Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Analisis Sentimen Kesehatan Mental pada Media Sosial X Menggunakan BERT dengan Pendekatan XAI Berbasis SHAP dan LIME

DOI: http://dx.doi.org/10.35889/progresif.v22i3.4070

Annisa Maulana Majid[1*], Ismasari Nawangsih[2], dan Karina Imelda[3]
Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
*e-mail *Corresponding Author*: annisa.maulanamajid@pelitabangsa.ac.id

Abstract

Mental health had become an important issue widely discussed on social media, creating the need for an automated method to identify mental health conditions based on textual data. This study aimed to develop a text classification model using Bidirectional Encoder Representations from Transformers and improve the transparency of prediction results through Shapley Additive Explanations and Local Interpretable Model-agnostic Explanations. The dataset consisted of 51,073 text records categorized into seven mental health classes. The research stages included data and text cleaning, label encoding, data splitting, tokenization, model training, evaluation, and result interpretation. The testing results showed that the model achieved an accuracy of 82% and a weighted average F1-score of 0.82. The interpretation results indicated that specific words and phrases contributed to class predictions. The findings demonstrated that the model performed text classification effectively, while both interpretation methods improved the transparency of the model's decision-making process.

Keywords: *Sentiment Analysis; BERT; XAI; SHAP; LIME*

Abstrak

Kesehatan mental telah menjadi isu penting yang banyak dibahas melalui media sosial sehingga diperlukan metode otomatis untuk mengidentifikasi kategori kondisi kesehatan mental berdasarkan teks. Penelitian ini bertujuan membangun model klasifikasi teks menggunakan *Bidirectional Encoder Representations from Transformers* (BERT) serta meningkatkan transparansi hasil prediksi melalui *Shapley Additive Explanations* (SHAP) dan *Local Interpretable Model-agnostic Explanations* (LIME). *Dataset* yang digunakan terdiri atas 51.073 teks dalam tujuh kategori kesehatan mental. Tahapan penelitian meliputi pembersihan data dan teks, pengodean label, pembagian data, tokenisasi, pelatihan model, evaluasi, serta interpretasi hasil. Hasil pengujian menunjukkan bahwa model memperoleh akurasi sebesar 82% dan nilai F1-*score* rata-rata tertimbang sebesar 0,82. Interpretasi menunjukkan bahwa frasa dan kata tertentu memberikan kontribusi terhadap prediksi kelas. Hasil penelitian membuktikan bahwa model mampu melakukan klasifikasi dengan kinerja yang baik, sedangkan kedua metode interpretasi meningkatkan transparansi keputusan model.

Kata kunci: *Analisis Sentimen; BERT; XAI; SHAP; LIME*

1. Pendahuluan

Kesehatan mental merupakan salah satu aspek penting yang memengaruhi kualitas hidup, produktivitas, serta kemampuan individu dalam menjalankan aktivitas sehari-hari [1]. Individu dengan kondisi kesehatan mental yang baik mampu mengenali potensi dirinya, mengelola tekanan hidup, bekerja secara produktif, dan berkontribusi dalam kehidupan sosial. Sebaliknya, gangguan kesehatan mental seperti depresi, kecemasan (*anxiety*), stres, gangguan bipolar, hingga kecenderungan bunuh diri dapat menurunkan kualitas hidup individu dan memberikan dampak yang lebih luas terhadap Masyarakat [2]. Di Indonesia, permasalahan

kesehatan jiwa masih menjadi tantangan kesehatan masyarakat yang memerlukan perhatian serius. Berdasarkan Riset Kesehatan Dasar (Riskesdas) 2018, prevalensi rumah tangga yang memiliki anggota dengan gangguan jiwa berat menunjukkan peningkatan dibandingkan hasil Riskesdas 2013 [3]. Selanjutnya, Survei Kesehatan Indonesia (SKI) 2023 melaporkan bahwa sekitar 4,0% rumah tangga memiliki anggota keluarga dengan gejala gangguan jiwa berat (skizofrenia/psikosis) [4]. Meskipun kedua survei tersebut menggunakan metode dan indikator yang berbeda sehingga hasilnya tidak dapat dibandingkan secara langsung, temuan tersebut menunjukkan bahwa permasalahan kesehatan jiwa di Indonesia masih memerlukan perhatian, khususnya dalam mendukung upaya deteksi dini dan penanganan yang berkelanjutan.

Seiring dengan meningkatnya penggunaan media sosial, masyarakat semakin terbuka dalam mengekspresikan perasaan, opini, dan pengalaman pribadi melalui platform digital. Salah satu media sosial yang memungkinkan pengguna menyampaikan kondisi emosional secara spontan melalui unggahan berbasis teks adalah X (Twitter). Unggahan tersebut dapat mengandung berbagai ekspresi, opini, dan pengalaman yang berpotensi memberikan informasi mengenai kondisi emosional penggunanya. Hal ini menjadikan media sosial sebagai salah satu sumber data yang berpotensi dimanfaatkan untuk mendukung deteksi dini kondisi kesehatan mental. Namun, karakteristik data media sosial yang berjumlah besar, tidak terstruktur, serta mengandung penggunaan singkatan, simbol, bahasa informal, dan berbagai bentuk ekspresi menyebabkan proses identifikasi dan pengelompokan sentimen secara manual menjadi kurang efektif. Selain itu, hubungan konteks antar kata dalam teks dapat memengaruhi makna suatu pernyataan sehingga diperlukan metode yang mampu memahami karakteristik bahasa tersebut secara lebih baik. Oleh karena itu, pendekatan *Natural Language Processing* (NLP), khususnya analisis sentimen, dapat digunakan untuk mengolah dan mengklasifikasikan data teks secara otomatis berdasarkan pola sentimen yang terkandung di dalamnya.

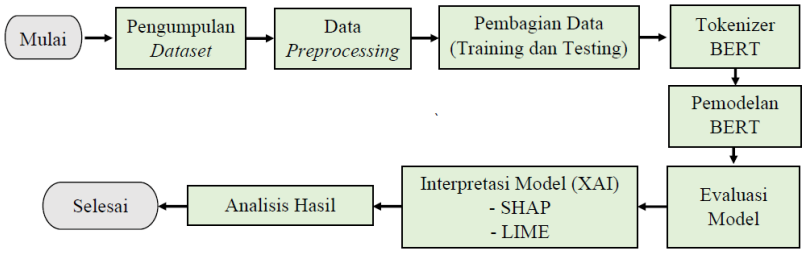
Berbagai penelitian telah dilakukan untuk mengembangkan metode analisis sentimen yang berkaitan dengan kesehatan mental pada data media sosial. Kharisma Rahayu dkk. melakukan klasifikasi terhadap 7.489 tweet menggunakan *Decision Tree*, *Random Forest*, *Naïve Bayes*, dan *K-Nearest Neighbor* (KNN), dengan akurasi masing-masing sebesar 95,1%, 95,7%, 91%, dan 85,8% [5]. Yuwan Jumaryadi dkk. menggunakan 5.209 data Twitter dengan algoritma *Support Vector Machine* (SVM), *Random Forest*, dan *Naïve Bayes*, yang menghasilkan akurasi masing-masing sebesar 78,69%, 75,04%, dan 70,44% [6]. Sementara itu, Abdul Rehman dkk. menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT) pada dataset *Sentiment Analysis for Mental Health* yang terdiri atas 51.074 data dan memperoleh akurasi sebesar 95,71% [7]. Penelitian-penelitian tersebut menunjukkan bahwa analisis sentimen kesehatan mental telah berkembang dari penggunaan machine learning konvensional menuju model berbasis Transformer yang memiliki kemampuan lebih baik dalam memahami konteks bahasa. BERT mampu menghasilkan representasi teks secara kontekstual dengan mempertimbangkan hubungan antar kata [8]. Meskipun demikian, penggunaan BERT dalam klasifikasi masih memiliki keterbatasan dari sisi interpretabilitas karena karakteristiknya sebagai model *black-box*, sehingga proses pengambilan keputusan model sulit dipahami secara langsung oleh pengguna [9]. Selain itu, penelitian terdahulu umumnya masih berfokus pada pengukuran performa klasifikasi menggunakan *Accuracy*, *Precision*, *Recall*, dan *F1-Score*, tanpa memberikan penjelasan yang memadai mengenai kata atau bagian teks yang berkontribusi terhadap keputusan model. Dengan demikian, terdapat research gap berupa kebutuhan untuk meningkatkan interpretabilitas model dalam analisis sentimen kesehatan mental berbasis BERT, khususnya untuk mengetahui faktor tekstual yang memengaruhi keputusan klasifikasi.

Penelitian ini menerapkan model *Bidirectional Encoder Representations from Transformers* (BERT) yang dipadukan dengan pendekatan *Explainable Artificial Intelligence* (XAI), yaitu *SHapley Additive exPlanations* (SHAP) dan *Local Interpretable Model-Agnostic Explanations* (LIME), pada data teks media sosial X. BERT digunakan untuk melakukan klasifikasi sentimen dengan memanfaatkan kemampuan representasi teks secara kontekstual [8], sedangkan LIME dan SHAP digunakan untuk memberikan interpretasi terhadap hasil prediksi dengan mengidentifikasi kata atau bagian teks yang berkontribusi terhadap keputusan klasifikasi [9]. Kinerja model dievaluasi menggunakan *Accuracy*, *Precision*, *Recall*, *dan F1-Score*, sedangkan interpretabilitas model dianalisis berdasarkan kontribusi fitur yang dihasilkan oleh SHAP dan LIME. Kebaruan penelitian ini terletak pada integrasi BERT dengan dua metode XAI, yaitu SHAP dan LIME, untuk klasifikasi tujuh kategori kondisi kesehatan mental pada data

teks media sosial X, sehingga penelitian tidak hanya berfokus pada performa klasifikasi, tetapi juga memberikan penjelasan mengenai kontribusi kata atau bagian teks terhadap keputusan model pada masing-masing kategori. Dengan pendekatan tersebut, penelitian ini diharapkan menghasilkan model klasifikasi yang memiliki performa baik sekaligus lebih transparan dan dapat dipahami dalam mendukung analisis serta upaya deteksi dini kondisi kesehatan mental.

2. Metodologi

Penelitian ini mengimplementasikan model BERT untuk menganalisis sentimen kesehatan mental pada media sosial X serta menerapkan pendekatan XAI menggunakan SHAP dan LIME. Alur tahapan penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

Tahap awal penelitian dimulai dengan pengumpulan data, data yang telah dikumpulkan kemudian melalui tahap *preprocessing* untuk membersihkan dan mempersiapkan data sebelum digunakan pada proses pemodelan. Selanjutnya data yang telah diproses dibagi menjadi data latih (*training)* dan data uji (*testing*) menggunakan metode *stratified split* untuk menjaga proporsi setiap kelas. Proses pemodelan dilakukan menggunakan BERT melalui tahapan tokenisasi, fine-tuning, dan klasifikasi sentimen. Kinerja model kemudian dievaluasi dengan confusion metrik *Accuracy*, *Precision*, *Recall*, dan F1-*Score*. Untuk meningkatkan interpretabilitas model, pendekatan *Explainable Artificial Intelligence* (XAI) diterapkan menggunakan SHAP dan LIME. Kedua metode tersebut digunakan untuk mengidentifikasi kata atau frasa yang memberikan kontribusi terbesar terhadap hasil prediksi sehingga proses pengambilan keputusan model dapat dipahami secara lebih transparan.

2.1 Pengumpulan *Dataset*

Penelitian ini menggunakan dataset publik *Sentiment Analysis for Mental Health* yang diperoleh melalui platform Kaggle. Dataset terdiri atas 53.043 data dengan tiga atribut, yaitu Unnamed: 0, statement, dan status. Atribut *statement* berisi teks yang merepresentasikan pernyataan terkait kondisi emosional maupun psikologis, sedangkan atribut status merupakan label yang menunjukkan kategori kondisi kesehatan mental pada setiap data. Adapun atribut Unnamed: 0 merupakan indeks baris. Dalam penelitian ini, atribut statement digunakan sebagai masukan model, sedangkan atribut status digunakan sebagai label kelas/kategori pada proses klasifikasi. Dataset terdiri atas tujuh kategori, yaitu *Normal, Depression, Suicidal, Anxiety, Bipolar, Stress*, dan *Personality Disorder*. Jumlah data pada masing-masing kategori disajikan pada Tabel 1.

Tabel 1. Distribusi Data Berdasarkan Label Kelas Kategori

No.	Status	Jumlah
1	*Normal*	16.351
2	*Depression*	15.404
3	*Suicidal*	10.653
4	*Anxiety*	3.888
5	*Bipolar*	2.877
6	*Stress*	2.669
7	*Personality Disorder*	1.201
	Total	53.043

2.2 Data Preprocessing

Data *preprocessing* merupakan tahap untuk mempersiapkan data sebelum digunakan dalam proses pemodelan *machine learning* [10]. Pada penelitian ini, *preprocessing* dilakukan menggunakan Python pada Google Colab yang meliputi *data cleaning*, *text preprocessing*, dan *label encoding*.

a. ***Data Cleaning***
   Data cleaning dilakukan untuk menangani data yang tidak diperlukan atau bermasalah sebelum digunakan dalam pemodelan [11]. Pada penelitian ini, atribut **Unnamed: 0** dihapus karena hanya berfungsi sebagai indeks. Selanjutnya, data dengan nilai kosong pada atribut **statement** dihapus menggunakan fungsi dropna(), kemudian dilakukan penghapusan data duplikat.
b. ***Text Preprocessing***
   Text preprocessing merupakan tahapan untuk membersihkan dan mempersiapkan data teks agar dapat digunakan dalam proses klasifikasi [12]. Pada penelitian ini, atribut **statement** dibersihkan dengan menghapus URL, *mention*, karakter khusus, tanda baca, dan spasi berlebih. Selanjutnya, teks diubah menjadi huruf kecil (*lowercase*) dan dilakukan *tokenization* menggunakan BERT Tokenizer sesuai dengan *vocabulary* model BERT
c. ***Label Encoding***
   Label encoding merupakan proses mengubah label kategorikal menjadi representasi numerik agar dapat digunakan dalam proses pembelajaran model [13]. Pada penelitian ini, label pada dataset dikonversi menjadi nilai numerik yang selanjutnya digunakan sebagai *target* dalam proses pelatihan dan evaluasi model BERT.

2.3 Pembagian Data

Pembagian data (*data splitting*) dilakukan setelah tahap preprocessing untuk menghasilkan data pelatihan, data validasi, dan data pengujian. *Dataset* sebanyak 51.073 data terlebih dahulu dibagi dengan rasio 80:20 menggunakan metode stratified split. Selanjutnya, data pelatihan dibagi kembali dengan rasio 90:10 untuk memperoleh 36.772 data pelatihan, 4.086 data validasi, dan 10.215 data pengujian. Data pelatihan digunakan untuk fine-tuning model BERT, data validasi digunakan untuk memantau proses pelatihan, sedangkan data pengujian digunakan untuk evaluasi kinerja model. menggunakan metrik *accuracy*, *precision, recall*, dan F1-*score*.

2.4 Tokenizer BERT

Tokenization BERT merupakan proses mengubah teks menjadi format masukan yang dapat dipahami oleh model BERT. *Tokenizer* memanfaatkan metode *WordPiece* untuk memecah teks menjadi token sesuai dengan vocabulary yang dimiliki model. Selanjutnya, *tokenizer* secara otomatis menambahkan token khusus, yaitu [CLS] pada awal teks, [SEP] pada akhir teks, serta [PAD] apabila panjang token lebih pendek dari batas yang ditentukan [14]. Pada penelitian ini, tokenisasi dilakukan setelah proses pembagian data training dan testing menggunakan *BertTokenizer* dari model bert-base-uncased. Selain itu, diterapkan proses padding dan truncation dengan *max_length* = 256 agar seluruh data memiliki panjang input yang seragam. Hasil tokenisasi berupa input IDs dan attention mask yang kemudian digunakan sebagai masukan pada proses fine-tuning model BERT untuk melakukan klasifikasi sentimen.

2.5 Pemodelan BERT

BERT adalah model pemrosesan bahasa alami yang mampu menghasilkan representasi teks kontekstual dengan mempertimbangkan informasi dari seluruh bagian kalimat. Kemampuan tersebut digunakan untuk mendeteksi konten toksik, mengenali kecenderungan sentimen, dan mengelompokkan topik kesehatan mental dalam komentar pengguna media sosial [15]. BERT memahami konteks kata secara dua arah (*bidirectional*), sehingga mampu menghasilkan representasi teks yang lebih kaya dan akurat. Kemampuan tersebut menjadikan BERT banyak digunakan dalam berbagai tugas NLP, seperti klasifikasi teks, analisis sentimen, dan deteksi topik [16]. Dengan kemampuan memahami hubungan kontekstual antar kata, BERT dapat melakukan pelabelan entitas secara otomatis pada teks [17].

Pada penelitian ini, digunakan model *bert-base-uncased* yang telah melalui tahap *pre-training*. Model tersebut kemudian dilakukan *fine-tuning* menggunakan *dataset* teks kesehatan mental. Data *training* yang telah melalui proses tokenisasi digunakan sebagai masukan pada

proses pelatihan. Selama proses *fine-tuning*, model mempelajari pola dan hubungan kontekstual antar kata untuk menghasilkan model yang mampu mengklasifikasikan teks ke dalam tujuh kelas, yaitu *Normal*, *Depression*, *Suicidal*, *Anxiety*, *Bipolar*, *Stress*, dan *Personality Disorder*. Data *testing* digunakan untuk mengevaluasi kinerja model setelah proses pelatihan selesai. BERT dibangun di atas arsitektur Transformer yang memanfaatkan mekanisme *Scaled Dot-Product Attention* untuk menghitung hubungan antar token dalam suatu kalimat [7]. Mekanisme tersebut dinyatakan dengan persamaan berikut [18]:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

Keterangan:
Q = *Query vector*.
K = *Key vector*.
V = *Value vector*.
d_k = dimensi *vektor Key* sebagai faktor normalisasi.
Softmax = fungsi untuk menghasilkan bobot perhatian.
Attention = keluaran mekanisme *self-attention* berupa representasi kontekstual token.

2.6 Evaluasi Model

Evaluasi model merupakan tahap untuk mengukur kinerja model BERT dalam mengklasifikasikan sentimen pada data testing. Pada penelitian ini, evaluasi dilakukan menggunakan *confusion matrix* yang membandingkan hasil prediksi model dengan label aktual. *Confusion matrix* digunakan sebagai dasar untuk menghitung berbagai metrik evaluasi yang digunakan dalam mengukur kinerja model klasifikasi [19]. Bentuk *confusion matrix* yang digunakan sebagai dasar perhitungan metrik evaluasi ditunjukkan pada Tabel 2.

Tabel 2. *Confusion Matrix*

		Kelas Aktual	
		Positif	**Negatif**
Kelas Prediksi	**Positif**	*True Positive* (TP)	(FP)
	Negatif	*False Negative* (FN)	True Negative (TN)

Persamaan Confusion Matrix sebagai berikut:

Accuracy merupakan metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data. Nilai *accuracy* diperoleh dari perbandingan jumlah prediksi yang benar terhadap seluruh data yang diuji [20].

$$Accuracy: \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

Precision merupakan metrik yang digunakan untuk mengukur ketepatan model dalam memprediksi data pada kelas positif. Nilai precision diperoleh dari perbandingan antara *True Positive* (TP) dengan jumlah *True Positive* (TP) dan *False Positive* (FP).

$$Precision: \frac{TP}{TP + F} \quad (3)$$

Recall merupakan metrik yang digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh data yang sebenarnya termasuk ke dalam kelas positif. Nilai recall dihitung berdasarkan perbandingan antara *True Positive* (TP) dengan jumlah *True Positive* (TP) dan *False Negative* (FN).

$$Recall: \frac{TP}{TP + FN} \quad (4)$$

F1-*score* merupakan metrik evaluasi yang menggabungkan nilai *precision* dan *recall* dalam bentuk rata-rata harmonik. Metrik ini digunakan untuk memberikan penilaian yang lebih seimbang.

$$F1:2 * \frac{Recall * Precision}{Recall + Precision} \quad (5)$$

2.7 Interpretasi Model XAI

Pada penelitian ini, interpretasi model dilakukan menggunakan pendekatan XAI dengan metode SHAP dan LIME. XAI adalah pendekatan yang bertujuan menjelaskan proses pengambilan keputusan model *Deep Learning* yang bersifat *black box*. Dengan menyediakan interpretasi terhadap hasil prediksi, XAI meningkatkan transparansi, kepercayaan, dan pemahaman pengguna terhadap sistem AI [21]. XAI dikembangkan untuk mengatasi keterbatasan model AI yang sulit diinterpretasikan (black box). Pendekatan ini membantu menjelaskan alasan di balik suatu prediksi dengan menunjukkan pengaruh setiap fitur terhadap keputusan model. XAI membantu pengguna memahami faktor-faktor yang memengaruhi hasil prediksi sehingga meningkatkan transparansi, kepercayaan, dan akuntabilitas sistem AI [22]. Dalam implementasinya, metode SHAP digunakan untuk menganalisis kontribusi fitur secara global, sedangkan LIME berfokus pada interpretasi lokal terhadap setiap hasil prediksi [23].

SHAP merupakan metode yang dikembangkan dari konsep *Shapley Value* untuk menjelaskan kontribusi setiap fitur terhadap hasil prediksi model. SHAP membuat perhitungan *Shapley Value* yang kompleks menjadi lebih efisien sehingga dapat diterapkan pada berbagai model *machine learning* dan *deep learning*, berikut persamaan SHAP [24].

$$\emptyset_i = \sum_{S \subseteq T \setminus \{i\}} \frac{|S|!\,(|T| - |S| - 1)!}{|T|!} \left(f_{S \cup i}(x_{S \cup \{i\}}) - f_S(x_S) \right) \quad (6)$$

Keterangan:

$\emptyset_i$ = nilai *Shapley (Shapley value)* untuk fitur ke-i
$f(x)$ = fungsi prediksi model (*prediction function*).
T = jumlah seluruh fitur (*total features*).
S = subset dari fitur yang tidak mencakup fitur ke-1

LIME adalah metode interpretasi yang menjelaskan prediksi model pada satu data tertentu dengan menggunakan model sederhana. LIME mengidentifikasi fitur-fitur yang paling memengaruhi hasil prediksi dan dapat digunakan pada berbagai jenis model AI, berikut persamaan LIME [23].

$$\xi(x) = arg\ min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (7)$$

Keterangan
f = model kompleks (*black-box model*) yang akan dijelaskan.
$g \in G$ = model yang dapat diinterpretasikan, misalnya model linear atau *decision tree*.
G = himpunan model yang dapat diinterpretasikan.
Ω(g) = kompleksitas model yang dapat diinterpretasikan.
X = instance atau data yang akan dijelaskan.
π_x = ukuran kedekatan (*proximity measure*)
$L(f, g, \pi_x)$ = fungsi kerugian (loss function) yang mengukur seberapa baik model sederhana mendekati prediksi model kompleks pada area lokal
$\xi(x)$ =penjelasan lokal (*local explanation*) yang dihasilkan oleh LIME.

SHAP digunakan untuk mengidentifikasi kontribusi setiap kata terhadap hasil prediksi model secara global maupun lokal, sedangkan LIME digunakan untuk memberikan penjelasan lokal terhadap prediksi pada suatu sampel data. Melalui kedua metode tersebut, proses

pengambilan keputusan oleh model BERT menjadi lebih transparan sehingga faktor-faktor yang memengaruhi hasil klasifikasi sentimen dapat dipahami dengan lebih baik

4. Hasil dan Pembahasan

Penelitian ini diimplementasikan menggunakan platform Google Colab dengan dukungan GPU NVIDIA Tesla T4 dan bahasa pemrograman Python. Proses pengolahan data, pelatihan model BERT, evaluasi kinerja, serta interpretasi hasil prediksi dilakukan menggunakan pustaka *Pandas, NumPy, Scikit-learn, PyTorch, Transformer*s, SHAP, dan LIME.

4.1 Hasil Data *Preprocessing*

Hasil *preprocessing* meliputi *data cleaning*, *text cleaning*, dan *label encoding*. Hasil proses *data cleaning* disajikan pada Tabel 3. Selanjutnya, hasil *text cleaning* dan *label encoding* ditampilkan pada Gambar 2 dan Gambar 3.

Tabel 3. Hasil *Data Cleaning*

Tahapan *Data Cleaning*	Jumlah Data	Data yang Berkurang
Data Awal	53.043	-
Setelah penghapusan data kosong (drop NA)	52.681	362
Setelah penghapusan data duplikat	51.073	1.608
Jumlah data akhir	**51.073**	**1.970**

Hasil *text cleaning* disajikan pada Gambar 2.

	statement	status	clean_text
0	oh my gosh	Anxiety	oh my gosh
1	trouble sleeping, confused mind, restless hear...	Anxiety	trouble sleeping confused mind restless heart ...
2	All wrong, back off dear, forward doubt. Stay ...	Anxiety	all wrong back off dear forward doubt stay in ...
3	I've shifted my focus to something else but I'...	Anxiety	ive shifted my focus to something else but im ...
4	I'm restless and restless, it's been a month n...	Anxiety	im restless and restless its been a month now ...
5	every break, you must be nervous, like somethi...	Anxiety	every break you must be nervous like something...
6	I feel scared, anxious, what can I do? And may...	Anxiety	i feel scared anxious what can i do and may my...
7	Have you ever felt nervous but didn't know why?	Anxiety	have you ever felt nervous but didnt know why

Gambar 2. Hasil *Text Cleaning*

Selanjutnya, hasil *label encoding* ditampilkan pada Gambar 3.

Label	Status	Jumlah Data
0	Anxiety	3617
1	Bipolar	2501
2	Depression	15087
3	Normal	16039
4	Personality disorder	895
5	Stress	2293
6	Suicidal	10641

Gambar 3. Hasil Label Encoding

Berdasarkan hasil tersebut, data telah melalui proses pembersihan data, pembersihan teks, dan pengubahan label ke dalam bentuk numerik sehingga siap digunakan pada tahap pembagian data dan tokenisasi BERT.

4.2 Hasil Pembagian Data

Hasil pembagian data menghasilkan 36.772 data pelatihan, 4.086 data validasi, dan 10.215 data pengujian. Rincian hasil pembagian data disajikan pada Tabel 4.

Tabel 4. Hasil Pembagian Data

Dataset	Jumlah Data
Data pelatihan (*training*)	36.772
Data validasi (*validation*)	4.086
Data pengujian (*testing*)	10.215
Total	**51.073**

Berdasarkan hasil tersebut, data pelatihan digunakan dalam proses *fine-tuning* model BERT, data validasi digunakan untuk memantau kinerja model selama pelatihan, sedangkan data pengujian digunakan untuk mengevaluasi kinerja akhir model.

4.3 Hasil Tokenisasi BERT

Hasil tokenisasi menunjukkan bahwa data teks telah berhasil dikonversi ke dalam bentuk numerik berupa input IDs dan *attention mask*. Proses tokenisasi menggunakan nilai *max_length* sebesar 256 dengan penerapan padding dan truncation. Contoh hasil tokenisasi disajikan pada Gambar 4.

```
Input IDs (20 token pertama):
[101, 2021, 2112, 1997, 2033, 10069, 2008, 2746, 2067, 2041, 2045, 5665, 2424, 1037, 5920, 2002, 2015, 5720, 2055, 2009]

Attention Mask (20 token pertama):
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

Gambar 4. Hasil Tokenisasi BERT

Berdasarkan hasil tersebut, data teks telah berhasil dikonversi ke dalam format numerik yang sesuai sebagai masukan pada proses pelatihan dan pengujian model BERT.

4.4 Hasil Pelatihan dan Evaluasi Model BERT

Hasil pelatihan model BERT selama tujuh epoch disajikan pada Gambar 5. Evaluasi dilakukan menggunakan nilai *training loss*, *validation loss*, *accuracy*, *precision*, *recall*, dan F1-*score*.

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	1.194168	0.652945	0.758199	0.804563	0.758199	0.762925
2	0.537222	0.573096	0.811062	0.832078	0.811062	0.812681
3	0.387103	0.572391	0.831865	0.840093	0.831865	0.831133
4	0.290787	0.630616	0.820362	0.834668	0.820362	0.821440
5	0.207865	0.675562	0.829418	0.831561	0.829418	0.829541
6	0.163184	0.733354	0.827460	0.829616	0.827460	0.827691
7	0.128592	0.755105	0.825991	0.829714	0.825991	0.826232

Gambar 5. Hasil Pelatihan Model BERT

Berdasarkan Gamabr 5, nilai training loss mengalami penurunan dari 1,194168 pada epoch pertama menjadi 0,128592 pada epoch ketujuh. Nilai accuracy tertinggi diperoleh pada epoch ketiga sebesar 0,831865, dengan nilai F1-*score* sebesar 0,831133. Hasil tersebut menunjukkan bahwa model BERT mampu mempelajari pola pada data pelatihan dengan baik. Selanjutnya, evaluasi model dilakukan menggunakan data pengujian untuk mengetahui kinerja model dalam mengklasifikasikan setiap kategori. Hasil evaluasi model disajikan pada Gambar 6.

```
================================================
Accuracy : 82%
================================================

Classification Report
```

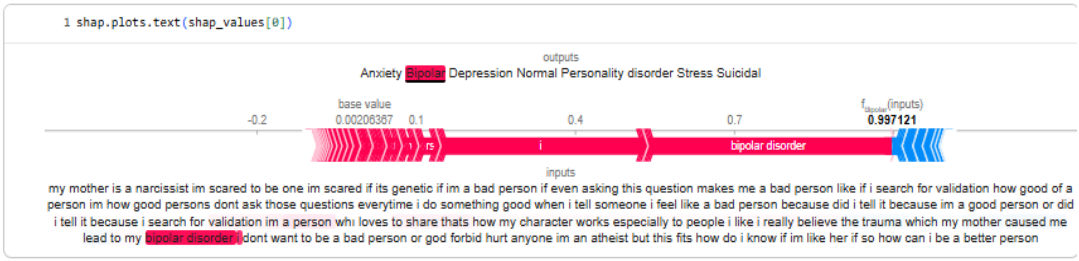
	precision	recall	f1-score	support
Anxiety	0.86	0.87	0.87	723
Bipolar	0.77	0.85	0.81	500
Depression	0.84	0.65	0.74	3018
Normal	0.95	0.96	0.95	3208
Personality disorder	0.65	0.68	0.66	179
Stress	0.69	0.75	0.72	459
Suicidal	0.67	0.84	0.75	2128
accuracy			0.82	10215
macro avg	0.78	0.80	0.78	10215
weighted avg	0.83	0.82	0.82	10215

Gambar 6. Hasil Evaluasi Model BERT

Berdasarkan Gambar 6, model BERT memperoleh nilai *accuracy* sebesar 82% pada data pengujian. Kelas Normal memperoleh nilai F1-*score* tertinggi sebesar 0,95, sedangkan kelas *Personality Disorder* memperoleh nilai F1-*score* terendah sebesar 0,66. Secara keseluruhan, nilai *weighted average* F1-score sebesar 0,82 menunjukkan bahwa model BERT memiliki kinerja yang baik dalam mengklasifikasikan tujuh kategori.

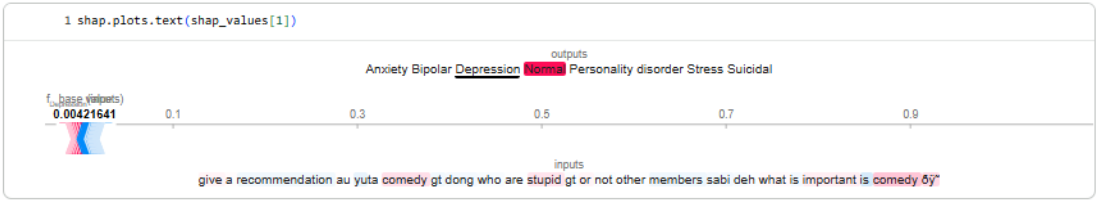
4.5 Hasil Interpretasi Menggunakan SHAP

Metode SHAP digunakan untuk menginterpretasikan kontribusi setiap kata terhadap hasil prediksi model BERT. Hasil interpretasi SHAP menunjukkan kata-kata yang memberikan pengaruh terhadap prediksi kelas pada masing-masing data. Dalam penelitian ini, interpretasi dilakukan pada dua sampel teks yang diprediksi sebagai kelas Bipolar dan Normal. Hasil interpretasi SHAP disajikan pada Gambar 7 dan Gambar 8.



Gambar 7. Hasil Interpretasi SHAP pada Prediksi Kelas Bipolar

Berdasarkan Gambar 7, model menghasilkan nilai prediksi sebesar 0,99 pada kelas Bipolar. Area berwarna merah terlihat lebih dominan dibandingkan area berwarna biru. Hal tersebut menunjukkan bahwa kontribusi kata atau bagian teks yang mendorong hasil klasifikasi menuju kelas Bipolar lebih besar daripada kontribusi yang mengurangi kecenderungan terhadap kelas tersebut. Kata atau frasa "*bipolar disorder*" memberikan kontribusi positif yang kuat terhadap hasil klasifikasi. Dominannya kontribusi positif tersebut menghasilkan nilai prediksi yang tinggi pada kelas Bipolar. Selain pada prediksi kelas Bipolar, interpretasi SHAP juga dilakukan pada sampel teks yang diprediksi sebagai kelas Normal. Hasil interpretasi tersebut disajikan pada Gambar 8.

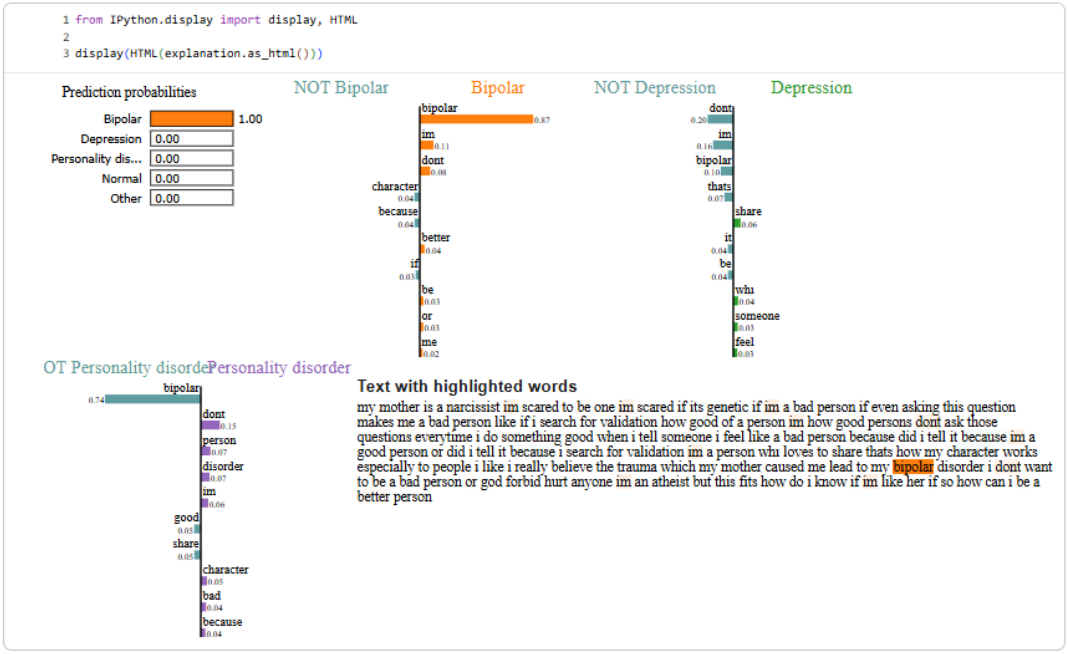


Gambar 8. Hasil Interpretasi SHAP pada Prediksi Kelas Normal

Berdasarkan Gambar 8, model memprediksi teks sebagai kelas Normal. Area berwarna merah menunjukkan kata atau bagian teks yang memberikan kontribusi positif dan mendorong hasil klasifikasi menuju kelas Normal, sedangkan area berwarna biru menunjukkan kontribusi yang mengurangi kecenderungan prediksi terhadap kelas tersebut. Kata "*comedy*" memberikan kontribusi positif terhadap hasil klasifikasi kelas Normal. Hasil interpretasi tersebut menunjukkan bahwa setiap kata pada teks memiliki kontribusi yang berbeda dalam menentukan hasil prediksi model.

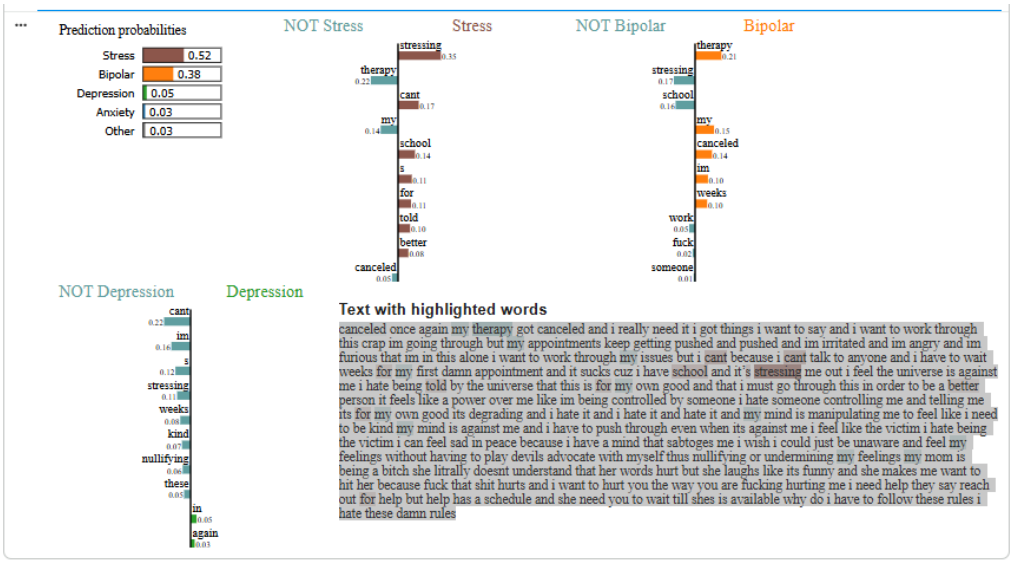
4.6 Hasil Interpretasi Menggunakan LIME

Metode LIME digunakan untuk menginterpretasikan kontribusi setiap kata terhadap hasil prediksi model BERT. Hasil interpretasi LIME menunjukkan kata-kata yang memberikan pengaruh terhadap prediksi kelas pada masing-masing data. Dalam penelitian ini, interpretasi dilakukan pada tiga sampel teks yang diprediksi sebagai kelas Bipolar, Stress, dan *Anxiety*. Hasil interpretasi LIME disajikan pada Gambar 9, Gambar 10, dan Gambar 11.



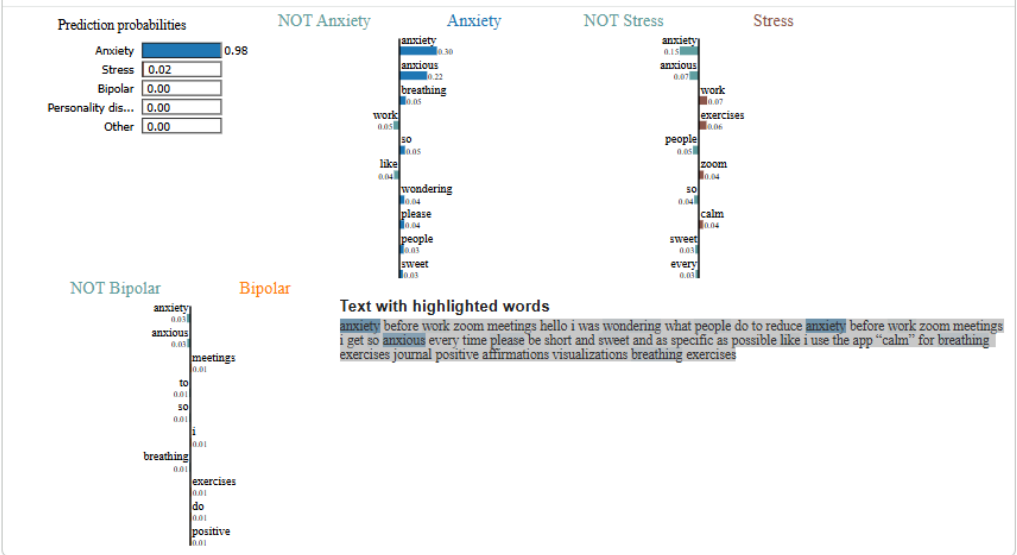
Gambar 9. Hasil Interpretasi LIME pada Prediksi Kelas Bipolar

Berdasarkan Gambar 9, model menghasilkan probabilitas sebesar 1,00 pada kelas Bipolar, sedangkan kelas *Depression*, *Personality Disorder*, Normal, dan *Other* masing-masing memperoleh probabilitas sebesar 0,00. Hasil tersebut menunjukkan bahwa model memiliki tingkat keyakinan yang sangat tinggi dalam mengklasifikasikan teks ke dalam kelas Bipolar. Pada visualisasi LIME, kata "bipolar" memberikan kontribusi paling besar terhadap hasil prediksi, diikuti oleh beberapa kata seperti "*im",* "*dont*", dan "*character*". Kata-kata tersebut memberikan kontribusi terhadap keputusan model dalam menentukan teks sebagai kelas Bipolar. Selanjutnya, hasil interpretasi LIME pada sampel teks yang diprediksi sebagai kelas Stress disajikan pada Gambar 10.



Gambar 10. Hasil Interpretasi LIME pada Prediksi Kelas Stress

Berdasarkan Gambar 10, model memberikan probabilitas sebesar 0,52 pada kelas Stress, sedangkan kelas Bipolar memperoleh probabilitas sebesar 0,38, *Depression* sebesar 0,05, Anxiety sebesar 0,03, dan Other sebesar 0,03. Hasil tersebut menunjukkan bahwa model memprediksi teks sebagai kelas Stress, tetapi memiliki tingkat keyakinan yang lebih rendah dibandingkan sampel pada kelas Bipolar dan *Anxiety*. Pada visualisasi LIME, kata "*stressing*" memberikan kontribusi paling besar terhadap prediksi kelas Stress, diikuti oleh kata "*therapy*", "*cant*", *"my*", *"school*", dan "*cancelled*". Sementara itu, adanya probabilitas yang cukup besar pada kelas Bipolar menunjukkan bahwa terdapat kata atau pola tertentu dalam teks yang juga memiliki karakteristik yang mendukung kelas tersebut. Hal ini menunjukkan bahwa LIME dapat membantu melihat adanya kedekatan karakteristik antar kelas pada suatu teks. Selanjutnya, hasil interpretasi LIME pada sampel teks yang diprediksi sebagai kelas *Anxiety* disajikan pada Gambar 11.



Gambar 11. Hasil Interpretasi LIME pada Prediksi Kelas *Anxiety*

Berdasarkan Gambar 11, model menghasilkan probabilitas sebesar 0,98 pada kelas Anxiety, sedangkan kelas Stress memperoleh probabilitas sebesar 0,02, sementara kelas Bipolar dan *Personality Disorder* memperoleh probabilitas sebesar 0,00. Hasil tersebut

menunjukkan bahwa model memiliki tingkat keyakinan yang tinggi dalam mengklasifikasikan teks ke dalam kelas Anxiety. Pada visualisasi LIME, kata "*anxiety*" memberikan kontribusi paling besar terhadap hasil prediksi, diikuti oleh kata "*anxious*" dan "*breathing*". Selain itu, kata "work", "*meetings*", "*people*", dan kata lainnya juga memberikan kontribusi terhadap hasil klasifikasi. Kata-kata yang berkaitan dengan kecemasan tersebut memperkuat keputusan model dalam menentukan teks sebagai kelas Anxiety.

Secara keseluruhan, hasil interpretasi LIME menunjukkan bahwa model BERT menggunakan kata-kata tertentu yang memiliki keterkaitan dengan karakteristik masing-masing kelas sebagai dasar dalam menghasilkan prediksi. LIME juga memperlihatkan bahwa tingkat keyakinan model dapat berbeda pada setiap sampel, seperti pada kelas Stress yang memiliki probabilitas 0,52 dan masih menunjukkan kecenderungan terhadap kelas Bipolar.

4.7 Analisis Kesalahan Klasifikasi

Hasil evaluasi menunjukkan bahwa model BERT belum mencapai tingkat akurasi maksimal, sehingga masih terdapat data yang tidak berhasil diklasifikasikan sesuai dengan label aktual. Dari 10.215 data pengujian, sebanyak 1.785 data mengalami kesalahan klasifikasi. Oleh karena itu, analisis kesalahan klasifikasi dilakukan untuk mengetahui pola dan faktor yang menyebabkan model menghasilkan prediksi yang berbeda dari label aktual. Beberapa sampel data mengalami kesalahan klasifikasi dipilih untuk dianalisis dan mewakili pola kesalahan yang ditemukan pada hasil pengujian. Hasil analisis kesalahan klasifikasi disajikan pada Tabel 5.

Tabel 5. Hasil Analisis Kesalahan Klasifikasi

Index	Label Aktual	Label Prediksi	Analisa Kesalahan
22	*Depression*	*Suicidal*	Teks memuat tekanan emosional yang berat serta istilah *suicide note*. Meskipun terdapat pernyataan bahwa penulis tidak ingin meninggal, keberadaan indikator tersebut dapat membuat teks memiliki kemiripan dengan kelas *Suicidal.*
25	*Suicidal*	*Depression*	Teks lebih banyak menggambarkan kehampaan, kesedihan, kehilangan emosi, dan kondisi emosional yang berkepanjangan sehingga memiliki kemiripan dengan karakteristik kelas *Depression*.
32	*Depression*	*Suicidal*	Teks secara eksplisit memuat istilah *suicidal ideations* dan pembahasan mengenai keinginan untuk memperoleh kedamaian, sehingga model lebih mengarah pada kelas *Suicidal.*
41	*Personality Disorder*	*Bipolar*	Teks membahas *anhedonia*, kurangnya dorongan, serta karakteristik psikologis yang dapat memiliki kemiripan dengan kategori kesehatan mental lainnya sehingga model memprediksi Bipolar.
45	*Suicidal*	*Depression*	Teks didominasi ungkapan kehilangan motivasi, keputusasaan, dan pertanyaan mengenai kehidupan sehingga memiliki kemiripan dengan karakteristik *Depression*.
48	*Anxiety*	*Depression*	Teks menyebutkan *Anxiety* dan *Depression*, tetapi pembahasan mengenai dampak kondisi terhadap pekerjaan, hubungan sosial, dan kehidupan sehari-hari lebih dominan sehingga model memprediksi *Depression*.
53	*Depression*	*Suicidal*	Teks memuat dorongan dan keinginan untuk melakukan bunuh diri sehingga karakteristik kuat yang mengarah pada kelas *Suicidal*.

Berdasarkan Tabel 5, kesalahan klasifikasi yang ditemukan menunjukkan bahwa sebagian besar kesalahan terjadi pada kelas yang memiliki karakteristik teks yang saling

berdekatan, terutama antara kelas Depression dan Suicidal. Beberapa teks pada kedua kelas sama-sama memuat ungkapan kesedihan, keputusasaan, tekanan emosional, kehilangan motivasi, serta pembahasan mengenai kematian atau bunuh diri. Kondisi tersebut menyebabkan batas karakteristik antarkelas menjadi kurang tegas, sehingga model dapat memprediksi kelas yang berbeda dari label aktual.

Selain kesalahan antara *Depression* dan *Suicidal*, ditemukan pula kesalahan klasifikasi pada pasangan kelas *Anxiety–Depression* dan *Personality Disorder–Bipolar*. Pada teks tertentu, satu dokumen dapat memuat lebih dari satu kondisi atau karakteristik psikologis sehingga informasi yang terdapat dalam teks tidak selalu merepresentasikan satu kategori secara eksklusif. Sebagai contoh, pada Index 48, teks menyebutkan *Anxiety* dan *Depression* secara bersamaan, tetapi pembahasan mengenai dampak kondisi terhadap pekerjaan, hubungan sosial, dan kehidupan sehari-hari lebih dominan sehingga model memprediksi kelas *Depression*. Hal ini menunjukkan bahwa kesalahan klasifikasi tidak hanya dipengaruhi oleh keberadaan kata tertentu, tetapi juga oleh konteks dan dominasi informasi yang terdapat dalam keseluruhan teks.

Temuan tersebut menunjukkan bahwa meskipun BERT mampu memahami hubungan kontekstual dalam teks, klasifikasi tujuh kategori kesehatan mental tetap menghadapi tantangan ketika karakteristik antar kelas memiliki kemiripan. Kesalahan prediksi yang ditemukan juga memperlihatkan bahwa teks kesehatan mental memiliki pola bahasa yang kompleks dan dapat memuat beberapa indikator kondisi psikologis secara bersamaan. Oleh karena itu, analisis kesalahan klasifikasi memberikan informasi tambahan mengenai kondisi ketika model mengalami kesulitan dalam membedakan kategori yang memiliki karakteristik linguistik serupa.

4.8 Pembahasan

Hasil penelitian menunjukkan bahwa model BERT mampu melakukan klasifikasi teks kesehatan mental dengan baik berdasarkan hasil evaluasi yang diperoleh. Model menghasilkan *accuracy* sebesar 82% dengan nilai *F1-score* sebesar 0,82. Hasil tersebut menunjukkan bahwa BERT mampu mempelajari pola bahasa yang terdapat dalam data dan menggunakannya untuk membedakan tujuh kategori kesehatan mental, yaitu *Anxiety*, Bipolar, *Depression*, Normal, *Personality Disorder*, *Stress*, dan *Suicidal*. Kemampuan tersebut berkaitan dengan karakteristik BERT sebagai model Transformer yang menghasilkan representasi kata secara kontekstual dengan mempertimbangkan hubungan antarkata dalam suatu teks [25]. Dengan demikian, hasil penelitian ini menunjukkan bahwa penggunaan BERT dapat menjadi pendekatan yang sesuai untuk melakukan klasifikasi teks kesehatan mental yang memiliki karakteristik bahasa yang kompleks.

Perbedaan performa antar kelas menunjukkan bahwa kemampuan model dalam mengenali setiap kategori tidak sama. Kelas Normal memperoleh nilai *F1-score* tertinggi sebesar 0,95, sedangkan kelas *Personality Disorder* memperoleh nilai *F1-score* terendah sebesar 0,66. Perbedaan tersebut dapat dipengaruhi oleh jumlah data pada masing-masing kelas serta karakteristik bahasa yang digunakan dalam teks. Kelas dengan jumlah data yang lebih banyak cenderung memiliki pola linguistik yang lebih banyak dipelajari oleh model, sedangkan kelas dengan jumlah data yang lebih sedikit dapat memiliki representasi pola yang lebih terbatas. Kondisi ini juga menunjukkan bahwa ketidakseimbangan data dapat menjadi salah satu faktor yang memengaruhi kemampuan model dalam mengenali kelas tertentu secara konsisten [26].

Selain evaluasi berdasarkan nilai metrik, hasil penelitian juga menunjukkan bahwa masih terdapat data yang mengalami kesalahan klasifikasi. Dari 10.215 data pengujian, sebanyak 1.785 data tidak berhasil diklasifikasikan sesuai dengan label aktual. Kesalahan tersebut menunjukkan bahwa meskipun model memiliki kinerja yang baik secara keseluruhan, masih terdapat teks yang memiliki karakteristik linguistik yang beririsan antar kelas. Kondisi ini terutama terlihat pada kelas *Depression* dan *Suicidal*, karena teks pada kedua kategori dapat sama-sama mengandung ungkapan kesedihan, kehampaan, keputusasaan, tekanan emosional, maupun pembahasan mengenai kehidupan [27].

Analisis terhadap beberapa sampel kesalahan klasifikasi memperlihatkan pola tersebut. Pada indeks 22, data dengan label aktual *Depression* diprediksi sebagai *Suicidal*. Teks tersebut membahas tekanan emosional, kesendirian, serta terdapat istilah suicide note. Keberadaan istilah tersebut memberikan karakteristik yang kuat terhadap kelas *Suicidal*, meskipun dalam teks terdapat pernyataan bahwa penulis tidak ingin meninggal. Sebaliknya, pada indeks 25,

data dengan label aktual *Suicidal* diprediksi sebagai *Depression*. Teks tersebut lebih dominan menggambarkan kehampaan, kesedihan, dan kehilangan emosi sehingga memiliki kemiripan dengan karakteristik kelas *Depression*. Temuan ini menunjukkan bahwa satu teks dapat memiliki karakteristik yang berdekatan dengan lebih dari satu kategori sehingga batas antar kelas tidak selalu bersifat tegas [28].

Kesalahan klasifikasi juga ditemukan pada indeks 32, yaitu data dengan label aktual *Depression* yang diprediksi sebagai *Suicidal.* Pada teks tersebut terdapat istilah *suicidal ideations* yang secara langsung berkaitan dengan kelas *Suicidal.* Kondisi serupa terlihat pada indeks 53, ketika data berlabel *Depression* diprediksi sebagai Suicidal karena teks mengandung pernyataan mengenai dorongan dan keinginan untuk melakukan bunuh diri. Hal tersebut menunjukkan bahwa keberadaan kata atau frasa tertentu dapat menjadi indikator linguistik yang kuat dalam proses klasifikasi. Namun, keputusan model tidak hanya dipengaruhi oleh satu kata, melainkan juga oleh konteks keseluruhan teks yang diproses oleh BERT.

Pada indeks 41, data dengan label aktual *Personality Disorder* diprediksi sebagai Bipolar. Teks tersebut membahas anhedonia, kurangnya dorongan, serta karakteristik psikologis tertentu yang dapat memiliki kemiripan dengan kategori kesehatan mental lainnya. Kesalahan ini menunjukkan bahwa beberapa kategori memiliki karakteristik linguistik yang sulit dibedakan hanya berdasarkan teks, terutama ketika suatu teks membahas gejala atau pengalaman psikologis yang dapat muncul pada lebih dari satu kondisi. Sementara itu, pada indeks 48, data dengan label aktual *Anxiety* diprediksi sebagai *Depression*. Teks tersebut secara eksplisit menyebutkan anxiety dan depression, tetapi pembahasan mengenai dampak kondisi terhadap pekerjaan, hubungan sosial, kehidupan sehari-hari, dan perasaan tertekan lebih dominan. Hal tersebut dapat menyebabkan model memilih kelas Depression. Temuan ini memperlihatkan bahwa kesalahan klasifikasi tidak selalu disebabkan oleh ketidakmampuan model mengenali istilah tertentu, tetapi juga dapat terjadi karena adanya kemiripan karakteristik linguistik antar kelas dalam suatu teks [29].

Hasil analisis kesalahan klasifikasi tersebut memperlihatkan bahwa salah satu faktor yang memengaruhi kesalahan prediksi adalah kemiripan karakteristik antar kelas. Teks kesehatan mental pada media sosial cenderung bersifat panjang, tidak terstruktur, menggunakan bahasa informal, serta dapat membahas beberapa kondisi emosional secara bersamaan. Karakteristik tersebut menyebabkan satu teks dapat mengandung indikator yang berkaitan dengan beberapa kategori sekaligus. Dengan demikian, kesalahan klasifikasi sebanyak 1.785 data tidak semata-mata menunjukkan kelemahan model, tetapi juga menunjukkan kompleksitas karakteristik data teks kesehatan mental yang digunakan dalam penelitian.

Untuk mengetahui alasan di balik keputusan model, penelitian ini selanjutnya menggunakan pendekatan *Explainable Artificial Intelligence* (XAI), yaitu SHAP dan LIME. Hasil interpretasi SHAP pada sampel kelas Bipolar menunjukkan bahwa frasa "*bipolar disorder*" memberikan kontribusi positif yang kuat sehingga menghasilkan nilai prediksi sebesar 0,99. Pada sampel kelas Normal, kata "*comedy*" memberikan kontribusi terhadap hasil klasifikasi. Hasil tersebut menunjukkan bahwa SHAP dapat memberikan informasi mengenai kontribusi fitur terhadap keputusan model [30].

Hasil interpretasi LIME menunjukkan pola yang serupa. Pada kelas Bipolar, kata "bipolar" menjadi salah satu fitur dengan kontribusi terbesar dan menghasilkan probabilitas prediksi sebesar 1,00. Pada kelas *Anxiety*, kata "*anxiety*", "*anxious*", dan "*breathing*" menjadi kata yang dominan sehingga model menghasilkan probabilitas sebesar 0,98. Sementara itu, pada kelas Stress, kata "*stressing*", "*therapy*", "*school*", dan "*cancelled*" memberikan kontribusi terhadap prediksi kelas Stress. Namun, probabilitas kelas Stress hanya sebesar 0,52, sedangkan kelas Bipolar memperoleh probabilitas sebesar 0,38. Hasil tersebut menunjukkan bahwa tingkat keyakinan model dapat berbeda pada setiap teks, terutama ketika teks memiliki karakteristik yang beririsan antar kelas. Penggunaan SHAP dan LIME memberikan sudut pandang yang saling melengkapi dalam memahami keputusan model. SHAP digunakan untuk menunjukkan kontribusi fitur terhadap hasil prediksi berdasarkan pendekatan nilai Shapley, sedangkan LIME memberikan penjelasan lokal terhadap prediksi pada sampel tertentu [23]. Kesamaan pola yang terlihat pada beberapa hasil interpretasi, seperti kuatnya kontribusi kata "bipolar" pada kelas Bipolar dan kata "*anxiety*" pada kelas *Anxiety*, menunjukkan bahwa keputusan model memiliki dasar linguistik yang dapat diinterpretasikan. Dengan demikian, penggunaan XAI dapat membantu mengurangi sifat black-box pada model BERT dan

memberikan pemahaman mengenai alasan suatu teks diklasifikasikan ke dalam kategori tertentu.

Jika dibandingkan dengan penelitian terdahulu, hasil penelitian ini menunjukkan bahwa penggunaan BERT mampu menghasilkan kinerja klasifikasi yang baik pada data kesehatan mental. Rahayu dkk. memperoleh *accuracy* sebesar 95,7% menggunakan *Random Forest*, sedangkan Jumaryadi dkk. memperoleh *accuracy* sebesar 78,69% menggunakan SVM [5][6]. Sementara itu, Rehman dkk. melaporkan *accuracy* sebesar 95,71% menggunakan BERT pada dataset *Sentiment Analysis for Mental Health* [7]. Penelitian ini memperoleh *accuracy* sebesar 82% dan *F1-score* sebesar 0,82. Meskipun nilai *accuracy* penelitian ini lebih rendah dibandingkan beberapa penelitian terdahulu, hasil tersebut tetap menunjukkan bahwa BERT mampu melakukan klasifikasi terhadap tujuh kategori kesehatan mental dengan kinerja yang baik. Perbedaan hasil tersebut dapat dipengaruhi oleh karakteristik *dataset*, jumlah kategori yang diklasifikasikan, distribusi data antar kelas, serta konfigurasi dan tahapan pelatihan model. Penelitian ini tidak hanya mengevaluasi kinerja klasifikasi, tetapi juga menganalisis kesalahan prediksi dan memberikan interpretasi menggunakan SHAP dan LIME. Dengan demikian, penelitian ini memiliki fokus yang berbeda dari penelitian terdahulu karena tidak berhenti pada pengukuran performa model, tetapi juga berupaya menjelaskan faktor linguistik yang memengaruhi keputusan BERT serta mengidentifikasi pola kesalahan klasifikasi.

Dari sisi kontribusi ilmiah, penelitian ini memberikan pendekatan yang menggabungkan kemampuan klasifikasi BERT dengan interpretabilitas XAI pada analisis teks kesehatan mental. Hasil interpretasi dapat membantu peneliti memahami pola bahasa yang digunakan model ketika membedakan kategori kesehatan mental. Selain itu, analisis kesalahan klasifikasi memberikan informasi tambahan mengenai karakteristik teks yang sulit dibedakan oleh model. Pendekatan tersebut dapat menjadi dasar untuk pengembangan penelitian selanjutnya, khususnya dalam meningkatkan kualitas data, menangani ketidakseimbangan kelas, serta mengembangkan metode klasifikasi yang lebih mampu membedakan kategori dengan karakteristik linguistik yang berdekatan.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, model masih menghasilkan 1.785 kesalahan klasifikasi dari 10.215 data pengujian sehingga menunjukkan bahwa model belum mampu memberikan prediksi yang sempurna. Kedua, ketidakseimbangan jumlah data antar kelas berpotensi memengaruhi kemampuan model dalam mengenali kelas dengan jumlah data yang lebih sedikit. Ketiga, interpretasi SHAP dan LIME dalam penelitian ini dilakukan pada beberapa sampel sehingga belum menggambarkan keseluruhan perilaku model pada seluruh data pengujian. Oleh karena itu, penelitian selanjutnya dapat mempertimbangkan penggunaan strategi penanganan ketidakseimbangan data, pengujian model Transformer lainnya, serta analisis XAI pada jumlah sampel yang lebih besar agar interpretasi perilaku model dapat dilakukan secara lebih komprehensif.

Secara keseluruhan, hasil penelitian menunjukkan bahwa BERT mampu melakukan klasifikasi teks kesehatan mental dengan kinerja yang baik, sedangkan SHAP dan LIME dapat digunakan untuk memberikan interpretasi terhadap keputusan model. Analisis kesalahan klasifikasi juga menunjukkan bahwa sebagian kesalahan terjadi karena adanya kemiripan karakteristik antar kategori, khususnya *Depression* dan *Suicidal*. Dengan demikian, kombinasi BERT dan XAI tidak hanya menghasilkan model klasifikasi, tetapi juga memberikan pemahaman mengenai faktor linguistik yang memengaruhi prediksi serta keterbatasan model dalam menghadapi teks kesehatan mental yang kompleks.

5. Simpulan

Penelitian ini menunjukkan bahwa model BERT mampu melakukan klasifikasi teks kesehatan mental ke dalam tujuh kelas, yaitu *Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress*, dan *Suicidal.* Hasil evaluasi pada 10.215 data pengujian menunjukkan bahwa model memperoleh nilai *accuracy* sebesar 82% dan F1-*score* sebesar 0,82. Hasil tersebut menunjukkan bahwa model BERT memiliki kinerja yang baik dalam mengklasifikasikan teks berdasarkan kategori kondisi kesehatan mental. Meskipun memperoleh kinerja yang baik, model masih mengalami kesalahan klasifikasi sebanyak 1.785 data. Hasil analisis terhadap beberapa sampel menunjukkan bahwa kesalahan prediksi terutama terjadi pada teks yang memiliki karakteristik linguistik yang beririsan antar kelas, khususnya antara kelas *Depression* dan *Suicidal*. Beberapa teks juga mengandung indikator dari lebih dari satu kategori sehingga model cenderung memilih kelas berdasarkan pola bahasa yang lebih dominan. Temuan ini

menunjukkan bahwa kompleksitas dan kemiripan karakteristik teks kesehatan mental menjadi salah satu faktor yang memengaruhi kesalahan klasifikasi.

Penerapan XAI menggunakan SHAP dan LIME berhasil menginterpretasikan keputusan model BERT. SHAP menunjukkan kontribusi kata atau frasa, seperti “bipolar disorder” terhadap prediksi kelas Bipolar, sedangkan LIME menunjukkan pengaruh kata seperti “bipolar”, “anxiety”, “anxious”, dan “breathing” pada sampel tertentu. Hasil ini menunjukkan bahwa SHAP dan LIME dapat meningkatkan transparansi dan interpretabilitas model BERT. Penelitian selanjutnya dapat membandingkan BERT dengan Transformer lainnya, menggunakan dataset yang lebih beragam dan seimbang, serta memperluas analisis XAI pada lebih banyak sampel.

Daftar Referensi

[1] D. L. Pardede *et al.*, “Kehidupan Sehat Dan Sejahtera : Analisis Peran Kesehatan Mentaldalam Peningkatan Produktivitas Generasi Z,” *J. Pengabdi. Kpd. Masy. MAJU UDA Univ.*, vol. 5, no. 3, pp. 43–52, 2024.

[2] D. K. Ningrum and A. M. Ismawardi, “Efektivitas Algoritma Kecerdasan Buatan Dalam Implementasi Kesehatan Mental : Systematic Literature Review,” *J. Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 689–698, 2025.

[3] Kementrian Kesehatan Republik Indonesia, “Hasil Riskesdas 2013,” *Expert Opin. Investig. Drugs*, vol. 7, no. 5, pp. 803–809, 2013, doi: 10.1517/13543784.7.5.803.

[4] Kemenkes, “Survei Kesehatan Indonesia 2023 (SKI),” *Kemenkes*, p. 235, 2023.

[5] K. Rahayu, V. Fitria, D. Septhya, R. Rahmaddeni, and L. Efrizoni, “Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 108–114, 2023, doi: 10.57152/malcom.v3i2.780.

[6] Y. Jumaryadi, R. Fajriah, U. Salamah, B. Priambodo, and A. Lystha, “Machine Learning Approaches to Sentiment Analysis of Mental Health Discussions on Platform X,” *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 13, no. 2, pp. 43–54, 2025, doi: 10.33558/piksel.v13i2.11350.

[7] A. Rehman, M. Ahmed, H. U. Khan, A. Bukhari, A. Daud, and H. Dawood, “Mental Health Sentiment Analysis: Exploring an Optimized BERT with Deep Encodings,” *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 26242–26248, 2025, doi: 10.48084/etasr.10469.

[8] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” *J. Big Data*, vol. 9, no. 1, pp. 1–21, 2022, doi: 10.1186/s40537-022-00564-9.

[9] V. Hassija *et al.*, “Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence,” *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[10] M. D. Salman, N. R. Pratama, and M. N. F. A, “Comparison of K-Means and K-Medoids Clustering Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure Perbandingan Kinerja Algoritma Clustering K-Means dan K-Medoids dalam Pengelompokan Sekolah di,” *Malcom*, vol. 5, no. 3, pp. 797–806, 2025, [Online]. Available: https://www.journal.irpi.or.id/index.php/malcom/article/view/1950

[11] A. A. Bastian, H. H. Handayani, D. Wahiddin, and T. Rohana, “Implementasi Algoritma Support Vector Regression dan Linear Regression Untuk Prediksi Harga Rumah,” *Progresif J. Ilm. Komput.*, vol. 20, no. 2, p. 961, 2024, doi: 10.35889/progresif.v20i2.2191.

[12] A. F. Y. Khainur, T. Yuares, M. H. Fathurrohman, Widianingsih, and C. Rozikin, “Analisis Komparatif Efektivitas Pipeline Data Cleaning Berbasis Aturan Dan Lemmatisasi Untuk Klasifikasi Sentimen,” *J. TIMES*, vol. 14, no. 2, pp. 141–149, 2025, doi: 10.51351/jtm.14.2.2025890.

[13] L. Denna, H. Monasari, and N. Pratiwi, “Klasifikasi Stunting Pada Balita Menggunakan Algoritma Decision Tree Pada Machine Learning,” *Decod. Pendidik. Teknol. Inf.*, vol. 5, no. 2, pp. 416–428, 2025.

[14] N. Karimah, “Multi-Aspect Sentiment Analysis Pada Review Film Menggunakan Metode Bidirectional Encoder Representations From Transformers (BERT),” *Komputika J. Sist.*

Komput., vol. 13, no. 1, pp. 63–72, 2024, doi: 10.34010/komputika.v13i1.11098.
[15] A. O. Ibitoye, O. O. Oladimeji, I. O. Olaleye, and O. N. Emuoyibofarhe, “A context-aware BERT framework for detecting and modeling the mental health impact of online toxic language,” *Data Sci. Manag.*, vol. 9, no. 2, p. 100175, 2026, doi: 10.1016/j.dsm.2025.12.002.
[16] S. L. Mirtaheri, A. Pugliese, N. Movahed, and R. Shahbazian, “A comparative analysis on using GPT and BERT for automated vulnerability scoring,” *Intell. Syst. with Appl.*, vol. 26, no. 0, p. 200515, 2025, doi: 10.1016/j.iswa.2025.200515.
[17] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, “A Novel Named Entity Recognition approach of Indonesian fake news using part of speech and BERT model on presidential election,” *Int. J. Inf. Manag. Data Insights*, vol. 5, no. 2, p. 100354, 2025, doi: 10.1016/j.jjimei.2025.100354.
[18] D. Chi, T. Huang, Z. Jia, and S. Zhang, “Research on sentiment analysis of hotel review text based on BERT-TCN-BiLSTM-attention model,” *Array*, vol. 25, no. 0, p. 100378, 2025, doi: 10.1016/j.array.2025.100378.
[19] A. R. Hanum *et al.*, “ANALISIS KINERJA ALGORITMA KLASIFIKASI TEKS BERT DALAM MENDETEKSI BERITA HOAKS,” *J. Teknol. dan Ilmu Komput.*, vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik2024118093.
[20] A. Ripa’i, F. Santoso, and F. Lazim, “Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT Asep,” vol. 8, no. 3, pp. 1749–1758, 2024.
[21] N. Al-Ansari, D. Al-Thani, and M. Bahameish, “Explaining in context: Perceived informativeness of Explainable Artificial Intelligence (XAI) in Arabic hate speech detection,” *Comput. Hum. Behav. Reports*, vol. 22, no. 0, p. 101083, 2026, doi: 10.1016/j.chbr.2026.101083.
[22] U. Lagap, S. Ghaffarian, S. Gelinas-Gagne, J. Jilma, Z. Liu, and Z. Luo, “Towards reliable deep learning for post-disaster damage Assessment: An XAI-based evaluation,” *Int. J. Disaster Risk Reduct.*, vol. 130, no. 0, p. 105839, 2025, doi: 10.1016/j.ijdrr.2025.105839.
[23] A. Budhkar, Q. Song, J. Su, and X. Zhang, “Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics,” *Comput. Struct. Biotechnol. J.*, vol. 27, no. 0, pp. 346–359, 2025, doi: 10.1016/j.csbj.2024.12.027.
[24] S. Deng, C. Aldrich, X. Liu, and F. Zhang, “Explainability in Reservoir Well-logging Evaluation: Comparison of Variable Importance Analysis with Shapley Value Regression, SHAP and LIME,” *IFAC-PapersOnLine*, vol. 58, no. 22, pp. 66–71, 2024, doi: 10.1016/j.ifacol.2024.09.292.
[25] H. A. P. Mitchella Sinta Larasati, Suryasatriya Trihandaru, “Sistem Otomatis Klasifikasi Bukti Pembayaran Menggunakan Ocr Dan Embedding Bert Dengan Pendekatan Multi-Model Pembelajaran Mesin”, *Pendas: Jurnal Ilmiah Pendidikan Dasar,* vol. 11, no. 01, pp. 94–106, 2026.
[26] C. Sintiya, G. H. Hutagaol, D. Bate`e, and S. Irviantina, “Evaluasi Teknik Resampling untuk Class Balancing dalam Analisis Sentimen Kesehatan Mental Berbasis Bi-LSTM,” *J. Sifo Mikroskil*, vol. 26, no. 2, pp. 257–274, 2025, doi: 10.55601/jsm.v26i2.1799.
[27] D. Phiri, F. Makowa, V. L. Amelia, Y. V. A. Phiri, L. P. Dlamini, and M. H. Chung, “Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis,” *J. Med. Internet Res.*, vol. 27, no. 0, pp. 1–17, 2025, doi: 10.2196/59002.
[28] İ. Baydili, B. Tasci, and G. Tasci, “Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection,” *Behav. Sci. (Basel).*, vol. 15, no. 3, p. 352, 2025, doi: 10.3390/bs15030352.
[29] F. A. Wagay and Jahiruddin, “An efficient approach to detecting mental illness on online social media platforms by using multidimensional textual features,” *Acta Psychol. (Amst).*, vol. 262, no. 0, p. 106191, 2026, doi: 10.1016/j.actpsy.2025.106191.
[30] T. Shaik, X. Tao, H. Xie, L. Li, N. Higgins, and J. D. Velásquez, “Towards Transparent Deep Learning in Medicine: Feature Contribution and Attention Mechanism-Based Explainability,” *Human-Centric Intell. Syst.*, vol. 5, no. 2, pp. 209–229, 2025, doi: 10.1007/s44230-025-00104-7.