Deteksi Dini Gangguan Kesehatan Mental di Media Sosial X menggunakan *IndoBERTweet* dan Teknik *Fine-Tuning* Adaptif

Desi Dwi Purwasih
Program Studi Informatika PJJ
Fakultas Informatika
Telkom University
Bandung, Indonesia
desidwipurwasih@student.telkomuniversity.ac.id

Fitriyani
Program Studi Informatika PJJ
Fakultas Informatika
Telkom University
Bandung, Indonesia
fitriyani@telkomuniversity.ac.id

Kesehatan mental saat ini menjadi isu global yang terus meningkat, terutama di kalangan remaja dan pengguna media sosial. *Platform* X merupakan salah satu *platform* untuk mengekspresikan opini serta pengalaman pribadi. Volume data tekstual dalam jumlah besar menyebabkan identifikasi gangguan mental manual kurang efektif. Oleh karena itu, penelitian ini menggunakan model IndoBERTweet untuk mendeteksi tiga kategori gangguan kesehatan mental, yaitu *Depression*, *Social Anxiety*, dan *Obsessive-Compulsive Disorder* (OCD), pada *tweet* berbahasa Indonesia. Dataset diperoleh melalui proses *crawling* dan diproses lebih lanjut melalui tahap *preprocessing*. Model dilatih menggunakan empat skema, yaitu *fine-tuning* standar dan adaptif, masing-masing dengan dan tanpa *undersampling*.Hasil penelitian yang telah dilakukan, model IndoBERTweet yang dilatih menggunakan teknik *fine-tuning* adaptif dengan penerapan *undersampling* menghasilkan kinerja paling seimbang antar kelas dengan akurasi 87,26% serta nilai *precision*, *recall*, dan *F1-score* yang relatif konsisten. Sementara itu, *fine-tuning* standar tanpa penerapan *undersampling* mencapai akurasi tertinggi sebesar 98,31%, namun menunjukkan kecenderungan bias terhadap kelas mayoritas. Hasil ini menunjukkan bahwa kombinasi *fine-tuning* adaptif dengan *undersampling* data lebih efektif untuk klasifikasi multi-kelas gangguan kesehatan mental.

Kata kunci — kesehatan mental, *fine tuning* adaptif, klasifikasi teks, IndoBERTweet, *undersampling*

I. PENDAHULUAN

Gangguan kesehatan mental merupakan isu global yang semakin mendapat perhatian dalam beberapa dekade terakhir. Kesehatan mental adalah kondisi psikologis dan emosional individu yang memungkinkan perkembangan fisik, intelektual, dan emosional secara optimal serta kemampuan beradaptasi terhadap tekanan sosial [1], [2]. *Menurut World Health Organization* (WHO), sekitar satu dari tujuh anak berusia 10–19 tahun mengalami gangguan kesehatan mental, termasuk depresi dan gangguan kecemasan [3]. Di Indonesia, hasil *Indonesia National Adolescent Mental Health Survey* (I-NAMHS) tahun 2022 menunjukkan bahwa sekitar 15,5 juta remaja atau 34,9% dari total populasi remaja mengalami gangguan kesehatan mental [4]. Tingginya prevalensi ini menunjukkan bahwa gangguan kesehatan mental merupakan persoalan signifikan yang membutuhkan penanganan menyeluruh. Namun, stigma negatif di masyarakat serta keterbatasan akses terhadap layanan kesehatan mental, khususnya di wilayah terpencil, masih menjadi tantangan utama dalam deteksi dan penanganan dini [5], [6]

Seiring perkembangan teknologi digital, media sosial telah menjadi bagian integral dalam kehidupan masyarakat, khususnya generasi muda. Melalui platform X, pengguna mengekspresikan pikiran, perasaan, dan pengalaman pribadi melalui unggahan teks, yang dapat mencerminkan kondisi psikologis mereka [6] Volume unggahan yang besar membuat identifikasi manual menjadi tidak praktis, sehingga analisis otomatis berbasis teknologi menjadi pendekatan yang diperlukan untuk mendukung deteksi dini gangguan kesehatan mental.

Berbagai penelitian telah memanfaatkan pendekatan *machine learning* dan *Natural Language Processing* (NLP) untuk mendeteksi gangguan kesehatan mental melalui analisis teks media sosial. Penelitian oleh Fikri Ilham dan Warih Maharani menggunakan model *BERT* untuk mendeteksi depresi dari *tweet* berbahasa Indonesia, dengan akurasi 71%, yang masih berada pada tingkat sedang dan membutuhkan optimasi lebih lanjut [7]. Penelitian lain menunjukkan bahwa penggunaan *IndoBERTweet* mampu meningkatkan akurasi hingga 82%, namun masih menghadapi kesalahan klasifikasi false positive yang berpotensi menimbulkan salah interpretasi dalam konteks medis [8]. Pendekatan lain menggunakan algoritma konvensional seperti *K-Nearest Neighbor* dan *Random Forest* menunjukkan performa memadai, dengan akurasi antara 83% hingga 90%, tetapi masih belum mempertimbangkan secara menyeluruh karakteristik bahasa informal dan kontekstual yang umum di media sosial [9]. Sementara itu, penelitian berbasis *Naive Bayes* dan analisis sentimen menunjukkan akurasi tinggi, namun terbatas pada klasifikasi sentimen dan belum diarahkan secara spesifik untuk deteksi gangguan kesehatan mental [10]. *IndoBERTweet* yang dilatih menggunakan *domain-adaptive pretraining* terbukti efektif untuk berbagai tugas NLP, namun penerapannya untuk klasifikasi gangguan kesehatan mental dengan pendekatan *fine-tuning* masih sangat terbatas [11].

Berdasarkan kajian terhadap penelitian sebelumnya, masih terdapat sejumlah celah penelitian yang belum terjawab secara komprehensif. Sebagian besar penelitian berfokus pada satu jenis gangguan kesehatan mental, khususnya depresi, sehingga belum mencakup klasifikasi

multi-kelas untuk gangguan lain seperti *social anxiety* dan *obsessive-compulsive disorder* (OCD). Selain itu, dominasi dataset berbahasa Inggris menyebabkan keterbatasan relevansi terhadap karakteristik bahasa Indonesia yang bersifat informal dan kontekstual. Dari sisi metodologi, penerapan pendekatan *adaptive fine-tuning* pada model *IndoBERTweet* untuk klasifikasi gangguan kesehatan mental berbasis teks berbahasa Indonesia masih relatif sedikit, serta permasalahan ketidakseimbangan data belum banyak ditangani secara optimal.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model klasifikasi teks berbasis *IndoBERTweet* dengan pendekatan *adaptive fine-tuning* untuk mendeteksi indikasi gangguan kesehatan mental secara multi-kelas, meliputi *Depression, Social Anxiety*, dan *Obsessive-Compulsive Disorder* (OCD), dari unggahan pengguna di platform X. Penelitian ini juga menerapkan teknik *undersampling* untuk mengatasi ketidakseimbangan distribusi data serta mengevaluasi performa model menggunakan berbagai metrik evaluasi, termasuk *accuracy*, *precision*, *recall*, dan *F1-score*. Diharapkan bahwa hasil penelitian ini dapat memberikan kontribusi pada pengembangan sistem deteksi dini gangguan kesehatan mental berbasis media sosial yang lebih akurat, kontekstual, dan representatif, sekaligus mendukung upaya peningkatan kesadaran serta penanganan kesehatan mental di Indonesia.

II. KAJIAN TEORI

A. Kesehatan Mental

Kesehatan mental merujuk pada kondisi keseimbangan psikologis, emosional, dan sosial yang memengaruhi proses berpikir, perasaan, serta perilaku individu dalam aktivitas sehari-hari. Faktor-faktor eksternal seperti tekanan akademik maupun lingkungan kerja dapat memengaruhi kondisi psikologis seseorang dan berpotensi menimbulkan gangguan mental apabila tidak ditangani secara tepat. Gangguan kesehatan mental dapat berdampak negatif terhadap produktivitas, motivasi, serta kualitas kesehatan secara keseluruhan, sehingga penanganan pencegahan dan intervensi yang berkelanjutan menjadi krusial [12].

Dalam konteks penelitian ini, deteksi gangguan kesehatan mental difokuskan pada tiga kategori utama, yaitu depresi, gangguan kecemasan sosial (*social anxiety*), dan *obsessive-compulsive disorder* (OCD). Depresi ditandai oleh perasaan sedih yang berkepanjangan dan menurunnya minat atau kesenangan pada aktivitas sehari-hari. Gangguan kecemasan sosial ditandai oleh rasa takut atau cemas yang berlebihan dalam interaksi sosial, sementara OCD ditandai oleh munculnya pikiran obsesif dan perilaku kompulsif yang berulang dan mengganggu fungsi sehari-hari.

B. Media Sosial

Media sosial merupakan platform yang memungkinkan individu untuk berinteraksi, berbagi informasi, serta menghasilkan konten secara mandiri. Platform X menyediakan ruang bagi pengguna untuk mengekspresikan pengalaman pribadi, opini, serta isu-isu sosial, termasuk terkait kesehatan mental [5]. Peran media sosial dalam konteks kesehatan mental bersifat ganda; di satu sisi, platform ini dapat meningkatkan kesadaran, mengurangi stigma, serta menyediakan dukungan sosial, namun di sisi lain, pengguna juga rentan mengalami dampak psikologis negatif seperti kecemasan dan depresi akibat praktik *cyberbullying,* perbandingan sosial, maupun penyebaran informasi yang tidak akurat.

Dalam berbagai penelitian, Platform X telah banyak dimanfaatkan sebagai sumber data karena berfungsi sebagai ruang diskusi isu-isu aktual dan memiliki fitur *trending topic* yang mencerminkan perhatian publik secara luas[13]. Struktur data yang relatif sederhana, dukungan antarmuka pemrograman aplikasi (API), serta aksesibilitas yang tinggi menjadikan platform ini sesuai untuk proses ekstraksi data dan analisis berbasis natural language processing (NLP) dalam upaya mendeteksi indikasi gangguan kesehatan mental[14].

C. *Deep Learning*

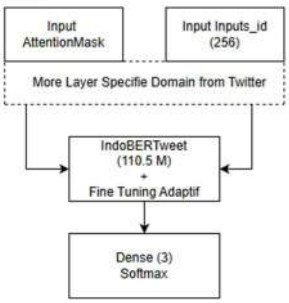
Deep Learning merupakan cabang dari *machine learning* yang memanfaatkan jaringan saraf tiruan berlapis untuk memproses dan mengekstraksi representasi data secara hierarkis, dari pola sederhana hingga kompleks. Pendekatan ini memungkinkan model untuk melakukan berbagai tugas seperti klasifikasi, prediksi, dan pengenalan pola [15].

D. *Natural Language Processing* (NLP)

Natural Language Processing merupakan cabang *Artificial Intelligence* yang berfokus pada pemrosesan dan analisis data berbasis bahasa alami secara otomatis. NLP menyediakan berbagai metode untuk mengekstraksi dan menganalisis teks media sosial secara efisien, sehingga memungkinkan pemahaman pola komunikasi dalam skala besar [16].

E. *IndoBERTweet*

IndoBERTweet merupakan model bahasa berbasis *BERT* yang dirancang untuk menganalisis teks berbahasa Indonesia pada platform X. Model ini diprapelatih menggunakan korpus berskala besar dari media sosial sehingga mampu menangani bahasa informal, singkatan, dan gaya percakapan khas media sosial. *IndoBERTweet* mengatasi permasalahan *vocabulary mismatch* melalui pendekatan adaptasi *embedding* kata baru, di mana strategi inisialisasi menggunakan rata-rata embedding subkata *BERT* terbukti mempercepat proses prapelatihan dan meningkatkan kinerja ekstrinsik model[5], [11]



GAMBAR 1
Arsitektur IndoBERTweet [8]

IndoBERTweet menerima *Input IDs* dan *Attention Mask* sebagai masukan utama untuk merepresentasikan token dan fokus perhatian model. Model ini dilengkapi lapisan domain-spesifik media sosial untuk memahami konteks seperti slang, emotikon, dan singkatan. Keluaran selanjutnya diproses

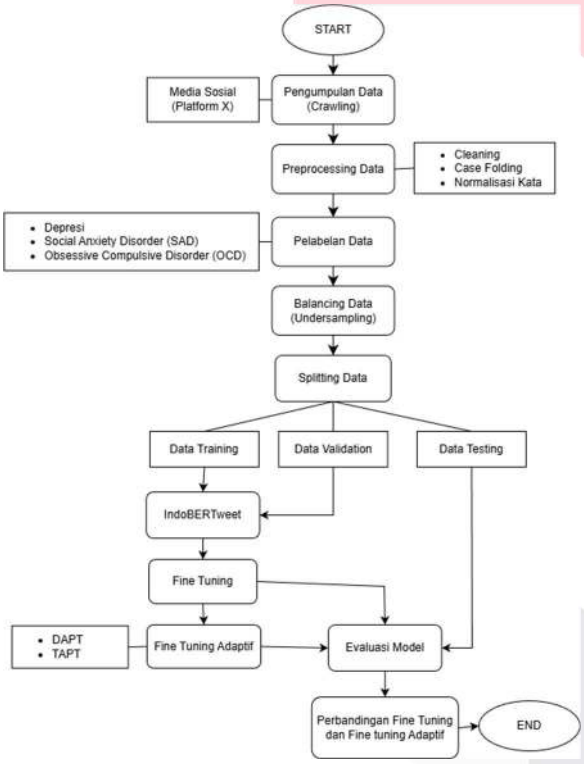
melalui lapisan *Dense* dan *Softmax* untuk menghasilkan klasifikasi ke dalam tiga kategori[8], [17]. Selain itu, *IndoBERTweet* menerapkan *fine-tuning* adaptif guna menyesuaikan bobot model dengan karakteristik dataset penelitian.

F. *Undersampling*

Undersampling merupakan Teknik untuk menangani ketidakseimbangan kelas dengan mengurangi jumlah sampel pada kelas mayoritas agar seimbang dengan kelas minoritas [14]. Teknik ini bertujuan mencegah bias model terhadap kelas mayoritas serta meningkatkan kemampuan prediksi pada kelas minoritas, sehingga pelatihan model *IndoBERTweet* menjadi lebih efektif.

III. METODE

Penelitian ini terdiri atas beberapa tahapan utama yang dirancang secara sistematis. Alur tahapan penelitian secara keseluruhan ditunjukkan pada GAMBAR 2.



GAMBAR 2
Flowchart Penelitian

A. Pengumpulan Dataset

Data penelitian diperoleh dari platform X melalui proses *crawling* menggunakan kata kunci yang berkaitan dengan kesehatan mental dalam bahasa Indonesia. Data dikumpulkan dalam bentuk unggahan teks dan disimpan dalam format .csv. Pada tahap ini, diperoleh sebanyak 46.096 unggahan.

GAMBAR 3
Hasil *Crawling* Data

B. *Preprocessing* Teks

Preprocessing dilakukan untuk meningkatkan kualitas data agar lebih terstruktur dan siap digunakan pada tahap pemodelan. Tahapan yang diterapkan meliputi *cleaning* data, case folding, dan normalisasi kata. *Cleaning* data bertujuan menghapus elemen noninformatif seperti *URL*, *username*, emoji, simbol, dan tanda baca tanpa mengubah makna teks. *Case folding* dilakukan dengan mengonversi seluruh huruf menjadi huruf kecil, sedangkan normalisasi kata digunakan untuk mengubah kata tidak baku, singkatan, dan slang ke dalam bentuk baku sesuai kaidah bahasa Indonesia.

Proses *stopword removal* dan *stemming* tidak diterapkan karena berpotensi menghilangkan informasi penting, khususnya kata negasi dan bentuk kata asli yang berpengaruh terhadap makna emosional teks. Selain itu, *IndoBERTweet* sebagai turunan *BERT* dilatih menggunakan bentuk kata asli, sehingga mempertahankan kata tanpa *stemming* menjaga kesesuaian dengan data pra-pelatihan dan meningkatkan representasi fitur [18].

C. Pelabelan dan Penyeimbangan Data

Pelabelan data dilakukan menggunakan pendekatan *lexicon-based* secara semi-manual dengan tiga kategori gangguan mental, yaitu *Depression*, *Social Anxiety*, dan *Obsessive-Compulsive Disorder* menggunakan daftar kata kunci yang disusun berdasarkan kajian literatur dan penelitian terdahulu [19]. Setiap kategori memiliki daftar kata kunci yang diberi bobot sesuai tingkat keparahan gejala. Skor pada setiap kategori dihitung berdasarkan frekuensi kemunculan kata kunci, dan label akhir ditentukan dari skor tertinggi. Tingkat keyakinan pelabelan diukur menggunakan *confidence score* sebagai indikator keandalan penentuan label.

$$Confidence = \frac{Score(\text{label terbaik})}{\sum score \text{ semua label}} \quad (1)$$

Hasil pelabelan divalidasi melalui analisis distribusi label, pemeriksaan konsistensi kata kunci, serta evaluasi *confidence score*. Selanjutnya, dataset diseimbangkan menggunakan teknik *undersampling* untuk mengatasi ketidakseimbangan kelas, sehingga setiap kelas memiliki jumlah sampel yang setara

D. Pembagian dan Transformasi Data

Dataset dibagi menjadi data *training*, *validation*, dan *testing* dengan rasio 80:10:10 menggunakan teknik *stratified splitting* untuk ntuk mempertahankan proporsi distribusi masing-masing kelas. Selanjutnya, label kelas dikonversi ke bentuk numerik melalui *label encoding* . Teks kemudian diproses menggunakan *AutoTokenizer* IndoBERTweet untuk

menghasilkan *input IDs* dan *attention mask* dengan panjang maksimum 128 token, sehingga data siap digunakan oleh model berbasis deep learning [20].

E. Pemodelan dan *Fine tuning*

Pemodelan dilakukan menggunakan *IndoBERTweet* dengan empat skema percobaan, yang merupakan kombinasi dari dua pendekatan *fine-tuning* dan penerapan teknik *undersampling*. eksperimen pertama adalah *fine-tuning* standar, yang dilakukan dengan menambahkan lapisan klasifikasi pada model dan melatihnya menggunakan data latih. Untuk mengurangi risiko *overfitting*, diterapkan *dropout* dan pembekuan sebagian lapisan encoder. Skema ini diuji pada dua kondisi: tanpa *undersampling*, di mana distribusi kelas tidak diubah, dan dengan *undersampling*, di mana jumlah sampel pada kelas mayoritas dikurangi untuk menyeimbangkan distribusi kelas.

eksperimen kedua adalah *fine-tuning* adaptif, yang dilakukan melalui dua tahap, yaitu *Domain-Adaptive Pretraining* (DAPT) dan *Task-Adaptive Pretraining* (TAPT) menggunakan skema *Masked Language Modeling*. Model yang telah melalui proses adaptasi kemudian digunakan untuk klasifikasi dengan pengaturan pelatihan yang sama seperti pada *fine-tuning* standar. Skema *fine-tuning* adaptif juga diuji pada kondisi tanpa dan dengan *undersampling*

F. Evaluasi Model

Evaluasi model dilakukan menggunakan *confusion matrix*, yang memungkinkan identifikasi jumlah prediksi benar dan salah pada setiap kelas [21]. Matrik ini digunakan untuk menghitung metrik akurasi, presisi, *recall*, dan *F1-score*. Metrik-metrik tersebut digunakan untuk menilai kemampuan model dalam mengklasifikasikan tiga kategori gangguan kesehatan mental secara seimbang dan akurat. Kemudian, perbandingan kinerja antara *fine-tuning* standar dan *fine-tuning* adaptif dilakukan berdasarkan nilai metrik evaluasi tersebut.

Confusion Matrix terdiri dari beberapa komponen utama:

TABEL 1
Komponen *Confusion Matrix*

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

Berdasarkan *confusion matrix*, beberapa metrik evaluasi dapat dihitung, yaitu akurasi, presisi, *recall,* dan *F1-score* untuk menilai kinerja model dalam mengklasifikasikan setiap kategori gangguan kesehatan mental secara akurat dan seimbang.

1. Akurasi digunakan untuk mengukur tingkat ketepatan model dalam memprediksi seluruh kelas dengan benar.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

2. Presisi menunjukkan ketepatan sistem dalam memberikan prediksi positif yang benar.

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

3. *Recall* mengukur kemampuan model dalam mengidentifikasi seluruh sampel yang benar-benar termasuk ke dalam kelas positif

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

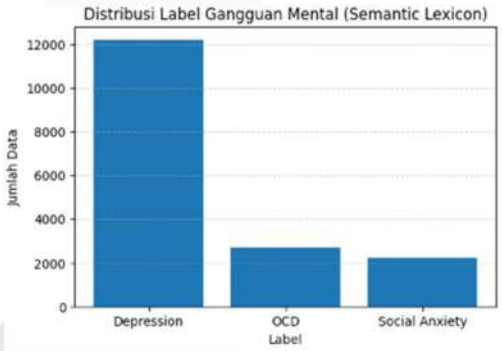
4. *F1-score* mengukur keseimbangan antara presisi dan recall melalui rata-rata harmonik, sehingga merepresentasikan performa model secara keseluruhan.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Presisi} \times Recall}{\text{Presisi} + Recall} \quad (5)$$

IV. HASIL DAN PEMBAHASAN

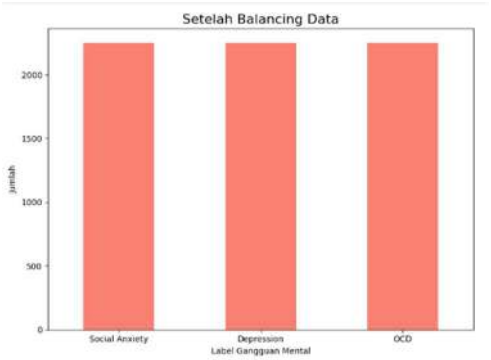
A. Hasil Pelabelan dan *Balancing* data

Setelah melalui tahap *preprocessing* dan pelabelan, jumlah data teks yang digunakan dalam penelitian ini berkurang menjadi 17.101 sampel. Analisis distribusi kelas menunjukkan dominasi kelas *Depression* sebesar 71,19%, diikuti oleh *Obsessive-Compulsive Disorder (OCD)* sebesar 15,67%, dan *Social Anxiety* sebesar 13,15%



GAMBAR 4
Hasil Pelabelan Data

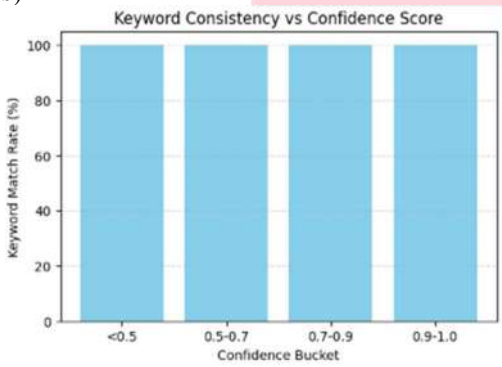
Dominasi kelas *Depression* mencerminkan karakteristik alami data media sosial, di mana pengguna lebih sering mengekspresikan kondisi emosional terkait depresi dibandingkan gangguan kesehatan mental lainnya. Distribusi ini menunjukkan bahwa dataset merepresentasikan pola ekspresi yang realistis, namun bersifat tidak seimbang antar kelas. Ketidakseimbangan tersebut berpotensi memengaruhi performa model klasifikasi, sehingga perlu menjadi perhatian khusus pada tahap pelatihan dan evaluasi. Untuk mengurangi dampak ketidakseimbangan dan meningkatkan kemampuan model dalam mengenali kelas minoritas, penelitian ini menerapkan teknik *data balancing* menggunakan metode *undersampling*, dengan mengurangi jumlah sampel pada kelas mayoritas agar distribusi antar kelas menjadi lebih seimbang. Hasil dari proses *undersampling* ditunjukkan pada Gambar 5, yang memperlihatkan distribusi kelas setelah penyeimbangan data.



GAMBAR 5
Hasil *Balancing* Data

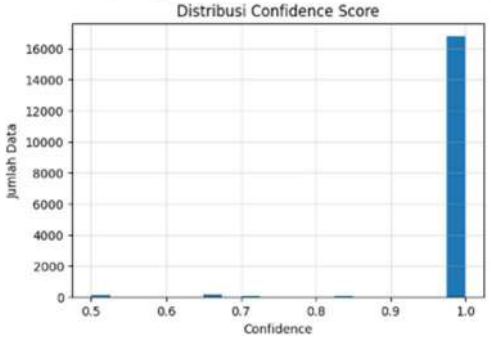
B. Evaluasi dan Kualitas Pelabelan

Evaluasi kualitas pelabelan dilakukan melalui analisis konsistensi kata kunci (*keyword consistency*), distribusi *confidence score*, serta evaluasi ambang batas (*threshold analysis*)



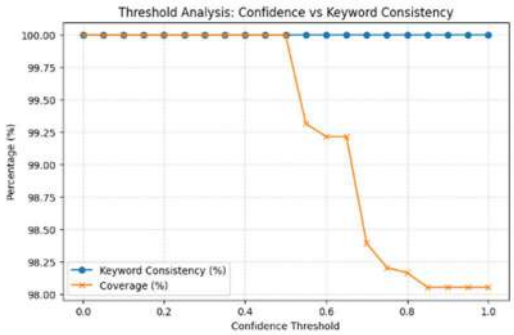
GAMBAR 6
Consistency vs Confidence bucket

Hasil analisis menunjukkan bahwa tingkat kecocokan kata kunci terhadap label mencapai 100% pada seluruh rentang *confidence,* menandakan tidak adanya inkonsistensi antara *lexicon* dan label yang dihasilkan.



GAMBAR 7
Distribusi *Confidence Score*

Distribusi *confidence score* menunjukkan bahwa sebagian besar data memiliki nilai yang mendekati 1,0, sementara data dengan tingkat *confidence* menengah hingga rendah jumlahnya relatif sedikit. Temuan ini mengindikasikan bahwa proses pelabelan semi-manual menghasilkan label dengan tingkat keyakinan yang tinggi, sehingga kualitas label yang digunakan untuk pelatihan model dapat dianggap andal dan representati.



GAMBAR 8
Threshold Analisis

Gambar 8 menunjukkan bahwa peningkatan ambang *confidence* mengakibatkan penurunan jumlah data yang digunakan, seiring dengan meningkatnya keandalan label. Hal ini menegaskan adanya *trade-off* antara kuantitas data dan kualitas pelabelan, di mana ambang *confidence* yang lebih tinggi menghasilkan label yang lebih dapat dipercaya meskipun cakupan data menjadi lebih terbatas.

Secara keseluruhan, hasil analisis mengindikasikan bahwa proses pelabelan stabil, konsisten, dan memiliki tingkat keandalan tinggi, sehingga dataset layak digunakan untuk pelatihan dan evaluasi model tanpa mengorbankan representativitas data secara signifikan.

C. Hasil Kinerja Model IndoBERTweet

Hasil dari empat percobaan disajikan berdasarkan skema pelatihan *fine-tuning* standar dan adaptif, baik dengan maupun tanpa penerapan teknik *undersampling*.

1. *Fine Tuning* Adaptif dengan *Undersampling*

Pada skenario ini, *fine-tuning* adaptif diterapkan dengan teknik undersampling untuk menyeimbangkan jumlah data antar kelas, sehingga model dapat mempelajari pola dari semua kelas secara proporsional.

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro
1	1.083000	1.009274	0.615727	0.629216	0.615556	0.616857
2	1.031700	0.884195	0.738872	0.756099	0.739015	0.740180
3	0.950900	0.769464	0.820475	0.824395	0.820463	0.821342
4	0.894300	0.676912	0.859050	0.861243	0.859028	0.859541
5	0.887800	0.642567	0.867953	0.869659	0.867950	0.868357

GAMBAR 9
Hasil Performa Percobaan 1

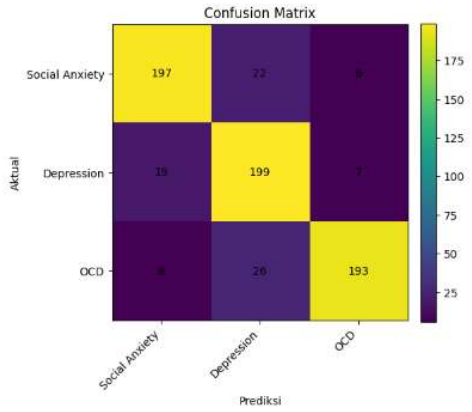
Hasil pelatihan menunjukkan penurunan konsisten pada *training loss* (1,0830 → 0,8878) dan *validation loss* (1,0092 → 0,6425), yang mengindikasikan bahwa fine-tuning adaptif berhasil mempelajari karakteristik dataset secara efektif tanpa menunjukkan tanda-tanda *overfitting.*

Classification Report

	precision	recall	f1-score	support
Social Anxiety	0.8874	0.8756	0.8814	225
Depression	0.8057	0.8844	0.8432	225
OCD	0.9369	0.8578	0.8956	225
accuracy			0.8726	675
macro avg	0.8766	0.8726	0.8734	675
weighted avg	0.8766	0.8726	0.8734	675

GAMBAR 10
Hasil Klasifikasi Percobaan 1

Evaluasi pada data uji menunjukkan akurasi sebesar 87,26%, dengan nilai *precision*, *recall,* dan *F1-score* yang relatif seimbang, menunjukkan bahwa model mampu mengklasifikasikan setiap kelas secara merata. Penerapan *undersampling* juga terbukti membantu mengurangi bias terhadap kelas mayoritas



GAMBAR 11
Confusion Matrix Percobaan 1

Hasil *confusion matrix* menunjukkan bahwa model berhasil mengklasifikasikan 197 sampel *Social Anxiety*, 199 sampel *Depression*, dan 193 sampel OCD dengan benar. Kesalahan klasifikasi sebagian besar terjadi pada kelas dengan pola teks yang serupa, terutama antara *Depression* dan OCD. Meskipun jumlah sampel *Depression* lebih dominan, selisih prediksi antar kelas tetap terkendali, menunjukkan bahwa model mampu mempertahankan distribusi klasifikasi yang seimbang di antara ketiga kelas.

2. *Fine Tuning* Standar dengan *Undersampling*

Pada skenario ini, *fine tuning* dilakukan tanpa mekanisme adaptif, namun tetap menggunakan teknik *undersampling* untuk menyeimbangkan distribusi data pada set pelatihan.

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro
1	1.080200	1.051462	0.467359	0.493084	0.467665	0.453114
2	1.062200	0.973671	0.541543	0.687238	0.542149	0.516956
3	0.963100	0.789504	0.743323	0.767092	0.743538	0.743116
4	0.816000	0.567617	0.827893	0.847758	0.828089	0.828032
5	0.746500	0.500986	0.838279	0.860919	0.838485	0.838639

GAMBAR 12
Hasil Performa Percobaan 2

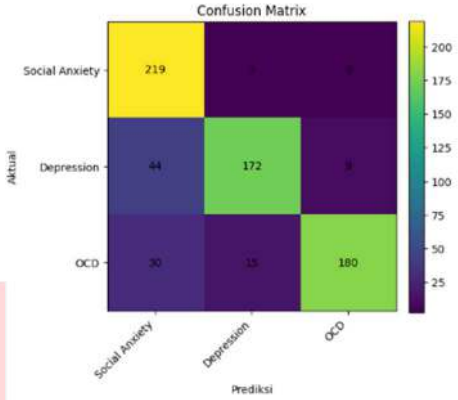
Hasil pelatihan selama lima epoch menunjukkan penurunan yang konsisten pada *training loss* (1,0802 → 0,7465) dan *validation loss* (1,0514 → 0,5009), mengindikasikan bahwa model berhasil mempelajari pola data secara stabil tanpa menunjukkan tanda-tanda *overfitting.*

Classification Report

	precision	recall	f1-score	support
Social Anxiety	0.7474	0.9733	0.8456	225
Depression	0.9053	0.7644	0.8289	225
OCD	0.9375	0.8000	0.8633	225
accuracy			0.8459	675
macro avg	0.8634	0.8459	0.8459	675
weighted avg	0.8634	0.8459	0.8459	675

GAMBAR 13
Hasil Klasifikasi Percobaan 2

Gambar 13 menunjukkan bahwa model mencapai akurasi 84,59%, dengan *precision* 86,34%, *recall* 84,59%, dan *F1-score* 84,59%. Hasil ini mengindikasikan bahwa model cukup selektif dalam melakukan prediksi, meskipun beberapa sampel masih mengalami kesalahan klasifikasi.



GAMBAR 14
Confusion Matix Percobaan 2

Hasil *confusion matrix* menunjukkan kelas *Social Anxiety* terklasifikasi dengan benar pada 219 data, *Depression* pada 172 data, dan OCD pada 180 data. Sebagian besar kesalahan prediksi terjadi pada kelas *Depression* yang salah diklasifikasikan sebagai *Social Anxiety*, menunjukkan kemiripan pola teks antara kedua kelas.

3. *Fine Tuning* Adaptif tanpa *Undersampling*

Dalam skenario ini, *fine-tuning* adaptif diterapkan tanpa teknik *undersampling*, mempertahankan distribusi data latih yang tidak seimbang. Skenario ini bertujuan mengevaluasi adaptasi model terhadap data asli dan dampaknya pada hasil klasifikasi

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro
1	0.821800	0.821189	0.711698	0.237232	0.333333	0.277189
2	0.791500	0.745031	0.711698	0.237232	0.333333	0.277189
3	0.768100	0.636102	0.757310	0.841012	0.438430	0.464061
4	0.752400	0.575846	0.809942	0.878274	0.561580	0.628490
5	0.758300	0.553498	0.836257	0.889871	0.626621	0.697993

GAMBAR 15
Hasil Performa Percobaan 3

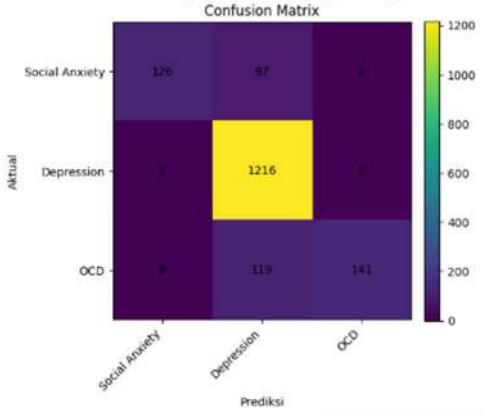
Pelatihan selama 5 epoch menunjukkan penurunan konsisten pada *training loss* (0,8218 → 0,7583) dan *validation loss* (0,8212 → 0,5535). Setelah epoch ke-3, kedua metrik cenderung stabil, menandakan model telah mencapai konvergensi dan mampu mempelajari data secara bertahap tanpa gejala *overfitting* meskipun distribusi kelas tidak seimbang.

Classification Report

	precision	recall	f1-score	support
Social Anxiety	0.9265	0.5600	0.6981	225
Depression	0.8492	0.9984	0.9177	1218
OCD	0.9860	0.5261	0.6861	268
accuracy			0.8667	1711
macro avg	0.9205	0.6948	0.7673	1711
weighted avg	0.8808	0.8667	0.8526	1711

GAMBAR 16
Hasil Klasifikasi Percobaan

Hasil *classification report* menunjukkan bahwa model cukup selektif dalam prediksi, mencapai akurasi 86,67% dan mampu mendeteksi kelas dominan seperti Depression dengan baik. *Recall* yang lebih rendah pada *Social Anxiety* dan OCD menandakan sebagian data tidak terdeteksi atau salah diklasifikasikan, terutama pada kelas dengan jumlah data lebih sedikit atau pola teks yang mirip.



GAMBAR 17
Confusion Matrix Percobaan 3

Confusion matrix menunjukkan bahwa model mampu memprediksi sebagian besar data dengan tepat. Kelas *Depression* memiliki jumlah prediksi benar tertinggi, yaitu 1.216 data, sejalan dengan jumlah data yang lebih besar pada kelas ini. Kelas *Social Anxiety* dan OCD masing-masing terklasifikasi dengan benar sebanyak 126 dan 141 data, meskipun beberapa prediksi kelas OCD tertukar ke kelas *Depression*. Temuan ini mengindikasikan bahwa tanpa penerapan undersampling, model cenderung memprediksi kelas mayoritas, namun prediksi pada kelas *Social Anxiety* dan OCD tetap signifikan, menunjukkan bahwa model tidak sepenuhnya bias terhadap satu kelas.

4. *Fine Tuning* Standar tanpa *Undersampling*

Pada percobaan ini, model dilakukan *fine-tuning* menggunakan dataset asli tanpa penerapan *undersampling*, sehingga distribusi data tetap tidak seimbang, dengan kelas *Depression* jauh lebih dominan dibanding *Social Anxiety* dan OCD. Percobaan bertujuan mengevaluasi performa model dalam kondisi standar serta mengamati dampak ketidakseimbangan data terhadap klasifikasi.

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro
1	0.662400	0.639283	0.715789	0.838347	0.342991	0.296923
2	0.211300	0.169127	0.970760	0.953449	0.958217	0.955409
3	0.169600	0.198010	0.971345	0.957220	0.961163	0.959126
4	0.079600	0.182756	0.974269	0.965491	0.960611	0.963009
5	0.182600	0.177076	0.977193	0.971786	0.962951	0.967228

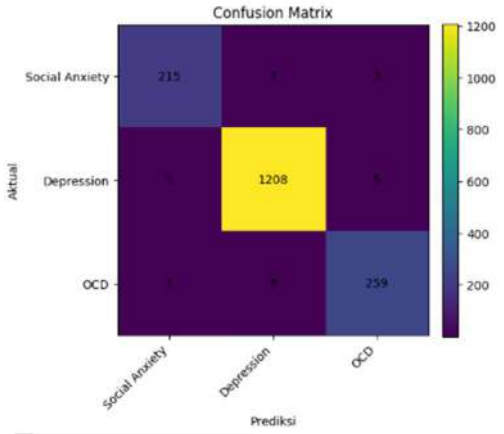
GAMBAR 18
Hasil Performa Percobaan 4

Pelatihan model selama 5 epoch menunjukkan *proses fine-tuning* yang stabil. *Training loss* menurun dari 0,6624 menjadi 0,1826, sedangkan *validation loss* turun lebih tajam, dari 0,6392 menjadi 0,1771. Penurunan loss ini diikuti peningkatan akurasi dari 71,58% pada awal pelatihan menjadi 97,72% pada epoch terakhir, sementara metrik makro menunjukkan model mampu belajar dengan baik dan menangkap pola pada ketiga kelas meskipun distribusi data tidak seimbang.

Classification Report	precision	recall	f1-score	support
Social Anxiety	0.9729	0.9556	0.9641	225
Depression	0.9877	0.9918	0.9898	1218
OCD	0.9700	0.9664	0.9682	268
accuracy			0.9831	1711
macro avg	0.9769	0.9713	0.9740	1711
weighted avg	0.9830	0.9831	0.9830	1711

GAMBAR 19
Hasil Klasifikasi Percobaan

Model berhasil mengklasifikasikan tiga kelas gangguan mental dengan baik menggunakan *fine-tuning* pada dataset asli tanpa *undersampling*, mencapai akurasi 98,31%. Kelas *Depression* menampilkan performa terbaik, sejalan dengan jumlah data yang paling banyak, sementara *Social Anxiety* dan OCD juga diklasifikasikan dengan baik, dengan F1-score masing-masing 0,9641 dan 0,9682. Nilai *macro average* dan *weighted average* menunjukkan bahwa model bekerja seimbang antar kelas meskipun distribusi data tidak seimbang.



GAMBAR 20
Confusion Matrix Percobaan 4

Berdasarkan *confusion matrix*, model menunjukkan performa terbaik pada kelas *Depression*, dengan 1.208 data diklasifikasikan dengan benar. Kelas *Social Anxiety* dan OCD masing-masing memiliki 215 dan 259 prediksi benar. Sejumlah data pada kedua kelas minoritas ini salah diklasifikasikan sebagai *Depression*, menunjukkan kecenderungan model memprediksi kelas mayoritas ketika menghadapi data ambigu. Temuan ini terkait dengan ketidakseimbangan distribusi data latih, yang menyebabkan model lebih terfokus pada karakteristik kelas *Depression*, sementara prediksi untuk *Social Anxiety* dan OCD relatif lebih rendah.

D. Evaluasi Kinerja Model *IndoBERTweet*

Berdasarkan hasil pengujian, empat skema *fine-tuning* menunjukkan perbedaan kinerja yang signifikan dalam klasifikasi tiga kelas gangguan kesehatan mental, yaitu *Depression, Social Anxiety,* dan *Obsessive-Compulsive Disorder* (OCD). *Fine-tuning* standar tanpa *undersampling*

menghasilkan nilai *accuracy* tertinggi, yakni 98,31%. Namun, performa yang tinggi ini cenderung dipengaruhi oleh ketidakseimbangan data, sehingga model lebih dominan dalam memprediksi kelas mayoritas (*Depression*) dan kurang mampu merepresentasikan kemampuan generalisasi pada seluruh kelas. Penerapan *undersampling* pada skema ini menurunkan *accuracy* menjadi 84,59% dan *F1-score* 84,59%, yang menandakan bahwa pengurangan data pada kelas mayoritas berdampak langsung terhadap kemampuan model dalam mengenali pola secara keseluruhan, terutama pada skema *fine-tuning* yang tidak adaptif terhadap distribusi data.

Pendekatan *fine-tuning* adaptif dengan *undersampling* menghasilkan kinerja yang lebih seimbang antar kelas, dengan *accuracy* 87,26%, *precision* 87,66%, *recall* 87,26%, dan *F1-score* 87,34%. Keselarasan nilai antar metrik menunjukkan bahwa model mampu mengklasifikasikan setiap kelas secara proporsional tanpa bias berlebihan terhadap kelas tertentu. Sebaliknya, *fine-tuning* adaptif tanpa *undersampling* menampilkan ketidakseimbangan antara *precision* (92,05%) dan *recall* (69,48%), sehingga *F1-score* menurun menjadi 76,73%. Hal ini mengindikasikan bahwa model lebih selektif dalam memberikan prediksi positif, tetapi kurang optimal dalam menangkap seluruh data yang relevan.

E. Implikasi

Hasil penelitian ini, memberikan beberapa implikasi penting. Pertama, penggunaan *fine-tuning* adaptif, terutama pada dataset tidak seimbang, lebih efektif untuk menghasilkan prediksi yang merata pada seluruh kelas, sehingga dapat mendukung deteksi gangguan kesehatan mental minoritas seperti *Social Anxiety* dan OCD. Kedua, penerapan teknik *undersampling* atau metode *balancing* data lainnya terbukti membantu mengurangi bias kelas mayoritas, namun perlu disesuaikan agar tidak mengurangi informasi penting dari data mayoritas. Ketiga, meskipun *accuracy* tinggi dapat dicapai pada *fine-tuning* standar tanpa *undersampling*, hasil ini cenderung menimbulkan bias terhadap kelas mayoritas, sehingga interpretabilitas dan keamanan model dalam konteks kesehatan mental menjadi terbatas.

Secara keseluruhan, kombinasi *fine-tuning* adaptif dengan *undersampling* menawarkan keseimbangan terbaik antara performa dan representativitas antar kelas, yang penting untuk implementasi sistem deteksi dini gangguan kesehatan mental berbasis media sosial secara akurat. Temuan ini juga menjadi dasar bagi penelitian selanjutnya untuk mengeksplorasi teknik *balancing* data lebih lanjut, integrasi model hybrid, dan analisis interpretabilitas untuk mendukung keputusan klinis dan intervensi kesehatan mental.

V. KESIMPULAN

Berdasarkan hasil penelitian, model *IndoBERTweet* mampu mengklasifikasikan unggahan media sosial terkait gangguan mental secara efektif dengan perbedaan kinerja yang signifikan antar skema *fine-tuning*. *Fine-tuning* standar menghasilkan akurasi yang tinggi, namun model cenderung memprediksi kelas mayoritas, yaitu *Depression*, sehingga kemampuan mengenali kelas minoritas seperti *Social Anxiety* dan OCD menjadi terbatas. Sebaliknya, *fine-tuning* adaptif meningkatkan sensitivitas model terhadap kelas minoritas, sehingga distribusi prediksi lebih seimbang meskipun akurasi total sedikit lebih rendah dibanding *fine-tuning* standar tanpa *undersampling*. Penerapan teknik undersampling pada *fine-tuning* adaptif terbukti efektif dalam mengurangi bias terhadap kelas mayoritas, sehingga semua kelas dapat dikenali secara proporsional, sementara pada *fine-tuning* standar, *undersampling* menurunkan performa karena mengurangi data penting dari kelas mayoritas yang diperlukan model untuk mempelajari pola keseluruhan. Analisis terhadap kelas mayoritas dan minoritas menunjukkan bahwa kelas *Depression* selalu memiliki jumlah prediksi benar terbanyak akibat dominasi jumlah data, sedangkan kelas *Social Anxiety* dan OCD lebih menantang untuk diklasifikasikan tanpa strategi balancing.

Kombinasi *fine-tuning* adaptif dengan *undersampling* memberikan keseimbangan terbaik dalam mendeteksi seluruh kelas. Selain itu, semua skema *fine-tuning* menunjukkan penurunan *training loss* dan *validation loss* yang stabil, mencerminkan proses pelatihan yang lancar dan bebas *overfitting*, serta kemampuan model belajar pola data secara bertahap sejalan dengan peningkatan metrik evaluasi

REFERENSI

[1] K. Dasar, K. Mental, and A. M. Tripertiwi, “EDUCAZIONE: Jurnal Multidisiplin Lembaga Penelitian Dan Publikasi Ilmiah (LPPI) Yayasan Almahmudi Bin Dahlan Website: https://j-educa.org/index.php/educazione.” [Online]. Available: https://j-educa.org/index.php/educazione

[2] D. Kusuma Ningrum and A. Maytsa Ismawardi, “EFEKTIVITAS ALGORITMA KECERDASAN BUATAN DALAM IMPLEMENTASI KESEHATAN MENTAL : SYSTEMATIC LITERATURE REVIEW,” 2025.

[3] World Health Organization, “Mental health of adolescents,” World Health Organization. Accessed: Jul. 28, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

[4] Kemenkes, “Pentingnya Kesehatan Mental bagi Remaja dan Cara Menghadapinya,” Kemenkes. Accessed: Jul. 28, 2025. [Online]. Available: https://ayosehat.kemkes.go.id/pentingnya-kesehatan-mental-bagi-remaja

[5] D. Perhatian, “ANALISIS SENTIMEN DAN PERILAKU PENGGUNA MEDIA SOSIAL TERHADAP ISU KESEHATAN MENTAL MENGGUNAKAN METODE NATURAL LANGUAGE PROCESSING (NLP) Analysis Of Sentiment And Behavior Of Social Media Users Towards Mental Health Issues Using The Natural Language Processing (NLP) Method,” *Jurnal Pengabdian Masyarakat (Kesehatan)*, vol. 6, no. 2, 2024.

[6] M. Fadhilla, R. Wandri, A. Hanafiah, P. R. Setiawan, Y. Arta, and S. Daulay, “Analisis Performa Algoritma Machine Learning Untuk Identifikasi Depresi Pada Mahasiswa,” *Journal of Informatics*

Management and Information Technology, vol. 5, no. 1, pp. 40–47, 2025, doi: 10.47065/jimat.v5i1.473.

[7] F. Ilham and W. Maharani, "Analyze Detection Depression In Social Media Twitter Using Bidirectional Encoder Representations from Transformers," *Journal of Information System Research (JOSH)*, vol. 3, no. 4, pp. 476–482, Jul. 2022, doi: 10.47065/josh.v3i4.1885.

[8] M. Fadhel, "Depression Detection of Users in Social Media X using IndoBERTweet," *Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, 2024, doi: 10.33395/v8i2.13354.

[9] N. Nurdiansyah, F. S. Febriyan, Z. G. D. Amanta, D. A. Saputra, and W. M. Baihaqi, "Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Random Forest," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 1–9, Nov. 2024, doi: 10.57152/malcom.v5i1.1537.

[10] O. S. D. Silaen, H. Herlawati, and R. Rasim, "Analisis Sentimen Mengenai Gangguan Bipolar Pada Twitter Menggunakan Algoritma Naïve Bayes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 6, no. 2, pp. 62–73, Nov. 2022, doi: 10.31603/komtika.v6i2.8198.

[11] F. Koto Jey Han Lau Timothy Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," Nov. 2021. doi: 10.18653/v1/2021.emnlp-main.833.

[12] Erzha Tri Setyo Rochman, Septiana Mukti, Nazwa Mahabatul Thigah Yanani, Fita Sinta Dewi, and Liss Dyah Dewi A, "Pengaruh Media Sosial terhadap Kesehatan Mental pada Anak Muda di Indonesia," *Student Research Journal*, vol. 2, no. 3, pp. 12–27, May 2024, doi: 10.55606/srjyappi.v2i3.1219.

[13] F. Aulia Girnanfa and D. A. Susilo, "Studi Dramaturgi Pengelolaan Kesan Melalui Twitter Sebagai Sarana Eksistensi Diri Mahasiswa di Jakarta," vol. 1, no. 1, pp. 58–73, 2022.

[14] A. Wandani, "Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN, Random Forest, dan Naive Bayes," 2021.

[15] Fujiyama, "Deep Learning dan Neural Networks: Masa Depan Kecerdasan Buatan," Universitas Telkom Surabaya. Accessed: Oct. 05, 2025. [Online]. Available: https://surabaya.telkomuniversity.ac.id/deep-learning-dan-neural-networks-masa-depan-kecerdasan-buatan/

[16] F. Hanafi, "Mengenal Natural Language Processing (NLP)," CodePolitan. Accessed: Oct. 05, 2025. [Online]. Available: https://www.codepolitan.com/blog/mengenal-natural-language-processing-nlp/

[17] 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," *Jurnal Sifo Mikroskil*, vol. 26, no. 2, Oct. 2025, doi: 10.55601/jsm.v26i2.1799.

[18] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 887–894, Aug. 2024, doi: 10.25126/jtiik.1148315.

[19] Aisyatin Kamila and Finanin Nur Indana, "Kajian Psikologis Mengenai Kesadaran Kesehatan Mental pada Serial 'Daily Dose of Sunshine,'" *WISSEN : Jurnal Ilmu Sosial dan Humaniora*, vol. 3, no. 3, pp. 80–91, Jun. 2025, doi: 10.62383/wissen.v3i3.947.

[20] R. A. Chaerudin[1] *et al.*, "Implementasi Algoritma Naïve Bayes Untuk Analisis Klasifikasi Survei Kesehatan Mental (Studi Kasus: Open Sourcing Mental Illness)," *JURNAL INFORMATIK Edisi ke-19*, Apr. 2023.

[21] G. Zahra, N. Fadhilah, R. A. Saputra, and A. H. Wibowo, "DETEKSI TINGKAT GANGGUAN KECEMASAN MENGGUNAKAN METODE RANDOM FOREST Anxiety Disorder Level Detection Using Random Forest Method," *https://jurnal.umt.ac.id/index.php/jt/article/view/10648/4974*, pp. 38–47, 2024, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index