada akhirnya dapat menimbulkan emosi (Sudianto, 2022). Menurut Vygotsky mengungkapkan istilah *perezhivanie* yang berarti bahwa emosi manusia terbentuk dari interaksi antara

karakteristik pribadi individu dengan situasi sosial yang dialaminya (Burkitt, 2021). Emosi pada manusia dapat dikategorikan secara psikologis berdasarkan jenis emosi, intensitas, dan faktor lainnya (Zidan et al., 2022). Dalam teori psikologi klasik, Paul Ekman mengemukakan bahwa terdapat enam hingga tujuh emosi dasar yang bersifat universal, yang dapat dikenali melalui ekspresi wajah dan tidak dipengaruhi oleh budaya maupun bahasa. Emosi-emosi tersebut meliputi kebahagiaan, kemarahan, ketakutan, kesedihan, kejijikan, keterkejutan, dan rasa (Junho & Cheongtag, 2023). Selain itu, Robert Plutchik mengembangkan model emosi berbentuk kerucut tiga dimensi yang dikenal dengan sebutan *Plutchik's Wheel of Emotions* yang divisualisasikan pada Gambar 2.1.

Gambar 2.1 *Plutchik's Wheel of Emotions*

(Mondal & Gokhale, 2020)

Plutchik's Wheel of Emotions merupakan model psikologis yang mengkategorikan emosi dasar dalam bentuk roda, yaitu: *joy, trust, fear, surprise,*

sadness, disgust, anger, dan *anticipation*. Setiap emosi dalam model ini memiliki tingkatan intensitas (*mild – basic – intense*), misalnya: *serenity – joy – ecstasy* untuk emosi *joy*. Plutchik juga menjelaskan bahwa emosi kompleks dapat terbentuk dari kombinasi dua emosi dasar. Sebagai contoh, *love* merupakan hasil gabungan dari *joy* dan *trust* (Mondal & Gokhale, 2020).

2.1.4 *Bilingual*

Bilingual atau bilingualisme merupakan kemampuan komunikasi yang melibatkan penggunaan dua bahasa secara aktif. Kemampuan *bilingual* ini dapat berkembang melalui berbagai situasi, seperti terbiasa menggunakan dua bahasa sejak lahir, tinggal di lingkungan *bilingual*, atau mempelajari bahasa kedua di kemudian hari (Babazade, 2025). Dalam konteks komunikasi digital, penggunaan dua bahasa ini memunculkan fenomena campur kode dan alih kode. Hal ini dipengaruhi oleh kebutuhan untuk menyesuaikan konteks di media sosial, mempercepat komunikasi, atau menunjukan ekspresi tertentu (Nordin, 2023).

2.1.5 Media Sosial

Media sosial merupakan alat interaksi secara virtual yang digunakan oleh individu maupun organisasi untuk berbagi dan bertukar informasi. Media sosial identik dengan mengunggah teks, gambar, atau video yang dilengkapi dengan fitur *like*, komentar, dan *share* (Fazry & Apsari, 2021). Media sosial juga dipahami sebagai media digital di internet yang memungkinkan pengguna untuk merepresentasikan diri, berinteraksi, dan berkolaborasi secara virtual. Fungsi media sosial juga mencakup pencarian informasi, mobilisasi sosial, dan sarana berbagi (Kartini et al., 2020).

Penggunaan media sosial ini memberikan dampak terhadap perilaku pengguna yang mempengaruhi kognitif, emosi, dan tindakan seseorang. Paparan informasi tanpa batas di media sosial dapat membentuk cara berpikir, reaksi emosional tertentu, bahkan mempengaruhi keputusan atau sikap pengguna di kehidupan nyata (Kartini et al., 2020).

2.1.6 Pre-Processing

Pre-processing adalah proses pembersihan data karena data mentah tidak terstruktur dan tidak dapat dianalisis (Rifaldi et al., 2023). *Pre-processing* merupakan tahap penting dalam *text mining* karena pada tahap ini karakter, kata, dan kalimat diidentifikasi dan dipersiapkan untuk tahap berikutnya (Aksoy et al., 2020). Pembersihan dan normalisasi data menjadi langkah yang krusial, karena kualitas data setelah *pre-processing* sangat mempengaruhi kinerja model yang digunakan (Siino et al., 2024).

2.1.6.1 *Data Balancing*

Data balancing merupakan proses untuk mengatasi ketidakseimbangan data (*data imbalance*), yaitu kondisi ketika jumlah data pada satu kelas (mayoritas) jauh lebih besar dibandingkan dengan kelas lainnya (minoritas). Kondisi ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas (Akbar & Hayaty, 2020). Beberapa pendekatan umum yang digunakan untuk menangani data imbalance adalah teknik *resampling*, seperti *oversampling* (menambah jumlah data pada kelas minoritas) dan *undersampling* (mengurangi jumlah data pada kelas mayoritas).

Random oversampling merupakan salah satu metode *oversampling* yang banyak digunakan karena kesederhanaannya, yaitu dengan menduplikasi data minoritas secara acak hingga jumlahnya setara dengan kelas mayoritas (Tyas et al., 2023). *Random undersampling* merupakan salah satu metode *undersampling* yang sering digunakan untuk mengurangi data pada kelas mayoritas secara acak hingga seimbang dengan kelas minoritas. Metode ini dapat mempercepat pemrosesan dan menghemat ruang penyimpanan karena mengurangi volume data, namun beresiko menghilangkan informasi penting yang dapat mempengaruhi performa klasifikasi (Untoro & Yusuf, 2023).

2.1.6.2 *Case Folding*

Case folding merupakan proses yang bertujuan untuk mengubah semua huruf kapital menjadi huruf kecil. Langkah ini penting dalam pemrosesan teks, terutama untuk menyamakan representasi data, sehingga kata dengan huruf besar dan kecil tidak dianggap berbeda oleh sistem. Proses tokenisasi dapat lebih optimal dengan menerapkan *case folding* karena mencegah terjadinya fragmentasi token, yaitu ketika sistem menghasilkan token yang berbeda untuk kata yang sebenarnya sama, hanya karena perbedaan dalam kapitalisasi huruf (Aripin et al., 2023).

2.1.6.3 *Noise Character Removal*

Proses *Noise Character Removal* adalah proses untuk menghapus tanda baca yang dianggap tidak relevan dalam tugas klasifikasi teks tertentu. Penghapusan tanda baca digunakan untuk menyederhanakan teks dalam tahap *pre-processing*. Tidak semua tanda baca sebaiknya dihapus karena beberapa di antaranya memiliki makna penting dalam teks, terutama dalam konteks analisis

emosi. Proses ini juga mencakup pembersihan angka dalam teks untuk menyederhanakan analisis, terutama ketika angka dianggap tidak relevan terhadap konteks tugas. Keputusan untuk menghapus atau mempertahankan tanda baca dan angka harus disesuaikan dengan konteks dan tujuan analisis (Siino et al., 2024).

2.1.6.4 *Text Normalization*

Text normalization adalah proses mengubah kata-kata informal seperti slang, singkatan, atau kata pendek menjadi bentuk yang merepresentasikan makna aslinya. Proses ini bertujuan untuk meningkatkan kinerja klasifikasi dengan memastikan teks dapat dianalisis secara lebih akurat, tanpa kehilangan informasi penting. Pada media sosial, di mana gaya penulisan informal sering digunakan, pengolahan kata-kata ini menjadi krusial untuk menghindari kesalahan interpretasi (Siino et al., 2024).

2.1.6.5 *Convert Emoji*

Emoji merupakan representasi grafis atau objek, wajah, atau simbol yang sering digunakan untuk mengekspresikan emosi atau opini pengguna dalam teks di media sosial (Siino et al., 2024). Proses *pre-processing* emoji dapat diterjemahkan kedalam teks untuk mempermudah analisis sentimen pada teks serta meningkatkan performa model dalam mendeteksi emosi (Yoo & Rayz, 2021). Emoji dapat diterjemahkan ke dalam teks menggunakan pustaka Python. Sebagai contoh, emoji 😊 diubah menjadi kata "*happy_face*".

2.1.6.6 *Tokenizing*

Tokenizing merupakan proses memecah teks menjadi unit-unit kecil yang disebut token. Pada model *XLM-R*, proses ini dilakukan langsung terhadap teks

mentah menggunakan teknik tokenisasi berbasis *Sentence Piece*. Pendekatan ini dapat menangani berbagai bahasa tanpa memerlukan alat tokenisasi khusus untuk setiap bahasa. Hasil tokenisasi berupa potongan-potongan subword memungkinkan model untuk tetap memahami kata-kata yang jarang atau belum pernah ditemui sebelumnya. Keuntungan utama dari penggunaan *Sentence Piece* ini adalah kesederhanaan dan konsistensi dalam tokenisasi, memungkinkan model bekerja langsung pada teks asli tanpa kehilangan informasi atau kompatibilitas antar bahasa. Model *XLM-R* dilatih dengan kosakata besar sebanyak 250.000 token, sehingga mampu mencakup banyak variasi linguistik dari 100 bahasa yang digunakan dalam proses pelatihan (Conneau et al., 2020).

2.1.6.7 *Label Encoding*

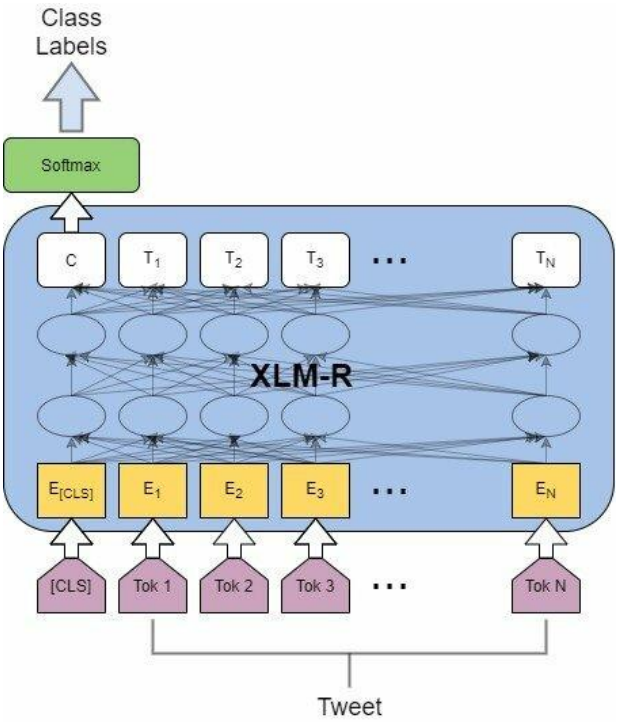
Label Encoding merupakan salah satu teknik yang digunakan dalam pemrosesan data kategorikal dengan cara mengubah nilai berupa teks menjadi representasi numerik yang unik untuk setiap kategori. Metode ini dikenal sederhana dan efisien, khususnya saat diterapkan pada data yang memiliki sifat ordinal. Sebagai contoh, kategori warna seperti Merah, Hijau, dan Biru dapat dikodekan menjadi 0, 1, dan 2. Keunggulan dari Label Encoding terletak pada kemampuannya dalam mengurangi dimensi data, meskipun teknik ini dapat menimbulkan asumsi adanya urutan atau hubungan peringkat antar nilai yang sebenarnya tidak relevan pada variabel kategori non-ordinal. Pemilihan metode ini perlu mempertimbangkan karakteristik data dan kemampuan algoritma pembelajaran mesin yang digunakan dalam menangani hubungan antar label (Herdian et al., 2024).

2.1.7 *XLM-RoBERTa*

XLM-RoBERTa atau *XLM-R* adalah model berbasis *transformer* yang dikembangkan Facebook AI (Aripin et al., 2023). Model ini merupakan pengembangan dari dua model sebelumnya yaitu XLM (*Cross-lingual Language Model*) dan *RoBERTa* (*Robustly Optimized BERT Pretraining Approach*), yang memanfaatkan teknik *Masked Language Modeling* (MLM*)* untuk mempelajari hubungan semantik antar kata dengan lebih efektif. Dilatih pada 100 bahasa menggunakan dataset *CommonCrawl* yang besar (lebih dari 2 terabyte), *XLM-RoBERTa* mengatasi keterbatasan model multilingual sebelumnya seperti *mBERT* dan XLM dengan memberikan performa lebih baik pada bahasa dengan sumber daya rendah (Conneau et al., 2020; Ranasinghe & Zampieri, 2020).

Pada model *XLM-RoBERTa*, setelah teks diubah menjadi token, setiap token akan diproses pada tahap *embedding,* yang mencakup *token embedding* dan *position embedding*. *Token embedding* mengubah token menjadi representasi vektor numerik agar dapat diproses secara matematis oleh model. Sementara itu, *position embedding* menyisipkan informasi posisi urutan token dalam input, sehingga model dapat memahami konteks berdasarkan struktur kalimat. Berbeda dari *BERT*, *XLM-R* tidak menggunakan *segment embedding* karena seluruh input dianggap satu segmen. Simbol khusus [CLS] ditempatkan di awal input untuk merepresentasikan keseluruhan urutan dalam tugas klasifikasi (Conneau et al., 2020; Aripin et al., 2023). Setelah *embedding* selesai, representasi token diproses melalui jaringan *XLM-R* yang terdiri dari beberapa lapisan *transformator*. Hasil akhir dari proses model ini adalah output klasifikasi yang menghasilkan label kelas melalui lapisan

softmax (Ranasinghe & Zampieri, 2020). Gambar arsitektur *XLM-RoBERTa* untuk klasifikasi teks dituangkan dalam Gambar 2.3.



Gambar 2.2 Arsitektur *XLM-RoBERTa*

(Ranasinghe & Zampieri, 2020)

Model *XLM-R* ini dapat mengidentifikasi makna suatu kata menggunakan mekanisme *self-attention*, yang bekerja dengan mempertimbangkan hubungan antar kata dalam satu kalimat secara menyeluruh, bukan hanya kata sebelumnya atau berikutnya, tetapi juga seluruh konteks kalimat. (Aripin et al., 2023). Mekanisme *self-attention* dapat dihitung menggunakan formula berikut pada Persamaan (2.1).

$$Attention(Q, K, V) = \text{softmax} + \left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (2.1)$$

Variabel Q (*query*), K (*key*), dan V (*value*) merupakan representasi *embedding* dari kata-kata dalam teks, variabel QK^T menghitung skor relevansi antara Q dan K. $\sqrt{d_k}$ merepresentasikan dimensi vektor K, yang digunakan sebagai faktor normalisasi untuk menjaga stabilitas komputasi. Fungsi *softmax* menormalisasi skor relevansi menjadi bobot probabilitas, dan V adalah vektor nilai yang dirata-ratakan menggunakan bobot hasil *softmax* untuk menghasilkan representasi kontekstual (Aripin et al., 2023). Lapisan *softmax* digunakan untuk menghasilkan probabilitas dari setiap label klasifikasi berdasarkan rumus pada Persamaan (2.2).

$$p(c|h) = softmax(Wh) \quad (2.2)$$

Variabel c merupakan label kelas yang diprediksi oleh model berdasarkan representasi urutan teks. h merupakan representasi akhir dari token [CLS], yang mencakup informasi dari seluruh urutan teks. Matriks w adalah parameter spesifik untuk tugas yang dioptimalkan selama proses *fine-tuning*. Fungsi *softmax* digunakan untuk mengonversi hasil ini menjadi probabilitas yang terdistribusi untuk setiap kelas (Ranasinghe & Zampieri, 2020). *Pseudocode* untuk model *XLM-RoBERTa* pada tugas klasifikasi emosi ini dituangkan dalam Algoritma 1.

Algoritma 1. *XLM-RoBERTa* for Emotion Classification

Input: Text Sentence
Start
1: Tokenize input sentence using *XLM-R* tokenizer
➔ Get **input_ids** and **attention_mask**
2: Embed input tokens
a. For each token in **input_ids**:
➔ Get token embedding vector
b. Add positional embeddings
3: Pass embeddings through *XLM-RoBERTa* layers
for each layer in *XLM-R*
for each attention head
a. Compute Q, K, V matrices
b. Calculate attention scores
c. Multiply attention score with V
end for
4: Extract [CLS] token output as sentence representation
5: Feed [CLS] vector token into a linear classification layer
6: Apply softmax to get emotion probabilities
7: Return the emotion with the highest probability
End
Output: Emotion Label

2.1.8 *Early Stopping*

Early stopping merupakan teknik regulasi yang digunakan dalam pelatihan model untuk mencegah *overfitting*. Teknik ini bekerja dengan cara memantau performa model pada data validasi selama proses pelatihan, dan menghentikannya ketika metrik validasi, seperti akurasi atau *loss*, tidak lagi menunjukkan perbaikan dalam sejumlah iterasi tertentu. Parameter *patience* menentukan berapa lama pelatihan bertahan tanpa peningkatan sebelum dihentikan. Penentuan nilai *patience* yang optimal bergantung pada karakteristik model dan dataset yang digunakan. Nilai *patience* yang terlalu kecil dapat menghentikan pelatihan terlalu cepat, sebaliknya nilai yang terlalu besar menyebabkan pelatihan berlarut tanpa manfaat signifikan. Hubungan *patience* dan jumlah *epoch* tidak linear, sehingga pemilihannya bersifat kontekstual (Hussein & Shareef, 2024).

Pada model besar dengan banyak parameter (*over-informative*), waktu optimal untuk menghentikan pelatihan biasanya lebih singkat karena model lebih cepat mencapai titik *overfitting*. Sebaliknya, dalam kondisi *under-informative* dengan fitur yang terbatas, pelatihan cenderung membutuhkan waktu lebih lama karena model memerlukan lebih banyak *epoch* untuk menangkap pola yang relevan (Shen et al., 2022). Penentuan nilai *patience* harus disesuaikan, karena tidak ada ukuran yang secara *universal* tepat untuk semua kasus. Pemilihannya perlu mempertimbangkan sifat data, kapasitas model, serta tujuan dari pelatihan itu sendiri, sehingga dapat dicapai keseimbangan antara performa dan efisiensi (Hussein & Shareef, 2024).

2.1.9 *Confusion Matrix*

Confusion matrix atau *error matrix* merupakan salah satu tabel yang digunakan untuk menggambarkan performa model dalam melakukan klasifikasi dengan membandingkan hasil prediksi model terhadap data yang sebenarnya. Matriks ini memiliki beberapa istilah, diantaranya: *True Positive* (TP), yaitu jumlah data positif yang diklasifikasikan dengan benar oleh model. *True Negative* (TN), yaitu jumlah data negatif yang juga diklasifikasikan dengan benar. Terdapat *False Positive* (FP), yaitu jumlah data yang seharusnya negatif tetapi salah diklasifikasikan sebagai positif, dan *False Negative* (FN), yaitu jumlah data positif yang justru salah diklasifikasikan sebagai negatif (Maghfiroh et al., 2023). Ilustrasi dari *confusion matrix* ini dituangkan dalam Tabel 2.1.

Pada konteks klasifikasi *multiclass*, konsep *confusion matrix* mengalami penyesuaian. Matriks ini memiliki dimensi N x N, di mana N merupakan jumlah

label kelas yang berbeda, misalnya C_0, C_1, … , C_N. Dalam konteks ini, karakterisasi istilah seperti TP, TN, FP, dan FN tidak sepenuhnya berlaku, karena masing-masing kelas memiliki peran yang berbeda dalam evaluasi performa.

Tabel 2.1 *Confusion Matrix*

		Predicted Class			
		C_1	C_2	…	C_N
Actual Class	C_1	$C_{1,1}$	FP	…	$C_{1,N}$
	C_2	FN	TP	…	FN
	…	…	…	…	…
	C_N	$C_{N,1}$	FP	…	$C_{N,N}$

Evaluasi dengan confusion matrix pada klasifikasi multiclass biasanya dilakukan dengan mempertimbangkan keseluruhan parameter untuk memberikan pengukuran yang lebih komprehensif (Markoulidakis et al., 2021). Rumus yang digunakan untuk menghitung nilai akurasi model berdasarkan *confusion matrix multiclass* terdapat dalam Persamaan (2.3).

$$accuracy = \frac{\sum_{i=1}^{N} TP(Ci)}{\sum_{i=1}^{N} \sum_{i=1}^{N} Ci,j} \tag{2.3}$$

Accuracy dalam rumus ini mengukur proporsi jumlah prediksi yang benar (*True Positive*) dibandingkan dengan total jumlah seluruh prediksi yang dilakukan oleh model. Sebagai contoh, jika model berhasil mengklasifikasikan teks dengan benar dalam berbagai kelas, maka nilai *accuracy* akan lebih tinggi.

2.2 *State-of-The-Art*

State of the art pada Tabel 2.2 menyajikan ringkasan penelitian terdahulu yang relevan dengan klasifikasi emosi pada teks, mencakup dataset, hasil, serta kelebihan dan kekurangannya.

Tabel 2.2 *State-of-the-Art* Penelitian

No	Nama Peneliti/Jurnal		Dataset	Hasil	Kelebihan	Kekurangan dan Gap Penelitian
1	**Peneliti (tahun)**	(Wiciaputra et al., 2021)	Judul berita Bahasa Indonesia dan Bahasa Inggris	Akurasi 93,3%	Menguji berbagai kombinasi dataset untuk menemukan konfigurasi optimal dalam klasifikasi teks multibahasa.	Performa model menurun saat diuji pada data Bahasa Indonesia dibanding Bahasa Inggris, menunjukkan perlunya eksplorasi *hyperparameter* untuk meningkatkan performa pada data Indonesia.
	Judul	*Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa*				
	Algoritma	*XLM-RoBERTa*				
2	**Peneliti (tahun)**	(Aripin et al., 2023)	Data twitter Bahasa Indonesia dengan 6 emosi dasar, serta 15 kombinasi emosi majemuk	Akurasi 95,56%	Menggunakan *fine-tuning* (*dropout* & *learning rate*) untuk mencegah *overfitting,* memperluas klasifikasi emosi kompleks.	Penelitian ini hanya berfokus pada Bahasa Indonesia dan belum menguji efektivitas model dalam skenario bilingual. Selain itu, belum dilakukan integrasi emoji yang berpotensi meningkatkan akurasi klasifikasi emosi.
	Judul	*Optimizing the Accuracy of the Semantic-Based Compound Emotion Classifications using the XLM-RoBERTa*				
	Algoritma	*XLM-RoBERTa*				

No	Nama Peneliti/Jurnal		Dataset	Hasil	Kelebihan	Kekurangan dan Gap Penelitian
3	**Peneliti (tahun)**	(Kannan et al., 2024)	Data twitter Bahasa Inggris dengan 4 emosi (WASSA-2017)	Akurasi 90.44%	Mengeksplorasi berbagai metode *machine learning* dan *ensemble* (SVM + XGBoost) untuk *multiclass emotion*	Penelitian hanya menggunakan data Bahasa Inggris sehingga belum mengkaji klasifikasi emosi secara *bilingual*, serta belum mengintegrasikan emoji sebagai fitur tambahan.
	Judul	*Emotion Categorization on Multiclass Textual Data using ML and Ensemble Techniques*				
	Algoritma	*Ensemble Model (SVM + XGBoost)*				
4	**Peneliti (tahun)**	(Najwa et al., 2025)	Data Media Sosial Bahasa Indonesia dengan 6 emosi	Akurasi 93.51% untuk data 6 emosi, dan Akurasi 91.57% untuk data 5 emosi (*anger,fear ,joy,sadness)*	Memanfaatkan *FastText* untuk *embedding* dan B-SMOTE untuk penyeimbangan kelas	Fokus data hanya Bahasa Indonesia, belum diuji dalam klasifikasi emosi *bilingual*. Penelitian ini juga belum mengintegrasikan fitur emoji yang dapat memperkaya representasi emosi dalam teks.
	Judul	*Emotion Classification In Social Media Posts Related To Telecommunication Services Using Bidirectional LSTM*				
	Algoritma	*BiLSTM*				
5	**Peneliti (tahun)**	(Marthen & Novita, 2024)	Dataset Bahasa Indonesia berisi data tweet publik, 6 label emosi.	Akurasi 83.64%	Menggunakan *Bayesian Optimization* untuk menemukan *hyperparameter* terbaik.	Penelitian ini sudah mengoptimasi *hyperparameter*, namun cakupan hanya sebatas Bahasa Indonesia dan tanpa eksplorasi kombinasi teks dan
	Judul	Optimasi *RoBERTa* dengan *Hyperparameter Tuning* untuk				

No	Nama Peneliti/Jurnal		Dataset	Hasil	Kelebihan	Kekurangan dan Gap Penelitian
		Deteksi Emosi berbasis Teks				emoji. Pengujian dalam skenario bilingual dengan kombinasi fitur tersebut perlu dilakukan untuk validasi lebih luas.
	Algoritma	*RoBERTa*				
6	**Peneliti (tahun)**	(Putra & Purwarianti, 2020)	Data Bahasa Indonesia dan Inggris untuk analisis *hate speech*	Akurasi 89,52%	Membandingkan pendekatan *feature-based* dan *fine-tuning*, menemukan bahwa fine-tuning penuh lebih unggul	Fokus penelitian adalah *hate speech*, bukan emosi. Selain itu, belum mengintegrasikan emoji dan belum menguji efektivitas pengoptimalan *hyperparameter* dalam teks *bilingual*.
	Judul	*Improving Indonesian Text Classification Using Multilingual Language Model*				
	Algoritma	*mBERT & XLM-RoBERTa*				
7	**Peneliti (tahun)**	(Setiawan et al., 2021)	Menggunakan 2 dataset InaMoodMeter ISEAR yang diterjemahkan kedalam Bahasa Indonesia dengan 5 kelas emosi.	*Fine-tuning BERT* akurasi 79%	Melakukan *fine-tuning* dengan *learning rate* 2e-5 dan *batch size* 32, serta eksperimen dengan berbagai jumlah *epochs* untuk menemukan konfigurasi optimal.	Dataset Bahasa Indonesia hasil terjemahan dari Bahasa Inggris, sehingga belum merepresentasikan karakteristik asli bahasa informal Indonesia, serta belum diuji pada data bilingual dan belum mengintegrasikan emoji.
	Judul	*Social Media Emotion Analysis in Indonesian Using Fine-Tuning BERT Model*				
	Algoritma	*Naïve Bayes, Support Vector Machine, Convolutional Neural Network, dan BERT*				

No	Nama Peneliti/Jurnal		Dataset	Hasil	Kelebihan	Kekurangan dan Gap Penelitian
8	**Peneliti (tahun)**	(Saleh Alzubaidi & Bourennani, 2021)	Dataset Bahasa Inggris dengan 970 emoji	Akurasi 85.89% dalam SVM kombinasi teks dan emoji	Menggabungkan teks dan emoji dalam *Vector Space Model*, membuktikan kombinasi teks dan emoji lebih akurat dibanding teks saja	Studi terbatas dua kelas (positif/negatif) dan hanya berbahasa Inggris. Perlu pengembangan menuju klasifikasi emosi dan bilingual dengan memanfaatkan model *transformer* terkini.
	Judul	*Sentiment Analysis of Tweets Using Emojis and Texts*				
	Algoritma	*Support Vector Machine* *Naïve Bayes* *Fast Large Margin*				

2.3 Matriks Penelitian

Tabel 2.3 menyajikan matriks perbandingan berbagai penelitian terdahulu yang ada pada Tabel 2.2 berdasarkan perbedaan bahasa dataset, tujuan penelitian dan algoritma yang digunakan.

Tabel 2.3 Matriks Penelitian

No	Penelitian	Data		Tujuan Penelitian				Algoritma										
		Indonesia	Inggris	Analisis Emosi	Optimasi *Hyperparameter*	Optimasi *Fine Tuning*	Analisis Pengaruh Emoji	*XLM-RoBERTa*	*RoBERTa*	*mBERT*	*BERT*	*XLM-T*	*Naïve Bayes*	*Support Vector Machine*	*CNN*	*SVM + XGBoost*	*BiLSTM*	*Fast Large Margin*
1	(Wiciaputra et al., 2021)	✓	✓					✓										
2	(Aripin et al., 2023)	✓		✓		✓		✓										
3	(Kannan et al., 2024)		✓	✓												✓		
4	(Najwa et al., 2025)	✓		✓													✓	
5	(Marthen & Novita, 2024)	✓		✓	✓				✓									
6	(Putra & Purwarianti, 2020)	✓	✓			✓		✓		✓								
7	(Setiawan et al., 2021)	✓		✓							✓		✓	✓	✓			
8	(Saleh Alzubaidi & Bourennani, 2021)		✓				✓						✓	✓				✓
9	*Our Research*	✓	✓	✓	✓		✓	✓										

Tabel 2.2 menunjukkan bahwa klasifikasi emosi berbasis teks telah banyak dikembangkan, namun masih memiliki keterbatasan. Penelitian menggunakan *XLM-RoBERTa* melakukan klasifikasi emosi majemuk, tetapi hanya fokus pada Bahasa Indonesia dan belum mengkaji integrasi emoji (Aripin et al., 2023). Penelitian lain mengkaji klasifikasi teks dua bahasa, namun bukan dalam konteks emosi dan tanpa eksplorasi *hyperparameter* (Wiciaputra et al., 2021; Putra & Purwarianti, 2020). *RoBERTa* juga digunakan untuk klasifikasi emosi Bahasa Indonesia, namun belum mengkombinasikan teks dan emoji (Najwa et al., 2025). Terdapat pula studi yang telah menggabungkan teks dan emoji, namun masih terbatas pada dua kelas sentimen, bukan klasifikasi emosi (Alzubaidi & Bourennani, 2021).

Berdasarkan gap yang ada pada penelitian sebelumnya, penelitian ini melakukan klasifikasi emosi berbasis teks dan emoji secara *bilingual*, yaitu pada data Bahasa Indonesia dan Bahasa Inggris dengan menggunakan model *XLM-RoBERTa*. Data yang digunakan terdiri dari teks Bahasa Indonesia dan Bahasa Inggris yang diproses secara terpisah (Najwa et al., 2025; S. Kannan et al., 2024). Penelitian ini juga mengeksplorasi variasi *hyperparameter* dengan fokus pada jumlah *epoch* yang telah terbukti berpengaruh terhadap peningkatan kinerja model (Marthen & Novita, 2024). Penelitian ini diharapkan dapat melengkapi kekurangan studi sebelumnya terkait klasifikasi emosi bilingual, integrasi fitur emoji, serta pengoptimalan model *XLM-RoBERTa* melalui eksplorasi *hyperparameter*.