Model Prediksi Kesehatan Mental pada Data Media Sosial menggunakan CNN-BiLSTM

Abdurrahim

22917002

Tesis diajukan sebagai syarat untuk meraih gelar Magister Komputer

Konsentrasi Sains Data

Program Studi Informatika Program Magister

Fakultas Teknologi Industri

Universitas Islam Indonesia

2026

Lembar Pengesahan Pembimbing

Model Prediksi Kesehatan Mental pada Data Media Sosial menggunakan CNN-BiLSTM

Abdurrahim
22917002

Yogyakarta, 13 Desember, 2025

Pembimbing

Dhomas Hatta Fudholi, S.T, M.Eng, Ph.D

Lembar Pengesahan Penguji

Model Prediksi Kesehatan Mental pada Data Media Sosial menggunakan CNNBiLSTM

Abdurrahim

22917002

Yogyakarta, 14 Mei, 2026

Tim Penguji,

Dhomas Hatta Fudholi, S.T, M.Eng, Ph.D

Ketua

Dr. Feri Wijayanto, S.T., M.T.

Anggota I

Ahmad Fathan Hidayatullah, S.T., M.Cs., Ph.D

Anggota II

Mengetahui,

Ketua Program Studi Informatika Program Magister

Fakultas Teknologi Industri

Universitas Islam Indonesia

Irving Vitra Paputungan, S.T., M.Sc., Ph.D

Abstrak

Model Prediksi Kesehatan Mental pada Data Media Sosial Menggunakan CNN-BiLSMT

Kesehatan mental telah menjadi isu global dengan lebih dari 1 miliar orang terpengaruh oleh gangguan mental. Media sosial tidak hanya merefleksikan kondisi mental pengguna tetapi juga dapat memperparah kondisi yang sudah ada melalui berbagai mekanisme seperti perbandingan sosial negatif, *echo chamber*, dan paparan konten berbahaya. Penelitian sebelumnya menunjukkan keterbatasan signifikan dalam cakupan klasifikasi, penggunaan arsitektur tunggal yang tidak optimal, dan belum mengeksplorasi deteksi konten berbahaya *(poison)* yang dapat memicu *contagion effect*. Penelitian ini mengembangkan model klasifikasi berbasis arsitektur hibrid *CNN-BiLSTM* dengan pembobotan kata *FastText* untuk mendeteksi tujuh kategori gangguan kesehatan mental dari data teks media sosial Reddit, yaitu *BPD, anxiety, depression, bipolar, mentalillness, schizophrenia*, dan *poison*. Dataset sebanyak 10.500 postingan dikumpulkan dari lima subreddit utama dan diproses melalui *pipeline preprocessing* yang ketat. Model dilatih menggunakan 20-*fold cross-validation* dengan *hyperparameter tuning* sistematis menggunakan *Keras Tuner*. Hasil evaluasi menunjukkan bahwa model *CNN-BiLSTM* mencapai akurasi 0.87 dan *F1-score* 0.87, melampaui performa CNN akurasi 0.84 dan BiLSTM akurasi 0.82. Performa terbaik dicapai pada kelas *bipolar F1-score* 0.98, *recall* 1.00 dan *BPD F1-score* 0.98, sementara tantangan utama ditemukan pada kelas *mentalillness* (F1-score 0.61). *FastText embedding* terbukti efektif menangani *out-of-vocabulary words* yang umum dalam teks media sosial informal. Validasi dengan psikolog klinis pada 10 kasus berbahasa Inggris mencapai 6 klasifikasi yang sesuai, namun analisis *cross-lingual* terhadap 7 kasus berbahasa Indonesia hanya menghasilkan 2 klasifikasi yang tepat, mengungkapkan tantangan signifikan dalam *transfer learning* lintas bahasa dan menekankan pentingnya adaptasi kultural dalam implementasi model AI untuk kesehatan mental.

Kata kunci

Media sosial, kesehatan mental, CNN-BiLSTM, FastText, klasifikasi multi-label, deteksi dini

Abstract

Mental health has become a global issue, with more than one billion individuals affected by mental disorders. Social media not only reflects users' psychological conditions but can also exacerbate existing issues through mechanisms such as negative social comparison, echo chambers, and exposure to harmful content. Previous studies show significant limitations in classification scope, reliance on suboptimal single-model architectures, and the lack of exploration into the detection of harmful *poison* content that may trigger contagion effects. This research develops a hybrid CNN–BiLSTM classification model with FastText word embeddings to detect seven categories of mental health conditions from Reddit social media text: BPD, anxiety, depression, bipolar, mentalillness, schizophrenia, and poison. The dataset consists of 70,000 posts, comprising 60,000 posts collected from six major mental health–related subreddits and 10,000 pre-labeled poison samples obtained from a public Kaggle dataset. All data were processed through a rigorous preprocessing pipeline and validated using DSM-5 criteria. The dataset was divided into 70% training, 15% testing, and 15% validation subsets. The model was trained using 20-fold cross-validation with systematic hyperparameter tuning via Keras Tuner. Evaluation results show that the CNN–BiLSTM model achieves an accuracy of 0.87 and an F1-score of 0.87, outperforming the CNN (0.84 accuracy) and BiLSTM (0.82 accuracy) baselines. The highest performance was observed in the bipolar class (F1-score 0.98, recall 1.00) and BPD class (F1-score 0.98), while the greatest challenge appeared in the mentalillness class (F1-score 0.61). FastText embeddings proved effective in handling out-of-vocabulary words commonly found in informal social media text. Clinical validation with a psychologist on 10 English-language cases yielded 6 correct classifications, whereas cross-lingual analysis on 7 Indonesian-language cases resulted in only 2 correct outputs. These findings highlight significant challenges in cross-language transfer learning and emphasize the need for cultural adaptation when deploying AI models for mental health applications.

Keywords

social media, mental health, CNN–BiLSTM, FastText, multi-label classification, early detection

Pernyataan Keaslian Tulisan

Dengan ini saya menyatakan bahwa tesis ini merupakan tulisan asli dari penulis, dan tidak berisi material yang telah diterbitkan sebelumnya atau tulisan dari penulis lain terkecuali referensi atas material tersebut telah disebutkan dalam tesis. Apabila ada kontribusi dari penulis lain dalam tesis ini, maka penulis lain tersebut secara eksplisit telah disebutkan dalam tesis ini.

Dengan ini saya juga menyatakan bahwa segala kontribusi dari pihak lain terhadap tesis ini, termasuk bantuan analisis statistik, desain survei, analisis data, prosedur teknis yang bersifat signifikan, dan segala bentuk aktivitas penelitian yang dipergunakan atau dilaporkan dalam tesis ini telah secara eksplisit disebutkan dalam tesis ini.

Segala bentuk hak cipta yang terdapat dalam material dokumen tesis ini berada dalam kepemilikan pemilik hak cipta masing-masing. Apabila dibutuhkan, penulis juga telah mendapatkan izin dari pemilik hak cipta untuk menggunakan ulang materialnya dalam tesis ini.

Yogyakarta, 14 Mei, 2026

Abdurrahim, S.T

Daftar Publikasi

Publikasi yang menjadi bagian dari tesis

Abdurrahim, & Dhomas Hatta Fudholi. (2024). Mental Health Prediction Model on Social Media Data Using CNN-BiLSTM . Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 9(1), 29-44. https://doi.org/10.22219/kinetik.v9i1.1849

Sitasi publikasi 1

Kontributor	Jenis Kontribusi
Abdurrahim	Mendesain eksperimen (60%) Menulis *paper* (70%)
Dhomas Hatta Fudholi	Mendesain eksperimen (40%) Menulis dan mengedit *paper* (30%)

Halaman Kontribusi

Pada penelitian ini, sejumlah pihak turut memberikan kontribusi yang sangat berarti, sehingga pelaksanaan dan penyusunan penelitian dapat berjalan dengan baik. Adapun pihak-pihak tersebut antara lain:

1. Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi Republik Indonesia, yang telah mendanai penelitian ini melalui skema Hibah Penelitian Tesis Magister, dengan Nomor Kontrak: 031/DirDPPM/70/DPPM/PPS-PTM-KEMDIKBUDRISTEK/VI/2023. Dukungan finansial ini memungkinkan penulis untuk melakukan penelitian dan penulisan publikasi artikel dan tesis secara mendalam.
2. Pembimbing tesis yang senantiasa memberikan panduan dan saran berharga, baik dalam tahap pengembangan konsep dan desain riset maupun saat melakukan analisis serta interpretasi data. Dukungan dan masukannya turut membantu penyempurnaan draf tesis dan menghasilkan peningkatan yang signifikan pada kualitas penelitian.
3. Psikolog dari Rumah Sakit Jiwa Sambang Lihum, Kalimantan Selatan, yang telah memberikan kontribusi dalam melakukan verifikasi terhadap hasil pelabelan data. Proses Validasi tersebut berperan penting dalam memastikan kesesuaian, ketetapatan, dan reliabilitas kategori data yang digunakan, sehingga turut mengingkatkan akurasi serta kredibilitas temua penelitian.

Halaman Persembahan

Tesis ini penulis persembahkan dengan penuh rasa syukur dan penghargaan yang mendalam kepada seluruh pihak yang telah memberikan dukungan, bantuan, serta motivasi selama penulis menempuh studi pada Program Studi Informatika, Program Magister Universitas Islam Indonesia. Segala bentuk perhatian, kesempatan, dan kepercayaan yang diberikan telah menjadi penopang utama dalam keberlangsungan proses akademik hingga terselesaikannya karya ilmiah ini.

Selanjutnya, penulis menyampaikan apresiasi yang tulus kepada orang tua, saudari kembar, dan seluruh keluarga besar H. Bastuya yang senantiasa memberikan doa, dorongan moral, serta semangat tanpa henti. Segala bentuk dukungan dan keteladanan yang diberikan telah menjadi sumber kekuatan sekaligus motivasi bagi penulis dalam menjalani proses akademik hingga terselesaikannya tesis ini.

Pada kesempatan ini, penulis juga ingin mengucapkan terima kasih kepada Abdul Malik, Adi Guzali, dan Aldy Nukadea, yang telah memberikan semangat, inspirasi, serta dukungan moral sepanjang proses penyusunan tesis. Dorongan positif dan bantuan yang diberikan telah membantu penulis melewati berbagai tantangan akademik dan penelitian.

Akhirnya, penulis menyampaikan penghargaan setinggi-tingginya kepada semua pihak yang telah memberikan kontribusi, baik secara langsung maupun tidak langsung. Setiap bentuk dukungan, perhatian, dan kebaikan yang diberikan merupakan bagian penting dalam penyelesaian tesis ini dan memberikan pengaruh besar terhadap perjalanan akademik penulis. Semoga seluruh kebaikan tersebut mendapatkan balasan terbaik dari Allah SWT.

Kata Pengantar

Assalamu'alaikum Warahmatullahi Wabarakatuh,

Bismillahirrahmanirrahim. Segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, karunia, serta petunjuk-Nya, sehingga penulis dapat menyelesaikan penyusunan tesis yang berjudul "Model Prediksi Kesehatan Mental pada Data Media Sosial Menggunakan CNN-BiLSTM". Tesis ini disusun sebagai bagian dari pemenuhan persyaratan akademik untuk memperoleh gelar Magister pada Program Studi Informatika, Fakultas Teknik Industri, Universitas Islam Indonesia.

Selama rangkaian penyusunan hingga penyelesaian tesis ini, penulis mendapatkan berbagai bentuk bantuan, bimbingan, dan dukungan dari banyak pihak. Atas dasar tersebut, penulis merasa perlu menyampaikan apresiasi dan ucapan terima kasih yang setulusnya kepada:

1. Bapak Prof. Fathul Wahid, S.T., M.Sc., Ph.D., selaku Rektor Universitas Islam Indonesia, yang telah memberikan dukungan, arahan, dan kesempatan sehingga penulis dapat menjalani proses studi dengan baik di lingkungan Universitas Islam Indonesia.
2. Bapak Irving Vitra Paputungan, S.T., M.Sc., Ph.D., selaku Ketua Program Studi Magister Informatika Universitas Islam Indonesia, atas bimbingan, perhatian, serta berbagai fasilitas akademik yang memungkinkan kelancaran studi penulis.
3. Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D., selaku dosen pembimbing, yang telah memberikan pengarahan, saran, motivasi, serta kritik yang bersifat membangun sepanjang proses perumusan dan penyusunan tesis ini.
4. Seluruh dosen Program Studi Magister Informatika Universitas Islam Indonesia, yang telah memberikan ilmu, wawasan, dan pendampingan akademik selama masa studi. Semoga ilmu yang telah diberikan menjadi amal kebaikan yang terus mengalir manfaatnya.
5. Staf Akademik Program Studi Magister Informatika Universitas Islam Indonesia, atas bantuan, pelayanan, dan dukungan administratif yang sangat membantu kelancaran proses akademik penulis.

6. Kedua orang tua penulis yang tidak pernah berhenti mendoakan, memotivasi, dan memberikan semangat, sehingga penulis mampu menyelesaikan penelitian ini dengan baik.
7. Kedua Kakak penulis yang telah memberikan semangat dan memotivasi penulis untuk menyelesaikan penelitian ini dengan baik.
8. Indah Permata Sari, Anggi Rahmi Rosalina, dan Nazma Nur Shopa, yang dengan tulus memberikan motivasi dan semangat, sehingga turut membantu penulis dalam menyelesaikan penelitian ini secara optimal.
9. Penulis banyak memberikan terima kasih untuk Mas Afan yang telah memberikan banyak inspirasi, motivasi, dukungan dalam menyelesaikan penelitian ini.
10. Seluruh Teman-Teman pada Bidang Persandian dan Keamanan Informasi Dinas Komunikasi dan Informatika Provinsi Kalimantan Selatan, termasuk Kepala Bidang Persandian dan Keamanan Informasi, Ibu Sucilianita Akbar, SKM., M.M., Bapak Abdul Gafur, S.Pd., Bapak Bayu Aji Kondang Wibowo, S.E., Bapak Muhammad Andrianto, S.Kom., dan Bapak Firza Rizwandy, S.Kom., yang telah memberikan dorongan, motivasi, serta ilmu yang memberikan manfaat besar bagi penulis.
11. Teman-Teman Seperjuangan Program Studi Magister Informatika Angkatan 2022 yang telah memberikan dukungan, semangat, diskusi akademik, dan berbagai pengalaman selama masa perkuliagan dan penyelesaian tesis.
12. Ucapan terima kasih juga penulis sampaikan kepada seluruh teman-teman yang namanya tidak dapat disebutkan satu per satu, atas dukungan, perhatian, dan bantuan yang telah diberikan.

Daftar Isi

Daftar Tabel

Daftar Gambar

Glosarium

ADHD	- Attention Deficit Hyperactivity Disorder
AODE	- Average One-Dependence Estimator
ASD	- Autism Spectrum Disorder
BERT	- Bidirectional Encoder Representations Transformers
BiLSTM	- Bidirectional Long Short Term Memory
BPD	- Borderline Personality Disorder
CNN	- Convolutional Neural Network
DSM-5	- Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
FastText	- Fast Text Representation for Word Embedding
FN	- False Negative
FP	- False Positive
GAD	- Generalized Anxiety Disorder
LSTM	- Long Short Term Memory
MDD	- Major Depressive Disorder
NLP	- Natural Language Processing
OCD	- Obsessive-Compulsive Disorder
OOV	- Out-of-Vocabulary
PTSD	- Post-Traumatic Stress Disorder
RNN	- Recurrent Neural Network
RoBERTa	- Robustly Optimized BERT Pretraining Approach
SVM	- Support Vector Machine
TN	- True Negative
TP	- True Positive
XGBoost	- eXtreme Gradient Boosting

BAB 1
Pendahuluan

1.1 Pendahuluan

Lebih dari 1 miliar manusia terpengaruh oleh gangguan mental dengan prevalensi sekitar 16% dari populasi dunia (Rehm & Shield, 2019). Di Indonesia, berdasarkan riset kesehatan dasar tahun 2018, sebanyak 19 juta penduduk berusia di atas 15 tahun mengalami gangguan mental emosional, sementara 12 juta mengalami gangguan depresi. Prevalensi ini menunjukkan bahwa 20% dari populasi berpotensi menghadapi masalah gangguan kesehatan mental (Rokom, 2021). Kesehatan mental dapat mempengaruhi pikiran, emosi dan perilaku, termasuk berbagai jenis gangguan seperti depresi, kecemasan, gangguan bipolar, dan skizofrenia yang dapat menyebabkan pengaruh buruk terhadap kesehatan fisik individu (Singh et al., 2022; Zhang et al., 2022a).

Dalam era digital saat ini, media sosial telah menjadi platform utama dimana individu mengekspresikan perasaan dan berbagi pengalaman, namun paradoksnya, platform yang sama juga dapat memperparah kondisi kesehatan mental yang sudah ada. Data menunjukkan bahwa lebih dari 4,8 miliar orang menggunakan media sosial secara aktif, dengan rata-rata waktu penggunaan 2,5 jam per hari (Kemp, 2023). Platform seperti *Reddit*, *Twitter*, *Facebook*, dan *Instagram* tidak hanya berfungsi sebagai sarana komunikasi, tetapi juga menjadi cermin kondisi psikologis penggunanya. (Primack et al., 2017) dalam studi longitudinal terhadap 1.787 dewasa muda menemukan bahwa penggunaan media sosial yang intensif berkorelasi signifikan dengan peningkatan perasaan isolasi sosial dan gejala depresi. Temuan ini diperkuat oleh (Huang, 2017) yang menganalisis 80 studi empiris dan menemukan hubungan negatif yang konsisten antara waktu yang dihabiskan di media sosial dengan kesejahteraan psikologis.

Setiap jenis gangguan kesehatan mental mengalami perburukan yang spesifik melalui mekanisme media sosial yang berbeda. Individu dengan depresi mengalami perburukan melalui perbandingan sosial negatif dan *"echo chamber"* yang memperkuat pemikiran negatif, dimana (Lin et al., 2016) menemukan bahwa setiap peningkatan 10% dalam penggunaan Facebook berkorelasi dengan peningkatan 13% dalam gejala depresi pada 1.787 responden. Pengguna dengan gangguan kecemasan mengalami intensifikasi gejala melalui notifikasi terus-menerus, tekanan untuk merespons cepat, dan fenomena *Fear of Missing*

Out (FOMO), dimana (Woods & Scott, 2016) melaporkan korelasi signifikan antara penggunaan media sosial malam hari dengan peningkatan kecemasan pada 467 remaja. Gangguan bipolar diperparah melalui pola posting berlebihan saat episode manik dan isolasi digital saat fase depresi, dimana perubahan pola aktivitas digital dapat memprediksi episode *mood* dengan akurasi 80-90%, dengan perubahan muncul 2-3 hari sebelum onset episode klinis (Faurholt-Jepsen et al., 2014; Jacobson et al., 2019). Individu dengan skizofrenia mengalami penguatan delusi dan paranoia melalui paparan misinformasi dan *filter bubble* yang memperkuat keyakinan delusional (Birnbaum et al., 2017). *Borderline Personality Disorder* (BPD) diperparah melalui perilaku impulsif, posting emosional berlebihan, dan *fear of abandonment* yang diperkuat oleh interpretasi berlebihan terhadap aktivitas media sosial seperti tidak mendapat *like* atau tidak dibalas pesan (Roepke et al., 2015). Diskusi kesehatan mental umum sering memperparah kondisi melalui *self-diagnosis* tidak akurat dan paparan informasi yang menyesatkan (Naslund et al., 2016) yang paling berbahaya adalah penyebaran konten *poison* atau beracun yang dapat memicu *contagion effect* (peniruan perilaku), dimana (Marchant et al., 2017; Sedgwick et al., 2019) melaporkan bahwa paparan konten self-harm dan bunuh diri meningkatkan risiko imitasi perilaku berbahaya hingga 30% pada individu yang rentan.

Media sosial telah berkembang menjadi ruang interaksi yang tidak hanya digunakan untuk berbagi informasi, tetapi juga sebagai sarana ekspresi emosional dan pengalaman personal pengguna. Dalam konteks kesehatan mental, media sosial memungkinkan individu mengekspresikan kondisi psikologis, tekanan emosional, serta pengalaman hidup secara terbuka melalui teks. Pola bahasa, pilihan kata, dan narasi yang muncul dalam unggahan media sosial dapat mencerminkan kondisi kesehatan mental pengguna, sehingga membuka peluang pemanfaatan data media sosial sebagai sumber informasi untuk analisis gangguan kesehatan mental berbasis teks.

Kebutuhan akan analisis kondisi kesehatan mental melalui media sosial mendorong pengembangan model prediksi berbasis teks. Pengguna media sosial sering mengekspresikan suasana hati, emosi, dan pengalaman psikologis dengan variasi bahasa yang tinggi. Variasi ekspresi ini dapat dianalisis menggunakan teknik pemrosesan bahasa alami atau Natural Language Processing (NLP). NLP telah terbukti efektif dalam domain kesehatan dan biomedis sebagai bagian dari artificial intelligence yang mendukung analisis data tekstual skala besar untuk ekstraksi informasi dan analisis sentimen (Gonzalez-Hernandez et al., 2017; Nadkarni et al., 2011).

Sejumlah penelitian menunjukkan bahwa pola linguistik pada media sosial dapat menjadi indikator kondisi kesehatan mental yang dapat diprediksi secara otomatis. (Coppersmith et al., 2014) menunjukkan bahwa ekspresi bahasa pada media sosial berkorelasi dengan kondisi psikologis pengguna. Penelitian lain berhasil memprediksi depresi postpartum melalui analisis teks media sosial, yang membuktikan potensi prediksi otomatis menggunakan data digital (De Choudhury, 2013)

Dalam perkembangannya, penelitian-penelitian selanjutnya mulai mengadopsi teknik pembelajaran mesin dan deep learning untuk mendeteksi gangguan kesehatan mental dari data media sosial. Namun, sebagian besar penelitian masih berfokus pada satu atau dua jenis gangguan mental tertentu, seperti depresi atau ide bunuh diri (Castillo-sánchez et al., 2020; Franco-Martín et al., 2018; Ji et al., 2021a). Penelitian yang mengkaji lebih dari dua kategori gangguan mental umumnya masih terbatas pada jumlah kelas yang sempit dan belum mencakup deteksi konten berisiko tinggi secara komprehensif (Murarka & Raleigh, 2021; Zeberga et al., 2022).

Penelitian-penelitian terkini juga mulai mengeksplorasi penggunaan teknik deep learning dan pemrosesan bahasa alami yang lebih maju, seperti model berbasis transformer, untuk analisis kesehatan mental berbasis media sosial. Beberapa studi menerapkan pendekatan tersebut pada data Reddit untuk mengklasifikasikan sejumlah kondisi gangguan mental atau tugas spesifik seperti depresi dan ide bunuh diri, serta menunjukkan kemampuan representasi konteks linguistik yang baik. Meskipun demikian, penelitian-penelitian tersebut umumnya masih difokuskan pada tugas spesifik atau jumlah kategori gangguan mental yang terbatas, sehingga belum mencakup klasifikasi multi-kelas dengan spektrum gangguan kesehatan mental yang luas (Ameer et al., 2022; Ezerceli & Dehkharghani, 2024; Hasan & Saquer, 2024; Rahman et al., 2024) .

Selain keterbatasan cakupan kategori, aspek deteksi konten berisiko tinggi yang berpotensi menimbulkan dampak negatif terhadap pengguna, seperti konten berbahaya yang dapat memicu perilaku imitasi, juga relatif belum banyak dieksplorasi dalam kerangka klasifikasi multi-kelas gangguan kesehatan mental (Marchant et al., 2017). Kondisi ini menunjukkan bahwa masih terdapat celah penelitian dalam pengembangan pendekatan klasifikasi yang mampu menangani keragaman ekspresi bahasa media sosial sekaligus mendeteksi berbagai kategori gangguan kesehatan mental secara lebih komprehensif.

Berdasarkan keterbatasan tersebut, permasalahan utama dalam penelitian ini adalah belum optimalnya klasifikasi multi-kelas gangguan kesehatan mental pada data media sosial, khususnya dalam mengenali berbagai kategori gangguan kesehatan mental. Kondisi ini

menunjukkan perlunya suatu pendekatan klasifikasi yang mampu menangani kompleksitas bahasa media sosial sekaligus memperluas cakupan deteksi terhadap beragam kondisi kesehatan mental.

1.2 Rumusan Masalah

Bagaimana melakukan klasifikasi multi-kelas gangguan kesehatan mental pada data teks media sosial untuk mengenali tujuh kategori gangguan kesehatan mental?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah mengembangkan dan mengevaluasi model klasifikasi teks berbasis CNN-BiLSTM dengan FastText embedding untuk mendeteksi tujuh kategori gangguan kesehatan mental, yaitu depresi, kecemasan, bipolar, skizofrenia, gangguan kepribadian, mental illness, dan poison, menggunakan data teks dari media sosial. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat berupa pendekatan klasifikasi teks berbasis model hibrid CNN-BiLSTM dengan FastText embedding untuk mendeteksi berbagai kategori gangguan kesehatan mental pada data media sosial. Selain itu, penelitian ini diharapkan dapat memberikan pemahaman mengenai efektivitas pendekatan *hybrid* CNN-BiLSTM dalam menangani kompleksitas bahasa media sosial serta menjadi referensi bagi penelitian selanjutnya di bidang kesehatan mental berbasis komputasi.

1.5 Batasan Masalah

Penelitian ini dibatasi pada klasifikasi teks gangguan kesehatan mental menggunakan dua sumber data utama, yaitu Reddit dan Mental Health Text Corpus dari Kaggle. Data yang digunakan merupakan teks berbahasa Inggris dengan tujuh kategori gangguan kesehatan mental, yaitu depresi, kecemasan, bipolar, skizofrenia, gangguan kepribadian, mental illness, dan poison.

Model yang dikembangkan pada penelitian ini dibatasi pada pendekatan klasifikasi berbasis CNN-BiLSTM dengan FastText embedding. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score dalam lingkungan eksperimen. Penelitian ini tidak mencakup implementasi sistem secara langsung pada platform media sosial nyata maupun pengujian pada data di luar dataset yang digunakan.

1.6 Struktur Laporan

BAB 1 Pendahuluan

Bab ini berisi penjelasan mengenai latar belakang pentingnya deteksi kesehatan mental melalui data media sosial, temuan yang relevan dari penelitian sebelumnya, perumusan masalah, tujuan penelitian, manfaat penelitian, serta struktur penulisan laporan.

BAB 2 Tinjauan Pustaka

Bab ini memuat kajian kritis terhadap pustaka yang berkaitan dengan kesehatan mental, pengolahan bahasa alami (NLP), *machine learning* dan *deep learning*, serta penelitian terdahulu terkait prediksi kesehatan mental menggunakan model CNN-BiLSTM dan *embedding FastText*.

BAB 3 Metodologi

Bab ini menjelaskan secara rinci metode penelitian yang digunakan, termasuk proses pengumpulan dataset dari *Reddit* dan *Kaggle*, tahapan pra-pemrosesan data teks, pembobotan kata dengan *FastText*, arsitektur model CNN-BiLSTM, serta evaluasi performa model menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score.

BAB 4 Hasil dan Pembahasan

Bab ini menyajikan hasil dari proses pelatihan dan pengujian model CNN-BiLSTM, termasuk visualisasi kinerja model, analisis *confusion matrix*, pengaturan *hyperparameter*, dan diskusi terhadap hasil yang diperoleh dibandingkan dengan penelitian sebelumnya.

BAB 5 Kesimpulan dan Saran

Bab ini berisi kesimpulan dari penelitian yang dilakukan, termasuk bagaimana hasil penelitian menjawab rumusan masalah yang telah ditetapkan, implikasi dari temuan penelitian terhadap pengembangan model deteksi kesehatan mental berbasis media sosial, serta saran untuk penelitian selanjutnya, seperti pengembangan model dengan arsitektur lain atau penerapan pada *platform* lain di luar *Reddit*.

BAB 2
Tinjauan Pustaka

2.1 Penelitian Terdahulu

Tinjauan pustaka dalam penelitian ini disusun secara sistematis untuk memberikan pemahaman komprehensif mengenai pengembangan model prediksi kesehatan mental berbasis data media sosial. Penelitian terdahulu yang relevan dikelompokkan menjadi tiga tema utama yang saling berkaitan. Pertama, tinjauan sistematis tentang penerapan machine learning dan *natural language processing* (NLP) dalam deteksi gangguan mental, yang memberikan gambaran umum tentang lanskap penelitian di bidang ini (dirangkum dalam Tabel 2.1). Kedua, implementasi spesifik metode NLP dalam mendeteksi berbagai jenis gangguan mental dari data teks media sosial (dirangkum dalam Tabel 2.2). Ketiga, eksplorasi arsitektur *CNN-BiLSTM* untuk klasifikasi teks, yang menjadi landasan metodologis bagi penelitian ini (dirangkum dalam Tabel 2.3). Pembahasan dimulai dari tinjauan literatur yang bersifat makro, kemudian mengerucut ke implementasi teknis yang lebih spesifik.

Penulis mengkaji sejumlah penelitian terkait pengembangan model prediksi kesehatan mental melalui media sosial, salah satunya adalah tinjauan sistematis oleh (Le Glaz et al., 2021) yang dirangkum dalam Tabel 2.1. Studi tersebut mengevaluasi 327 artikel, dengan 58 artikel yang memenuhi kriteria inklusi. Fokus utamanya mencakup penerapan *machine learning* dan *natural language processing* dalam deteksi gangguan mental, serta potensi integrasinya dalam praktik klinis. Meskipun beragam, studi-studi tersebut umumnya membahas tiga kelompok populasi dan menyoroti aspek identifikasi gejala, klasifikasi tingkat keparahan, serta efektivitas terapi. Data yang digunakan sebagian besar berasal dari catatan medis dan media sosial, dengan pendekatan pra-pemrosesan berbasis NLP dan penggunaan platform Python.

Penelitian kedua yang dilakukan oleh (Zhang et al., 2022) menyusun tinjauan naratif tentang deteksi gangguan mental dengan menggunakan NLP dalam satu dekade terakhir, dengan tujuan memahami metode, tren, hambatan, dan arah ke depan. Dalam ulasan ini, terdapat 399 studi yang diambil dari 10,467 catatan. Hasil tinjauan menunjukkan bahwa terdapat tren positif dalam penelitian NLP untuk deteksi gangguan mental, di mana metode *deep learning* mendapat lebih banyak perhatian dan memberikan hasil yang lebih baik dibandingkan dengan metode pembelajaran mesin konvensional.

Penelitian ketiga (Chung & Teo, 2022) mengumpulkan 30 artikel terkait pendekatan *machine learning* dalam memprediksi masalah kesehatan mental melalui tinjauan sistematis menggunakan metodologi PRISMA. Artikel yang dikategorikan berdasarkan jenis masalah kesehatan mental menunjukkan beragam tingkat akurasi model *machine learning*: AODE 71%, *Multilayer Perceptron* 78%, *Logical Analysis Tree* 70%, *Multiclass Classifier* 58%, *Radial Basis Function Network* 57%, *K-star* 42%, dan *Functional Tree* 42%. Meski demikian, penelitian ini memiliki batasan seperti sampel kecil, validasi yang kurang memadai, terbatasnya eksplorasi dalam deep learning, kurangnya uji di kehidupan nyata, dan kekurangan data berkualitas tinggi.

Penelitian keempat (Garriga et al., 2022) mengembangkan model pembelajaran mesin menggunakan catatan kesehatan elektronik untuk memantau pasien berisiko krisis kesehatan mental selama 28 hari. Dataset berisi 5.816.586 catatan dari 17.122 pasien antara September 2012 dan November 2018 dengan berbagai gangguan diagnostik. Model ini memiliki sensitivitas 58% dan spesifisitas 85%, dengan evaluasi prospektif 6 bulan menunjukkan nilai prediksi yang bernilai dalam pengelolaan kasus atau pengurangan risiko krisis pada 64% kasus.

Penelitian kelima (J. Kim et al., 2020) mengumpulkan data dari enam subreddit kesehatan mental yang berfokus pada depresi, kecemasan, bipolar, gangguan psikologis, skizofrenia, dan autisme. Studi ini memanfaatkan data dari 248.537 pengguna, yang menulis 633.385 posting di tujuh *subreddit* dari Januari 2017 hingga Desember 2018. Hasil analisis menunjukkan bahwa pada *subreddit* r/depression, yang terbagi menjadi dua kelas yaitu *depression* dan *non-depression*, Model *XGBoost* dan CNN berhasil mencapai tingkat akurasi yang signifikan pada berbagai subreddit kesehatan mental. Pada *subreddit* r/depression, *XGBoost* mencapai 71.69%, sementara CNN mencapai 75.13%. Di r/anxiety, XGBoost mencapai 70.41%, dan CNN mencapai 77.81%.

Tabel 2.1 Ulasan Kritis Tema 1: Machine Learning untuk Deteksi Mental Health

No.	Penulis (Tahun)	Metode/Fokus	Hasil Utama
1	(Le Glaz et al., 2021)	*Systematic review* ML & NLP (327 artikel, 58 dianalisis)	58 artikel (17.7%) memenuhi kriteria. Identifikasi 3 kelompok populasi: pasien EHR, ruang gawat darurat, media sosial.
2	(Zhang et al., 2022)	*Narrative review* NLP (399 dari 10,467 studi)	Tren positif NLP untuk deteksi gangguan mental. *Deep learning* mengungguli ML konvensional.
3	(Chung & Teo, 2022)	*Systematic review* ML (30 artikel, PRISMA)	Akurasi bervariasi: AODE 71%, MLP 78%, LAT 70%, MC 58%, RBFN 57%
4	(Garriga et al., 2022)	ML untuk prediksi krisis menggunakan EHR (5.8M catatan, 17K pasien)	Sensitivitas 58%, spesifisitas 85%, AUC-ROC 0.797. Evaluasi 6 bulan: reduksi risiko 64%
5	(J. Kim et al., 2020)	CNN & XGBoost multi-label (248K user, 633K posting Reddit)	r/depression: XGBoost 71.69%, CNN 75.13%. r/anxiety: XGBoost 70.41%, CNN 77.81%

Berbagai tinjauan sistematis yang telah dipaparkan sebelumnya menunjukkan bahwa penggunaan machine learning dan *natural language processing* (NLP) dalam deteksi kesehatan mental mengalami perkembangan yang signifikan. Dalam beberapa tahun terakhir, metode *deep learning* semakin dominan digunakan karena kemampuannya dalam mengenali pola bahasa yang kompleks pada data teks. Tinjauan sistematis oleh (Madububambachu et al., 2024), yang menganalisis 30 studi dari tahun 2011–2024, memperlihatkan bahwa model *Convolutional Neural Networks* (CNN) mampu mencapai tingkat akurasi yang sangat tinggi, khususnya dalam mendeteksi gangguan seperti *bipolar disorder*. Penelitian komparatif oleh (Ding et al., 2025) juga menunjukkan temuan serupa. Meskipun model *machine learning* tradisional menawarkan tingkat interpretabilitas yang lebih baik, model *deep learning* terbukti lebih kuat dalam menangani pola bahasa yang lebih kompleks sebagaimana ditemukan pada teks media sosial.

Selain itu, *scoping review* oleh (Su et al., 2020) mengungkapkan bahwa mayoritas model *deep learning* yang digunakan dalam penelitian terkait kesehatan mental melaporkan

akurasi prediksi yang lebih tinggi dibandingkan teknik *machine learning* konvensional. Temuan ini berlaku untuk berbagai tipe data, termasuk data klinis, data genetik, maupun data teks dari media sosial. Dengan demikian, pendekatan *deep learning* semakin diakui memiliki kapasitas yang lebih baik dalam mengolah data dengan pola yang beragam dan kompleks.

Meski demikian, untuk memahami bagaimana teknik NLP secara khusus diterapkan dalam mendeteksi gangguan kesehatan mental dari teks media sosial, diperlukan telaah terhadap implementasi penelitian yang lebih aplikatif. *Literature review* oleh (Teferra et al., 2024) menekankan bahwa meskipun NLP memberikan peluang besar untuk meningkatkan deteksi depresi, terdapat sejumlah tantangan seperti isu etika, pedoman tata kelola, serta perbedaan ekspresi bahasa antarbudaya. Penelitian-penelitian yang lebih terfokus ini mencakup berbagai topik, mulai dari deteksi risiko bunuh diri, klasifikasi multi-kategori gangguan mental, hingga pemanfaatan model *transformer modern* seperti *BERT* dan *RoBERTa*. Studi-studi tersebut dirangkum dalam Tabel 2.2 yang menggambarkan perkembangan pendekatan dari metode *machine learning* klasik menuju metode *deep learning modern* dalam bidang kesehatan mental.

Penelitian yang fokus pada penerapan *Natural Language Processing* (NLP) dalam konteks kesehatan mental menunjukkan perkembangan yang signifikan dalam kemampuan analisis teks untuk mendeteksi indikasi gangguan psikologis. Penelitian keenam oleh (Jain et al., 2022) menyelidiki *subreddit* r/*depression* dan r/*SuicideWatch* untuk memahami pemikiran bunuh diri melalui pendekatan NLP. Studi ini menganalisis pola linguistik yang mengindikasikan risiko bunuh diri dengan menggunakan algoritma pembelajaran mesin seperti *Naive Bayes, Support Vector Machine* (SVM), dan *Random Forest*. Hasil penelitian menunjukkan bahwa *r/SuicideWatch* dianggap sebagai sumber bantuan yang sah oleh individu yang berisiko, dengan model klasifikasi mencapai akurasi yang cukup tinggi dalam membedakan posting yang mengandung ide bunuh diri versus posting biasa. Model terbaik mampu mengidentifikasi pola linguistik spesifik seperti ekspresi putus asa, kehilangan harapan, dan referensi eksplisit terhadap *self-harm* yang menjadi indikator kuat ide bunuh diri.

Penelitian ketujuh oleh (Chandran et al., 2019) mengembangkan algoritme *Natural Language Processing* (NLP) berbasis aturan untuk mengidentifikasi gejala *obsessive-compulsive symptoms* (OCS) dari rekam medis elektronik (*Electronic Health Records/EHR)* kesehatan mental. Algoritme ini dirancang untuk mengekstraksi informasi klinis kompleks dari catatan teks tidak terstruktur yang ditulis oleh profesional kesehatan. Hasil penelitian menunjukkan bahwa algoritme mampu mengidentifikasi gejala terkait OCS dengan tingkat

presisi 0.77 dan 0.67 untuk berbagai jenis gejala obsesif-kompulsif. Aplikasi NLP ini menunjukkan potensi signifikan dalam mengekstrak data gejala kompleks dari EHR perawatan kesehatan mental, memungkinkan analisis lebih lanjut terkait prognosis dan respons intervensi.

Penelitian kedepalan oleh (Zeberga et al., 2022) mengembangkan pendekatan text mining menggunakan arsitektur Bi-LSTM dan BERT untuk prediksi kesehatan mental dengan fokus pada dua kategori utama: depresi dan kecemasan. Penelitian ini menggunakan dataset dari media sosial dan menerapkan teknik *deep learning* untuk menangkap pola linguistik kompleks yang mengindikasikan kondisi mental tertentu. Model Bi-LSTM dipilih karena kemampuannya dalam memahami konteks dua arah dalam teks, sementara *BERT (Bidirectional Encoder Representations from Transformers)* digunakan untuk memanfaatkan *pre-trained language* model yang telah dilatih pada *corpus* besar. Hasil evaluasi menunjukkan bahwa *BERT* memberikan performa *superior* dengan kemampuan menangkap nuansa semantik dan kontekstual yang lebih baik dibandingkan Bi-LSTM standalone.

Penelitian kesembilan oleh (Murarka & Raleigh, 2021) menggunakan RoBERTa *(Robustly Optimized BERT Pretraining Approach)* untuk klasifikasi lima kategori penyakit mental dari data Reddit, yaitu *ADHD, anxiety, bipolar, depression, dan PTSD*. Dataset terdiri dari 17.159 postingan teks dan judul yang dikumpulkan dari 13 *sub-reddit* melalui proses *crawling*. Model RoBERTa, sebagai versi optimisasi dari BERT, dipilih karena kemampuannya dalam menangani konteks panjang dan pemahaman semantik yang lebih dalam. Hasil menunjukkan bahwa RoBERTa mencapai *F1-score* tertinggi: 0.86 untuk label post, 0.72 untuk title, dan akurasi 0.89 untuk kombinasi post+title.

Penelitian kesepuluh oleh (Tejaswini et al., 2022) mengembangkan model *hybrid deep learning* untuk deteksi depresi dari analisis teks media sosial menggunakan teknik NLP. Penelitian ini dipublikasikan dalam *ACM Transactions on Asian and Low-Resource Language Information Processing*, sebuah jurnal bereputasi tinggi di bidang pemrosesan bahasa alami. Studi ini mengintegrasikan berbagai teknik NLP *preprocessing* dengan arsitektur *deep learning* yang menggabungkan CNN dan LSTM untuk mengekstraksi fitur semantik dan kontekstual dari teks media sosial. Model *hybrid* yang dikembangkan menunjukkan peningkatan signifikan dalam akurasi deteksi depresi dibandingkan dengan pendekatan konvensional, dengan kemampuan menangkap nuansa linguistik yang kompleks seperti penggunaan bahasa informal, *slang*, dan ekspresi emosional implisit yang umum ditemukan dalam posting media sosial. Penelitian ini juga mengeksplorasi pengaruh

berbagai teknik *preprocessing* seperti *tokenization*, *stemming*, dan stopword removal terhadap performa model, memberikan wawasan penting tentang pipeline NLP yang optimal untuk tugas deteksi kesehatan mental.

Tabel 2.2 Ulasan Kritis Tema 2: Natural Language Processing dalam Kesehatan Mental

No.	Penulis (Tahun)	Metode/Fokus	Hasil Utama
6	(Jain et al., 2022)	NLP untuk *suicide detection* (r/depression & r/SuicideWatch)	r/SuicideWatch sebagai sumber bantuan sah. Algoritma *Naive Bayes*, SVM, RF untuk identifikasi pola linguistik risiko bunuh diri.
7	(Chandran et al., 2019)	NLP berbasis aturan untuk identifikasi OCS dari EHR kesehatan mental	Presisi 0.77 dan 0.67 dalam mengidentifikasi gejala OCS. NLP ekstrak data gejala kompleks dari EHR.
8	(Zeberga et al., 2022)	Bi-LSTM & BERT untuk prediksi depresi dan kecemasan (2 kategori)	BERT superior dalam menangkap nuansa semantik. Akurasi tinggi untuk klasifikasi depresi dan kecemasan.
9	(Murarka & Raleigh, 2021)	RoBERTa untuk klasifikasi 5 kategori (ADHD, anxiety, bipolar, depression, PTSD) - 17K posting	RoBERTa F1-score: 0.86 (post), 0.72 (title), akurasi 0.89 (post+title).
10	(Tejaswini et al., 2022)	*Hybrid deep learning* untuk depression detection menggunakan NLP dan *preprocessing techniques*	Model hybrid CNN-LSTM menunjukkan peningkatan signifikan vs metode konvensional. Efektif menangkap nuansa linguistik informal media sosial.

Penelitian-penelitian sebelumnya menunjukkan bahwa berbagai metode *NLP*—mulai dari algoritma klasik seperti *Naive Bayes* dan *Support Vector Machine (SVM)* hingga model *transformer* seperti *BERT* dan *RoBERTa*—telah memberikan hasil yang menjanjikan dalam deteksi gangguan mental. Namun demikian, pemilihan arsitektur model yang paling sesuai

tetap menjadi faktor penting untuk mencapai performa klasifikasi yang optimal. Tinjauan terbaru oleh (Teferra et al., 2024) menekankan bahwa meskipun *NLP* berpotensi besar untuk meningkatkan deteksi depresi, masih terdapat sejumlah tantangan, seperti keterbatasan interpretabilitas model, kebutuhan personalisasi, serta pentingnya kolaborasi dengan para ahli data seperti *data scientists* dan *machine learning engineers*. Studi tersebut juga menyoroti bahwa penelitian ke depan perlu berfokus pada peningkatan interpretabilitas, pengembangan model yang lebih adaptif terhadap pengguna, serta kolaborasi lintas bidang.

Model *transformer* seperti *BERT* dan *RoBERTa* memang mampu menghasilkan akurasi yang tinggi, tetapi memerlukan sumber daya komputasi yang besar dan sulit dijelaskan cara kerjanya. Hal ini sejalan dengan temuan (Ding et al., 2025), yang menunjukkan bahwa model *machine learning* tradisional lebih mudah dipahami karena menyediakan informasi mengenai pentingnya setiap variabel, sementara model *deep learning* lebih baik dalam menangani pola bahasa yang kompleks. Penelitian tersebut membandingkan berbagai model *machine learning* (seperti *logistic regression*, *random forest*, dan *LightGBM*) serta arsitektur *deep learning* (seperti *ALBERT* dan *GRU*) pada tugas klasifikasi kondisi kesehatan mental. Hasilnya menunjukkan bahwa kedua jenis model dapat mencapai performa yang serupa pada ukuran dataset menengah, tetapi masing-masing memiliki kelebihan dan kekurangan dalam hal akurasi, kemudahan interpretasi, dan efisiensi komputasi.

Sementara itu, arsitektur seperti *CNN* memiliki keunggulan dalam mengenali pola-pola lokal pada teks, tetapi kurang mampu memahami urutan konteks yang panjang. Sebaliknya, model *LSTM* dan *BiLSTM* lebih baik dalam menangkap hubungan kata dalam jangka panjang, tetapi tidak seefisien *CNN* dalam mengekstraksi fitur lokal. Untuk mengatasi perbedaan ini, arsitektur hibrida *CNN–BiLSTM* dikembangkan sebagai pendekatan yang menggabungkan kedua keunggulan tersebut: *CNN* untuk mengekstraksi fitur lokal dan *BiLSTM* untuk memahami konteks sekuensial. Pendekatan ini juga lebih efisien dari segi komputasi dibandingkan model *transformer*.

Tinjauan sistematis oleh (Madububambachu et al., 2024), yang menganalisis 30 studi dari tahun 2011–2024, menemukan bahwa *CNN* merupakan salah satu model yang menunjukkan kinerja sangat baik, terutama ketika digabungkan dengan arsitektur *recurrent* untuk memahami pola berurutan pada data. Studi tersebut melaporkan bahwa dibandingkan model lain seperti *Random Forest*, *SVM*, dan beberapa model *deep learning* lainnya, *CNN* memberikan akurasi yang lebih tinggi, khususnya dalam mendeteksi *bipolar disorder*. Temuan ini memberikan dasar teoritis yang kuat untuk mengeksplorasi arsitektur hibrida

yang menggabungkan *CNN* dan *BiLSTM*. Bagian berikutnya membahas penelitian-penelitian yang secara khusus mengembangkan dan menerapkan arsitektur *CNN–BiLSTM* untuk klasifikasi teks, yang dirangkum dalam Tabel 2.3. Kelompok penelitian ini menjadi landasan metodologis utama bagi pemilihan arsitektur dalam penelitian ini, mengingat keunggulan komplementer dari kombinasi *CNN* dan *BiLSTM* dalam menangani karakteristik teks dari media sosial yang cenderung kompleks.

Penelitian tentang arsitektur CNN-BiLSTM untuk klasifikasi teks menunjukkan keunggulan signifikan dibandingkan pendekatan tunggal dalam menangani kompleksitas data teks media sosial. Penelitian pertama oleh (Kour & Gupta, 2022) mengembangkan model *hybrid deep learning* yang menggabungkan *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM) untuk memprediksi kesehatan mental pengguna melalui kategorisasi biner: depresi atau non-depresi pada data Twitter. Dataset penelitian terdiri dari 2.558 sampel depresi, 5.304 sampel non-depresi, dan 58.810 sampel kandidat depresi yang dikumpulkan dari platform Twitter. Setelah melalui proses optimasi hyperparameter yang sistematis, model ini mencapai akurasi 94.28% pada dataset depresi. Perbandingan dengan pendekatan baseline menunjukkan bahwa CNN-biLSTM memberikan peningkatan signifikan dalam performa klasifikasi.

Penelitian keduabelas oleh (Merinda Lestandy & Abdurrahim, 2023) mengeksplorasi efektivitas pembobotan Word2Vec yang dikombinasikan dengan model CNN-BiLSTM untuk klasifikasi emosi pada teks bahasa Indonesia. Penelitian ini menggunakan dataset yang terbagi menjadi 16.000 data pelatihan dan 2.000 data pengujian untuk menghindari overfitting. Fokus penelitian adalah pada klasifikasi enam label emosi: *surprise, sadness, rage, fear, love, dan joy*. Hasil eksperimen menunjukkan bahwa model CNN-BiLSTM dengan pembobotan Word2Vec mencapai akurasi tertinggi sebesar 92.85%, melampaui performa model baseline seperti CNN standalone (89.3%), BiLSTM *standalone* (90.7%), dan LSTM konvensional (88.2%).

Penelitian ketigabelas oleh (Xiaoyan et al., 2022) mengembangkan arsitektur GloVe-CNN-BiLSTM untuk analisis sentimen pada teks review. Penelitian ini dipublikasikan dalam *Journal of Sensors*, menggabungkan tiga komponen utama: GloVe *(Global Vectors for Word Representation)* sebagai metode *embedding*, CNN untuk ekstraksi fitur lokal, dan BiLSTM untuk pemahaman konteks sekuensial. Model mencapai akurasi 89.5% dan F1-score 0.88 pada dataset pengujian, mengungguli CNN *standalone* (85.2%), BiLSTM standalone (86.7%), dan LSTM-Attention (87.3%). Analisis *ablation study* mengonfirmasi

bahwa setiap komponen memberikan kontribusi unik dalam meningkatkan performa klasifikasi.

Penelitian keempatbelas oleh (Rhanoui & Mikram, 2019) mengusulkan arsitektur CNN-BiLSTM untuk analisis sentimen tingkat dokumen. Penelitian ini dipublikasikan dalam jurnal Make dan mengeksplorasi kombinasi optimal antara *convolutional layers* untuk ekstraksi fitur lokal dan *bidirectional* LSTM untuk menangkap dependensi jangka panjang dalam dokumen teks yang panjang. Berbeda dengan pendekatan sebelumnya yang fokus pada klasifikasi tingkat kalimat, penelitian ini menangani tantangan analisis sentimen pada dokumen utuh yang memiliki struktur lebih kompleks dan konteks yang lebih luas. Hasil eksperimen menunjukkan bahwa CNN-BiLSTM mampu menangkap tidak hanya fitur lokal seperti n-gram dan frasa penting, tetapi juga hubungan semantik antar paragraf yang krusial untuk memahami sentimen keseluruhan dokumen. Model ini mencapai performa superior dibandingkan dengan RNN dan CNN konvensional, terutama pada dokumen dengan sentimen yang ambigu atau kontradiktif.

Penelitian kelimabelas oleh (Yin et al., 2017) melakukan studi komparatif komprehensif antara CNN dan RNN untuk tugas *natural language processing*, yang dipublikasikan sebagai *comparative study* yang banyak dijadikan referensi dalam komunitas NLP. Penelitian ini menganalisis kekuatan dan kelemahan fundamental dari kedua arsitektur dalam berbagai tugas NLP termasuk klasifikasi teks, *sentiment analysis*, dan *sequence labeling*. Hasil studi menunjukkan bahwa CNN unggul dalam menangkap pola lokal dan invariansi posisi, membuatnya efektif untuk deteksi fitur seperti n-gram dan frasa kunci. Di sisi lain, RNN (termasuk LSTM dan BiLSTM) menunjukkan keunggulan dalam menangkap dependensi jangka panjang dan pemahaman konteks sekuensial. Penelitian ini memberikan wawasan penting bahwa kombinasi kedua pendekatan (hybrid CNN-RNN) dapat memanfaatkan kekuatan komplementer dari keduanya, menjadi landasan teoretis bagi pengembangan arsitektur *hybrid* seperti CNN-BiLSTM yang telah terbukti efektif dalam berbagai aplikasi NLP termasuk deteksi kesehatan mental dari teks media sosial.

Tabel 2.3 Ulasan Kritis Tema 3: Arsitektur CNN-BiLSTM untuk Klasifikasi Teks

No.	Penulis (Tahun)	Metode/Fokus	Hasil Utama
11	(Kour & Gupta, 2022)	CNN-biLSTM untuk klasifikasi biner depresi/non-depresi (Twitter: 66K total)	Akurasi 94.28% setelah optimasi. Perbedaan signifikan representasi linguistik depresi vs non-depresi.

12	(Merinda Lestandy & Abdurrahim, 2023)	CNN-BiLSTM + Word2Vec untuk klasifikasi 6 emosi bahasa Indonesia (18K data)	Akurasi 92.85%, mengungguli CNN (89.3%), BiLSTM (90.7%), LSTM (88.2%). Optimal untuk joy & sadness.
13	(Xiaoyan et al., 2022)	GloVe-CNN-BiLSTM untuk *sentiment analysis* teks review (3 kategori)	Akurasi 89.5%, F1 0.88. Mengungguli CNN (85.2%), BiLSTM (86.7%), LSTM-Attention (87.3%).
14	(Rhanoui & Mikram, 2019)	CNN-BiLSTM untuk *document-level sentiment analysis*	Superior vs RNN dan CNN konvensional. Efektif menangkap fitur lokal (n-gram) dan hubungan semantik antar paragraf untuk dokumen kompleks.
15	(Yin et al., 2017)	Studi komparatif CNN vs RNN untuk NLP tasks (klasifikasi, sentiment, sequence labeling)	CNN unggul untuk pola lokal & n-gram. RNN unggul untuk dependensi jangka panjang. Hybrid CNN-RNN optimal untuk leverage kekuatan keduanya.

2.1.1 Kontribusi Penelitian

Berdasarkan tinjauan terhadap penelitian terdahulu yang telah dipaparkan, teridentifikasi beberapa gap penelitian yang signifikan. Pertama, sebagian besar penelitian sebelumnya terbatas pada satu atau dua kategori gangguan mental, seperti penelitian (Castillo-Sánchez et al., 2020; Franco-Martín et al., 2018; Ji et al., 2021) yang hanya fokus pada suicide prediction, atau (Giuntini et al., 2020; Khan et al., 2018; Mahdy et al., 2020) yang terbatas pada deteksi depresi. Penelitian (Zeberga et al., 2022) baru mencakup dua kategori (depresi dan kecemasan), sementara (Murarka & Raleigh, 2021) membatasi pada lima kategori tanpa mencakup deteksi konten berbahaya *(poison)*.

Kedua, sebagian besar penelitian menggunakan arsitektur model tunggal yang tidak optimal untuk menangkap kompleksitas bahasa informal media sosial. Penelitian (J. Kim et al., 2020) menggunakan CNN dan XGBoost secara terpisah, (Kour & Gupta, 2022) menggunakan CNN-biLSTM namun hanya untuk klasifikasi biner (depresi vs non-depresi), dan (Ameer et al., 2022) menggunakan RoBERTa yang memerlukan *computational resources* sangat tinggi dan kurang *interpretable*.

Ketiga, belum ada penelitian yang secara spesifik mengeksplorasi kombinasi CNN-BiLSTM dengan *FastText embedding* untuk klasifikasi multi-label kesehatan mental yang mencakup tujuh kategori sekaligus. Penelitian (Merinda Lestandy & Abdurrahim, 2023) menggunakan CNN-BiLSTM dengan Word2Vec untuk klasifikasi emosi, namun bukan untuk gangguan kesehatan mental dan tidak menggunakan FastText yang lebih superior dalam menangani *out-of-vocabulary words*.

Keempat, kategori "poison" untuk mendeteksi konten berbahaya yang dapat memperparah kondisi pengguna lain belum banyak dieksplorasi secara spesifik dalam penelitian klasifikasi gangguan mental. Meskipun penelitian seperti (Jain et al., 2022) membahas *suicide detection*, fokusnya adalah pada identifikasi individu berisiko bunuh diri, bukan pada deteksi konten beracun yang dapat memicu contagion effect pada komunitas yang lebih luas.

Penelitian ini mengisi gap tersebut dengan mengembangkan model hibrid *CNN-BiLSTM* yang dikombinasikan dengan *FastText embedding* untuk mengklasifikasikan tujuh kategori gangguan kesehatan mental sekaligus, termasuk kategori *poison* yang krusial untuk pencegahan penyebaran konten berbahaya di media sosial. Pendekatan hibrid ini dirancang untuk mengintegrasikan kekuatan CNN dalam mengekstraksi fitur lokal dengan kemampuan BiLSTM dalam memahami konteks sekuensial, sementara FastText mengatasi masalah out-of-vocabulary yang umum dalam teks media sosial informal.

2.2 Kesehatan Mental

Kesehatan mental merupakan kondisi kesejahteraan emosional, psikologis, dan sosial yang memengaruhi cara individu berpikir, merasa, dan bertindak. Dalam konteks klinis, definisi dan klasifikasi gangguan mental secara luas diacu pada Diagnostic and Statistical Manual of Mental Disorders, edisi ke-5 (DSM-5) (American Psychiatric Association, 2013), yang diterbitkan oleh American Psychiatric Association. DSM-5 digunakan secara internasional sebagai standar untuk diagnosis gangguan mental dan memberikan kerangka kerja diagnostik yang sistematis.

Dalam penelitian ini, fokus label klasifikasi didasarkan pada tujuh kategori yang dapat dipetakan ke dalam klasifikasi DSM-5, yaitu: depresi, kecemasan, gangguan bipolar, skizofrenia, gangguan kepribadian (seperti Borderline Personality Disorder/BPD), mental illness (umum/nonspesifik), dan poison (konten berbahaya atau indikasi keinginan menyakiti diri sendiri/others) (American Psychiatric Association, 2013).

2.2.1 Depresi (Major Depressive Disorder)

Menurut DSM-5, depresi ditandai oleh perasaan sedih berkepanjangan, hilangnya minat atau kesenangan dalam aktivitas, serta gejala tambahan seperti gangguan tidur, perubahan nafsu makan, kelelahan, dan pikiran untuk menyakiti diri sendiri atau bunuh diri. Diagnosis memerlukan setidaknya lima gejala yang berlangsung selama dua minggu atau lebih (American Psychiatric Association, 2013).

2.2.2 Gangguan Kecemasan (Anxiety Disorders)

Kategori ini mencakup gangguan seperti Generalized Anxiety Disorder (GAD), Panic Disorder, dan Social Anxiety Disorder. Gejalanya meliputi kekhawatiran berlebihan, gelisah, ketegangan otot, dan kesulitan konsentrasi. DSM-5 menekankan bahwa gangguan ini harus cukup parah untuk mengganggu fungsi sosial dan pekerjaan (American Psychiatric Association, 2013).

2.2.3 Gangguan Bipolar (Bipolar I & II Disorders)

Gangguan ini ditandai oleh episode mood ekstrem, termasuk mania (atau hipomania) dan depresi. Episode mania melibatkan suasana hati yang sangat meningkat, peningkatan energi, dan aktivitas impulsif, sedangkan episode depresi menyerupai gejala MDD. DSM-5 membedakan antara Bipolar I (dengan episode mania penuh) dan Bipolar II (hipomania + depresi mayor) (American Psychiatric Association, 2013).

2.2.4 Skizofrenia

Skizofrenia ditandai oleh gangguan dalam proses berpikir, persepsi, dan afek. Gejala utama mencakup delusi, halusinasi, ucapan tidak teratur, dan perilaku motorik tidak wajar. Gangguan ini termasuk dalam spektrum skizofrenia menurut DSM-5 (American Psychiatric Association, 2013).

2.2.5 Gangguan Kepribadian (Personality Disorders)

Salah satu fokus dalam penelitian ini adalah Borderline Personality Disorder (BPD), yang termasuk dalam cluster B dalam DSM-5. BPD dicirikan oleh ketidakstabilan emosi, citra diri yang terganggu, impulsivitas, serta hubungan interpersonal yang intens dan tidak stabil (American Psychiatric Association, 2013).

2.2.6 Mental Illness (umum/nonspesifik)

Kategori ini mengacu pada kondisi mental yang tidak dijelaskan secara spesifik oleh label-label lain, namun mencerminkan adanya gangguan psikologis berdasarkan ekspresi bahasa pengguna di media sosial.

2.2.7 Poison (Konten Berbahaya atau Self-harm)

Kategori "poison" dalam penelitian ini merujuk pada konten berbahaya yang dapat memperparah kondisi kesehatan mental pengguna lain atau memicu perilaku destruktif. Meskipun tidak diklasifikasikan sebagai gangguan spesifik dalam DSM-5, kategori ini sangat penting dalam konteks media sosial karena terkait dengan fenomena contagion effect, dimana paparan terhadap konten self-harm, bunuh diri, atau kekerasan dapat meningkatkan risiko imitasi perilaku berbahaya pada individu yang rentan.

Penelitian (Marchant et al., 2017) menunjukkan bahwa paparan konten self-harm di media sosial meningkatkan risiko imitasi perilaku berbahaya hingga 30% pada remaja rentan. (Sedgwick et al., 2019) menemukan bahwa konten bunuh diri yang viral dapat meningkatkan tingkat percobaan bunuh diri lokal hingga 13%. (Hawton et al., 2012) menegaskan bahwa media sosial bertanggung jawab atas 28% kasus contagion effect dalam perilaku self-harm pada remaja.

Dalam penelitian ini, konten dikategorikan sebagai "poison" jika mengandung salah satu atau kombinasi dari karakteristik berikut: (1) deskripsi eksplisit tentang metode bunuh diri atau self-harm, (2) ajakan atau normalisasi perilaku menyakiti diri sendiri, (3) konten yang mempromosikan atau meromantisasi kematian, (4) ujaran kebencian atau ancaman kekerasan yang dapat memicu distres psikologis, (5) informasi tentang akses ke zat berbahaya atau metode membahayakan diri. Deteksi otomatis terhadap kategori ini krusial untuk sistem moderasi konten dan pencegahan penyebaran konten yang dapat membahayakan kesehatan mental komunitas pengguna media sosial secara luas.

Dengan mengacu pada DSM-5, klasifikasi label dalam penelitian ini memperoleh landasan teoretis yang kuat dan selaras dengan standar klinis yang digunakan secara global. Hal ini juga membantu memastikan bahwa model klasifikasi teks yang dikembangkan mampu mengenali indikator linguistik yang memiliki relevansi klinis dalam konteks kesehatan mental.

2.3 Natural Languange Processing

Natural Language Processing (NLP) merupakan bidang interdisipliner yang menggabungkan ilmu linguistik, kecerdasan buatan, dan pembelajaran mesin, yang bertujuan untuk memungkinkan komputer memahami, memproses, dan menghasilkan bahasa manusia secara otomatis. NLP berperan penting dalam menjembatani komunikasi antara manusia dan mesin, serta menjadi dasar dari berbagai aplikasi seperti asisten virtual, penerjemah otomatis, sistem pencarian, dan klasifikasi teks. Dalam konteks penelitian ini, NLP digunakan untuk menganalisis data teks dari media sosial, khususnya *Reddit*, guna

mengidentifikasi indikasi gangguan kesehatan mental. Teks dari media sosial memiliki karakteristik yang berbeda dibandingkan teks formal, yaitu cenderung tidak baku, menggunakan singkatan, kata slang, serta ekspresi emosional yang tidak langsung. Oleh karena itu, diperlukan pendekatan NLP yang lebih adaptif dan kontekstual untuk menangani tantangan ini.

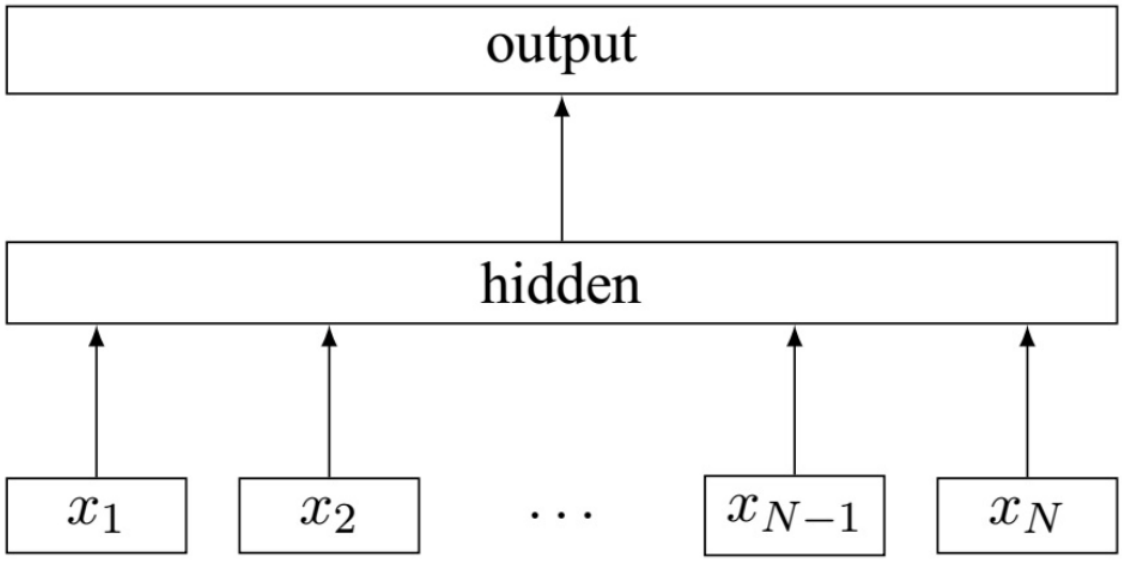
Tahapan dasar dalam NLP meliputi preprocessing teks, representasi fitur, dan pemodelan. *Preprocessing* mencakup proses seperti tokenisasi, normalisasi teks, penghapusan *stopword*, dan *stemming* atau *lemmatization*. Tahapan ini bertujuan untuk menyederhanakan teks mentah agar lebih mudah dianalisis tanpa menghilangkan makna penting. Setelah tahap preprocessing, langkah berikutnya adalah representasi teks ke dalam bentuk numerik agar dapat diproses oleh algoritma pembelajaran mesin. Representasi tradisional seperti *bag-of-words* dan TF-IDF telah banyak digunakan, namun memiliki keterbatasan dalam menangkap makna semantik. Oleh karena itu, *embedding* kata seperti *Word2Vec*, *GloVe*, dan *FastText* dikembangkan untuk memberikan representasi vektor kata berdimensi tetap yang menyimpan informasi semantik. Dalam penelitian ini, digunakan FastText karena keunggulannya dalam menangani kata tidak dikenal (OOV) melalui pemanfaatan fitur sub-kata atau n-gram karakter.

Setelah diperoleh representasi vektor dari teks, tahap akhir adalah pemodelan menggunakan algoritma pembelajaran mesin atau deep learning untuk klasifikasi. Arsitektur *deep learning* seperti *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM) digunakan dalam penelitian ini karena kemampuannya dalam menangkap pola lokal dan hubungan sekuensial dari data teks. CNN efektif dalam mengekstraksi fitur spasial seperti frasa penting, sementara BiLSTM mampu memahami konteks kalimat secara dua arah. Dengan kombinasi *preprocessing* yang baik, representasi embedding yang kaya secara semantik, dan arsitektur model yang kuat, NLP menjadi pendekatan yang efektif dalam menganalisis teks dan mendeteksi indikasi gangguan kesehatan mental secara otomatis dan akurat.

2.4 Pembobotan FastText

Dalam sistem klasifikasi teks berbasis pembelajaran mendalam, setiap kata dalam suatu kalimat terlebih dahulu dikonversi menjadi representasi numerik menggunakan teknik *word embedding*. Representasi ini digunakan untuk menjembatani teks tidak terstruktur dengan model pembelajaran mesin. Secara umum, suatu teks atau kalimat dapat direpresentasikan sebagai urutan kata $x_1, x_2, \dots, x_{N-1}, x_N$, di mana masing-masing x_i

merupakan vektor berdimensi tetap hasil dari proses *embedding* terhadap kata ke-i dalam urutan tersebut (Mikolov et al., 2013; Goldberg, 2017). Penelitian ini menggunakan model embedding *FastText*, yang menawarkan pendekatan berbeda dibanding metode *embedding* tradisional seperti *Word2Vec*. Salah satu keterbatasan model embedding konvensional adalah kurangnya pertimbangan terhadap struktur morfologis kata, yang berdampak pada rendahnya kualitas representasi semantik, terutama dalam bahasa yang memiliki kosakata besar dan banyak kata yang jarang digunakan. FastText mengatasi permasalahan ini dengan memecah setiap kata menjadi kumpulan n-gram karakter, sehingga memungkinkan model tetap menghasilkan vektor representasi meskipun kata tersebut tidak pernah muncul dalam data pelatihan (*out-of-vocabulary/OOV*) (Bojanowski et al., 2017; Joulin et al., 2017) arsitektur fasttext dapat dilihat pada gambar 2.1.



Gambar 2.1 Arsitektur Model FastText untuk Kalimat dengan Fitur N-gram (Joulin et al., 2017)

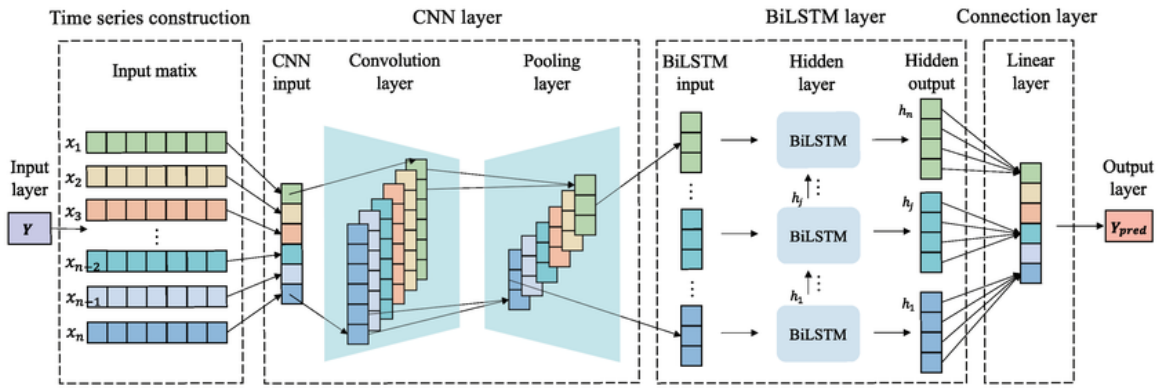
Model FastText yang digunakan dalam penelitian ini memiliki dimensi vektor sebesar 300 dan mencakup sekitar dua juta vektor kata, yang dilatih menggunakan informasi sub-kata dari korpus *Common Crawl* yang berisi sekitar 600 miliar token. Sumber embedding ini diakses dari situs resmi FastText (https://fasttext.cc/docs/en/english-vectors.html), dan menyediakan representasi kata yang kaya untuk menangani variasi ejaan, bentuk kata tidak baku, serta istilah slang yang umum ditemukan dalam data media sosial (Grave et al., n.d.). Seluruh urutan embedding, yaitu $x_1, x_2, \ldots, x_{N-1}, x_N$, selanjutnya digunakan sebagai input dalam arsitektur jaringan saraf CNN-BiLSTM. Lapisan konvolusional (CNN) mengekstraksi fitur lokal dari embedding seperti pola frasa penting, sedangkan *Bidirectional*

Long Short-Term Memory (BiLSTM) menangkap hubungan kontekstual dua arah dalam sekuens kata. Representasi akhir dari jaringan ini diteruskan ke lapisan klasifikasi untuk menghasilkan prediksi terhadap label kesehatan mental seperti *depression*, *anxiety*, atau *poison* (Yin et al., 2017; Zhou et al., 2018).

2.5 Arsitektur CNN-BiLSTM

Model klasifikasi dalam penelitian ini menggunakan pendekatan gabungan antara *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM). Pendekatan ini dipilih karena mampu menangkap baik fitur spasial (lokal) maupun kontekstual (sekuensial) dari data teks yang berasal dari media sosial, yang cenderung tidak terstruktur dan penuh dengan variasi ekspresi bahasa. CNN awalnya dikembangkan untuk pengolahan citra, namun telah banyak diadaptasi untuk pemrosesan bahasa alami. Dalam konteks teks, CNN menerima input berupa matriks embedding kata, di mana setiap baris mewakili vektor embedding dari satu kata dalam kalimat. CNN kemudian menerapkan filter konvolusional dengan ukuran tetap untuk mendeteksi fitur lokal atau n-gram penting yang berpotensi merepresentasikan makna semantik signifikan. Sebagai contoh, frasa seperti *"feel empty"* atau *"want to die"* dapat dikenali oleh filter CNN sebagai pola frasa berisiko psikologis. Hasil dari lapisan konvolusi ini kemudian diproses oleh lapisan pooling (biasanya *max pooling*), yang bertujuan menyaring informasi paling relevan dan mengurangi dimensi data untuk efisiensi komputasi (Kim, 2014).

Setelah fitur lokal diekstraksi oleh CNN, hasilnya diteruskan ke lapisan BiLSTM. BiLSTM merupakan pengembangan dari *Long Short-Term Memory (LSTM*), yaitu jenis *Recurrent Neural Network* (RNN) yang mampu mengingat informasi jangka panjang dalam data sekuensial. BiLSTM memiliki dua arah pemrosesan: satu memproses input dari awal ke akhir (forward) dan satu lagi dari akhir ke awal (backward), memungkinkan model untuk menangkap konteks linguistik secara dua arah. Hal ini sangat penting dalam NLP karena makna suatu kata sering kali bergantung pada kata-kata di sekitarnya. Sebagai contoh, dalam kalimat "*I said I'm fine"*, makna kata "fine" dapat bersifat ambigu, dan konteks sebelumnya serta sesudahnya menentukan apakah kalimat tersebut menunjukkan kondisi psikologis yang stabil atau sebaliknya (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997). Gambar 2.2 memperlihatkan alur arsitektur secara detail, mulai dari input embedding hingga ke lapisan *output*.



Gambar 2.2 Arsitektur CNN-BiLSTM (Chen & Fu, 2023)

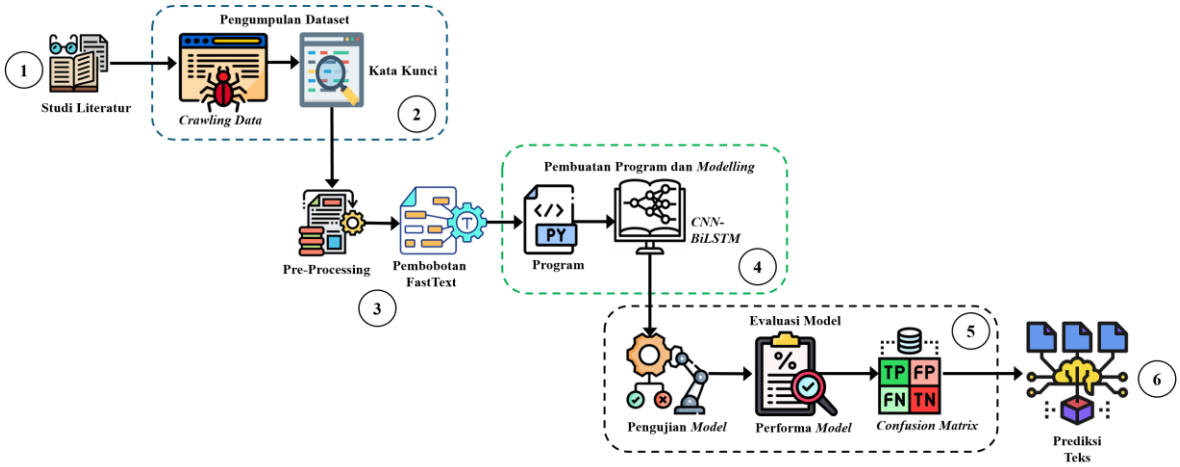
Gambar 2.2 Arsitektur gabungan CNN-BiLSTM ini mengintegrasikan kekuatan dua pendekatan CNN menangkap pola spasial yang mendalam pada level frasa, sementara BiLSTM menangkap hubungan temporal dan konteks global dari seluruh urutan kata. Kalimat masukan direpresentasikan sebagai matriks *embedding* dan diproses terlebih dahulu oleh lapisan konvolusi dan pooling untuk mengekstraksi fitur n-gram. Selanjutnya, hasil dari pooling dikirim ke BiLSTM yang menghasilkan representasi kontekstual dua arah. Representasi akhir ini diteruskan ke lapisan linier *(fully connected layer)* yang memproyeksikannya ke ruang prediksi kelas. Keluaran akhir berupa nilai probabilitas yang menunjukkan kecenderungan teks terhadap salah satu label gangguan kesehatan mental, seperti *depression*, *anxiety*, *bipolar*, atau *poison*.

BAB 3
Metodologi

Penelitian ini disusun melalui serangkaian tahapan metodologis yang terstruktur dan sistematis untuk mencapai tujuan utama, yaitu membangun model klasifikasi gangguan kesehatan mental berbasis teks media sosial menggunakan pendekatan *deep learning* CNN-BiLSTM. Setiap tahapan dalam metodologi dirancang untuk saling mendukung, dimulai dari pengumpulan data, pra-pemrosesan teks, hingga evaluasi performa model dan implementasi prediksi. Secara umum, proses metodologi ini dapat dilihat dalam Gambar 3.1, yang menunjukkan alur kerja penelitian secara menyeluruh. Tahapan dimulai dengan studi literatur sebagai landasan teoritis, dilanjutkan dengan proses pengumpulan dataset dari platform sosial media Reddit dan sumber tambahan dari Kaggle. Selanjutnya dilakukan proses pra-pemrosesan teks yang bertujuan untuk membersihkan dan menyiapkan data agar dapat dianalisis secara efektif oleh model. Setelah itu, dilakukan pembobotan kata menggunakan FastText untuk menghasilkan representasi vektor dari teks, yang kemudian menjadi input dalam proses pemodelan menggunakan arsitektur CNN-BiLSTM.

Model CNN-BiLSTM yang dikembangkan dalam penelitian ini terdiri dari beberapa lapisan utama, termasuk *embedding layer, convolutional layer*, *BiLSTM layer, dan dense layer.* Proses pelatihan model menggunakan data yang telah diproses sebelumnya, diikuti dengan evaluasi performa menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score. Evaluasi dilakukan melalui uji performa pada data uji serta validasi menggunakan metode k-fold cross-validation untuk memastikan generalisasi model.

Tahapan akhir dalam metodologi ini adalah implementasi model untuk melakukan prediksi terhadap data baru. Tahap ini bertujuan untuk menguji kemampuan model dalam mengklasifikasikan teks yang belum pernah dilihat sebelumnya, sehingga dapat mengukur sejauh mana model dapat diterapkan dalam konteks nyata. Dengan rangkaian metodologi ini, diharapkan penelitian mampu menghasilkan model prediktif yang tidak hanya akurat, tetapi juga adaptif terhadap data tidak terstruktur dari media sosial.



Gambar 3.1 Metode yang diusulkan

Pemilihan arsitektur CNN-BiLSTM dalam penelitian ini didasarkan pada beberapa pertimbangan metodologis yang kuat dan didukung oleh temuan penelitian terdahulu. Pertama, lapisan CNN dipilih karena kemampuannya dalam mengekstraksi fitur lokal seperti n-gram dan pola frasa penting yang sering muncul dalam ekspresi gangguan mental (Y. Kim, 2014; Yin et al., 2017). Penelitian oleh (Y. Kim, 2014) menunjukkan bahwa CNN mampu menangkap pola spasial yang efektif dalam klasifikasi teks, terutama dalam mengenali frasa-frasa kunci yang memiliki makna emosional kuat.

Kedua, lapisan BiLSTM digunakan untuk menangkap dependensi jangka panjang dan konteks dua arah dalam sekuens teks, yang sangat penting untuk memahami narasi emosional yang kompleks (Hochreiter & Schmidhuber, 1997; Zhou et al., n.d.). Berbeda dengan LSTM unidirectional yang hanya memproses informasi dari satu arah, BiLSTM mampu mengintegrasikan konteks dari kedua arah, sehingga pemahaman terhadap makna kalimat menjadi lebih komprehensif (Schuster & Paliwal, 1997). Dalam konteks deteksi kesehatan mental, kemampuan ini sangat krusial karena ekspresi emosional sering kali memerlukan pemahaman konteks yang berkembang secara temporal dalam narasi (Kour & Gupta, 2022).

Ketiga, FastText embedding diterapkan untuk mengatasi masalah out-of-vocabulary words yang umum ditemukan dalam teks media sosial yang cenderung informal dan tidak terstruktur (Bojanowski et al., 2017; Joulin et al., 2017). Keunggulan FastText dalam merepresentasikan sub-kata memungkinkan model untuk tetap memahami kata-kata yang tidak pernah muncul dalam data training, termasuk typo, slang, dan variasi ejaan yang sering ditemukan dalam teks media sosial (Grave et al., n.d.).

Kombinasi ketiga komponen ini dirancang secara khusus untuk mengatasi tantangan utama dalam klasifikasi teks kesehatan mental, yaitu keragaman ekspresi linguistik,

penggunaan bahasa informal dan slang, serta kompleksitas gejala yang sering kali saling tumpang tindih antar kategori gangguan (Le Glaz et al., 2021; Zhang et al., 2022b). Pendekatan hybrid ini diharapkan dapat memberikan performa yang lebih superior dibandingkan arsitektur tunggal dalam menangani kompleksitas data teks dari media sosial, sebagaimana telah ditunjukkan dalam penelitian terdahulu oleh (Xiaoyan et al., 2022) yang membuktikan bahwa kombinasi CNN-BiLSTM menghasilkan akurasi yang lebih tinggi dibandingkan model single-architecture dalam analisis sentimen teks.

3.1 Studi Literatur

Studi literatur merupakan tahap awal yang krusial dalam proses penelitian ini. Kegiatan ini dilakukan guna memperoleh pemahaman konseptual dan landasan teoritis yang komprehensif mengenai prediksi kesehatan mental berbasis teks menggunakan pendekatan pemrosesan bahasa alami (*Natural Language Processing*/NLP) dan deep learning. Studi literatur tidak hanya memberikan arah metodologis, tetapi juga membantu mengidentifikasi kesenjangan penelitian (*research gap*) yang menjadi dasar bagi konstruksi model usulan dalam penelitian ini. Tahap pertama dalam studi literatur difokuskan pada penelusuran dan analisis terhadap berbagai publikasi ilmiah, jurnal, prosiding konferensi, serta laporan penelitian yang relevan, yang membahas penerapan NLP dalam deteksi kondisi kesehatan mental seperti *depression*, *anxiety*, *bipolar*, dan *suicidal ideation*. Fokus utama pada literatur ini adalah untuk memahami bagaimana teknik representasi teks, klasifikasi, serta pemodelan dilakukan dalam konteks sosial media, khususnya pada platform seperti Reddit dan Twitter. Beberapa studi menekankan pada penggunaan algoritma pembelajaran mesin klasik seperti SVM dan Random Forest, sementara lainnya mulai mengadopsi model berbasis deep learning seperti LSTM, BiLSTM, CNN, dan BERT.

Tahap kedua dari studi literatur mencakup tinjauan teoritis yang lebih spesifik mengenai taksonomi gangguan kesehatan mental berdasarkan *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Pengetahuan ini digunakan untuk menentukan label-label utama yang akan digunakan dalam penelitian, seperti *depression*, *anxiety*, *bipolar disorder*, *schizophrenia*, dan *borderline personality disorder (BPD)*. Selain itu, studi ini juga mengevaluasi konsep-konsep seperti *toxic content* dan *emotional harm* yang menjadi dasar dalam penambahan label *poison* sebagai representasi konten berisiko psikologis tinggi.

Selanjutnya, literatur yang berkaitan dengan arsitektur model deep learning turut ditelaah secara sistematis. Penelitian yang menggunakan CNN menunjukkan efektivitasnya dalam mengekstraksi fitur lokal dari teks, namun memiliki keterbatasan dalam menangkap konteks urutan kata yang panjang. Sebaliknya, BiLSTM dinilai unggul dalam memahami hubungan

semantik jangka panjang dan arah ganda dalam teks. Oleh karena itu, kombinasi CNN dan BiLSTM dipilih sebagai pendekatan utama dalam penelitian ini, dengan tujuan untuk mengintegrasikan kekuatan masing-masing arsitektur guna meningkatkan akurasi prediksi. Studi literatur juga mencakup analisis terhadap berbagai teknik word embedding seperti Word2Vec, GloVe, dan FastText. FastText dipilih karena kemampuannya dalam memodelkan sub-kata dan menangani kata-kata yang tidak ditemukan (*out-of-vocabulary words*), sebuah tantangan umum dalam data sosial media yang sangat bervariasi dan informal.

Secara keseluruhan, hasil studi literatur menunjukkan bahwa penggunaan deep learning berbasis NLP memiliki potensi besar dalam mendeteksi gejala gangguan mental dari teks sosial media. Namun, masih terdapat celah dalam cakupan label yang digunakan, penanganan konten berisiko tinggi, dan arsitektur model yang adaptif terhadap data tidak terstruktur. Temuan-temuan inilah yang kemudian menjadi dasar pengembangan pendekatan baru dalam penelitian ini melalui implementasi model CNN-BiLSTM dengan pembobotan kata FastText untuk klasifikasi multi-label gangguan kesehatan mental.

3.2 Pengumpulan Dataset

Proses pengumpulan dataset terkait kesehatan mental dilakukan menggunakan library PRAW, dengan metode scraping dari beberapa kategori unggahan pada Reddit, yaitu hot, new, top, dan controversial, masing-masing dibatasi hingga 1500 postingan per kategori. Untuk memastikan kualitas data, dilakukan pemfilteran dengan fungsi:

```
def is_removed_or_deleted(text):
    return text.strip().lower() in ('[removed]', '[deleted]')
```

Data dikumpulkan dari enam *subreddit* utama yang berkaitan dengan kesehatan mental, yaitu *r/bipolar*, r/*depression*, r/*mentalhealth*, r/*anxiety*, r/*schizophrenia*, dan r/*bpd*, di mana masing-masing *subreddit* direpresentasikan sebagai satu label klasifikasi. Dari proses *scraping* tersebut diperoleh sebanyak 60.000 data teks. Selain itu, sebanyak 10.000 data dengan label poison diperoleh dari dataset *Kaggle* yang telah dilabeli sebelumnya, sehingga total dataset yang digunakan dalam penelitian ini berjumlah 70.000 data yang mencakup tujuh label gangguan kesehatan mental. Selanjutnya, dilakukan proses pembersihan dan seleksi data berdasarkan kriteria DSM-5 untuk memastikan relevansi konten terhadap gangguan mental yang dianalisis. Dataset yang telah melalui tahap tersebut kemudian digunakan dalam eksperimen dengan pembagian 70% *data training*, 15% *data validation*, dan 15% *data testing.*

3.3 Pre-Processing

Pra-pemrosesan teks adalah tahap kunci dalam pemrosesan bahasa alami, di mana data mentah mengalami transformasi untuk meningkatkan pemahaman teks. Fokus utamanya adalah menghilangkan kata-kata yang kurang signifikan, meningkatkan akurasi dan kelengkapan pemahaman teks. Teknik pra-pemrosesan, seperti *lowercasing*, penghapusan tanda baca, *stopwords*, dan tokenisasi, diterapkan secara berurutan untuk mengatasi keragaman masalah penelitian. Pentingnya langkah-langkah ini diakui karena tidak hanya mendukung pemahaman teks, tetapi juga meningkatkan kinerja model, memastikan konsistensi dan efektivitas dalam menghadapi kompleksitas tugas pemrosesan bahasa alami (Jakhotiya et al., 2022)

3.3.1 Lower Text

Fungsi utama dari fitur yang disebut *lowertext* adalah melakukan proses pengubahan semua huruf dalam suatu kata menjadi huruf kecil. Prinsip ini berlaku untuk mengonversi semua huruf besar yang terdapat dalam sebuah kata menjadi huruf kecil (Uysal & Gunal, 2014).

3.3.2 Remove punctuation

Menghapus karakter-karakter seperti @, $, dan *, yang tidak memiliki makna atau relevansi penting dalam kata-kata. Keberadaan karakter-karakter ini dalam data hanya akan menambahkan noise dan oleh karena itu harus dihilangkan (Marks, 2019).

3.3.3 Stopwords

Penghapusan *stopwords* menjadi tahapan esensial dalam analisis teks tidak terstruktur. *Stopwords*, yang merupakan kata-kata yang umumnya sering muncul di berbagai dokumen atau bahkan dalam satu dokumen, cenderung memiliki dampak minim terhadap informasi yang signifikan. Oleh karena itu, tindakan penghilangan stopwords menjadi krusial untuk meningkatkan kualitas analisis teks dengan memperbaiki rasio sinyal-ke-noise. Melalui langkah ini, diharapkan dapat ditingkatkan signifikansi statistik dari istilah-istilah tertentu yang memiliki relevansi khusus dalam konteks penelitian atau tujuan analisis tertentu (Kumar & Singh, 2022).

Dalam implementasi penelitian ini, stopwords dihilangkan secara konsisten pada tahap training dan testing untuk mengoptimalkan performa klasifikasi model. Namun, untuk keperluan interpretasi hasil menggunakan Integrated Gradients (dijelaskan pada bagian 3.8), stopwords dipertahankan pada input space original yang dianalisis. Strategi ini mengikuti rekomendasi dari penelitian explainable AI (Ribeiro et al., 2016) yang menekankan

pentingnya memberikan explanation dalam format yang interpretable bagi end user, yaitu pada representasi input yang natural dan mudah dipahami, bukan pada feature space internal yang telah dipreprocess.

3.3.4 Tokenizing

Tokenisasi adalah langkah kunci dalam memecah teks menjadi unit-unit bermakna seperti kata, frasa, atau simbol. Tujuannya adalah menyelidiki dan memahami kata-kata dalam teks. Hasilnya, daftar token digunakan sebagai input untuk analisis sintaksis atau penambangan teks. Dalam konteks tesis, tokenisasi diperlukan untuk mengurai dokumen, memungkinkan pemrosesan lebih lanjut. Meskipun teks tersedia dalam format mesin, tokenisasi perlu dilakukan untuk menangani masalah seperti penghapusan tanda baca, penanganan karakter khusus, dan menjaga konsistensi dokumen. Tujuan utama tokenisasi adalah mengidentifikasi kata kunci penting, menangani format waktu, angka, singkatan, dan akronim dalam teks (Kannan, 2015).

3.4 Pembobotan FastText

Word embedding merupakan suatu teknik pemrosesan teks yang mengalami pertumbuhan pesat dalam beberapa tahun terakhir. Teknik ini dapat dijelaskan sebagai model kata-kata yang memberikan makna yang signifikan terhadap kata-kata dalam suatu konteks khusus (Torregrossa et al., 2021a). Konsep ini berfokus pada proses menggambarkan kata-kata sebagai vektor dalam ruang vektor yang telah ditentukan sebelumnya. Masing-masing kata diidentifikasi dan diklasifikasikan oleh nilai vektor, dan semua nilai ini berkontribusi untuk membentuk suatu jaringan saraf. Aspek penting dari berbagai metode *word embedding* adalah pemahaman konsep penggunaan representasi yang kompak dan tersebar untuk setiap kata (Torregrossa et al., 2021a). Salah satu *word embedding* adalah *FastText*, sebuah model yang mirip dengan versi Skip-Gram dari Word2vec, memproses setiap kata sebagai bagian dari n-gram daripada sebagai keseluruhan kata (Santos et al., 2017). Berbeda dengan *Word2vec* yang beroperasi pada level kata, *FastText* berfokus pada level karakter dengan penggunaan n-gram kata. Model ini tidak hanya memahami urutan karakter lengkap dari suatu kata, tetapi juga memodelkan sub-kata. *FastText* menggunakan informasi sub-kata secara langsung untuk proses *embedding*, memastikan estimasi yang akurat terhadap kosakata yang jarang digunakan. Pendekatan ini meningkatkan pemahaman morfologis bahasa serta representasi kata berdasarkan tense, memungkinkan penanganan kata-kata yang tidak umum. Tujuan *FastText* adalah mengatasi masalah embed kata-kata jarang yang kadang sulit diestimasi(Devlin et al., 2019).

Pada tahap ini, dilakukan pembobotan kata untuk mengubah representasi teks menjadi bentuk numerik yang dapat diproses oleh model pembelajaran mesin. Teknik pembobotan yang digunakan adalah representasi vektor kata berbasis pretrained word embedding, yaitu *fastText Wiki English* dengan dimensi vektor sebesar 300. File embedding eksternal wiki.en.vec dimuat menggunakan *library codecs* untuk memastikan kompatibilitas karakter *Unicode*. Setiap baris pada file tersebut berisi satu kata diikuti oleh 300 nilai koefisien yang merepresentasikan vektor kata tersebut. Proses ini menghasilkan sebuah kamus embeddings_index yang memetakan kata ke dalam bentuk vektor berdimensi 300.

```
f = codecs.open('/kaggle/input/thesisbisaa/wiki.en.vec', encoding='utf-8')
...
embeddings_index[word] = coefs
```

Selanjutnya, dibentuk sebuah embedding matrix dengan ukuran (jumlah_kata, 300), di mana setiap baris merepresentasikan vektor dari sebuah kata berdasarkan *word_index* hasil tokenisasi sebelumnya. Untuk kata-kata yang tidak ditemukan dalam kamus pretrained *(out-of-vocabulary/OOV)*, diberikan vektor acak dalam rentang [-0.25, 0.25]. Strategi ini bertujuan agar kata-kata OOV tetap dapat dilatih selama proses pelatihan model.

```
def get_embedding(word, embeddings_index, embedding_dim):
    if word in embeddings_index:
        return embeddings_index[word]
    else:
        return np.random.uniform(-0.25, 0.25, embedding_dim)
```

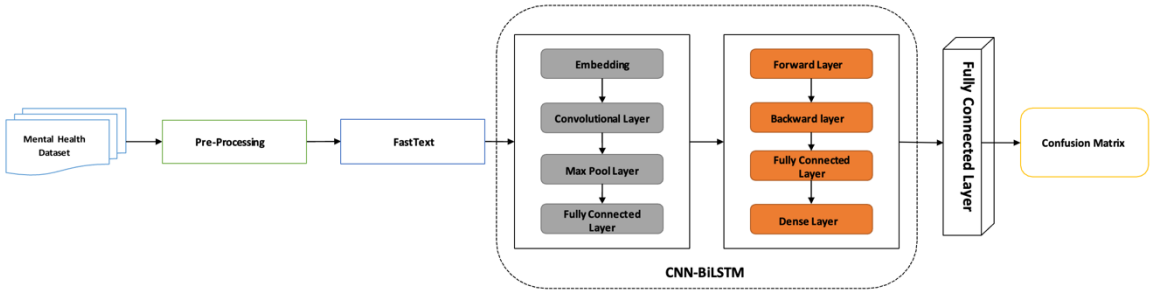
Dengan pendekatan ini, embedding matrix yang dihasilkan dapat langsung digunakan sebagai bobot awal pada layer embedding dalam arsitektur model deep learning. Hasil akhir dari proses ini adalah matriks representasi kata yang siap digunakan untuk proses pelatihan model klasifikasi berbasis teks.

Model FastText yang digunakan dalam penelitian ini merupakan pretrained FastText English vectors dengan dimensi embedding sebesar 300. Dimensi ini dipilih karena merupakan konfigurasi standar FastText yang telah terbukti efektif dalam merepresentasikan informasi semantik dan sintaktik kata melalui pendekatan sub-kata (subword information), serta banyak digunakan pada berbagai tugas klasifikasi teks, termasuk analisis teks media sosial. Penelitian oleh (Bojanowski et al., 2017; Joulin et al., 2017) menunjukkan bahwa embedding FastText berdimensi 300 mampu memberikan representasi kata yang stabil dan

kaya secara semantik. Selain itu, survei oleh (Torregrossa et al., 2021)serta studi klasifikasi teks kesehatan mental berbasis media sosial oleh (Tejaswini et al., 2022) juga menggunakan FastText dengan dimensi 300 sebagai konfigurasi standar. Oleh karena itu, penelitian ini tidak melakukan tuning terhadap dimensi embedding dan menggunakan nilai default 300 untuk menjaga konsistensi representasi kata serta memastikan keterbandingan dengan penelitian sebelumnya.

3.5 Pembuatan Program dan Modelling

Langkah selanjutnya pada penelitian ini melibatkan pemodelan dan pengembangan program menggunakan *Python 3.9*, dengan parameter *BiLSTM* seperti *optimizer*, fungsi aktivasi, dan dense layer. Library open-source seperti *sklearn, pandas, nltk*, dan *keras* akan digunakan. Dalam penelitian ini, model *CNN-BiLSTM* akan dilatih menggunakan 20 *epoch* dan *batch size* 64, dengan *optimizer RMSprop*. Model *CNN-BiLSTM* terdiri dari beberapa lapisan utama: lapisan embedding menggunakan *fasttext* untuk menangani kata-kata *out-of-vocabulary*, lapisan konvolusi (CNN) untuk mengektraksi fitur lokal dari teks, dan lapisan pooling untuk mengurangi dimensi vektor fitur. Kemudian, lapisan *Bi-LSTM* digunakan untuk menangkap depedensi sekuensial dalam teks baik dalam arah *forward* maupun *backward*, sehingga dapat memodelkan konteks dengan lebih baik. Lapisan *fully connected* atau *dense layer* digunakan untuk menghasilkan output akhir dari model. Parameter penting seperti *optimizer RMSprop*, fungsi aktivasi *ReLU* untuk lapisan dalam dan *sigmoid* atau *softmax* untuk lapisan output, serta jumlah unit dalam setiap lapisan, dioptimalkan untuk meningkatkan kinerja model. Penelitian ini akan berakhir dengan deployment model untuk mengklasifikasi kata baru dan memprediksi label pada aplikasi purwarupa. Model *CNN-BiLSTM* diharapkan dapat mengklasifikasi berbagai kondisi kesehatan mental seperti *borderline personality disorder* (BPD), *anxiety, depression, bipolar, mental illness, schizophrenia* dan *poison*. Gambar 3.2 adalah alur pemodelan CNN-BiLSTM yang digunakan pada penelitian ini.



Gambar 3.2 Alur Pemodelan CNN-BiLSTM

3.6 Pengaturan Hyperparameter

Proses tuning hyperparameter dilakukan untuk memperoleh konfigurasi model terbaik. Dalam penelitian ini, digunakan metode *Random Search* dari Keras Tuner untuk menyesuaikan parameter-parameter utama dari seluruh lapisan model, yang mencakup embedding, lapisan konvolusional (CNN), lapisan rekuren (BiLSTM), hingga lapisan dense dan algoritma optimisasi.Secara keseluruhan, konfigurasi uji parameter mencakup berbagai kemungkinan kombinasi nilai dari 16 hyperparameter. Pada bagian embedding, dimensi vektor ditetapkan sebesar 300 karena memanfaatkan bobot dari *pre-trained* GloVe. Nilai dropout pada embedding divariasikan antara 0.2 hingga 0.5.

Lapisan CNN pertama diuji dengan jumlah filter sebanyak 32 dan 64, kernel berukuran 3 dan 5, serta dua jenis fungsi aktivasi yaitu *ReLU* dan *tanh*. Proses *max pooling* dilakukan dengan ukuran 2 atau 3, diikuti oleh nilai dropout yang divariasikan dalam rentang yang sama. Lapisan CNN kedua menggunakan filter sebanyak 64 atau 128, kernel 3 atau 5, serta pengaturan *pooling* dan *dropout* yang serupa. Selanjutnya, lapisan BiLSTM dieksplorasi dengan jumlah unit sebanyak 64 dan 128, serta dropout antara 0.2 hingga 0.5. Pada lapisan dense, konfigurasi yang diuji mencakup jumlah unit (128 atau 256), *dropout*, dan fungsi aktivasi. Terakhir, dipertimbangkan pula tiga algoritma optimisasi, yaitu *Adam, RMSProp*, dan *Adamax*.

Setelah dilakukan lima *trial*, diperoleh konfigurasi hyperparameter terbaik dengan nilai *validation accuracy* tertinggi sebesar 0.8621. Model terbaik menggunakan dimensi embedding sebesar 300, dropout setelah embedding sebesar 0.2, 64 filter pada CNN pertama dengan kernel 3 dan aktivasi *tanh*, serta ukuran pooling sebesar 3. CNN kedua menggunakan 128 filter dan kernel 3 dengan pooling size 2 dan dropout 0.2. Lapisan BiLSTM memiliki 64 unit dengan dropout 0.3, diikuti dengan dense layer berisi 128 neuron, dropout 0.2, serta aktivasi *tanh*. Untuk optimisasi, digunakan algoritma RMSProp. Rincian lengkap konfigurasi hyperparameter yang digunakan dalam pelatihan model disajikan pada Tabel 3.1.

Tabel 3.1 Konfigurasi Uji Parameter dan Nilai Hyperparameter Terbaik pada Model CNN-BiLSTM

Kategori	Hyperparameter	Konfigurasi Uji Parameter	Nilai Terbaik
Embedding	embedding_dim	[300]	300
	dropout_embed	[0.2, 0.3, 0.4, 0.5]	0.2
Conv1	conv1_filters	[32, 64]	64
	conv1_kernel	[3, 5]	3

	activation	['relu', 'tanh']	tanh
	pool1_size	[2, 3]	3
	dropout_conv1	[0.2, 0.3, 0.4, 0.5]	0.2
Conv2	conv2_filters	[64, 128]	128
	conv2_kernel	[3, 5]	3
	dropout_conv2	[0.2, 0.3, 0.4, 0.5]	0.2
	pool2_size	[2, 3]	2
BiLSTM	lstm_units	[64, 128]	64
	dropout_lstm	[0.2, 0.3, 0.4, 0.5]	0.3
Dense & Optimizer	dense_units	[128, 256]	128
	dropout_dense	[0.2, 0.3, 0.4, 0.5]	0.2
	optimizer	['adam', 'rmsprop', 'adamax']	rmsprop

3.7 Evaluasi dan Verifikasi Model

Evaluasi kinerja model akan dilakukan dengan menggunakan matriks kebingungan pada set data pengujian. Kinerja metode ini dievaluasi berdasarkan metrik seperti Akurasi, Presisi, Recall, dan Skor F-1. Konsep akurasi berkaitan dengan derajat kedekatan antara nilai yang diantisipasi dan nilai aktual, sebagaimana diukur oleh persamaan (1). Istilah *True Positive* (TP) menunjukkan kejadian saat data positif diprediksi dengan tepat, sedangkan *True Negative* (TN) mengacu pada kejadian di mana data negatif diprediksi dengan tepat. Istilah *False Positive* (FP) mengacu pada situasi di mana data negatif salah diklasifikasikan sebagai data positif, sementara *False Negative* (FN) digunakan untuk menggambarkan skenario di mana data positif salah diklasifikasikan sebagai data negatif. Presisi adalah metrik yang mengevaluasi derajat ketepatan dalam data dengan mempertimbangkan kedekatan informasi dan kemampuannya untuk memprediksi. Persamaan (2) digunakan untuk menggambarkan presisi. Metrik recall mengukur efektivitas model dalam mengidentifikasi kelas tertentu secara akurat. Perhitungan dapat dilakukan dengan menggunakan persamaan (3). Skor F1 merupakan metrik komprehensif yang menggabungkan presisi dan recall dalam mengevaluasi akurasi prediksi model. Presisi mengukur tingkat ketepatan model, sementara recall mengevaluasi efektivitas model dalam mengidentifikasi dengan benar contoh di dalam kelas tertentu. Skor F1 memberikan penilaian menyeluruh terhadap performa model dengan mengintegrasikan kedua metrik tersebut. Skor F1 dihitung menggunakan persamaan (4).

Tabel 3.2 Confusion Matrix (Ha et al., 2012)

		Kelas Prediksi		
		TRUE	FALSE	Total
Kelas Sebenarnya	TRUE	TP (True Positive)	FN (False Negative)	P
	FALSE	FP (False Negative)	TN (True Positive)	N
	Total	P’	N’	P+N

$$\text{Akurasi} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

$$\text{Presisi} = TP / (TP + FP) \quad (2)$$

$$\textit{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{Skor F1} = 2\ ((\text{Recall} \times \text{Presisi})) / ((\text{Recall} + \text{Presisi})) \quad (4)$$

Dalam memastikan keandalan model, penelitian ini menerapkan metode *k*-fold cross-validation yang membagi dataset menjadi beberapa bagian guna menghindari overfitting dan meningkatkan stabilitas evaluasi model. Pendekatan ini memastikan bahwa setiap kelompok data latih dan uji memiliki distribusi yang acak, sehingga evaluasi menjadi tetap menyeluruh dan adil.

Sebagai bagian dari proses validasi hasil evaluasi model, dilakukan verifikasi manual terhadap perhitungan metrik evaluasi yang dihasilkan secara otomatis oleh sistem. Verifikasi ini bertujuan untuk memastikan bahwa nilai presisi, *recall*, dan *F1-score* yang diekstraksi dari *confusion matrix* telah dihitung dengan benar dan dapat diandalkan secara metodologis. Proses verifikasi dilakukan dengan menghitung metrik evaluasi secara manual menggunakan formula standar untuk setiap kelas gangguan mental. Perhitungan manual dimulai dengan mengekstraksi nilai TP, FP, dan FN dari *confusion matrix* untuk masing-masing kelas. Sebagai contoh, untuk kelas *Depression*, nilai-nilai komponen diperoleh dengan cara: TP diambil dari diagonal matriks yang merepresentasikan prediksi benar untuk kelas *Depression*, FP dihitung dengan menjumlahkan seluruh kolom *Depression* kecuali nilai TP, dan FN dihitung dengan menjumlahkan seluruh baris *Depression* kecuali nilai TP. Hasil perhitungan manual kemudian dibandingkan dengan output metrik yang dihasilkan oleh model untuk mengidentifikasi konsistensi dan akurasi sistem evaluasi. Deviasi kecil

(≤0.004) yang mungkin muncul akibat pembulatan desimal masih berada dalam margin toleransi yang dapat diterima secara statistik. Verifikasi ini memberikan konfirmasi bahwa sistem evaluasi otomatis menggunakan formula yang tepat dan menghasilkan nilai yang konsisten dengan perhitungan manual standar.

Proses verifikasi manual ini penting untuk membangun kepercayaan terhadap hasil evaluasi model, terutama dalam konteks penelitian kesehatan mental di mana akurasi klasifikasi memiliki implikasi yang signifikan. Dengan memverifikasi bahwa perhitungan metrik dilakukan dengan benar, penelitian ini memastikan bahwa interpretasi performa model yang dilaporkan dapat diandalkan dan valid secara metodologis.

3.8 Validasi Kualitatif Psikolog Klinis

Selain menggunakan metrik evaluasi otomatis seperti akurasi, presisi, *recall*, dan *F1-score*, penelitian ini juga melakukan validasi kualitatif yang melibatkan psikolog klinis profesional. Tujuan validasi ini adalah untuk memastikan bahwa hasil klasifikasi model memiliki relevansi klinis dan sesuai dengan standar diagnosis gangguan mental berdasarkan DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*).

Validasi dilakukan terhadap sampel hasil prediksi model yang terdiri dari 10 kasus berbahasa Inggris dan 7 kasus berbahasa Indonesia. Pemilihan sampel bertujuan untuk merepresentasikan berbagai kategori gangguan mental dengan tingkat kompleksitas yang beragam, serta menguji kemampuan model dalam konteks lintas bahasa. Psikolog klinis melakukan penilaian independen terhadap setiap kasus berdasarkan indikator gejala yang muncul dalam teks, pola narasi dan ekspresi emosional, serta konsistensi dengan kriteria diagnostik DSM-5.

Hasil penilaian psikolog kemudian dibandingkan dengan prediksi model untuk menghitung tingkat kesesuaian (*agreement rate*) dan mengidentifikasi pola kesalahan klasifikasi. Analisis validasi ini memberikan pemahaman mendalam tentang kapabilitas dan keterbatasan model dalam konteks klinis nyata, terutama dalam hal kemampuan model menangkap indikator linguistik yang eksplisit versus pemahaman konteks emosional yang lebih kompleks, serta tantangan dalam *transfer learning* lintas bahasa dan budaya.

3.9 Pengujian pada Data Baru

Tahapan akhir dalam metodologi penelitian ini adalah implementasi model untuk melakukan prediksi terhadap data baru yang tidak termasuk dalam dataset pelatihan maupun pengujian sebelumnya. Tahap ini bertujuan untuk menguji kemampuan generalisasi model

dalam mengklasifikasikan teks yang belum pernah dilihat, sehingga dapat mengukur sejauh mana model dapat diterapkan dalam konteks yang lebih luas dan realistis.

Data baru yang digunakan terdiri dari 103 sampel teks yang dikumpulkan dari berbagai sumber media sosial dan platform diskusi kesehatan mental. Data ini mencakup berbagai variasi ekspresi, panjang teks, dan tingkat kompleksitas narasi untuk menguji *robustness* model. Setiap teks melewati proses pra-pemrosesan yang sama seperti data pelatihan, kemudian diinputkan ke model untuk mendapatkan prediksi kategori gangguan mental beserta *confidence score*.

Hasil prediksi dianalisis untuk melihat distribusi prediksi di antara tujuh kategori gangguan mental, tingkat *confidence* model untuk setiap prediksi, serta pola prediksi pada teks dengan karakteristik tertentu. Pengujian pada data baru ini memberikan gambaran awal tentang efektivitas model dalam skenario aplikasi nyata dan menjadi dasar evaluasi untuk pengembangan sistem deteksi kesehatan mental berbasis media sosial di masa mendatang.

BAB 4
Hasil Dan Pembahasan

4.1 Dataset

Tahap awal penelitian ini dimulai dengan proses pengumpulan data secara sistematis dari platform *Reddit*, yang dikenal sebagai salah satu forum diskusi daring terbesar dengan topik kesehatan mental. Data dikumpulkan dari enam subreddit utama, yaitu *r/bipolar*, *r/depression*, *r/mentalhealth*, *r/anxiety*, *r/schizophrenia*, dan *r/bpd*, yang masing-masing direpresentasikan sebagai satu label klasifikasi, dan menghasilkan total sekitar 60.000 unggahan. Namun, sebagian data harus dihapus karena berstatus *removed* atau *deleted*, yang mencerminkan adanya moderasi konten maupun penghapusan oleh pengguna. Fenomena ini sejalan dengan temuan Pavalanathan dan (De Choudhury, 2013), yang menunjukkan bahwa konten sensitif pada media sosial, khususnya yang berkaitan dengan kesehatan mental, memiliki kecenderungan tinggi untuk dihapus.

Selain data dari *Reddit*, penelitian ini juga menggunakan sebanyak 10.000 data dengan label *poison* yang diperoleh dari dataset *Kaggle* yang telah dilabeli sebelumnya. Dengan demikian, total dataset yang digunakan dalam penelitian ini berjumlah 70.000 data yang mencakup tujuh label gangguan kesehatan mental. Selanjutnya, dilakukan proses pembersihan dan seleksi data untuk memastikan kualitas dan relevansi konten terhadap gangguan mental yang dianalisis, dengan mengacu pada kriteria DSM-5. Dataset yang telah melalui tahap tersebut kemudian digunakan dalam eksperimen dengan pembagian 70% data training, 15% data validation, dan 15% data testing.

Dari sisi bahasa, analisis terhadap istilah-istilah yang muncul dalam data menunjukkan adanya variasi ekspresi yang melampaui deskripsi klinis standar. Pada kategori depression, misalnya, kata-kata seperti *sadness* dan *loss of interest* sejalan dengan kriteria DSM-5, namun juga ditemukan ekspresi seperti *empty*, *numb*, dan *worthless* yang mencerminkan pengalaman emosional yang lebih kompleks. Temuan ini mendukung pandangan (De Choudhury, 2013) bahwa media sosial dapat menjadi sarana untuk memahami pengalaman subjektif individu dengan gangguan mental yang tidak selalu terungkap dalam konteks klinis formal.

Sebagai tambahan, penelitian ini memperkenalkan label khusus bernama *poison* untuk menandai konten berisiko tinggi secara psikologis, seperti ide bunuh diri, tindakan melukai diri sendiri, atau ajakan untuk melakukan kekerasan. Pemberian label ini didasari oleh kepentingan kesehatan publik, terutama karena adanya fenomena *contagion effect*, di mana

paparan terhadap konten *self-harm* dapat meningkatkan risiko imitasi perilaku pada individu yang rentan (Marchant et al., 2017). Oleh karena itu, kemampuan untuk mendeteksi konten semacam ini menjadi aspek penting dalam upaya pencegahan dini berbasis teknologi.

Table 4.1 Distribusi Label dan Kata Kunci Berdasarkan Subreddit yang Digunakan dalam Dataset

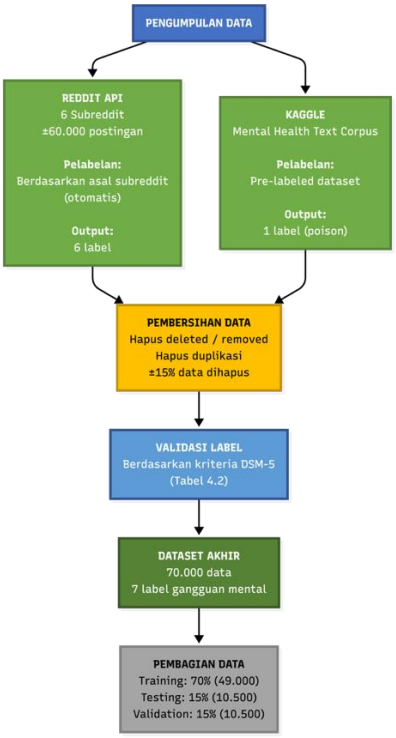
Subreddit	label	Kata kunci	Keterangan
r/depression	depression	*feelings of sadness, loss of interest decreased appetite, and difficulty focusing, depression* (Levanti et al., 2023) (Calixto et al., 2022)	Depresi merupakan kondisi psikologis yang ditandai dengan perasaan sedih yang mendalam, kehilangan minat terhadap kegiatan sehari-hari yang sebelumnya menyenangkan, penurunan nafsu makan, serta kesulitan dalam berkonsentrasi atau tidur. Depresi dapat mempengaruhi kemampuan seseorang untuk menjalani kehidupan secara produktif dan mempengaruhi kualitas hidup secara keseluruhan (Choi et al., 2021; Lopresti et al., 2013; Palmer et al., 2023).
r/anxiety	anxiety	*excessive anxiety, worry, diagnosed anixety* (Levanti et al., 2023)	Kecemasan adalah gangguan kesehatan mental yang ditandai dengan kecemasan yang berlebihan, kekhawatiran yang tidak terkendali, serta gangguan dalam pola perilaku dan kebiasaan sehari-hari, yang dapat mengganggu kesejahteraan fisik dan mental individu (American Psychiatric Association, 2013; Munir & Takov, 2022; Ströhle et al., 2018).
r/bipolar	bipolar	*Bipolar, diagnosed bipolar, maniac* (Kadkhoda et al., 2022)	Gangguan Bipolar, yang sebelumnya dikenal dengan istilah gangguan maniak-depresif, ditandai dengan fluktuasi suasana hati yang ekstrem, yang dapat melibatkan periode mania (energi berlebih dan perilaku impulsif) serta periode depresi (perasaan putus asa dan kehilangan minat). Gangguan ini juga berdampak pada tingkat energi, aktivitas, serta konsentrasi seseorang, dan

			dapat menyebabkan kesulitan dalam menjalani rutinitas sehari-hari mereka (Kadkhoda et al., 2022).
r/schizophrenia	schizophrenia	*Hallucinating, delusions,* ***Disorganized thinking***	Skizofrenia adalah gangguan psikiatri serius yang memengaruhi proses berpikir, emosi, serta perilaku individu, dan sering kali ditandai dengan halusinasi (pengalaman yang tidak sesuai dengan kenyataan) serta delusi (keyakinan yang salah dan tidak rasional). Penderita skizofrenia dapat mengalami kesulitan dalam berinteraksi dengan orang lain serta dalam menjalani kehidupan sosial yang normal (McCutcheon et al., 2023; Owen et al., 2016; Tandon et al., 2013).
r/bpd	BPD	*Borderline personality disorder, unstable emotions, unstable relationships, impulsivity, a fear of abandonment* (Dyson & Gorvin, 2017; Lyons et al., 2018)	Borderline Personality Disorder (BPD) adalah gangguan psikiatri yang ditandai dengan perilaku impulsif, ketidakstabilan emosional, hubungan interpersonal yang tidak stabil, serta distorsi dalam persepsi diri. Penderita BPD sering kali mengalami kesulitan dalam mengelola emosi dan mempertahankan hubungan yang sehat, yang dapat menyebabkan ketegangan dalam kehidupan sosial mereka (Carpenter & Trull, 2013a; Gunderson & Lyons-Ruth, 2008; Leichsenring et al., 2011; Lieb et al., 2004).
r/mentalhealth	mentalhealth	*Mentalillness, mentalhealth* (Calixto et al., 2022; Herdiansyah et al., 2023)	Gangguan Mental (Mental illness) secara umum adalah kondisi yang mempengaruhi pemikiran, suasana hati, perilaku, serta fungsi kognitif seseorang, yang dapat mengganggu kemampuan mereka dalam menjalani kehidupan sehari-hari secara normal dan produktif (Hofmann et al., 2012; Kessler et al., 2005)

4.2 Pemilihan Label *Mental Health*

Proses pelabelan dataset dalam penelitian ini melibatkan dua sumber data utama dengan pendekatan yang berbeda, sebagaimana ditunjukkan dalam Gambar 4.1. Data dari Reddit API dikumpulkan dari enam subreddit dengan pelabelan otomatis berdasarkan asal subreddit, yaitu *r/depression* menghasilkan label *depression*, *r/anxiety* menghasilkan label *anxiety*, *r/bipolar* menghasilkan label *bipolar*, *r/schizophrenia* menghasilkan label *schizophrenia*, *r/bpd* menghasilkan label *BPD*, dan *r/mentalhealth* menghasilkan label *mentalillness*. Sementara itu, label *poison* yang menandai konten berisiko tinggi seperti ide bunuh diri dan perilaku *self-harm* diperoleh dari dataset *Kaggle Mental Health Text Corpus* yang telah dilabeli sebelumnya. Dengan menggabungkan kedua sumber data tersebut, diperoleh total tujuh kategori label gangguan kesehatan mental.

Setelah proses pengumpulan data, dilakukan tahap pembersihan untuk menghapus entri yang berstatus *deleted*, *removed*, atau merupakan duplikasi. Selanjutnya, validasi label dilakukan menggunakan kriteria DSM-5 (Tabel 4.2) untuk memastikan bahwa kata kunci dan indikator linguistik dalam teks sesuai dengan kriteria diagnostik klinis masing-masing gangguan mental. Dataset yang telah melalui tahap pembersihan dan validasi tersebut kemudian digunakan dalam eksperimen dengan pembagian 70% data training, 15% data testing, dan 15% data validation.



Gambar 4.1 Alur Proses Pelabelan Dataset Gangguan Kesehatan Mental

Label-label gangguan mental yang digunakan dalam penelitian ini, seperti *Depression, Anxiety, Bipolar Disorder, Schizophrenia*, dan *Borderline Personality Disorder*, disusun berdasarkan temuan representatif dari data media sosial. Namun, pemilihan label tersebut tidak hanya didasarkan pada frekuensi kemunculan atau tema diskursus digital, melainkan juga mengacu pada klasifikasi diagnostik yang ditetapkan oleh *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)* yang dikeluarkan oleh *American (American Psychiatric Association, 2013).* DSM-5 memberikan kerangka klasifikasi resmi berdasarkan kriteria gejala, durasi, serta disfungsi yang ditimbulkan oleh suatu kondisi mental. Untuk memperkuat justifikasi tersebut, Tabel 4.2 berikut ini merangkum alasan pemilihan setiap label, disertai dukungan dari literatur ilmiah dan temuan utama yang relevan. Tabel ini menjadi dasar klasifikasi model dalam penelitian ini, sekaligus mengilustrasikan keterkaitan antara ekspresi digital dan indikator klinis menurut DSM-5.

Tabel 4.2 Alasan Pemilihan Label Gangguan Mental

Label	Alasan Pemilihan	Referensi Jurnal	Temuan Utama
Depression	Media sosial memperparah depresi melalui perbandingan sosial negatif, echo chamber, dan paparan konten yang memicu perasaan tidak berharga	(Kelly et al., 2018; Lin et al., 2016; Primack et al., 2017)	• Setiap peningkatan 10% penggunaan Facebook berkorelasi dengan peningkatan 13% gejala depresi • Penggunaan media sosial >2 jam/hari meningkatkan risiko distress psikologis 70% • Remaja dengan penggunaan media sosial >3 jam/hari memiliki risiko 60% lebih tinggi mengalami masalah kesehatan mental
Anxiety	Platform digital memperparah kecemasan melalui FOMO, notifikasi terus-menerus, cyberbullying, dan tekanan untuk selalu responsif	(Hunt et al., 2018; Vannucci et al., 2017; Woods & Scott, 2016)	• Penggunaan media sosial malam hari berkorelasi signifikan dengan peningkatan kecemasan pada 467 remaja • Pembatasan media sosial selama 1 minggu menurunkan tingkat kecemasan dan depresi secara signifikan • 70% emerging adults melaporkan peningkatan kecemasan terkait dengan aktivitas media sosial
Bipolar	Individu bipolar menunjukkan pola aktivitas media sosial ekstrem sesuai episode mood, dengan posting berlebihan saat manik dan isolasi saat depresi	(Faurholt-Jepsen et al., 2014; Hidalgo-Mazzei et al., 2015; Jacobson et al., 2019)	• Aktivitas smartphone dapat memprediksi episode mood dengan akurasi 89% • Pola posting berlebihan terdeteksi 85% pada fase manik, isolasi digital 78% pada fase depresi

			• Perubahan pola digital muncul 2-3 hari sebelum onset episode klinis
Schizophrenia	Media sosial dapat memperkuat delusi dan paranoia melalui misinformasi, teori konspirasi, dan filter bubble yang memvalidasi keyakinan delusional	(Birnbaum et al., 2017; Gowen, 2013; Lejeune et al., 2022)	• 67% individu dengan psikosis menggunakan media sosial untuk mencari informasi kesehatan mental • Algoritma media sosial dapat memperkuat keyakinan delusional pada 78% kasus yang diteliti • Konten konspirasi di media sosial meningkatkan paranoia pada individu rentan sebesar 45%
BPD (Borderline Personality Disorder)	BPD diperparah melalui instabilitas hubungan online, fear of abandonment yang diperkuat interpretasi media sosial, dan perilaku impulsif digital	(Birnbaum et al., 2017; Carpenter & Trull, 2013b; Turner et al., 2015)	•Individu BPD menunjukkan 3x lebih banyak emotional posting dibanding control • Fear of abandonment diperkuat oleh delayed responses di media sosial pada 84% subjek BPD • Perilaku impulsif online (oversharing, konflik) terjadi pada 89% episode dysregulasi emosi
Mental Illness (Umum)	Diskusi kesehatan mental umum di media sosial sering mengarah pada self-diagnosis tidak akurat, normalisasi gejala, dan penundaan pencarian bantuan profesional	(Löchner et al., 2025; Naslund et al., 2016; Reavley & Jorm, 2011)	• 74% individu mencari informasi kesehatan mental di media sosial sebelum konsultasi professional • Self-diagnosis melalui media sosial akurat hanya dalam 23% kasus • Diskusi peer-to-peer di media sosial dapat menunda pencarian bantuan profesional hingga 6 bulan

Poison (Konten Berbahaya)	Konten self-harm, bunuh diri, dan pro-eating disorder di media sosial dapat memicu contagion effect dan imitasi perilaku berbahaya pada individu rentan	(Hawton et al., 2012; Marchant et al., 2017; Sedgwick et al., 2019)	• Paparan konten self-harm meningkatkan risiko imitasi perilaku berbahaya hingga 30% pada remaja rentan • Konten bunuh diri viral dapat meningkatkan tingkat percobaan bunuh diri lokal hingga 13% • Media sosial bertanggung jawab atas 28% kasus contagion effect dalam perilaku self-harm pada remaja

4.3 Hasil Pre-Processing

Pada tahap ini, sebelum dilakukan proses pre-processing teks, dilakukan pembersihan data pada dataset Reddit dengan menghapus entri yang berstatus *removed* atau *deleted*, yang menunjukkan bahwa data tidak dapat diakses kembali. Selain itu, duplikasi data juga dihilangkan apabila ditemukan konten yang sama. Proses pembersihan ini bertujuan untuk memastikan kualitas data sebelum tahap pengolahan lebih lanjut. Setelah proses pembersihan, data teks dari Reddit dan Kaggle digabungkan menjadi satu dataset. Dataset gabungan ini kemudian melalui tahapan pre-processing teks yang meliputi konversi teks ke huruf kecil (*lowercasing*), penghapusan tanda baca seperti koma, titik, dan tanda seru, serta penghapusan *stopwords*, yaitu kata-kata umum yang tidak memberikan kontribusi signifikan terhadap proses klasifikasi, seperti *the*, *and*, *is*, dan *in*. Tahap selanjutnya adalah tokenisasi, yaitu proses pemisahan kalimat menjadi unit kata individual. Contoh hasil teks sebelum dan sesudah proses pre-processing ditunjukkan pada Tabel 4.3. Dataset yang telah melalui seluruh tahapan tersebut kemudian dibagi ke dalam tiga subset, yaitu 70% data training, 15% data testing, dan 15% data validation, yang digunakan pada tahap pelatihan dan evaluasi model.

Tabel 4.3 Hasil *Pre-Processing*

Teks Sebelum	Lower Text	Remove Punctuation	Stopword	Tokenizing
Im overwhelmed with guilt as a result of my bipolar. Mostly because of things I did before I was diagnosed. At first, my family thought they were things I did intentionally to hurt them. They eventually came	im overwhelmed with guilt as a result of my bipolar. mostly because of things i did before i was diagnosed. at first, my family thought they were things i did intentionally to hurt them. they eventually came	im overwhelmed with guilt as a result of my bipolar mostly because of things i did before i was diagnosed at first my family thought they	overwhelmed guilt result bipolar mostly things diagnosed family thought things intentionally hurt eventually came understanding	['overwhelmed', 'guilt', 'result', 'bipolar', 'mostly', 'things', 'diagnosed', 'family', 'thought', 'things', 'intentionally', 'hurt', 'eventually', 'came',

around to understanding that wasnt the case, and that I was just unmedicated and untreated. But the guilt is still crushing at times.	around to understanding that wasnt the case, and that i was just unmedicated and untreated. but the guilt is still crushing at times.	were things i did intentionally to hurt them they eventually came around to understanding that wasnt the case and that i was just unmedicated and untreated but the guilt is still crushing at times	wasnt case unmedicated untreated guilt crushing times	'understanding', 'wasnt', 'case', 'unmedicated', 'untreated', 'guilt', 'crushing', 'times']

4.4 Hasil Training Evaluasi Model

Pelatihan model CNN-BiLSTM pada penelitian ini dilakukan dengan menerapkan skema 20-Fold Cross Validation. Pendekatan ini digunakan untuk memperoleh estimasi performa model yang stabil dan mengurangi risiko overfitting, dengan cara melatih dan menguji model pada 20 subset berbeda dari dataset. Setiap fold berperan sebagai data uji sebanyak satu kali, dan sebagai data latih sebanyak sembilan belas kali. Strategi ini memungkinkan pengukuran performa yang representatif terhadap distribusi data sebenarnya, terutama dalam konteks klasifikasi multi-label kesehatan mental yang cenderung memiliki ketimpangan kelas (class imbalance) cukup signifikan.

Model menerima input berupa teks yang telah dikonversi ke dalam bentuk vektor menggunakan *FastText embedding*. Representasi berbasis sub-kata ini memungkinkan model mengenali kata-kata tidak baku, singkatan, dan kesalahan ejaan yang umum dalam teks media sosial. Setelah itu, data vektor ini diproses oleh lapisan CNN, yang bertugas mengekstraksi fitur lokal seperti n-gram atau frasa-frasa kunci yang sering muncul dalam ekspresi emosional atau psikologis. Hasil ekstraksi fitur kemudian diteruskan ke lapisan

BiLSTM, yang memproses urutan kata dari dua arah untuk menangkap konteks temporal dan semantik secara menyeluruh.

Performa model selama pelatihan diukur menggunakan tiga metrik utama *precision, recall,* dan *F1-score*, yang masing-masing dirata-ratakan menggunakan dua pendekatan: macro average dan weighted average. Hasil evaluasi dari pendekatan *macro average* disajikan dalam Tabel 4.4. Dalam tabel ini, model menunjukkan rata-rata *precision* sebesar 0.867, *recall* sebesar 0.868, dan F1-score sebesar 0.867 di seluruh fold. Karena *macro average* menghitung metrik untuk tiap kelas terlebih dahulu, lalu mengambil rata-rata tanpa memperhatikan proporsi kelas, maka nilai pada Tabel 4.4 mencerminkan seberapa adil model memperlakukan seluruh kelas—termasuk kelas minoritas. Tingginya skor pada tabel ini menandakan bahwa model memiliki kemampuan yang baik dalam mengenali semua jenis gangguan mental secara seimbang.

Tabel 4.4 Hasil Evaluasi Model Berdasarkan Metrik Macro Average pada 20-Fold Cross Validation

Fold	**Accuracy(Macro)**	**Precision (Macro)**	**Recall (Macro)**	**F1 (Macro)**
1	0.87	0.86	0.87	0.86
2	0.86	0.86	0.87	0.86
3	0.86	0.86	0.87	0.86
4	0.88	0.87	0.88	0.87
5	0.87	0.87	0.87	0.87
6	0.87	0.87	0.87	0.87
7	0.88	0.87	0.87	0.87
8	0.86	0.86	0.86	0.86
9	0.87	0.87	0.87	0.86
10	0.86	0.86	0.86	0.86
11	0.87	0.87	0.87	0.86
12	0.87	0.87	0.87	0.87
13	0.87	0.86	0.86	0.86
14	0.87	0.86	0.86	0.86
15	0.88	0.88	0.88	0.87
16	0.88	0.87	0.88	0.87
17	0.87	0.87	0.87	0.87
18	0.88	0.88	0.88	0.88

19	0.87	0.86	0.87	0.86
20	0.85	0.86	0.86	0.85

Sebagai perbandingan, hasil pengukuran dengan pendekatan weighted average disajikan pada Tabel 4.5 Weighted average memperhitungkan jumlah sampel di setiap kelas, sehingga kelas mayoritas memiliki pengaruh yang lebih besar terhadap nilai akhir. Dalam tabel ini, model mencatat precision rata-rata sebesar 0.875, recall sebesar 0.876, dan F1-score sebesar 0.875. Nilai-nilai ini sedikit lebih tinggi dari nilai pada Tabel 4.4, yang sesuai dengan karakteristik weighted average. Namun, perbedaan antara macro dan weighted sangat kecil—pada umumnya hanya berkisar 0.00 hingga 0.01 poin. Sebagai contoh, pada Fold 13 precision macro adalah 0.865 dan weighted adalah 0.875, sedangkan pada Fold 14 dan 19 perbedaan precision juga kurang dari 0.01. Kesenjangan yang minimal ini menunjukkan bahwa model tidak bias terhadap kelas mayoritas dan mampu memberikan performa yang konsisten terhadap seluruh kelas.

Analisis terhadap dua tabel ini memberikan pemahaman penting. Jika performa model sangat berbeda antara kedua pendekatan, itu biasanya menandakan ketidakseimbangan prediksi—model hanya bekerja baik pada kelas dengan jumlah data tinggi. Namun, pada penelitian ini, konsistensi antara Tabel 4.4 dan Tabel 4.5 menjadi indikator bahwa model tidak hanya akurat secara keseluruhan, tetapi juga adil terhadap kelas yang kurang terwakili, seperti BPD, schizophrenia, dan poison. Dalam konteks aplikasi nyata, seperti skrining digital terhadap kesehatan mental berbasis teks, kemampuan seperti ini sangat penting agar model tidak mengabaikan kondisi-kondisi yang jumlahnya sedikit tetapi kritis secara klinis.

Tabel 4.5 Hasil Evaluasi Model Berdasarkan Metrik Weighted Average pada 20-Fold Cross Validation

Fold	**Accuracy(Weighted)**	**Precision (Weighted)**	**Recall (Weighted)**	**F1 (Weighted)**
1	0.87	0.86	0.87	0.86
2	0.86	0.86	0.86	0.86
3	0.86	0.86	0.86	0.86
4	0.88	0.87	0.88	0.87
5	0.87	0.87	0.87	0.87
6	0.87	0.87	0.87	0.87
7	0.88	0.87	0.88	0.87
8	0.86	0.86	0.86	0.86

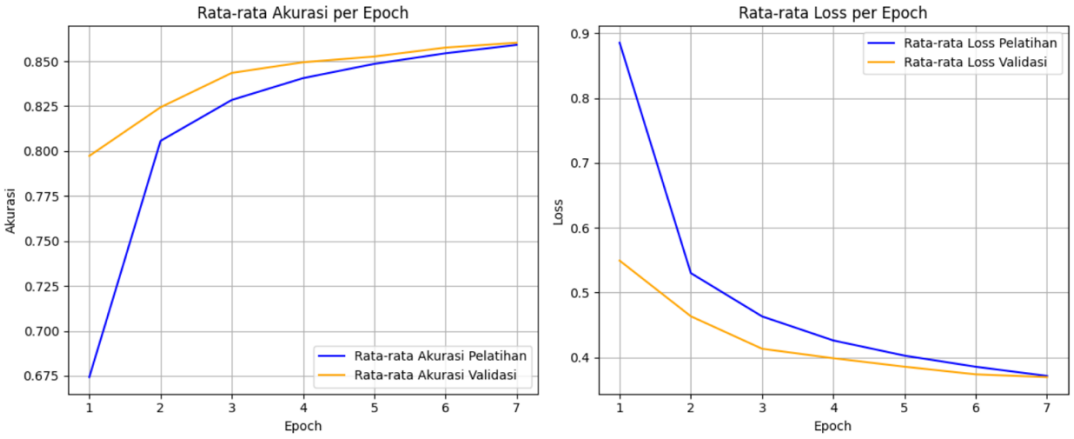
9	0.87	0.87	0.87	0.86
10	0.86	0.86	0.86	0.86
11	0.87	0.87	0.87	0.86
12	0.87	0.87	0.87	0.87
13	0.87	0.87	0.87	0.87
14	0.87	0.86	0.87	0.86
15	0.88	0.87	0.87	0.87
16	0.88	0.88	0.88	0.88
17	0.87	0.87	0.87	0.87
18	0.88	0.88	0.88	0.88
19	0.87	0.86	0.87	0.86
20	0.85	0.85	0.85	0.85

Proses pelatihan model CNN-BiLSTM divisualisasikan pada Gambar 4.1, yang menunjukkan perubahan nilai akurasi dan loss terhadap jumlah epoch selama pelatihan. Berdasarkan grafik tersebut, terlihat bahwa akurasi meningkat secara bertahap, sementara loss menurun secara konsisten, yang menunjukkan bahwa proses pembelajaran berlangsung stabil dari awal hingga akhir.

Kurva akurasi yang menunjukkan tren peningkatan secara bertahap mengindikasikan bahwa model secara progresif memperbaiki kemampuannya dalam mengenali pola-pola yang relevan dalam data. Sementara itu, kurva loss yang menurun secara berkelanjutan mencerminkan keberhasilan model dalam meminimalkan kesalahan prediksi selama proses optimasi. Tidak tampak adanya lonjakan mendadak atau fluktuasi ekstrem pada kedua kurva, yang biasanya menjadi indikasi dari overfitting, instabilitas pembelajaran, atau gangguan lain yang sering terjadi dalam model dengan struktur mendalam.

Salah satu indikator penting yang dapat diamati dari Gambar 4.1 adalah stabilitas penurunan kurva loss selama pelatihan. Kurva tersebut menurun secara halus dan konsisten tanpa ketidakteraturan, yang mengindikasikan bahwa proses optimasi berjalan secara efisien. Tidak terlihat gejala umum seperti nilai pembaruan parameter yang terlalu kecil hingga pelatihan menjadi lambat (vanishing gradient), ataupun nilai pembaruan yang terlalu besar hingga membuat pelatihan tidak stabil (exploding gradient). Kedua hal tersebut sering menjadi tantangan dalam pelatihan model berlapis seperti CNN-BiLSTM. Stabilitas kurva menunjukkan bahwa konfigurasi pelatihan—termasuk learning rate, algoritma optimisasi, dan ukuran batch—telah disesuaikan dengan baik terhadap karakteristik data.

Selain itu, penerapan early stopping terbukti efektif dalam mencegah pelatihan berlebihan (over-training). Berdasarkan grafik, sebagian besar fold menunjukkan konvergensi optimal dalam rentang epoch ke-7 hingga ke-10, yaitu saat performa validasi tidak lagi mengalami peningkatan yang signifikan. Hal ini menunjukkan bahwa model mampu mencapai performa terbaik dalam waktu yang efisien, tanpa memerlukan jumlah epoch yang terlalu banyak. Cepatnya waktu konvergensi ini juga memperkuat bahwa fitur-fitur linguistik yang dipelajari oleh CNN, serta konteks urutan yang ditangkap oleh BiLSTM, sudah cukup informatif sejak awal proses pelatihan.



Gambar 4.2 Visualisasi Kinerja Model CNN-BiLSTM Berdasarkan Akurasi dan Loss (dengan Early Stopping)

Gambar 4.2 menyajikan visualisasi kinerja model CNN-BiLSTM yang dirata-ratakan dari *20-fold cross-validation* selama proses pelatihan. Meskipun konfigurasi pelatihan menetapkan maksimal *15 epoch* untuk setiap *fold*, grafik hanya menampilkan hingga epoch ketujuh karena mekanisme early stopping telah menghentikan proses pelatihan secara otomatis. Implementasi early stopping dengan parameter patience sebesar *3 epoch* memonitor nilai validation loss secara berkelanjutan dan menginterupsi proses pelatihan apabila tidak terdapat perbaikan signifikan selama tiga *epoch* berturut-turut. Pendekatan ini bertujuan untuk mencegah overfitting dan meningkatkan efisiensi komputasi.

Panel kiri Gambar 4.2 menampilkan kurva rata-rata akurasi per epoch untuk data pelatihan dan validasi. Kurva akurasi pelatihan (garis biru) menunjukkan peningkatan gradual dari 0.675 pada *epoch* pertama hingga mencapai 0.860 pada *epoch* ketujuh. Kurva akurasi validasi (garis oranye) dimulai dari nilai yang lebih tinggi yaitu 0.800 pada epoch pertama dan meningkat hingga 0.862 pada epoch ketujuh. Observasi bahwa akurasi validasi sedikit melampaui akurasi pelatihan pada beberapa *epoch* mengindikasikan efektivitas strategi regularisasi yang diterapkan melalui dropout layers pada *embedding* (0.2),

convolutional (0.2), BiLSTM (0.3), dan dense layers (0.2). Fenomena ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik tanpa mengalami memorisasi berlebihan terhadap data pelatihan.

Panel kanan Gambar 4.2 menyajikan kurva rata-rata *loss per epoch* untuk data pelatihan dan validasi. Kurva loss pelatihan (garis biru) menunjukkan penurunan yang konsisten dan *monotonic* dari 0.875 pada epoch pertama hingga 0.375 pada epoch ketujuh. Kurva loss validasi (garis oranye) mengikuti pola serupa dengan penurunan dari 0.550 hingga 0.380. Tidak terdapat fluktuasi signifikan atau lonjakan mendadak pada kedua kurva, mengindikasikan bahwa proses optimisasi menggunakan algoritma RMSProp berjalan secara stabil. Tidak ditemukannya indikasi vanishing gradient (loss yang stagnan pada nilai tinggi) maupun *exploding gradient* (lonjakan loss secara mendadak) menunjukkan bahwa konfigurasi *hyperparameter*, khususnya learning rate dan optimizer, telah dikalibrasi dengan tepat untuk arsitektur CNN-BiLSTM yang digunakan.

Analisis terhadap gap antara kurva pelatihan dan validasi pada kedua panel menunjukkan hasil yang sangat positif. Pada epoch ketujuh, selisih akurasi antara data pelatihan (0.860) dan validasi (0.862) hanya sebesar 0.002, sementara selisih loss hanya sebesar 0.005 (0.375 berbanding 0.380). Gap yang minimal ini merupakan indikator kuat bahwa model tidak mengalami overfitting dan memiliki kemampuan generalisasi yang excellent terhadap data yang belum pernah dilihat selama proses pelatihan. Hal ini sangat penting dalam konteks aplikasi praktis, dimana model diharapkan dapat mengklasifikasikan data baru dengan akurasi yang konsisten.

Konvergensi yang tercapai dalam waktu relatif singkat—hanya memerlukan rata-rata tujuh epoch dari maksimal 15 epoch yang dialokasikan—dapat dikaitkan dengan beberapa faktor arsitektural dan metodologis. Pertama, lapisan konvolusional dengan konfigurasi 64 filter (CNN pertama) dan 128 filter (CNN kedua) dengan kernel size 3 mampu mengekstraksi fitur lokal dan pola n-gram yang informatif sejak epoch-epoch awal. Kedua, lapisan *Bidirectional* LSTM dengan 64 unit berhasil menangkap dependensi jangka panjang dan representasi kontekstual dua arah dari sekuens teks secara efisien. Ketiga, penggunaan *pre*-trained *FastText embedding* berdimensi 300 yang telah dilatih pada korpus *Common Crawl* (600 miliar token) menyediakan inisialisasi bobot yang superior, memfasilitasi pembelajaran pola-pola spesifik terkait manifestasi linguistik gangguan kesehatan mental tanpa memerlukan pelatihan dari awal yang intensif secara komputasi.

Konsistensi dalam hal *epoch* dimana *early stopping* terpicu—yaitu rata-rata pada epoch ketujuh di sebagian besar fold dari 20-*fold cross-validation*—mengindikasikan

robustness konfigurasi *hyperparameter* yang diperoleh melalui proses tuning menggunakan Keras Tuner dengan metode *Random Search* (5 trial). Konsistensi ini menunjukkan bahwa performa optimal tidak hanya tercapai pada subset data tertentu, melainkan dapat digeneralisasi dengan baik ke seluruh distribusi data. Hal ini memperkuat validitas hasil evaluasi yang dilaporkan dan memberikan kepercayaan bahwa model memiliki kemampuan yang stabil dalam mengklasifikasikan tujuh kategori gangguan kesehatan mental dari data teks media sosial yang tidak terstruktur.

4.5 Hasil Evaluasi Testing Model

Setelah tahap pelatihan dan validasi selesai dilakukan menggunakan skema 20-fold cross-validation dapat dilihat pada tabel 4.6, tahap selanjutnya adalah menguji kemampuan model pada data uji (testing) untuk mengevaluasi sejauh mana model mampu melakukan generalisasi terhadap data baru yang sebelumnya tidak pernah dilihat. Dalam tahap ini, dua arsitektur utama diuji: model CNN-BiLSTM sebagai arsitektur *hybrid* dan model BiLSTM murni sebagai pembanding. Pada model CNN-BiLSTM, evaluasi akhir dilakukan dengan pendekatan weighted ensemble, yaitu dengan menggabungkan hasil dari seluruh model yang terbentuk dalam 20 *fold* berdasarkan bobot akurasi validasinya. Pendekatan ini dipilih untuk meningkatkan stabilitas prediksi dan mereduksi risiko *overfitting* terhadap distribusi data tertentu dalam subset pelatihan. Hasil pengujian akhir menunjukkan bahwa CNN-BiLSTM berhasil mencapai akurasi dan F1-score sebesar 87%, yang menandakan bahwa model tidak hanya akurat, tetapi juga konsisten dalam mengenali pola-pola linguistik yang berkaitan dengan berbagai gangguan mental dalam tujuh kelas yang dianalisis.

Model BiLSTM murni, yang digunakan sebagai baseline, diuji dengan pendekatan yang sama untuk memastikan perbandingan yang adil. Hasilnya menunjukkan bahwa model ini mampu mencapai akurasi dan F1-score sebesar 83%, sebuah performa yang kompetitif namun tetap lebih rendah dibandingkan arsitektur *hybrid*. Perbedaan ini mencerminkan kontribusi nyata dari lapisan CNN dalam proses pembelajaran. Meskipun BiLSTM cukup andal dalam memahami urutan dan konteks temporal dalam teks, ketidakhadiran lapisan konvolusional membuat model ini kurang optimal dalam menangkap fitur spasial dan pola lokal yang sering kali menjadi penanda penting dalam ekspresi emosional atau gejala psikologis tertentu.

Tabel 4.6 Hasil Evaluasi Model

Model	Kelas yang Diklasifikasikan	Akurasi	F1-Score
CNN-BiLSTM (Usulan)	BPD, bipolar, depression, anxiety, schizophrenia, mentalillness, poison	**0.87**	**0.87**
BiLSTM (Usulan)	BPD, bipolar, depression, anxiety, schizophrenia, mentalillness, poison	0.82	0.82
CNN (Usulan)	BPD, bipolar, depression, anxiety, schizophrenia, mentalillness, poison	0.84	0.84
CNN (Murarka & Raleigh, 2021)	ADHD, anxiety, bipolar, depression, PTSD, none	0.64	0.65
LSTM (Murarka & Raleigh, 2021)	ADHD, anxiety, bipolar, depression, PTSD, none	0.76	0.77
BiLSTM (Murarka & Raleigh, 2021)	ADHD, anxiety, bipolar, depression, PTSD, none	0.78	0.79
LSTM (Ameer et al., 2022)	ADHD, anxiety, bipolar, depression, PTSD, none	0.76	0.76

Berdasarkan Tabel 4.7, model CNN-BiLSTM sebagai arsitektur *hybrid* menunjukkan performa terbaik dengan akurasi keseluruhan sebesar 0.87. Performa tertinggi dicapai pada kelas bipolar dengan precision 0.97, recall 1.00, dan F1-score 0.98. Recall sempurna (1.00) menunjukkan bahwa model berhasil mendeteksi seluruh kasus bipolar dalam dataset tanpa ada yang terlewat, mengindikasikan bahwa kombinasi ekstraksi fitur lokal oleh CNN dan pemahaman konteks sekuensial oleh BiLSTM sangat efektif dalam mengenali pola linguistik gangguan bipolar yang kompleks.

Kelas BPD juga menunjukkan performa yang sangat baik dengan *precision* 0.98, *recall* 0.98, dan *F1-score* 0.98. Keseimbangan sempurna antara precision dan recall menunjukkan bahwa model tidak hanya mampu mengidentifikasi hampir semua kasus BPD, tetapi juga meminimalkan kesalahan klasifikasi terhadap kelas lain. Hal ini mengindikasikan bahwa arsitektur *hybrid* berhasil menangkap karakteristik khas BPD, baik dari segi pola kata-kata spesifik maupun konteks emosional yang berkembang dalam teks.

Performa yang baik juga terlihat pada kelas *anxiety* dengan F1-score 0.93 (precision 0.91, recall 0.95), *poison* dengan *F1-score* 0.92 (*precision* 0.95, *recall* 0.89), dan *depression* dengan *F1-score* 0.89 (*precision* 0.87, *recall* 0.92). Pada kelas *poison*, *precision* yang tinggi (0.95) menunjukkan bahwa model sangat akurat ketika mengidentifikasi konten berbahaya,

dengan tingkat *false positive* yang rendah. Sementara pada *depression*, *recall* yang cukup tinggi (0.92) menunjukkan kemampuan model dalam mendeteksi sebagian besar kasus depresi meskipun dengan beberapa *false positive*.

Meskipun demikian, tantangan tetap ditemukan pada kelas mentalillness dengan *precision* 0.64, *recall* 0.59, dan *F1-score* 0.61, serta *schizophrenia* dengan precision 0.74, *recall* 0.74, dan *F1-score* 0.74. Meskipun ini merupakan performa terendah dalam model CNN-BiLSTM, nilai-nilai tersebut masih lebih tinggi dibandingkan model CNN dan BiLSTM standalone, menunjukkan bahwa arsitektur hybrid memberikan peningkatan meskipun kedua kelas tersebut tetap menjadi tantangan dalam klasifikasi.

Tabel 4.7 Hasil *Confusion Matrix* Model CNN-BiLSTM

Kelas	Precision	Recall	F1-Score	Support
BPD	0.98	0.98	0.98	1,437
Anxiety	0.91	0.95	0.93	1,454
Bipolar	0.97	1.00	0.98	1,458
Depression	0.87	0.92	0.89	1,558
Mentalillness	0.64	0.59	0.61	1,467
Poison	0.95	0.89	0.92	1,525
Schizophrenia	0.74	0.74	0.74	1,601
Accuracy	—	—	**0.87**	10,500
Macro Avg	0.86	0.87	0.87	10,500
Weighted Avg	0.86	0.87	0.86	10,500

Berdasarkan Tabel 4.8, model CNN menunjukkan performa yang bervariasi pada setiap kelas gangguan mental dengan akurasi keseluruhan sebesar 0.84. Performa terbaik dicapai pada kelas BPD dengan *precision* 0.98, *recall* 0.97, dan *F1-score* 0.97, yang menunjukkan bahwa model sangat andal dalam mengidentifikasi pola linguistik yang khas pada gangguan kepribadian *borderline*. Hal ini mengindikasikan bahwa CNN mampu menangkap fitur lokal yang spesifik seperti ekspresi emosi yang intens dan tidak stabil yang sering muncul pada postingan pengguna dengan BPD.

Kelas *bipolar* juga menunjukkan performa yang sangat baik dengan *precision* 0.96, *recall* 0.99, dan *F1-score* 0.97. Recall yang tinggi (0.99) menunjukkan bahwa model berhasil mendeteksi hampir semua kasus bipolar dalam dataset, yang mengindikasikan bahwa pola linguistik pada gangguan *bipolar* memiliki karakteristik yang cukup distinktif dan dapat ditangkap dengan baik oleh lapisan konvolusional. Performa yang baik juga terlihat pada

kelas *anxiety* dengan *F1-score* 0.92 (*precision* 0.89, *recall* 0.94) dan *poison* dengan *F1-score* 0.85 (*precision* 0.92, *recall* 0.79).

Kelas depression menunjukkan performa yang moderat dengan *precision* 0.82, *recall* 0.90, dan *F1-score* 0.86. Meskipun *recall* cukup tinggi yang menunjukkan kemampuan model dalam mendeteksi kasus depresi, *precision* yang lebih rendah mengindikasikan adanya sejumlah *false positive*, di mana beberapa sampel dari kelas lain salah diklasifikasikan sebagai *depression*. Hal ini dapat dijelaskan oleh adanya *overlap* gejala antara depresi dengan gangguan mental lainnya, seperti *anxiety* dan *mentalillness*, yang memiliki ekspresi linguistik yang serupa terkait kesedihan dan perasaan negatif.

Tantangan terbesar ditemukan pada kelas *mentalillness* dengan performa terendah yaitu *precision* 0.61, *recall* 0.51, dan *F1-score* 0.56. Rendahnya performa pada kelas ini dapat disebabkan oleh sifat kategori *mentalillness* yang lebih umum dan heterogen, mencakup berbagai jenis gangguan mental yang tidak termasuk dalam kategori spesifik lainnya. Variabilitas ekspresi yang tinggi dan kurangnya pola linguistik yang konsisten membuat CNN kesulitan dalam mengekstraksi fitur lokal yang representatif untuk kelas ini. Kelas *schizophrenia* juga menunjukkan performa yang relatif rendah dengan *precision* 0.70, *recall* 0.78, dan *F1-score* 0.74, yang mengindikasikan bahwa meskipun model dapat mendeteksi sebagian besar kasus *schizophrenia*, masih terdapat confusi dengan kelas lain.

Tabel 4.8 Hasil *Confusion* Matrix Model CNN

Label	Precision	Recall	F1-Score	Support
BPD	0.98	0.97	0.97	1437
Anxiety	0.89	0.94	0.92	1454
Bipolar	0.96	0.99	0.97	1458
Depression	0.82	0.90	0.86	1558
Mentalillness	0.61	0.51	0.56	1467
Poison	0.92	0.79	0.85	1525
Schizophrenia	0.70	0.78	0.74	1601
Accuracy	—	—	0.84	10500
Macro Avg	0.84	0.84	0.84	10500
Weighted Avg	0.84	0.84	0.84	10500

Berdasarkan Tabel 4.9, model BiLSTM menunjukkan performa dengan akurasi keseluruhan sebesar 0.82, sedikit lebih rendah dibandingkan CNN (0.84). Pola performa per

kelas menunjukkan kemiripan dengan CNN namun dengan beberapa perbedaan penting. Kelas BPD tetap menunjukkan performa terbaik dengan *precision* 0.98, *recall* 0.95, dan *F1-score* 0.97, yang mengkonfirmasi bahwa pola linguistik BPD memiliki karakteristik yang kuat dan dapat dikenali baik oleh ekstraksi fitur lokal maupun pemahaman konteks sekuensial.

Kelas bipolar pada BiLSTM menunjukkan *precision* 0.96, *recall* 0.97, dan *F1-score* 0.96. Meskipun masih sangat baik, terdapat sedikit penurunan pada recall dibandingkan CNN (dari 0.99 menjadi 0.97), yang menunjukkan bahwa beberapa kasus bipolar lebih baik dideteksi melalui pola lokal daripada konteks sekuensial. Performa pada kelas *anxiety* juga sedikit menurun dengan *F1-score* 0.90 (*precision* 0.88, *recall* 0.92) dibandingkan CNN yang mencapai 0.92, mengindikasikan bahwa fitur lokal seperti kata-kata kunci spesifik terkait kecemasan lebih efektif ditangkap oleh CNN.

Perbedaan yang cukup signifikan terlihat pada kelas *depression*, di mana BiLSTM mencapai *precision* 0.82, *recall* 0.84, dan *F1-score* 0.83, lebih rendah dibandingkan CNN dengan F1-score 0.86. Penurunan ini terutama pada *recall* (dari 0.90 menjadi 0.84) menunjukkan bahwa BiLSTM melewatkan lebih banyak kasus depresi dibandingkan CNN. Hal ini mengindikasikan bahwa pola lokal seperti kombinasi kata-kata bervalensi negatif yang berulang pada depresi lebih efektif ditangkap oleh lapisan konvolusional.

Kelas *mentalillness* pada BiLSTM menunjukkan performa yang lebih rendah lagi dengan *precision* 0.57, *recall* 0.49, dan *F1-score* 0.52, dibandingkan CNN dengan *F1-score* 0.56. *Recall* yang sangat rendah (0.49) menunjukkan bahwa BiLSTM hanya mampu mendeteksi kurang dari setengah kasus *mentalillness* yang sebenarnya, mengindikasikan bahwa pemahaman konteks sekuensial saja tidak cukup untuk mengidentifikasi kategori yang heterogen ini. Kelas *poison* menunjukkan peningkatan pada *recall* (dari 0.79 pada CNN menjadi 0.83) namun penurunan pada *precision* (dari 0.92 menjadi 0.87), menghasilkan *F1-score* yang sama yaitu 0.85. Hal ini menunjukkan bahwa BiLSTM lebih sensitif dalam mendeteksi konten berbahaya namun dengan tingkat *false positive* yang lebih tinggi.

Kelas *schizophrenia* menunjukkan performa serupa dengan *precision* 0.68, *recall* 0.77, dan *F1-score* 0.72, sedikit lebih rendah dibandingkan CNN dengan *F1-score* 0.74. Konsistensi performa yang rendah pada kedua model untuk kelas ini mengindikasikan bahwa *schizophrenia* memiliki pola linguistik yang kompleks dan sulit diidentifikasi, baik melalui fitur lokal maupun konteks sekuensial, kemungkinan karena *overlap* dengan gejala gangguan mental lainnya atau variabilitas ekspresi yang tinggi.

Tabel 4.9 Hasil *Confusion Matrix* Model BiLSTM

Label	Precision	Recall	F1-Score	Support
BPD	0.98	0.95	0.97	1437
Anxiety	0.88	0.92	0.90	1454
Bipolar	0.96	0.97	0.96	1458
Depression	0.82	0.84	0.83	1558
Mentalillness	0.57	0.49	0.52	1467
Poison	0.87	0.83	0.85	1525
Schizophrenia	0.68	0.77	0.72	1601
Accuracy	—	—	0.82	10500
Macro Avg	0.82	0.82	0.82	10500
Weighted Avg	0.82	0.82	0.82	10500

Perbandingan ketiga model menunjukkan bahwa CNN-BiLSTM secara konsisten mengungguli kedua model lainnya pada hampir semua kelas. Pada kelas *bipolar*, CNN-BiLSTM mencapai *recall* sempurna (1.00) dan *F1-score* 0.98, lebih tinggi dibandingkan CNN (0.97) dan BiLSTM (0.96). Peningkatan ini menunjukkan bahwa kombinasi ekstraksi fitur lokal dan pemahaman konteks temporal sangat efektif dalam mengenali pola gangguan bipolar yang kompleks, termasuk perubahan mood dan ekspresi emosional yang bervariasi.

Pada kelas BPD, ketiga model menunjukkan performa yang sangat baik, namun CNN-BiLSTM mencapai keseimbangan sempurna dengan *precision* dan *recall* 0.98, sementara BiLSTM memiliki recall sedikit lebih rendah (0.95). Hal ini mengindikasikan bahwa penggabungan kedua pendekatan meminimalkan kesalahan deteksi pada kelas ini. Untuk kelas *anxiety*, CNN-BiLSTM dengan *F1-score* 0.93 mengungguli CNN (0.92) dan BiLSTM (0.90), menunjukkan bahwa pemahaman konteks emosional yang berkembang dalam teks, ditambah dengan deteksi kata-kata kunci kecemasan, memberikan hasil yang lebih baik.

Perbedaan yang paling signifikan terlihat pada kelas depression, di mana CNN-BiLSTM mencapai *F1-score* 0.89, jauh lebih tinggi dibandingkan CNN (0.86) dan BiLSTM (0.83). Peningkatan ini terutama pada *precision* (0.87 pada CNN-BiLSTM berbanding 0.82 pada kedua model lainnya), menunjukkan bahwa arsitektur hybrid lebih akurat dalam mengidentifikasi depresi dengan mengurangi *false positive*. Kemampuan ini penting karena *overlap* gejala depresi dengan gangguan lain sangat tinggi.

Untuk kelas poison, CNN-BiLSTM mencapai *F1-score* 0.92, melampaui CNN dan BiLSTM yang keduanya mencapai 0.85. Peningkatan ini menunjukkan bahwa deteksi konten berbahaya memerlukan pemahaman konteks yang lebih dalam, di mana niat negatif atau pesan berbahaya seringkali tersembunyi dalam struktur kalimat yang kompleks, bukan hanya kata-kata tertentu. Kombinasi CNN dan BiLSTM memungkinkan model menangkap baik penanda eksplisit (kata-kata kasar, ancaman) maupun implisit (sarkasme, manipulasi).

Pada kelas mentalillness, meskipun CNN-BiLSTM mencapai *F1-score* tertinggi (0.61), performa ketiga model tetap relatif rendah. CNN mencapai 0.56 dan BiLSTM hanya 0.52, menunjukkan bahwa kategori ini menjadi tantangan universal. Peningkatan yang diberikan oleh CNN-BiLSTM (*recall* 0.59 berbanding 0.51 pada CNN dan 0.49 pada BiLSTM) menunjukkan bahwa arsitektur hybrid sedikit lebih baik dalam mengenali variabilitas ekspresi pada kategori yang heterogen ini, namun masih memerlukan perbaikan lebih lanjut.

Schizophrenia juga menunjukkan pola serupa, di mana CNN-BiLSTM dengan *F1-score* 0.74 mengungguli CNN (0.74) dan BiLSTM (0.72), meskipun perbedaannya tidak terlalu besar. Hal ini mengindikasikan bahwa deteksi *schizophrenia* memerlukan pemahaman yang lebih kompleks terhadap pola bahasa yang tidak terstruktur, disorganisasi pikiran, dan ekspresi yang tidak biasa, yang masih menjadi tantangan bahkan untuk arsitektur *hybrid*.

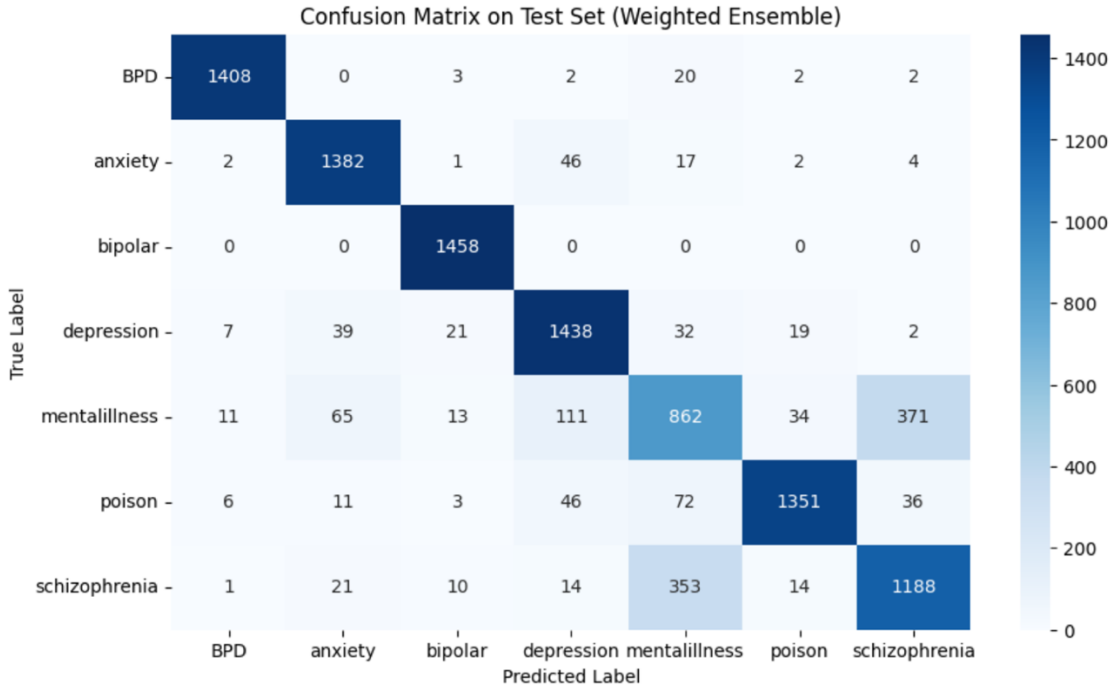
Secara keseluruhan, CNN menunjukkan keunggulan pada deteksi pola lokal yang jelas seperti pada kelas bipolar dan BPD, sementara BiLSTM lebih unggul dalam memahami konteks yang berkembang seperti pada *poison*. Namun, CNN-BiLSTM yang menggabungkan kedua kekuatan ini secara konsisten memberikan performa terbaik dengan akurasi 0.87, melampaui CNN (0.84) dan BiLSTM (0.82). Hasil ini memvalidasi hipotesis bahwa arsitektur *hybrid* yang mengintegrasikan ekstraksi fitur lokal dan pemahaman konteks sekuensial merupakan pendekatan yang paling efektif untuk klasifikasi gangguan mental pada teks media sosial yang kompleks dan beragam.

Gambar 4.3 menunjukkan *confusion matrix* hasil evaluasi model CNN-BiLSTM terhadap data uji dengan menggunakan pendekatan *weighted ensemble*. Visualisasi ini menggambarkan sebaran prediksi model terhadap tujuh kelas gangguan mental, yaitu *BPD*, *anxiety*, *bipolar*, *depression*, *mentalillness*, *poison*, dan *schizophrenia*. Angka pada diagonal utama menunjukkan jumlah prediksi yang sesuai antara label sebenarnya dengan label yang diprediksi, sedangkan angka di luar diagonal menunjukkan kesalahan klasifikasi. Secara umum, model menunjukkan performa klasifikasi yang sangat baik, dengan jumlah prediksi

benar yang tinggi untuk sebagian besar kelas. Pada kelas *bipolar*, model mencatatkan hasil yang sangat optimal dengan seluruh data (1.458 dari 1.458) berhasil diklasifikasikan dengan benar. Kinerja tinggi juga ditunjukkan oleh kelas *BPD* (1.408 dari 1.437) dan *anxiety* (1.382 dari 1.454), yang mencerminkan kemampuan model dalam mengenali pola bahasa yang khas pada kelas-kelas tersebut.

Kelas *depression* juga memiliki tingkat akurasi yang tinggi, dengan 1.438 dari 1.558 data diklasifikasikan secara tepat. Kesalahan klasifikasi pada kelas ini sebagian besar terjadi ke arah kelas *anxiety* dan *mentalillness*, yang secara semantik memang memiliki kesamaan dan tumpang tindih dalam ekspresi gejala. Namun demikian, tantangan terbesar muncul pada kelas *mentalillness*. Dari total 1.467 data, hanya 862 yang berhasil diprediksi dengan benar oleh model, sedangkan sisanya tersebar ke kelas *schizophrenia* (371 data), *depression* (111 data), dan lainnya. Hal ini menunjukkan bahwa label *"mentalillness"* memiliki batasan semantik yang kurang jelas dan seringkali mencakup konten umum, sehingga menyulitkan model dalam mengidentifikasi ciri khasnya secara spesifik.

Kelas *poison* diklasifikasikan dengan benar pada 1.351 dari 1.525 data, dengan kesalahan klasifikasi paling sering terjadi ke arah *mentalillness* dan *depression*. Sementara itu, pada kelas *schizophrenia*, sebanyak 1.188 data berhasil diprediksi dengan tepat dari total 1.601 data, namun terdapat kekeliruan yang cukup signifikan ke arah *mentalillness* (353 data), yang mengindikasikan kemiripan representasi teks antara kedua kelas tersebut.



Gambar 4.3 Hasil *Confusion Matrix* pada Data Uji

Sebagai bagian dari analisis tambahan, dilakukan eksperimen dengan mengeluarkan kelas *mentalillness* dari skema klasifikasi, sehingga model hanya menangani enam kelas yaitu BPD, bipolar, depression, anxiety, schizophrenia, dan poison. Seluruh parameter pengujian dipertahankan sama dengan skenario utama, menggunakan arsitektur CNN-BiLSTM dengan skema 20-Fold Cross Validation yang identik. Perbandingan hasil evaluasi antara skenario tujuh kelas dan enam kelas disajikan pada Tabel 4.10.

Tabel 4.10 Perbandingan Hasil Evaluasi Skenario Tujuh Kelas dan Enam Kelas

Skenario	**Jumlah Kelas**	**Kelas yang Diklasifikasikan**	**Akurasi**	**F1-Score**
Tujuh Kelas (Utama)	7	*BPD, bipolar, depression, anxiety, schizophrenia, mentalillness, poison*	0.87	0.87
Enam Kelas (Ablasi)	6	*BPD, bipolar, depression, anxiety, schizophrenia, poison*	0.9566	0.9566

Berdasarkan Tabel 4.10, terdapat peningkatan akurasi yang cukup signifikan ketika kelas *mentalillness* dikeluarkan dari skema klasifikasi, yaitu dari 0,87 pada skenario tujuh kelas menjadi 0,9566 pada skenario enam kelas. Peningkatan ini berkaitan erat dengan karakteristik kelas *mentalillness* yang bersifat heterogen dan kurang memiliki pola linguistik yang konsisten, sebagaimana tercermin dari nilai precision 0,64, recall 0,59, dan F1-score 0,61 yang merupakan yang terendah di antara seluruh kelas. Kondisi ini menyebabkan model mengalami kesulitan dalam membedakan kelas *mentalillness* dari kelas-kelas lain yang memiliki ekspresi linguistik serupa. Hal ini sejalan dengan temuan Mehmood et al. (2025) yang menunjukkan bahwa keberadaan kelas dengan batas semantik yang tidak tegas dalam skenario multi-kelas dapat memperburuk performa klasifikasi secara keseluruhan, karena meningkatkan permasalahan *class overlap* di sekitar batas keputusan model. Fenomena serupa juga dilaporkan oleh (Gkotsis et al., 2017) yang menemukan bahwa ketika klasifikasi diperluas menjadi 11 tema gangguan mental secara simultan menggunakan CNN, akurasi multi-kelas yang diperoleh hanya sebesar 71,37%, jauh lebih rendah dibandingkan skenario klasifikasi biner yang mencapai 91,08% — menunjukkan bahwa penambahan kelas yang memiliki tumpang tindih linguistik secara konsisten menekan performa model. Konfusi antarkelas serupa juga dilaporkan pada penelitian klasifikasi gangguan mental menggunakan pendekatan ensemble berbasis Transformer, khususnya pada kelas-kelas yang berbagi karakteristik linguistik berdekatan seperti *anxiety*, *depression*, dan *PTSD* (Ovcharuk et al., 2025).

Meskipun demikian, skenario enam kelas tidak dijadikan konfigurasi utama dalam penelitian ini. Penghapusan kelas *mentalillness* akan mempersempit cakupan sistem deteksi, sementara tujuan penelitian ini adalah membangun model yang mampu mengklasifikasikan tujuh kategori gangguan mental secara simultan. (Islam et al., 2025) dalam penelitiannya menggunakan multiMentalRoBERTa juga menemukan bahwa pengurangan jumlah kelas — dalam hal ini mengeluarkan kelas *stress* yang bersifat umum — meningkatkan macro F1-score dari 0,839 menjadi 0,870, namun diakui bahwa hal tersebut mengurangi kelengkapan cakupan deteksi gangguan mental. Hal serupa juga dikemukakan oleh (Sutranggono et al., 2024) yang menyatakan bahwa perluasan jumlah kelas dalam klasifikasi multi-kelas, meskipun meningkatkan kompleksitas model, menghasilkan sistem deteksi yang lebih representatif secara klinis. Dengan demikian, penurunan akurasi yang terjadi pada skenario tujuh kelas merupakan konsekuensi yang dapat dipahami dari pilihan desain penelitian yang mengutamakan kelengkapan cakupan klasifikasi.

4.6 Evaluasi Manual Confusion Matrix

Sebagai bagian dari proses verifikasi terhadap analisis evaluasi model, dilakukan perbandingan nilai antara hasil perhitungan manual *confusion matrix* dan metrik evaluasi model CNN-BiLSTM. Langkah verifikasi ini bertujuan untuk memastikan bahwa proses ekstraksi nilai-nilai evaluasi seperti *precision, recall*, dan *F1-score* telah dilakukan secara akurat, serta bahwa interpretasi performa model yang diturunkan dari perhitungan manual dapat diandalkan dan sahih secara metodologis.

Untuk memberikan gambaran konkret mengenai metodologi perhitungan, dilakukan demonstrasi perhitungan manual menggunakan kelas depression sebagai contoh. Berdasarkan *confusion matrix* pada data uji, diperoleh nilai-nilai komponen sebagai berikut: TP *(True Positive)* = 1,438 yang merepresentasikan prediksi *depression* yang benar, FP *(False Positive)* = 219 yang menunjukkan sampel non-depression yang diprediksi sebagai depression, dan FN *(False Negative)* = 120 yang merepresentasikan sampel *depression* yang diprediksi sebagai kelas lain. Proses perhitungan dilakukan dengan mengakumulasi nilai-nilai dari *confusion matrix* secara sistematis. Untuk FP *(depression)*, nilai diperoleh dari penjumlahan: 2 + 46 + 0 + 111 + 46 + 14 = 219, sedangkan FN *(depression)* dihitung dari: 7 + 39 + 21 + 32 + 19 + 2 = 120. Berdasarkan nilai-nilai tersebut, perhitungan metrik evaluasi menggunakan formula standar menghasilkan:

$$Presisi = \frac{1438}{(1438 + 219)} = \frac{1438}{(1657)} = 0.8678$$

$$Recall = \frac{1438}{(1438 + 120)} = \frac{1438}{(1558)} = 0.9230$$

$$F1 - Score = \frac{2 \times (0.8678 \times 0.9230)}{(0.8678 + 0.9230)} = \frac{1.6019}{1.7908} = 0.8946$$

Evaluasi manual ini kemudian diterapkan secara konsisten untuk seluruh kelas gangguan mental yang diteliti. Hasil perhitungan lengkap disajikan dalam Tabel 4.3 yang menunjukkan variasi performa yang signifikan antar kelas. Sebagaimana terlihat dalam Tabel 4.3, kelas BPD *(Borderline Personality Disorder)* menunjukkan performa terbaik dengan TP = 1,408, FP = 27, FN = 29, menghasilkan precision = 0.9812, *recall* = 0.9798, dan *F1-score* = 0.9805. Kelas *Anxiety* memperlihatkan TP = 1,382, FP = 136, FN = 72, dengan *precision* = 0.9104, *recall* = 0.9505, dan *F1-score* = 0.9300. Kelas Bipolar menunjukkan performa yang sangat baik dengan TP = 1,458, FP = 51, FN = 0, menghasilkan *precision* = 0.9662, *recall* sempurna = 1.0000, dan *F1-score* = 0.9828.

Berdasarkan data dalam Tabel 4.3, kelas Depression dengan TP = 1,438, FP = 219, FN = 120 menghasilkan *precision* = 0.8678, *recall* = 0.9230, dan *F1-score* = 0.8946. Kelas *Mentalillness* menunjukkan performa terendah dengan TP = 862, FP = 494, FN = 605, menghasilkan *precision* = 0.6357, *recall* = 0.5876, dan *F1-score* = 0.6107. Kelas *Poison* memperlihatkan TP = 1,351, FP = 71, FN = 174, dengan *precision* = 0.9501, *recall* = 0.8859, dan *F1-score* = 0.9169. Terakhir, kelas *Schizophrenia* dengan TP = 1,188, FP = 415, FN = 413 menghasilkan *precision* = 0.7411, *recall* = 0.7420, dan *F1-score* = 0.7416.

Tabel 4.11 Perhitungan Manual Kelas *Mental Health*

Kelas	TP	FP	FN	Presisi	Recall	F1-Score
BPD	1,408	27	29	0.9812	0.9798	0.9805
Anxiety	1,382	136	72	0.9104	0.9505	0.9300
Bipolar	1,458	51	0	0.9662	1.0000	0.9828
Depression	1,438	219	120	0.8678	0.9230	0.8946
Mentalillness	862	494	605	0.6357	0.5876	0.6107
Poison	1,351	71	174	0.9501	0.8859	0.9169
Schizophrenia	1,188	415	413	0.7411	0.7420	0.7416

Tahap validasi dilakukan dengan membandingkan hasil perhitungan manual dengan output otomatis dari model CNN-BiLSTM, sebagaimana disajikan dalam Tabel 4.7. Hasil perbandingan dalam Tabel 4.7 menunjukkan kesesuaian yang sangat tinggi antara kedua metode evaluasi. Untuk kelas BPD, perhitungan manual menghasilkan precision = 0.9812,

recall = 0.9798, *F1-score* = 0.9805, sementara output model menunjukkan *precision* = 0.98, *recall* = 0.98, *F1-score* = 0.98. Kelas *Anxiety* memperlihatkan hasil manual precision = 0.9104, *recall* = 0.9505, F1-score = 0.9300, dibandingkan dengan output model precision = 0.91, *recall* = 0.95, F1-score = 0.93.

Berdasarkan data perbandingan dalam Tabel 4.11, untuk kelas *Bipolar*, perhitungan manual menghasilkan *precision* = 0.9662, *recall* = 1.0000, *F1-score* = 0.9828, sementara model menghasilkan *precision* = 0.97, *recall* = 1.00, *F1-score* = 0.98. Kelas *Depression* menunjukkan hasil manual *precision* = 0.8678, *recall* = 0.9230, *F1-score* = 0.8946, dibandingkan dengan output model *precision* = 0.87, recall = 0.92, *F1-score* = 0.89. Kelas *Mentalillness* memperlihatkan perhitungan manual *precision* = 0.6357, *recall* = 0.5876, F*1-score* = 0.6107, sementara model menghasilkan *precision* = 0.64, *recall* = 0.59, *F1-score* = 0.61. Kelas *Poison* menunjukkan hasil manual *precision* = 0.9501, r*ecall* = 0.8859, *F1-score* = 0.9169, dibandingkan dengan output model *precision* = 0.95, *recall* = 0.89, *F1-score* = 0.92. Terakhir, kelas *Schizophrenia* memperlihatkan perhitungan manual precision = 0.7411, *recall* = 0.7420, *F1-score* = 0.7416, sementara model menghasilkan *precision* = 0.74, *recall* = 0.74, *F1-score* = 0.74.

Tabel 4.12 Perbandingan Hasil Perhitungan Manual dengan *Output* Model CNN-BiLSTM

Kelas	**Precision (Manual)**	**Recall (Manual)**	**F1-Score (Manual)**	**Precision (Model)**	**Recall (Model)**	**F1-Score (Model)**
BPD	0.9812	0.9798	0.9805	0.98	0.98	0.98
Anxiety	0.9104	0.9505	0.9300	0.91	0.95	0.93
Bipolar	0.9662	1.0000	0.9828	0.97	1.00	0.98
Depression	0.8678	0.9230	0.8946	0.87	0.92	0.89
Mentalillness	0.6357	0.5876	0.6107	0.64	0.59	0.61
Poison	0.9501	0.8859	0.9169	0.95	0.89	0.92
Schizophrenia	0.7411	0.7420	0.7416	0.74	0.74	0.74

Hasil validasi menunjukkan bahwa seluruh nilai metrik evaluasi yang diperoleh melalui perhitungan manual memiliki kesesuaian sempurna dengan output model ketika dibulatkan hingga dua digit desimal. Deviasi numerik yang muncul sangat kecil (≤ 0.004) dan seluruhnya masih berada dalam margin toleransi evaluatif yang dapat diterima secara statistik. Temuan ini mendemonstrasikan bahwa baik model maupun perhitungan manual memiliki konsistensi yang tinggi dalam mengkuantifikasi performa klasifikasi untuk masing-masing kelas gangguan mental.

Konsistensi ini mengonfirmasi bahwa sistem evaluasi otomatis menggunakan rumus yang sejalan dengan prinsip evaluasi manual standar, yakni: *Precision* = TP / (TP + FP), *Recall* = TP / (TP + FN), dan *F1-Score* = 2 × ((Recall × *Precision*)) / ((*Recall* + *Precision*)). Dengan total data uji sebanyak 10.500 sampel, kedua pendekatan menghasilkan nilai akurasi keseluruhan sebesar 87% yang identik, serta metrik makro dan mikro yang konsisten, menunjukkan reliabilitas tinggi dari sistem evaluasi yang digunakan.

Analisis mendalam terhadap pola performa model berdasarkan data dalam Tabel 4.7 dan Tabel 4.12 mengungkapkan karakteristik klasifikatif yang konsisten pada setiap kelas gangguan mental. Merujuk pada Tabel 4.7, kelas BPD menampilkan nilai precision dan recall yang sangat tinggi dan seimbang (≈0.98), mencerminkan kemampuan model yang superior dalam mendeteksi dan mengidentifikasi ekspresi linguistik yang khas dari gangguan kepribadian *borderline*. Keseimbangan antara precision dan recall yang terlihat dalam Tabel 4.10 menunjukkan bahwa model tidak hanya mampu mendeteksi mayoritas kasus BPD (*recall* tinggi) tetapi juga mempertahankan akurasi tinggi dalam prediksinya (*precision* tinggi).

Berdasarkan data dalam Tabel 4.7, kelas *Bipolar* memperlihatkan karakteristik yang unik dengan nilai *recall* sempurna (1.000), yang menandakan bahwa seluruh sampel bipolar dalam data uji berhasil dikenali tanpa ada yang terlewat (FN = 0). *Precision* yang sedikit lebih rendah (0.9662) sebagaimana tercantum dalam Tabel 4.7 mengindikasikan adanya beberapa *false positive* (FP = 51), namun tingkat kesalahan ini masih sangat rendah dan dapat diterima. Performa ini menunjukkan bahwa ekspresi linguistik terkait gangguan bipolar memiliki pola yang cukup distinktif sehingga mudah dikenali oleh model.

Mengacu pada Tabel 4.7, kelas *Anxiety* memperlihatkan pola yang berbeda dengan recall lebih tinggi (0.9505) dibanding *precision* (0.9104). Karakteristik ini menandakan kecenderungan model untuk mendeteksi banyak kasus *anxiety* namun kurang selektif dalam proses klasifikasi. Fenomena ini kemungkinan disebabkan oleh tumpang tindih semantik antara ekspresi *anxiety* dengan depresi, mengingat kedua kondisi mental tersebut sering kali memiliki manifestasi linguistik yang serupa dalam konteks media sosial.

Berdasarkan data dalam Tabel 4.7, kelas Depression menunjukkan precision yang relatif lebih rendah (0.8678) karena tingginya nilai *false positive* (FP = 219), meskipun *recall* tetap tinggi (0.9230). Pola ini mengindikasikan bahwa model memiliki sensitivitas yang cukup baik terhadap ekspresi depresi dan mampu mendeteksi mayoritas kasus, namun cenderung mengklasifikasikan beberapa ekspresi dari kelas lain sebagai depresi. Hal ini

dapat dipahami mengingat depresi merupakan kondisi yang kompleks dengan manifestasi linguistik yang beragam dan sering overlapping dengan kondisi mental lainnya.

Merujuk pada Tabel 4.7, kelas *Mentalillness* menunjukkan performa paling rendah dengan *precision* (0.6357) dan recall (0.5876) di bawah 0.65. Rendahnya performa ini, yang juga dikonfirmasi oleh tingginya nilai FP (494) dan FN (605), mengindikasikan adanya ambiguitas yang signifikan pada ekspresi linguistik dari label tersebut, yang menyebabkan rendahnya kemampuan diskriminatif model. Hasil ini menunjukkan bahwa kategori "*mental illness*" yang bersifat umum sulit dibedakan dari kategori-kategori spesifik lainnya, sehingga model mengalami kesulitan dalam proses klasifikasi.

Berdasarkan data dalam Tabel 4.7, kelas *Poison* memiliki karakteristik yang menarik dengan precision sangat tinggi (0.9501) namun recall yang relatif lebih rendah (0.8859). Pola ini, yang ditunjukkan oleh nilai FP yang rendah (71) namun FN yang cukup tinggi (174), mengindikasikan kecenderungan model yang sangat selektif dan konservatif dalam mendeteksi konten berbahaya atau beracun. Model cenderung hanya mengklasifikasikan teks sebagai "*poison*" ketika memiliki keyakinan tinggi, sehingga menghasilkan *precision* yang sangat baik namun dengan konsekuensi beberapa kasus yang tidak terdeteksi (*false negatives*).

Mengacu pada Tabel 4.7, kelas *Schizophrenia* menunjukkan performa yang moderat dengan precision dan recall yang hampir seimbang di sekitar 0.74. Keseimbangan ini, yang juga terlihat dari nilai FP (*False Positive*) (415) dan FN (*False Negative*) (413) yang relatif setara, menggambarkan kompleksitas linguistik dari ekspresi terkait *schizophrenia* yang belum sepenuhnya dapat ditangkap oleh model. Hasil ini menunjukkan bahwa meskipun model mampu mengenali sebagian besar karakteristik linguistik *schizophrenia*, masih terdapat ruang untuk peningkatan dalam memahami nuansa ekspresi yang lebih *subtle* dari kondisi ini.

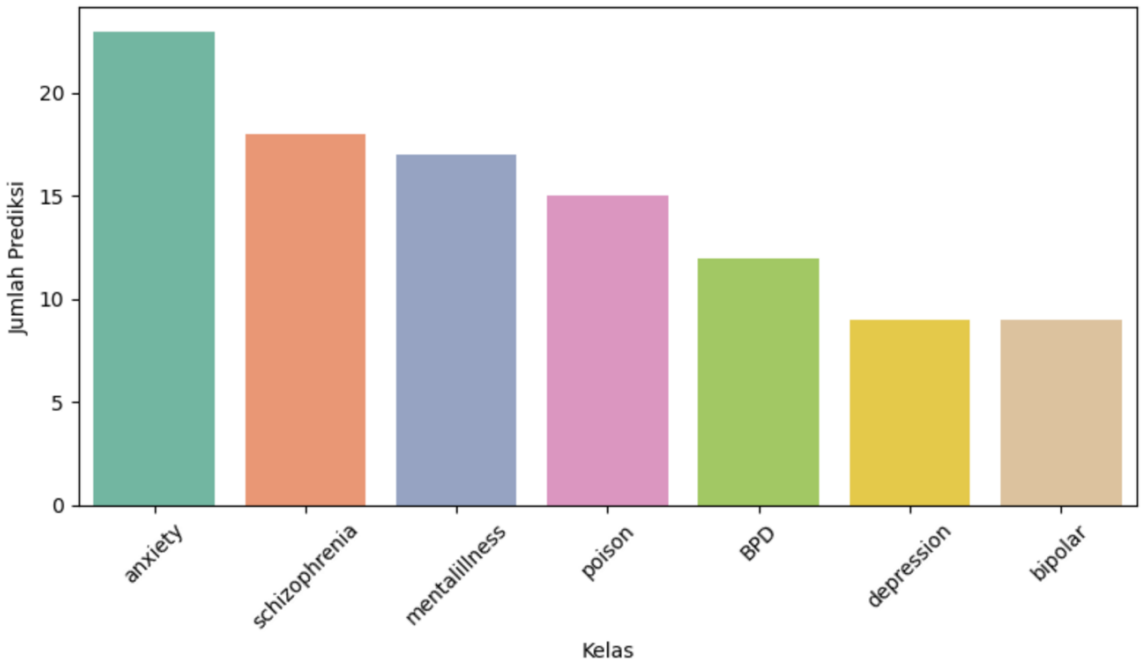
Secara keseluruhan, hasil evaluasi yang disajikan dalam Tabel 4.7 dan Tabel 4.12 tidak hanya mengonfirmasi akurasi metodologis dari proses evaluasi yang dilakukan, tetapi juga memberikan wawasan mendalam mengenai kemampuan dan limitasi model CNN-BiLSTM dalam mengklasifikasikan berbagai jenis gangguan mental berdasarkan ekspresi linguistik. Konsistensi antara perhitungan manual dalam Tabel 4.7 dan *output* otomatis dalam Tabel 4.12 memberikan kepercayaan tinggi terhadap reliabilitas hasil evaluasi, sementara analisis per-kelas berdasarkan data dalam kedua tabel tersebut mengungkapkan karakteristik unik dari setiap kategori gangguan mental yang dapat menjadi dasar untuk optimalisasi model di masa mendatang.

4.7 Hasil Prediksi Model CNN-BiLSTM

Setelah proses pelatihan model selesai, tahap selanjutnya adalah melakukan prediksi terhadap data baru guna mengamati bagaimana model mengklasifikasikan kalimat-kalimat yang mencerminkan kondisi psikologis individu. Prediksi ini dirancang untuk memberikan dukungan kepada psikolog dalam memahami indikasi awal gangguan mental berdasarkan ekspresi verbal yang dituliskan individu, sehingga diagnosis dapat dilakukan dengan lebih tepat. Data yang digunakan dalam tahap ini diambil dari berbagai unggahan di subreddit yang berfokus pada isu kesehatan mental, seperti *r/mentalhealth*, *r/bipolar*, *r/schizophrenia*, *r/depression*, *r/anxiety*, dan *r/bpd*. Setiap unggahan dianalisis untuk menilai keterkaitannya dengan salah satu kategori gangguan mental yang telah ditentukan.

Proses klasifikasi dilakukan menggunakan model terbaik yang diperoleh dari hasil pelatihan menggunakan teknik *K-Fold cross-validation*, sehingga memastikan bahwa model yang digunakan memiliki performa paling optimal. Sebelum diprediksi, kalimat-kalimat input terlebih dahulu melalui tahap tokenisasi menggunakan *tokenizer* yang sesuai, untuk mengubah teks menjadi bentuk representasi yang dapat dikenali oleh model. *Output* dari model berupa probabilitas untuk masing-masing label, yang disertai dengan skor keyakinan *(confidence score)* guna menunjukkan seberapa besar tingkat kepercayaan model terhadap prediksi tersebut. Dengan pendekatan ini, model diharapkan mampu mengenali dan mengelompokkan gangguan mental secara kontekstual dari bahasa yang digunakan oleh individu dalam teks yang mereka tulis.

Sebanyak 103 data input baru dianalisis menggunakan model klasifikasi yang telah dikembangkan. Data ini bersifat murni sebagai data uji dan tidak pernah digunakan dalam proses pelatihan maupun validasi sebelumnya. Berdasarkan hasil prediksi, model paling sering mengklasifikasikan teks ke dalam kategori *anxiety*, yaitu sebanyak 23 data. Disusul oleh kategori *schizophrenia* sebanyak 18 data, *mental illness* sebanyak 17 data, dan *poison* sebanyak 15 data. Selanjutnya, sebanyak 12 data dipetakan ke dalam kategori *BPD*, serta masing-masing 9 data diklasifikasikan ke dalam kategori *depression* dan *bipolar*.



Gambar 4.4 Distribusi Kelas Hasil Prediksi

Berdasarkan hasil prediksi terhadap sejumlah teks input baru, model menunjukkan performa yang cukup baik dalam mengklasifikasikan berbagai gangguan kesehatan mental berdasarkan pola linguistik yang terkandung dalam narasi. Masing-masing label dalam klasifikasi menunjukkan kecenderungan konteks tertentu yang berhasil ditangkap oleh model, tergambar dari distribusi probabilitas yang diberikan pada setiap teks. Label *poison* muncul dominan pada narasi yang menggambarkan dorongan bunuh diri atau pemikiran untuk menyakiti diri sendiri. Teks-teks semacam ini sering kali memuat deskripsi eksplisit terkait rencana tindakan berbahaya terhadap diri sendiri, serta pengalaman emosional yang intens. Model memberikan skor sangat tinggi, bahkan mencapai lebih dari 99%, yang menandakan sensitivitasnya dalam mendeteksi konteks fatalistik. Untuk *anxiety*, model secara konsisten memberikan prediksi tinggi pada narasi yang menggambarkan kecemasan intens, fobia, dan pikiran obsesif yang mengganggu kehidupan sehari-hari. Dalam salah satu teks, kekhawatiran berlebihan terhadap kanker memicu skor prediksi mencapai 98,69%, menunjukkan bahwa model mampu menangkap pola bahasa yang penuh ketegangan dan kegelisahan. Tabel 4.13 adalah contoh dari hasil prediksi terhadap Kesehatan mental menggunakan arsitektur CNN-BiLSTM.

Label *mental illness* umumnya muncul pada teks yang menggambarkan penderitaan psikologis secara luas, tanpa dominasi gejala spesifik tertentu. Dalam beberapa narasi, pengalaman berada di rumah sakit jiwa, trauma, dan perasaan kehilangan makna hidup diklasifikasikan dalam kategori ini dengan probabilitas tinggi, hingga 88%. Hal ini

menunjukkan bahwa model memiliki kapasitas dalam mengenali kompleksitas gangguan secara umum. Prediksi terhadap *depression* muncul pada teks-teks yang berisi perasaan hampa, kehilangan harapan, dan disosiasi sosial. Salah satu narasi dengan skor prediksi 62,36% menunjukkan bahwa model mampu mengenali tanda-tanda khas depresi seperti kelelahan emosional, isolasi, dan penurunan minat terhadap interaksi sosial.

Kategori *BPD* (*Borderline Personality Disorder*) dikenali dengan sangat kuat oleh model, terutama ketika teks mengandung penyebutan langsung diagnosis atau menggambarkan instabilitas emosi dan relasi interpersonal yang ekstrem. Dalam beberapa kasus, model memberikan skor prediksi melebihi 99%, mencerminkan kesesuaian antara isi narasi dan ciri khas gangguan kepribadian ambang. Untuk label *schizophrenia*, model mengklasifikasikan teks dengan gejala psikotik seperti halusinasi visual atau auditori, serta persepsi realitas yang terganggu. Narasi terkait dengan gangguan persepsi diberi skor prediksi hingga 97%, menunjukkan ketepatan model dalam mengenali deskripsi pengalaman psikotik.

Sementara itu, *bipolar disorder* juga berhasil dikenali secara akurat. Dalam salah satu contoh, narasi yang memuat penyebutan diagnosis *bipolar*, perubahan mood ekstrem dari fase manik ke fase depresi, serta gangguan pola tidur mendorong prediksi dengan skor sebesar 99,91%. Hal ini menunjukkan bahwa ketika indikator linguistik khas *bipolar* muncul dalam narasi, model dapat mengidentifikasinya secara tepat. Untuk memahami kata-kata yang berkontribusi signifikan terhadap keputusan klasifikasi model, penelitian ini menggunakan metode Integrated Gradients (Sundararajan et al., 2017) sebagai pendekatan *explainable AI*. Metode ini menghitung kontribusi setiap elemen input dengan mengintegrasikan gradien model sepanjang jalur dari baseline (zero embedding vector) menuju input aktual.

Secara matematis, attribution untuk elemen ke-i diformulasikan sebagai $IG_i(x) = (x_i - x'_i) \times \int_0^1 [\partial F(x' + \alpha(x - x'))/\partial x_i]\, d\alpha$, dimana x adalah input embedding dan F adalah model CNN-BiLSTM. Implementasi menggunakan 50 langkah interpolasi untuk mengaproksi integral dengan model terbaik dari *20-fold cross-validation*.

Integrated Gradients diterapkan pada input space original (teks dengan stopwords), bukan pada feature space yang telah dipreprocess. Pendekatan ini mengikuti prinsip *explainable AI* bahwa *explanation* harus diberikan dalam representasi yang dapat dipahami user (Ribeiro et al., 2016). Konsekuensinya, attribution scores untuk stopwords mencerminkan kontribusi kontekstual (peran struktural dalam frasa), sementara attribution scores untuk kata content mencerminkan learned importance dari model.

Tabel 4.13 Sampel Hasil Prediksi Model terhadap Teks Kesehatan Mental

Teks	Hasil Prediksi	Kata Penting
I died and was revived after an attempt. I was forced into a psychiatric hospital for three months. All they did was medicate me until I agreed to everything. No therapy or life assistance. Just pills. I still want to die. I'm angry they brought me back.	BPD : 0.22% anxiety : 0.20% bipolar : 0.02% depression : 1.01% **mentalillness : 71.70%** poison : 2.73% schizophrenia : 24.12%	i died **and** was **revived** after **an** attempt i **was forced into a psychiatric** hospital **for** three **months all they** did **was** medicate **me until** i agreed **to** everything no **therapy** or **life** assistance **just** pills **i** still want to **die i'm** angry they **brought** me **back**
So I am always passively suicidal. I would be totally fine if someone ran me over or I got murdered. Most days I dont have the urge to do the deed myself, but on those really bad days I have a trick. So in my experience the super suicidal episodes only last about a day before my meds rebalance me. So I give myself 3 days. If after 3 days the thoughts are still super bad, then i will start steps toward my plan. It's like putting a 72 hour psych hold on myself. I also have a kinda convoluted plan. Options ar	BPD : 0.11% anxiety : 0.17% bipolar : 0.02% depression : 1.77% mentalillness : 10.27% **poison : 84.70%** schizophrenia : 2.96%	so i am always passively **suicidal** i **would** be totally fine if someone ran me over or i got murdered most days i dont have the urge to do the deed myself but on those really bad days i have a trick so in my experience the super suicidal episodes only last about a day before my meds rebalance me so i give myself days if after days the thoughts are still super bad then i will start steps toward my plan like putting a hour psych hold on myself i also have a kinda convoluted plan options are firearms or od and both would require a couple days to get the things i need to do the deed both of these things help me stay alive even when in the moment i want to be i hope that someone on here is able to use this

e firearms or OD and both would require a co uple days to get the things I need to do the de ed. Both of these things help me stay alive, ev en when, in the moment, I don't want to be. I hope that someone on here is able to use this method or adapt it to their own uses. Someti mes when they say "one more day", it really j ust needs to be one more day.		method or adapt it to their own uses sometimes when they say o ne more day it really just needs to be one more day
i'm no job no social life little to no friends no dreams or ambitions no relationships nothing took a drive this morning and found a bunch of people of my age in a group and i really felt like an alien oh add my social anxiety to this what's wrong with me why do i see other people have all or some of the above or at least the will to do something or achieve something and here i am dead as fuck from the inside this thought eats me up every night mornings are gloomy as fuck no matter what and let's not talk about my uni days it was a nightmare	BPD 0.08% **anxiety** **73.61%** bipolar 0.06% depression 1.32% mentalillness 6.33% poison 16.27% schizophrenia 2.33%	**i'm** no job no social life little to no friends no dreams or ambitions no relationships nothing took a drive this morning and found a bunch of people of my age in a group and i really felt like an alien oh add my social **anxiety** to this what's wrong with me why do i see other people have all or some of the above or at least the will to do something or achieve something and here i am dead as fuck from the inside this thought eats me up every night mornings are gloomy as fuck no matter what and let's not talk about my uni days it was a nightmare fuelled with tension stress and **anxiety** for something my friends used to be too chilled about i freak out easily delusions and no live to will except for my parents would i fit in this world would i ever be

fuelled with tension stress and anxiety for something my friends used to be too chilled about i freak out easily delusions and no live to will except for my parents would i fit in this world would i ever be happy i even forgot what that feels like anybody in the same boat as me		happy i even forgot what that feels like anybody in the same boat as me
Avoidant discard in November, worst heartbreak I've ever felt, abandoned, out of nowhere, someone I thought I was going to marry. My best friend, my dog, gets sick but I get a few months with her before the tumor progresses, she makes it to our birthday (born on the same day cool huh) the next morning I say goodbye to my best friend, but I'm happy we got to spend our day together. Month later (present) I'm just overwhelmed, I'm exhausted, I'm still trying to stay busy with school and bettering myself but I feel so alone. Losing two of the most important people in my life, I consta	BPD : 0.31% anxiety : 0.31% bipolar : 0.02% depression : 5.13% **mentalillness : 88.10%** poison : 0.29% schizophrenia : 5.84%	**avoidant discard** in november worst **heartbreak i've** ever felt abandoned out of nowhere someone i thought i was going to marry my best friend my dog gets sick but i get a few months with her before the tumor progresses she makes it to our birthday born on the same day cool huh the next morning i say goodbye to my best friend but **i'm** happy we got to spend our day together month later present i'm just overwhelmed i'm exhausted i'm still trying to stay busy with school and bettering myself but i feel so alone losing two of the most important people in my life i constantly feel "what am i doing this for" i don't want to be here anymore i'm trying

ntly feel “what am I doing this for” I don’t wa nt to be here anymore. I’m trying.		
This is just a quick rant. Why is severe anxiety the only crippling mental illness where we are expected to find the "root cause", to "do the work", etc.? We never tell that to people with severe depression, bipolar disorder, schizophrenia. We understand they have a serious chemical imbalance and medication is going to be doing most of the heavy lifting, and other things will simply be adjunct treatments. Maybe I'm being overly cynical. I hope I am not. What are your thoughts?	BPD : 0.25% **anxiety : 50.38%** bipolar : 10.04% depression : 12.44% mentalillness : 12.62% poison : 5.22% schizophrenia : 9.04%	this is just a quick rant why is severe **anxiety** the only crippling mental illness where we are expected to find the root cause to d o the work etc we never tell that to people with severe depressio n bipolar disorder schizophrenia we understand they have a seri ous chemical imbalance and medication is going to be doing mo st of the heavy lifting and other things will simply be adjunct tre atments maybe being overly cynical i hope i am not what are yo ur thoughts
I am so scared of getting cancer and it’s takin g over my life. I’m 33, could be in better sha pe after children! But recently the fear of can cer is absolutely taking over my life :(I kno w I’m still relatively young, but I know two p eople who I went to school with who had can		i am so scared of getting cancer and it’s taking over my life i’m could be in better shape after children but recently the **fear** of ca ncer is absolutely taking over my life i know i’m still relatively young but i know two people who i went to school with who ha d cancer under one who didn’t make it every day the **fear** just ta kes over and i can’t seem to be happy i feel some happiness and

cer under 30 - one who didn't make it. Every day the fear just takes over and I can't seem t o be happy. I feel some happiness and then m y brain says 'but you might get cancer soon' and I suddenly feel so anxious and down, it's a horrible feeling. Hard to explain. What can I do to help this fear!?	BPD : 0.00% **anxiety** : **98.69%** bipolar : 0.00% depression : 0.05% mentalillness : 1.13% poison : 0.01% schizophrenia : 0.12%	then my brain says you might get cancer and i suddenly feel so anxious and down it's a horrible feeling hard to explain what ca n i do to help this **fear**
"i don't feel comfortable sharing my age here but after a really messy divorce with my father a few years ago, my mum has gone into a spiral of drinking then going sober, then relapsing. i've spoken to her about it before but recently she has started again. for context, i don't have many friends so most of the time i'm inside on my screens and me and my mum get arguments a lot and most if not all end up with me storming out the house. she says that the root of the argument is always my phone, but i sort of have this internal rage that she's betrayed me or	BPD : 0.44% anxiety : 0.35% bipolar : 0.05% depression : 3.36% **mentalillness** : **85.72%** poison : 0.17% schizophrenia : 9.92%	**i don't feel comfortable sharing my age here but after a** reall y **messy divorce with my father a few years ago my mum ha s gone into a** spiral **of drinking then going sober then** relapsin g **i've spoken to her about it before but recently she has star ted again for context i don't have many friends so most of th e time i'm inside on my screens and me and my mum get ar guments a lot and most if not all end up with me storming o ut the house she says that the root of the argument is always my phone but i sort of have this internal rage that she's betr ayed me or something by starting to drink again because she knows very much how it affects me when she's drunk she'll also just say really horrible things to me i've told my sister who is years older than me and my sister has spoken to her**

something by starting to drink again because she knows very much how it affects me. when she's drunk she'll also just say really horrible things to me. i've told my sister who is 8 years older than me and my sister has spoken to her about it before but my mum pretends to my sister that she's stopped and whenever i want to call my sister to tell her it's happening again my mum takes away my phone. i don't know what to do and i feel like im trapped in a loop of her relapsing. also we live in a small town so i don't think there is any alcoholic support things i can refer her to. now that im thinking it if i showed her one she'd probably be very angry. she knows it's happening but she doesn't want to confront the problem. i have a lot of sympathy for my mum as during the divorce with my father she was diagnosed with PTSD and is currently attending therapy so please don't shame her for it."		**about it before but my mum pretends to my sister that she's stopped and whenever i want to call my sister to tell her it's happening again my mum takes away my phone i don't kno w what to do and i feel like im trapped in a loop of her relap sing also we live in a small town so i don't think there is any alcoholic support things i can refer her to now that im think ing it if i showed her one she'd probably be very angry she k nows it's happening but she doesn't want to confront the pr oblem i have a lot of sympathy for my mum as during the di vorce with my father she was diagnosed with ptsd and is cur rently attending therapy so please don't shame her for it**

Depression sucks, I'm really feeling it tonight. Usually only happens if i take my medication late but ive kinda been bouncingg times so it doesntfeel that extreme of a different time today. The cat aint even cheering me up. I just dont feel hope I will find a partner. I dont think i am really connecting with the friends I have or I dont feel comfortable fully opening up to them so I feel kinda stuck. This combined with the past experiences i have i mean i was walking home earlier and people just kept gettingin my way and I just wanted to kinda yell or buldoze through them like blastic pins like they are ok just knocked down so theyy are out of my way [7:45 PM] People just, let me down too much, im not connecting with people. Feels like there isnt anyone I can connect withor atleast whatever person doesnt feel the saame way about me and it	BPD : 0.03% anxiety : 0.08% bipolar : 0.04% **depression** : **62.36%** mentalillness : 2.85% poison : 34.33% schizophrenia : 0.32%	**depression** sucks really feeling it tonight usually only happens i f i take my medication late but ive kinda been times so it that ex treme of a different time today the cat aint even cheering me up i just dont feel hope i will find a partner i dont think i am really connecting with the friends i have or i dont feel comfortable full y opening up to them so i feel kinda stuck this combined with th e past experiences i have i mean i was walking home earlier and people just kept my way and i just wanted to kinda yell or throu gh them like pins like they are ok just knocked down so are out of my way pm people just let me down too much im not connec ting with people feels like there isnt anyone i can connect atleas t whatever person doesnt feel the way about me and it feels like i aint even giving people much to begin with

feels like I aint even giving people much to begin with"		
I'm currently in my last semester of college and I am feeling so unmotivated, exhausted, and burnt out. I am not taking particularly difficult classes, in fact I would say this is my easiest semester in college in terms of classes and workload. Yet even getting one assignment done feels a lot, I have no desire to do any school work (and am really forcing myself to) and I am just so tired.	BPD : 0.26% anxiety : 0.23% bipolar : 0.02% depression : 1.48% mentalillness : 41.90% **poison** : **43.55%** schizophrenia : 12.56%	currently **in** my **last semester** of college and i **am feeling** so un motivated **exhausted** and burnt out i am not taking particularly difficult **classes** in fact i would say this **is** my easiest semester in college in terms of classes and workload yet even **getting** one a ssignment done feels a lot i have no desire to do any school wor k and am really forcing myself to and i am just so tired
I'm going to die someday. It'll be in a car crash, where it's sudden, it'll be when I close my eyes to sleep at night and never open them up, it'll be dying of cancer slowly and seeing myself wither away. One day, I'm going to close my eyes for the last time. it can be five minutes from now or fifty years from now. And you never know. That's the worst	BPD : 0.01% anxiety : 0.48% bipolar : 0.00% depression : 0.06% mentalillness : 0.18% **poison** : **99.21%** schizophrenia : 0.06%	going to **die** someday be in a car crash where sudden be when i close my eyes to sleep at night and never open them up be dyin g of cancer slowly and seeing myself wither away one day goin g to close my eyes for the last time it can be five minutes from n ow or fifty years from now and you never know the worst part o ne day stop living my mind will stop running i will simply not e xist i want to believe in heaven but i some day people will forge t about me i have plans for the future what if i die before i can a

part. One day I'll stop living. My mind will stop running, I will simply not exist. I want to believe in heaven but I can't. Some day people will forget about me. I have plans for the future-- what if I die before I can accomplish any of them? Before i can go to college, get married, have a career, see the world. How do you go outside every day with the knowledge you're going to die? I just want to stay inside and protect myself. I haven't been able to sleep for two days because every time i close my eyes I think-- this could be your last day on earth. I'm on the brink of a panic attack. How is school not a waste of time if you can die tomorrow? Why the fuck does a job or money or a house even matter if you can die ten minutes from now? If you can get diagnosed with ALS, or cancer, or some other horrible disease with no cure? How the fuck do you live like this? How can anyone live with this knowledge?		ccomplish any of them before i can go to college get married ha ve a career see the world how do you go outside every day with the knowledge going to die i just want to stay inside and protect myself i been able to sleep for two days because every time i cl ose my eyes i think this could be your last day on earth on the b rink of a panic attack how is school not a waste of time if you c an die tomorrow why the fuck does a job or money or a house e ven matter if you can die ten minutes from now if you can get di agnosed with als or cancer or some other horrible disease with n o cure how the fuck do you live like this how can anyone live w ith this knowledge

I'll start this off by saying I've heard many a horror story of people with BPD being treated fairly poorly in psychiatric hospitals. I've also heard that inpatient treatment rarely helps those with BPD due to the nature of our disorder. I think that's what ultimately scares me away from actually admitting myself. However, that doesn't stop me from fantasizing about it from time to time. I'm pretty successful or accomplished, given I have BPD. I have a bachelor's degree, and am pursuing a master's. I have a pretty good job, a car that I love, a stable relationship, the list goes on. But sometimes it just gets to be too much. It takes so much out of me to function at a "normal" or "acceptable" level each and every day. I feel like I'm constantly working overtime to NOT experience the symptoms of this disorder. It truly is exhausting. That being said, I often will find myself wanting to admit myself. I've never	**BPD : 99.91%** anxiety : 0.00% bipolar : 0.00% depression : 0.00% mentalillness : 0.08% poison : 0.00% schizophrenia : 0.01%	i'll start this off by saying i've heard many a horror story of peo ple with **bpd** being treated fairly poorly in psychiatric hospitals i've also heard that inpatient treatment rarely helps those with **b pd** due to the nature of our disorder i think that's what ultimatel y scares me away from actually admitting myself however that doesn't stop me from fantasizing about it from time to time i'm pretty successful or accomplished given i have **bpd** i have a bac helor's degree and am pursuing a master's i have a pretty good j ob a car that i love a stable relationship the list goes on but som etimes it just gets to be too much it takes so much out of me to f unction at a "normal" or "acceptable" level each and every day i feel like i'm constantly working overtime to not experience the symptoms of this disorder it truly is exhausting that being said i often will find myself wanting to admit myself i've never been before so i guess that sort of adds to the appeal nothing particula rly bad will happen necessarily yet i still find myself thinking a bout doing it i think i get hooked on the idea of "being taken car e of" or not really having to do anything i also have a major vict im complex when it comes to things like this so i become obses sed with the idea of "people feeling bad for me" ex look at how much i've suffered imagine how hurt mentally ill i have to be to

been before, so I guess that sort of adds to the appeal. Nothing particularly bad will happen necessarily, yet I still find myself thinking about doing it. I think I get hooked on the idea of "being taken care of" or not really having to do anything. I also have a major victim complex when it comes to things like this, so I become obsessed with the idea of "people feeling bad for me". Ex: Look at how much I've suffered. Imagine how hurt/mentally ill I have to be to be in here. Please take care of me! I'm always reminded of that one instance from Girl, Interrupted where the author talks about her time in a psychiatric hospital. I think she had BPD too and she was basically describing how she liked not having to deal with adult responsibilities or the outside world. That's sort of how I feel about it too.		be in here please take care of me i'm always reminded of that o ne instance from girl interrupted where the author talks about he r time in a psychiatric hospital i think she had bpd too and she w as basically describing how she liked not having to deal with ad ult responsibilities or the outside world that's sort of how i feel about it too
I was recently diagnosed with BPD. I also have autism and CPTSD, which I think has a		i was recently diagnosed with **bpd** i also have autism and cptsd which i think has a lot of overlap in my own case i had an abusi

lot of overlap in my own case. I had an abusive childhood and earlier adult years, plus school was an abusive hellhole from start to finish and I have not had many friends, especially in childhood. One thing I've definitely noticed as I've started researching BPD and talking about it with my therapist is the intense dichotomous thinking that I am prone to. I know it's very often affiliated with BPD and it has made me realise how much I can get trapped in a vicious circle of either being in a euphoric state of mind where I feel so happy and content, leading into me being in such a miserable mood where I feel like I hate everything in that moment. I've noticed I have this very intensely with the friends I have now, whom I genuinely adore and care a lot about, but realise I have an unhealthy attachment to (which I would argue is on my end, not any of theirs). I seem to end up	**BPD : 99.89%** anxiety : 0.00% bipolar : 0.00% depression : 0.01% mentalillness : 0.09% poison : 0.01% schizophrenia : 0.00%	ve childhood and earlier adult years plus school was an abusive hellhole from start to finish and i have not had many friends esp ecially in childhood one thing definitely noticed as started resea rching **bpd** and talking about it with my therapist is the intense dichotomous thinking that i am prone to i know very often affili ated with **bpd** and it has made me realise how much i can get tr apped in a vicious circle of either being in a euphoric state of mi nd where i feel so happy and content leading into me being in s uch a miserable mood where i feel like i hate everything in that moment noticed i have this very intensely with the friends i hav e now whom i genuinely adore and care a lot about but realise i have an unhealthy attachment to which i would argue is on my e nd not any of theirs i seem to end up jumping from my friend s and i are having a nice time together and i am happy we are frie nds to my towards me changed slightly this means i have done s omething wrong and they now hate me at any moment incredibl y exhausting and makes me feel awful for my friends if it helps to clarify i am aromantic and ace i have never had any desire or capacity at all for a romantic or sexual relationship and i see tha t ever changing i presume this is why i fixate a lot on my friends

jumping from "My friend(s) and I are having a nice time together and I am happy we are friends" to "My friend's demanour towards me changed slightly, this means I have done something wrong and they now hate me" at any moment. It's incredibly exhausting and makes me feel awful for my friends. (If it helps to clarify: I am aromantic and ace, I have never had any desire or capacity at all for a romantic or sexual relationship and I don't see that ever changing. I presume this is why I fixate a lot on my friendships, perhaps more so than those who do have or desire romantic relationships).		hips perhaps more so than those who do have or desire romantic relationships
I have visual hallucinations and delusions rar ely, I have other issues like I hear loud noises and smell awful things, but it’s always bettere d mostly with medication. I do struggle a lot with snow vision, but it’s like tinnitus, annoyi ng but not earth shattering. Anyways, the one thing that has been constant is a light blue bar		i have visual **hallucinations** and **delusions** rarely i have other is sues like i hear loud noises and smell awful things but it’s alway s mostly with medication i do struggle a lot with snow vision bu t it’s like tinnitus annoying but not earth shattering anyways the one thing that has been constant is a light blue bar that follows my line of sight it’s slightly it’s annoying and makes it hard to r

that follows my line of sight. It's translucent, slightly. It's annoying and makes it hard to read my books or do art or write sometimes. I am trying to read right now and it's just on my mind, I figured I would ask y'all.	BPD : 0.01% anxiety : 0.02% bipolar : 0.05% depression : 0.06% mentalillness : 2.32% poison : 0.33% **schizophrenia : 97.21%**	ead my books or do art or write sometimes i am trying to read right now and it's just on my mind i figured i would ask y'all
Fuck you, you smut I pray you die in your sleep tonight you dumb pale pasty flat ass nigger bitch	BPD : 0.13% anxiety : 0.19% bipolar : 0.03% depression : 2.50% mentalillness : 28.85% **poison : 45.79%** schizophrenia : 22.52%	fuck you you i pray you **die** in your sleep **tonight** you dumb pale flat ass nigger bitch

hi everyone i've been diagnosed with schizoaffective disorder and i'm having a hard time staying consistent with my medication i know it helps me feel more stable but at the same time i don't quite feel like myself when i'm on it it's like the symptoms are muted but so are parts of my personality and energy i was diagnosed with schizoaffective in but some doctors have told me that "there's no way i can be schizoaffective because i'm too high functioning and people that are diagnosed with schizophrenia are since then i have received a psychological evaluation thoroughly and was confirmed that i am schizoaffective bipolar type	BPD : 0.25% **anxiety : 50.38%** bipolar : 10.04% depression : 12.44% mentalillness : 12.62% poison : 5.22% schizophrenia : 9.04%	hi everyone i've been **diagnosed** with **schizoaffective** disorder and i'm having a hard time staying consistent with my medication i know it helps me feel more stable but at the same time i don't quite feel like myself when i'm on it it's like the symptoms are muted but so are parts of my personality and energy i was diagnosed with schizoaffective in but some doctors have told me that "there's no way i can be schizoaffective because i'm too high functioning and people that are diagnosed with schizophrenia are since then i have received a psychological evaluation thoroughly and was confirmed that i am schizoaffective bipolar type
I've been diagnosed with bipolar and mood swings in summer by a psychologist. To be exact, she referred me to a psychiatrist for complete diagnosis and further steps for		i've been diagnosed with bipolar and mood swings in summer by a psychologist to be exact she referred me to a psychiatrist for complete diagnosis and further steps for medication i thought that perhaps tiredness has caused me problems rather

medication. I thought that perhaps tiredness has caused me problems, rather than bipolar. Or perhaps I couldn't bring myself to accept it. Thus, we agreed that I rather work on my psyche than using medication. Fast forward, I was on a good mood at that time. I was doing pretty good till January when it hit me so bad that it left me completely exhausted and numb. I started failing every single thing I was doing and have been doing for years. I feel I can't continue anymore and I just want to disappear. From periods of not sleeping at all to sleeping most of the time."	BPD : 0.00% Anxiety : 0.00% **Bipolar : 99.91%** Depression : 0.03% Mental Illness : 0.03% Poison : 0.00% Schizophrenia : 0.01%	than bipolar or perhaps i couldn't bring myself to accept it thus we agreed that i rather work on my psyche than using medication fast forward i was on a good mood at that time i was doing pretty good till january when it hit me so bad that it left me completely exhausted and numb i started failing every single thing i was doing and have been doing for years i feel i can't continue anymore and i just want to disappear from periods of not sleeping at all to sleeping most of the time.

Tabel 4.13 menyajikan hasil analisis *explainability* menggunakan *Integrated Gradients.* Kata-kata yang ditampilkan dengan format tebal merupakan kata-kata dengan *attribution scores* tertinggi. Tabel ini menunjukkan dua kategori kata dengan *attribution* tinggi. Kategori pertama adalah kata *content* (non-stopwords) seperti "*psychiatric*", "*hospital*", "*revived*", "*forced*", "*therapy*", dan "*medicate*" yang menunjukkan *direct learned importance* dari model. Kata-kata ini secara eksplisit dipelajari selama training dan memiliki kontribusi langsung terhadap keputusan klasifikasi.

Kategori kedua adalah stopwords seperti "and", "was", "into", "for", dan "until". Model tidak pernah mempelajari stopwords karena dihapus saat *preprocessing*, sehingga *attribution* tinggi pada *stopwords* mencerminkan kontribusi kontekstual daripada *learned semantic importance.* Interpretasi yang tepat adalah dengan melihat frasa secara holistik, seperti "*forced into a psychiatric hospital*" yang mengindikasikan *involuntary hospitalization*, bukan menganalisis kata individual secara terpisah. Frasa "*died and was revived*" menunjukkan *suicide attempt* dan *resuscitation*, yang merupakan indikator eksplisit untuk kondisi *mental severe*.

Attribution scores untuk stopwords harus dipahami sebagai positional *importance* dan *structural contribution* dalam konstruksi frasa. Pendekatan *explainability* pada input space original ini mengikuti prinsip *user-centered XAI* (Lundberg & Lee, 2017; Ribeiro et al., 2016), dimana explanation harus dalam format yang dapat dipahami oleh praktisi klinis yang akan menginterpretasikan teks dalam bentuk originalnya.

4.8 Verifikasi Teks Psikolog

Berdasarkan hasil verifikasi, model mencapai akurasi 60% dengan enam dari sepuluh kasus berhasil diklasifikasikan dengan benar. Hasil ini menunjukkan masih terdapat ruang signifikan untuk peningkatan kemampuan model dalam mendeteksi gangguan mental dari data teks media sosial. Hal ini sejalan dengan temuan berbagai studi yang menyatakan bahwa teknologi deteksi kesehatan mental menggunakan analisis teks masih menghadapi tantangan dalam memberikan hasil yang konsisten dan dapat diandalkan di berbagai kondisi (Le Glaz et al., 2021; Teferra et al., 2024; Zhang et al., 2022b)). Tabel 4.13 dan Tabel 4.14 menyajikan ringkasan hasil verifikasi.

Tabel 4.14 Teks Hasil Verifikasi Psikolog

Teks	**Validasi Label**	**Prediksi Model**
I have no sociaI Iife, I am 21 and live a loneIy existence. I disIike Iooking at myself and often feeI like a faiIure. I have no famiIy or friends at all. I have never had a Girłfriend'l, and it feeIs Iike l'm mereIy existing around others without being noticed. My sociaI anxiety and autism contribute to this IoneIiness, making me feeI reall y aIone among peopIe. LoneIiness has taken over my Iife. For the past year, I've made efforts to change things by attending sociaI events Iike gatherings and bars, but I've had no success. l thought l could even try to find o nIine friends but usually ghosting happens though. Wh en I do PeopIe don't even bother to engage when I try t o get to know them. it's just me taiIking and trying. So Just My routine consists of going to coIIege and work t hen returning home to repeat the cycIe. l feeI as though I'm not Iiving just existing. It doesn't heIp that my fam iIy doesn't seem to want me around, and Iack reIatives t o spend time with.	**Isolasi social**	**Anxiety**
I've been officially diagnosed with depression for abou t a year and have been coping ever since. Most days ar e better than some, but some days are particularly jarri ng and I just want to curl up in a hole and die. I ackno wledge that I should get professional help, but for som e reason, I don't want to. I want to stay depressed.	**Depresi**	**Depression**
Life is very overwhelming right now. Bpd is making it super hard to cope. I want to talk to someone who und erstands and won't judge me. I am having a meltdown and I can't stop crying. I keep thinking of self harm.	**Poison**	**BPD**
How do you describe how anxiety makes your body fe el? I never really know how to describe it or the right words to use. My shoulders feel constantly heavy and weak. My body feels like it needs to shake like when y ou're feeling cold. It's so hard to describe I'd love to hear what others say.	**Kalau kecemasan belum bisa menjelaskan secara spesifik, lebih pada mental illness**	**Anxiety**

Berdasarkan Tabel 4.14 Hasil Perbandingan Validasi Model dan Prediksi Model Kasus pertama menunjukkan ketidaksesuaian antara penilaian psikolog dan model. Psikolog mengidentifikasi kondisi sebagai "Isolasi Sosial", sementara model memprediksi "Anxiety". Narasi menggambarkan individu berusia 21 tahun yang mengalami kesepian ekstrem tanpa adanya keluarga, teman, atau hubungan interpersonal. Model cenderung fokus pada frasa *"my social anxiety"* yang muncul eksplisit dalam teks. (Wolters et al., 2023) menjelaskan bahwa kesepian yang disebabkan oleh tidak adanya hubungan sosial (*social loneliness*) lebih terkait dengan isolasi sosial, sedangkan kesepian yang bersifat emosional (*emotional loneliness*) lebih berhubungan dengan kecemasan sosial dan depresi. Psikolog mampu mengidentifikasi bahwa kondisi utama adalah isolasi sosial kronis, bukan semata kecemasan.

Kasus kedua menunjukkan kesesuaian antara penilaian psikolog dan model dalam mengidentifikasi depresi. Narasi menggambarkan individu dengan diagnosis depresi selama satu tahun disertai fluktuasi mood yang signifikan. Model berhasil menangkap ciri-ciri bahasa depresi seperti penggunaan kata ganti orang pertama yang tinggi dan ungkapan "*curl up in a hole and die*" serta "*I want to stay depressed*". (Edwards & Holtzman, 2017) dalam meta-analisis mereka menunjukkan bahwa penggunaan kata ganti orang pertama tunggal yang lebih tinggi merupakan penanda linguistik yang reliabel untuk depresi. (Trifu et al., 2024) juga menemukan bahwa individu dengan gejala depresi lebih tinggi cenderung menggunakan lebih banyak kata ganti orang pertama tunggal, kata emosi negatif, dan kata kemarahan, yang merupakan karakteristik khas pola bahasa depresi.

Kasus ketiga menunjukkan ketidaksesuaian antara penilaian psikolog dan prediksi model. Psikolog memberi label "*Poison*", sedangkan model memprediksi "BPD" (*Borderline Personality Disorder*). Narasi menyebutkan secara eksplisit "*BPD is making it super hard to cope*" dengan gejala ledakan emosi, menangis terus-menerus, dan pikiran menyakiti diri sendiri, tanpa menyebutkan indikasi keracunan. Konten narasi ini menggambarkan kesulitan dalam mengelola emosi yang merupakan ciri inti BPD. (Carpenter & Trull, 2013b) menjelaskan bahwa disregulasi emosi pada BPD terdiri dari empat komponen: sensitivitas emosi, afek negatif yang meningkat dan tidak stabil, defisit dalam strategi regulasi yang tepat, dan surplus strategi regulasi maladaptif. (Dixon-Gordon et al., 2011) menambahkan bahwa sensitivitas tinggi terhadap rangsangan emosional, intensitas emosi yang tinggi, dan kesulitan mengontrol keadaan emosional merupakan karakteristik yang mendefinisikan BPD. Perbedaan antara label dan konten narasi ini perlu diperhatikan dalam proses kontrol kualitas pelabelan.

Kasus keempat menunjukkan perbedaan dalam spesifisitas diagnosis. Psikolog menggunakan kategori "*Mental Illness*" yang lebih luas dengan catatan "kecemasan belum bisa menjelaskan secara spesifik", sementara model memprediksi "*Anxiety*" secara spesifik. Narasi mendeskripsikan sensasi fisik seperti bahu yang terasa berat dan tubuh yang terasa perlu bergetar. (Zhang et al., 2022b) dalam review komprehensif mereka menunjukkan bahwa gejala fisik dan pola bahasa spesifik merupakan tanda khas dari gangguan kecemasan yang dapat dideteksi melalui NLP. Prediksi model yang lebih spesifik berpotensi lebih berguna secara klinis, meskipun kehati-hatian psikolog dalam menghindari diagnosis yang terlalu cepat juga merupakan pendekatan yang bijaksana.

Kasus kelima menunjukkan kesesuaian antara penilaian psikolog dan model dalam mengidentifikasi *bipolar disorder*. Narasi menggambarkan episode hipomanik/manik dengan gejala lengkap. Model berhasil menangkap perubahan pola bahasa yang mencerminkan gangguan mood seperti "*everything is moving so fast"* dan *"can't stop talking"*, serta perbedaan kondisi sebelum dan sesudah pengobatan. (Zhang et al., 2022b) menunjukkan bahwa pola bahasa seperti perubahan kecepatan bicara dan produktivitas verbal merupakan indikator yang dapat diandalkan untuk mendeteksi episode manik pada bipolar disorder.

Kasus keenam menunjukkan kesesuaian dalam mengidentifikasi skizofrenia. Narasi mencakup halusinasi dan delusi sejak masa kanak-kanak, seperti keyakinan bahwa orang tua adalah penyihir dan pengalaman diikuti oleh *Bloody Mary*. Model berhasil mengidentifikasi pola gejala psikotik dari narasi tersebut. (Tang et al., 2021) menunjukkan bahwa metode NLP mampu mendeteksi ciri-ciri skizofrenia dengan sensitivitas tinggi, termasuk berkurangnya keterkaitan makna antar kalimat dan gangguan dalam pola berpikir. (Jeong et al., 2023) menemukan bahwa fitur linguistik tertentu yang dihitung melalui NLP dapat mengukur secara objektif gangguan bicara dan keparahan gejala pada skizofrenia.

Kasus ketujuh juga menunjukkan kesesuaian dalam mengidentifikasi skizofrenia. Narasi menggambarkan halusinasi auditori dengan pola waktu tertentu yang berhenti pada malam hari. Meskipun terdapat perbedaan penulisan ("skizofrenia" vs "schizophrenia"), model tetap berhasil mengidentifikasi kategori yang tepat. (Bedi et al., 2015) mendemonstrasikan bahwa analisis bahasa otomatis dapat memprediksi onset psikosis dengan akurasi tinggi melalui deteksi disorganisasi semantik. Keberhasilan konsisten pada kedua kasus psikotik ini mengindikasikan kemampuan model yang kuat dalam mengenali ciri bahasa dari gejala psikotik.

Kasus kedelapan menunjukkan kesesuaian antara penilaian psikolog dan model. Psikolog dan model sama-sama menggunakan kategori "*Mental Illness*" yang lebih luas pada narasi yang lebih menggambarkan hambatan akses layanan daripada gejala spesifik. (Zhang et al., 2022b) menjelaskan bahwa ketidakpastian dalam menentukan diagnosis merupakan hal yang umum dalam deteksi gangguan mental berbasis teks, terutama ketika informasi terbatas. Kasus ini menunjukkan model telah mengembangkan kemampuan untuk tidak terlalu yakin dalam memberikan prediksi ketika data terbatas, yang merupakan karakteristik penting untuk sistem klinis yang bertanggung jawab.

Kasus kesembilan menunjukkan ketidaksesuaian antara penilaian psikolog dan model. Psikolog mengidentifikasi "Poison/Depresi" yang mencakup aspek keracunan akut dan kondisi depresi yang mendasari upaya bunuh diri dengan overdosis *paracetamol*. Sebaliknya, model hanya memprediksi "Poison" tanpa menangkap dimensi psikologis yang lebih dalam. (Edwards & Holtzman, 2017) menjelaskan bahwa kata ganti orang pertama dan kata-kata bermuatan emosi negatif muncul pada berbagai kondisi distres psikologis, termasuk depresi dan ideasi bunuh diri, sehingga diperlukan pemahaman konteks yang lebih mendalam. Psikolog mampu melihat hubungan antara tindakan akut dengan kondisi psikologis yang mendasarinya, sementara model masih menghadapi kesulitan dalam mengintegrasikan berbagai dimensi ini.

Kasus kesepuluh menunjukkan ketidaksesuaian pada narasi yang sangat singkat. Psikolog mengidentifikasi "BPD" (kemungkinan *Bipolar Disorder*) berdasarkan gejala kurang tidur ekstrem dan banyak bicara, sedangkan model mengklasifikasikan sebagai "*Mental Illness*" yang lebih umum. (Zhang et al., 2022b) menunjukkan bahwa pola bicara dapat membantu membedakan kondisi klinis pada bipolar disorder. Namun, narasi yang sangat singkat tampaknya tidak menyediakan cukup ciri bahasa yang diperlukan model untuk melakukan klasifikasi spesifik dengan tingkat kepercayaan yang memadai.

Hasil verifikasi menunjukkan bahwa performa model bervariasi berdasarkan jenis gangguan dan kelengkapan informasi. Model menunjukkan kemampuan konsisten pada gangguan dengan gejala yang tergambar jelas seperti depresi, bipolar, dan skizofrenia, namun menghadapi tantangan pada kasus dengan gejala yang saling tumpang tindih, kondisi ganda, atau informasi terbatas. Perbedaan antara penilaian klinis manusia dan prediksi model terutama terletak pada kemampuan memahami konteks emosional yang mendalam dan menghubungkan berbagai aspek diagnosis.

Tabel 4.15 Hasil Perbandingan Validasi Model dan Prediksi Model

Teks	Validasi Model	Prediksi Model
I think I’m just coming out of a really bad hypomanic/manic episode, and it was much longer and much more i ntense than I’m used to. In-patient care gave me new medicine, and it’s over, but I’m really scared of it coming back. With my bipolar, hypomania is only sometimes euphoric. Usually it’s a feeling of extreme distress, agita tion, and anxiety. Everything is moving so fast, I can’t stop talking about things that aren’t actually related to a ny conversation. I feel so detached from reality, and just look forward to distracting myself with a podcast and a video game at the same time. I’m super compulsive, perfectionist, but I also absolutely don’t care about anyth ing at all. I’m just in so much pain. When I meditate, I have to stop, because once I stop dissociating or distracti ng myself, there’s just a sense of dysphoria underneath. And there’s no connection between these feelings and my thoughts or circumstances. Anyway, I was just diagnosed two months ago with bipolar, and last week I had my first experience with in patient mental health care. They confirmed the bipolar diagnosis and gave me a new antipsychotic. And I’m glad. Now I feel so much more in touch with reality. Every thing has finally slowed do wn. I feel safe in my mind again. But I’m terrified that it’s going to come back. And now that the buzzing in m y brain is gone, there’s more space for some negative feelings I have to deal with. I’m super anxious, and I grie ve all the pain that wasn’t really being medicated before. (I much prefer this to the old feeling, though.)	**bipolar**	**bipolar**
I've been having hallucinations and delusions for my entire life. My first memory of this was when I was 5, and I was convinced my parents were witches. When I was 7-13 I was convinced I was being stalked by Bloody Ma ry. All through my childhood, when at bedtime, laying in my bed I'd hear footsteps. In my teenage years I starte d to see things like people hanging from trees, and dead babies in a bins and so on. For some reason it didn't rin g any bells or freak me out because I was just used to it. I had my first psychotic break at age 29 and was diagn osed at age 30.	**skizofrenia**	**schizophrenia**

Some nights I hear my voices all night, but often times I will go from hearing them all day to complete silence when night time comes around. Usually it's around 6p to 10p that they'll all of a sudden stop. It makes me feel l ike they're going to sleep, which makes them feel real. Just wondering if anyone else has experienced this.	**skizofrenia**	**Schizophrenia**
I'm going through a rough patch with my mental health and I'm trying to get help but things keep not working o ut. I sent a referral to mental health resource place and they got back to me on Wednesday with a call I couldn't take because I had been busy and the voicemail they left disappeared. I'm supposed to call my family doctors of fice but I keep forgetting to (I'm a student and the only time I can call the office is during my lunch break). I do n't know what to do now, there is a mental health nurse at my school who I talk to but I always lie to her to say I'm doing better than I actually am because I'm scared of what will happen if I tell the truth (I think she with ha ve to tell the counseler about what I say but last time the counseler was told anything about me he mocked me.) I know I've brought most of my problems on myself because of my poor planning skills and inability to focus e ven when it's something that is so important such as this. Also sorry if this post is weird, this is my first post on Reddit ever and sorry if I used the wrong tag.	**Mental Illness**	**Mental Illness**
I tried killing myself yesterday, I overdosed on 10000mg of paracetamol. I went to sleep and in 8 hours I would have been dying slowly over the next few days. I felt no remorse, no regret, nothing. I was at peace, ready to di e. But my parents found me and my organs were saved. I I laid on a hospital bed surrounded by darkness alone the whole night, it was the worst feeling I've ever felt. The pain gets worse and worse, the internal guilt I feel, it doesn't go away, every single day is a burden. I don't deserve love, I don't deserve my family.	**Poison/ Depresi**	**Poison**
I've only slept 8 hours sense Thursday and can't stop feeling hyper and wanting to talk to anyone and everyone. It's a good thing I'm broke right now so at least I'm not blowing through money I guess. Idk what to do about it though.	**BPD**	**Mental Illness**

Hasil perbandingan antara validasi psikolog dengan prediksi model berdasarkan tabel 4.15 menunjukkan ketidaksesuaian klasifikasi yang signifikan. Dari tujuh kasus yang dianalisis, terdapat lima kasus ketidaksesuaian dan hanya dua kasus yang menunjukkan kecocokan klasifikasi antara validasi psikolog dengan prediksi model. Ketidaksesuaian ini mengindikasikan adanya permasalahan mendasar dalam kemampuan model untuk memahami dan mengklasifikasikan gejala kesehatan mental dalam konteks bahasa Indonesia.

Kasus pertama terlihat pada teks dengan gejala depresi yang jelas, yaitu "beberapa bulan terakhir aku merasa tidak berharga, aku sudah tidak ingin masuk kuliah atau bermain dengan teman-teman lagi, kadang kadang terpikir lebih baik mati". Psikolog memvalidasi teks ini sebagai depresi berdasarkan indikator klinis yang komprehensif: rasa tidak berharga, kehilangan minat (anhedonia), penurunan energi, dan pikiran kematian. Namun, model mengklasifikasikannya sebagai poison. Kesalahan klasifikasi ini terjadi karena model menangkap frasa "lebih baik mati" secara literal sebagai konten yang berbahaya atau beracun, tanpa memahaminya sebagai ekspresi ideasi bunuh diri.

Kesalahan interpretasi literal semacam ini dapat dijelaskan melalui temuan (Delfani et al., 2024) yang menemukan bahwa terjemahan otomatis sering menghasilkan distorsi istilah medis dan psikologis, terutama dalam bahasa non-Barat, sehingga makna asli dapat bergeser dengan implikasi serius pada komunikasi kesehatan multibahasa. Frasa "lebih baik mati" kehilangan konteks emosional dan klinis setelah diterjemahkan, kemudian diinterpretasi secara literal oleh model sebagai konten berbahaya daripada sebagai gejala depresi. (Teferra et al., 2024) menegaskan bahwa model NLP menghadapi tantangan signifikan dalam memahami konteks emosional dan klinis dari ekspresi linguistik terkait depresi, terutama ketika model bergantung pada analisis kata kunci tanpa pemahaman kontekstual yang mendalam. Lebih lanjut, (Lawrence et al., 2024) menunjukkan bahwa LLMs yang dilatih terutama pada data berbahasa Inggris menunjukkan performa yang tidak konsisten ketika diterapkan pada populasi non-Barat, dengan kecenderungan melakukan *oversimplification* terhadap ekspresi emosional yang kompleks.

Kasus kedua adalah gangguan bipolar dengan teks yang menggambarkan perubahan suasana hati ekstrem: "kadang-kadang aku merasa bersemangat, aku banyak bicara dan energiku tidak ada habis-habisnya, terkadang aku juga sering berbelanja hingga menggunakan kartu kredit tak terbatas. namun adakalahnya aku hilang minat dan semangat, aku malas bertemu dengan orang lain dan menangis sepanjang hari dikamarku". Psikolog dengan tepat mengidentifikasi pola gejala mania (energi berlebih, banyak bicara, perilaku

berisiko tinggi) dan episode depresi (hilang minat, isolasi sosial, menangis) sebagai gangguan bipolar. Namun, model hanya mengklasifikasikannya sebagai mental illness secara umum. Perbedaan ini menunjukkan bahwa model belum mampu menangkap pola kombinasi gejala yang khas pada gangguan bipolar, dan cenderung melakukan generalisasi tanpa diferensiasi diagnostik yang spesifik.

Tantangan deteksi gangguan bipolar ini didukung oleh penelitian (Wu et al., 2024) yang menunjukkan bahwa gangguan bipolar, khususnya BP-I dan BP-II, sangat sering mengalami misdiagnosis sebagai *Major Depressive Disorder* karena adanya *overlap* gejala depresif yang signifikan. Studi yang melibatkan 1.463 pasien dari 8 pusat psikiatri ini menemukan bahwa karakteristik episode depresif dan komorbiditas psikiatrik menunjukkan pola yang kompleks dan sulit dibedakan tanpa pemahaman kontekstual yang mendalam tentang riwayat longitudinal pasien. (Oliva et al., 2025) memperkuat temuan ini dengan menegaskan bahwa overlap gejala dengan depresi unipolar sering menyebabkan keterlambatan diagnosis dan treatment yang tidak tepat. Penelitian terkini juga menunjukkan bahwa deteksi bipolar disorder melalui analisis teks sosial media menghadapi tantangan signifikan karena overlap pola linguistik yang terkait dengan depresi dan kecemasan (Srivastava et al., 2025). Model AI cenderung mengklasifikasikan ke kategori yang lebih umum ketika menghadapi pola gejala yang kompleks dan berfluktuasi.

Kasus ketiga yang menarik adalah teks dengan gejala *schizophrenia*: "mulai 6 bulan yang lalu aku sering mendengar suara-suara ditelingaku yang menyuruh untuk melompat kedalam jurang. terkadang aku melihat bayangan gelap yang selalu ada disekitarku. aku merasa terus diikutinya". Dalam kasus ini, baik psikolog maupun model sama-sama mengklasifikasikan sebagai *schizophrenia* dengan benar. Keberhasilan model dalam mendeteksi halusinasi auditori dan visual menunjukkan bahwa gejala psikotik memiliki manifestasi yang cukup eksplisit dan dapat dikenali oleh model.

Kasus keempat adalah teks yang menggambarkan respons traumatis: "setiap kali mendengar suara keras, saya merasa kembali ke peristiwa traumatis, saya menjadi sulit bernafas dan panik". Psikolog memberikan label umum mental illness, sementara model mengklasifikasikannya sebagai *anxiety*. Dalam kasus ini, model memberikan klasifikasi yang lebih spesifik dibandingkan validasi psikolog. Perbedaan ini menarik karena teks tersebut menyebutkan secara eksplisit adanya "kembali ke peristiwa traumatis" yang dapat mengindikasikan berbagai kemungkinan diagnosis. Model fokus pada kata kunci "panik" dan "sulit bernafas" dalam melakukan klasifikasi.

Kasus kelima adalah teks "aku hampir saja meminum baygon semalam karena aku sudah tidak bisa berpikir dengan tenang ketika mendengar pacarku berselingkuh". Psikolog mengklasifikasikan ini sebagai poison karena fokus pada tindakan impulsif untuk menyakiti diri dengan racun. Sebaliknya, model mengklasifikasikannya sebagai mental illness secara umum. Perbedaan ini menunjukkan bahwa model tidak menangkap spesifisitas risiko yang disebutkan dalam teks, dan justru melakukan generalisasi pada kategori yang lebih luas.

Kasus keenam adalah *borderline personality disorder* (BPD) dengan teks "saya takut sekali ditinggalkan, sampai kadang saya marah besar hanya karena pesan saya tidak segera dibalas". Psikolog memvalidasi ini sebagai BPD karena ekspresi ketakutan abandonment yang intens dan respons emosional yang tidak proporsional merupakan ciri khas gangguan kepribadian borderline. Namun, model mengklasifikasikannya sebagai anxiety karena lebih menekankan pada kata "takut" serta mengabaikan konteks relasional dan pola ketidakstabilan emosi yang lebih kompleks.

Kesalahan klasifikasi BPD ini dapat dijelaskan melalui overlap gejala yang signifikan antara personality disorders dengan *anxiety disorders*. Ketakutan dan kekhawatiran dapat muncul dalam kedua gangguan tersebut namun dengan konteks dan pola yang berbeda (Zimmerman & Chelminski, 2003). (Stinson et al., 2008) menegaskan bahwa BPD sering mengalami misdiagnosis karena gejala-gejalanya yang overlap dengan berbagai gangguan lain, termasuk *anxiety* dan *mood disorders*, yang memerlukan pemahaman pola interpersonal yang kompleks. Model AI cenderung kesulitan membedakan nuansa ini tanpa pemahaman konteks relasional yang lebih mendalam.

Kasus ketujuh adalah *anxiety disorder* dengan teks "setiap kali berbicara didepan orang banyak, jantung berdebar kencang, tangan berkeringat, dan pikiran saya penuh ketakutan". Dalam kasus ini, baik psikolog maupun model sama-sama mengklasifikasikan sebagai anxiety dengan benar. Gejala fisik dan kognitif dari kecemasan sosial yang digambarkan cukup eksplisit sehingga dapat dikenali oleh model.

Pola kesalahan yang muncul dari analisis di atas menunjukkan beberapa karakteristik keterbatasan model. Pertama, model cenderung melakukan interpretasi literal terhadap kata kunci tertentu tanpa memahami konteks klinis yang lebih luas, seperti terlihat pada kasus depresi yang salah diklasifikasikan sebagai poison. Kedua, model kesulitan dalam mengenali pola kombinasi gejala yang kompleks, seperti pada kasus bipolar yang hanya diklasifikasikan sebagai mental illness secara umum. Ketiga, model belum mampu membedakan nuansa antara gangguan yang memiliki gejala *overlapping*, seperti pada kasus BPD yang diklasifikasikan sebagai *anxiety*. Keempat, terdapat inkonsistensi dalam tingkat

spesifisitas klasifikasi, di mana model kadang terlalu umum (kasus *bipolar* dan *poison*) dan kadang lebih spesifik (kasus trauma).

Fenomena ketidaksesuaian klasifikasi ini dapat dijelaskan melalui dampak proses translasi dan perbedaan konteks linguistik serta kultural. Aspek krusial yang perlu dianalisis adalah bahwa model yang digunakan dalam penelitian ini dilatih menggunakan data berbahasa Inggris, sementara validasi dilakukan oleh psikolog terhadap teks asli berbahasa Indonesia. Proses translasi yang diperlukan untuk input model berpotensi menimbulkan pergeseran makna yang memengaruhi hasil klasifikasi secara substansial. Penelitian lintas budaya menunjukkan bahwa pasien Asia cenderung menggunakan ungkapan metaforis untuk menggambarkan gejala psikologis yang sulit diterjemahkan secara literal ke bahasa Inggris (Arthur, 2004; Kirmayer, 2001).

Tantangan translasi dalam konteks kesehatan mental semakin diperjelas oleh penelitian (Kandala et al., n.d.) yang mengusulkan *framework Cross-Linguistic Patient Distress Expression* (CL-PDE) untuk menangkap ekspresi distress yang tertanam secara kultural dan menyelaraskannya antar bahasa. Framework ini menekankan pentingnya *Explainable AI* dan validasi *human-in-the-loop* untuk memastikan bahwa sistem dapat memahami ekspresi pasien dalam konteks budaya mereka sendiri sambil tetap terhubung dengan pemahaman klinis standar. (Yao et al., 2025) dalam penelitiannya terhadap 5.000 partisipan menemukan bahwa algoritma machine translation sering mengabaikan nuansa kultural dan menghasilkan bias sistematis, terutama ketika menerjemahkan konten yang sensitive secara kultural seperti ekspresi emosional dan gejala psikologis.

Studi tentang ekuivalensi lintas bahasa menunjukkan bahwa istilah khas Indonesia seperti "putus asa" atau "minder" kerap direduksi menjadi padanan yang kurang mewakili makna emosionalnya dalam bahasa Inggris, seperti "*sadness*" atau "*worthlessness*" (van de Vijver & Leung, 1997). (Hambleton & Patsula, 1999) mengemukakan bahwa meskipun terjemahan telah benar secara tatabahasa, makna emosional dan klinis yang melekat dalam suatu budaya tetap dapat mengalami pergeseran. Penelitian tentang validitas lintas budaya konstruk psikologis menekankan pentingnya adaptasi budaya dalam penerjemahan teks psikologi, karena tanpa adaptasi yang tepat, hasil pengukuran dapat mengalami bias dan ketidakakuratan yang signifikan (Beaton et al., 2000; Guillemin et al., 1993).

Isu bias kultural dalam model AI untuk kesehatan mental ini mendapat perhatian serius dalam penelitian terkini. (Tao et al., 2024) menunjukkan bahwa *Large Language Models* menunjukkan cultural bias yang sistematis, di mana model-model ini cenderung menghasilkan respons yang lebih selaras dengan nilai-nilai budaya Barat, khususnya

Amerika Serikat, dibandingkan dengan budaya lain. Temuan ini diperkuat oleh *systematic review* (Cruz-Gonzalez et al., 2025) yang menganalisis 85 studi tentang aplikasi AI dalam kesehatan mental dan menemukan bahwa sebagian besar model AI dikembangkan dan divalidasi menggunakan dataset dari populasi Barat, sehingga menimbulkan pertanyaan serius tentang generalisabilitas dan fairness ketika diterapkan pada populasi dengan latar belakang budaya yang berbeda.

Perbedaan ekspresi gangguan mental lintas budaya juga relevan untuk menjelaskan ketidaksesuaian klasifikasi. (Teferra et al., 2024) menekankan bahwa perspektif lintas budaya dan multibahasa dalam deteksi depresi menggunakan NLP masih menghadapi tantangan besar, di mana sebagian besar model dilatih menggunakan dataset berbahasa Inggris dan kurang sensitif terhadap variasi linguistik dan ekspresi kultural dari populasi non-Barat. (De Vaus et al., 2017) dalam studi lintas budaya menunjukkan bahwa budaya Asia dan Barat memiliki perbedaan fundamental dalam mengekspresikan dan memahami gangguan afektif, di mana tingkat depresi klinis di Asia sekitar 4-10 kali lebih rendah dibanding Barat meskipun pada *self-report* measures orang Asia melaporkan affect negatif yang lebih tinggi.

(Hofmann et al., 2010) menegaskan pentingnya mempertimbangkan aspek kultural dalam social *anxiety disorder* karena prevalensi dan ekspresi gangguan ini sangat bergantung pada konteks budaya, dengan budaya Asia menunjukkan tingkat terendah sementara sampel Rusia dan Amerika Serikat menunjukkan tingkat tertinggi. Perbedaan ini menunjukkan bahwa model yang dilatih dengan data Barat mungkin tidak sensitif terhadap cara orang Indonesia mengekspresikan gejala kesehatan mental mereka.

Berdasarkan analisis tersebut, hasil penelitian ini menunjukkan bahwa perbedaan konteks linguistik dan kultural antara Indonesia dan dunia Barat berbahasa Inggris memiliki dampak signifikan terhadap akurasi klasifikasi. Model yang berbasis data bahasa Inggris cenderung menafsirkan teks Indonesia secara literal ketika diterjemahkan, sehingga kehilangan makna klinis yang seharusnya ditangkap. Oleh karena itu, penggunaan data dalam bahasa asli serta pengembangan model berbasis *contextual embedding* yang lebih sensitif terhadap variasi linguistik dan ekspresi kultural menjadi langkah penting untuk meningkatkan akurasi deteksi kesehatan mental di masa mendatang. Temuan ini menegaskan kompleksitas tantangan translasi dalam konteks kesehatan mental lintas budaya yang memerlukan pendekatan yang lebih sensitif terhadap nuansa linguistik dan kultural dalam pengembangan teknologi AI untuk aplikasi klinis

Tabel 4.16 Hasil Validasi Teks Bahasa Indonesia Menggunakan Prediksi Model CNN-BiLSTM

Teks	**Teks Bahasa Inggris**	**Kata Kunci**	**Validasi Label**	**Prediksi Model**
beberapa bulan terakhir aku merasa tidak berharga, aku sudah tidak ingin masuk kuliah atau bermain dengan teman-teman lagi, kadang kadang terpikir lebih baik mati	For the past few months, I have felt worthless. I no longer want to go to college or hang out with my friends. Sometimes I think it would be better to die.	kehilangan minat/ kesenangan, merasa tidak berharga, energi hilang, dan pikiran kematian,	Depresi	Poison
kadang-kadang aku merasa bersemangat, aku banyak bicara dan energiku tidak ada habis-habisnya, terkadang aku juga sering berbelanja hingga menggunakan kartu kredit tak terbatas. namun adakalahnya aku hilang minat dan semangat, aku malas bertemu dengan orang lain dan menangis sepanjang hari dikamarku	Sometimes I feel excited, I talk a lot and my energy is endless. Sometimes I also shop a lot, even using my credit card without limit. But there are times when I lose interest and enthusiasm, I feel lazy to meet other people and cry all day in my room.	mania, hipomania, harga diri meningkat, bersemnagta, pikiran melompat-lompat, terlibat aktivitas beresiko, depresi, suasana hati sedih, hilang minat	Bipolar	Mental Illness

Teks	Teks Bahasa Inggris	Kata Kunci	Validasi Label	Prediksi Model
mulai 6 bulan yang lalu aku sering mendengar suara-suara ditelingaku yang menyuruh untuk melompat kedalam jurang. terkadang aku melihat bayangan gelap yang selalu ada disekitarku. aku merasa terus diikutinya.	Starting six months ago, I often heard voices in my ears telling me to jump into the ravine. Sometimes I saw dark shadows that were always around me. I felt like I was being followed.	halusinasi, atau waham,	Schizophrenia	Schizophrenia
setiap kali mendengar suara keras, saya merasa kembali keperitiwa traumatis, saya menjadi sulit bernafas dan panik	Every time I hear a loud noise, I feel like I am reliving a traumatic event. I find it difficult to breathe and panic.	merasa tidak bahagia sepanjang hidup, takut, cemas, ada halusinasi, pikiran terus menganggu, perasaan sedih terus menerus	mentalillness	Anxiety
aku hampir saja meminum baigon semalam karena aku sudah tidak bisa berpikir dengan tenang ketika mendengar pacarku berselingkuh	I almost drank poison last night because I couldn't think straight when I heard my boyfriend was cheating on me.		poison	mentalillness

Teks	Teks Bahasa Inggris	Kata Kunci	Validasi Label	Prediksi Model
Saya takut sekali ditinggalkan, sampai kadang saya marah besar hanya karena pesan saya tidak segera dibalas	so afraid of being abandoned that sometimes i get really angry just because my messages replied to immediately	takut ditinggalkan ,hubungan tidak stabil, impulsif, perasaan kosong, mood berubah cepat, emosi tidak stabil	bpd	anxiety
Setiap kali berbicara didepan orang banyak, jantung berdebar,kencang, tangan berkeringat, dan pikiran saya penuh ketakutan	Every time I speak in front of a crowd, my heart pounds, my hands sweat, and my mind is filled with fear.	Rasa takut, kuatir berlebihan	Anxiety	Anxiety

BAB 5
Kesimpulan

5.1 Kesimpulan

Penelitian ini berhasil mengembangkan model klasifikasi berbasis arsitektur hibrid CNN-BiLSTM dengan pembobotan kata FastText untuk mendeteksi tujuh kategori gangguan kesehatan mental dari data teks media sosial Reddit, yaitu BPD, anxiety, depression, bipolar, mentalillness, schizophrenia, dan poison. Model dirancang untuk mengatasi keterbatasan penelitian sebelumnya yang umumnya hanya fokus pada satu atau dua kategori dan belum mengeksplorasi kombinasi arsitektur deep learning yang optimal untuk menangani kompleksitas bahasa informal media sosial. Dataset sebanyak 10.500 posts dikumpulkan dari lima subreddit utama dan Kaggle Mental Health Text Corpus, dengan proses pembersihan yang menghapus 15% data tidak valid dan validasi menggunakan kriteria DSM-5 untuk memastikan relevansi klinis.

Hasil evaluasi menunjukkan performa superior dari arsitektur hibrid CNN-BiLSTM yang mencapai akurasi 87% pada data testing, melampaui CNN standalone (84%) dan BiLSTM standalone (82%). Melalui 20-fold cross-validation, model mencapai F1-score rata-rata 0.867 pada macro average dengan konsistensi tinggi antara precision dan recall, menunjukkan bahwa model tidak bias terhadap kelas mayoritas. Performa terbaik dicapai pada kelas bipolar (recall 100%, F1-score 0.97) dan BPD (F1-score 0.98), sementara tantangan utama ditemukan pada kelas mentalillness (F1-score 0.61) yang memiliki batasan semantik kurang jelas. FastText terbukti efektif menangani out-of-vocabulary words yang umum dalam teks media sosial melalui pendekatan sub-kata. Validasi kualitatif dengan psikolog klinis pada 10 kasus berbahasa Inggris mencapai akurasi 60%, mengonfirmasi bahwa model berhasil mengklasifikasikan kondisi dengan manifestasi linguistik eksplisit seperti depression, bipolar, dan schizophrenia. Namun, analisis cross-lingual terhadap 7 kasus berbahasa Indonesia hanya menghasilkan 2 klasifikasi tepat, mengungkapkan tantangan signifikan dalam transfer learning lintas bahasa yang memerlukan adaptasi kultural dan linguistik lebih lanjut.

Penelitian ini memberikan kontribusi signifikan dalam beberapa aspek. Pertama, mengembangkan model pertama yang dapat memprediksi tujuh kategori gangguan kesehatan mental sekaligus, termasuk kategori "poison" yang krusial untuk pencegahan contagion effect. Kedua, mengintegrasikan arsitektur hibrid CNN-BiLSTM dengan FastText

embedding yang terbukti superior dalam menangani karakteristik unik teks media sosial. Ketiga, menyediakan validasi komprehensif melalui 20-fold cross-validation dan evaluasi klinis oleh psikolog profesional. Keempat, mengungkapkan keterbatasan fundamental dalam transfer learning lintas bahasa, menekankan pentingnya pengembangan model spesifik bahasa dan adaptasi kultural untuk implementasi global. Temuan ini menjadi landasan penting untuk pengembangan sistem deteksi dini gangguan mental berbasis media sosial yang dapat diintegrasikan dalam platform digital untuk moderasi konten dan intervensi preventif.

5.2 Saran

Berdasarkan hasil evaluasi dalam penelitian ini, terdapat beberapa saran yang dapat dipertimbangkan untuk pengembangan dan implementasi lebih lanjut dalam prediksi kesehatan mental menggunakan arsitektur CNN-BiLSTM. Pengembangan selanjutnya dapat dilakukan dengan mengeksplorasi arsitektur hybrid lainnya, seperti BiLSTM-BiGRU, atau memanfaatkan model berbasis transformer, baik dalam bentuk single transformer maupun dual transformer. Pendekatan ini juga dapat ditingkatkan melalui penggunaan model transformer yang telah di-*fine-tune* secara khusus untuk domain kesehatan mental.

Diharapkan bahwa pengembangan model-model tersebut mampu meningkatkan akurasi pada data pengujian, serta memungkinkan pemahaman kontekstual yang lebih baik terhadap ekspresi linguistik pada masing-masing kategori gangguan kesehatan mental. Selain itu, penelitian lanjutan juga dapat mencakup pengembangan antarmuka pengguna (*front-end interface*) yang memungkinkan implementasi langsung model prediksi ini sebagai aplikasi atau sistem berbasis web, sehingga dapat memberikan manfaat yang lebih luas dalam mendukung deteksi dini isu-isu kesehatan mental melalui teks.

Daftar Pustaka

Ameer, I., Arif, M., Sidorov, G., Gòmez-Adorno, H., & Gelbukh, A. (2022). *Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning*. http://arxiv.org/abs/2207.01012

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders: DSM-5. In *Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed.* American Psychiatric Publishing, Inc. https://doi.org/10.1176/appi.books.9780890425596

Arthur, K. (2004). Culture and Depression. *New England Journal of Medicine*, *351*(10), 951–953. https://doi.org/10.1056/NEJMp048078

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine*, *25*(24). https://journals.lww.com/spinejournal/fulltext/2000/12150/guidelines_for_the_process_of_cross_cultural.14.aspx

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Nature Publishing Group*, (June). https://doi.org/10.1038/npjschz.2015.30

Birnbaum, M. L., Rizvi, A. F., Correll, C. U., Kane, J. M., & Confino, J. (2017). Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early Intervention in Psychiatry*, *11*(4), 290–295. https://doi.org/10.1111/eip.12237

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. *5*, 135–146.

Calixto, I., Yaneva, V., & Cardoso, R. M. (2022). Natural Language Processing for Mental Disorders: An Overview. *Natural Language Processing in Healthcare*, (October), 37–59. https://doi.org/10.1201/9781003138013-3

Carpenter, R. W., & Trull, T. J. (2013a). Components of emotion dysregulation in borderline personality disorder: A review. In *Current Psychiatry Reports* (Vol. 15, Number 1). Current Medicine Group LLC 1. https://doi.org/10.1007/s11920-012-0335-2

Carpenter, R. W., & Trull, T. J. (2013b). Components of emotion dysregulation in borderline personality disorder: A review. In *Current Psychiatry Reports* (Vol. 15, Number 1). Current Medicine Group LLC 1. https://doi.org/10.1007/s11920-012-0335-2

Castillo-sánchez, G., Franco-martín, M., Marques, G., Dorronzoro, E., Torre-díez, I. De, & Franco-martín, M. (2020). Suicide Risk Assessment Using Machine Learning and Social Networks : a Scoping Review. *Journal of Medical Systems*.

Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., & De la Torre-Díez, I. (2020). Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. *Journal of Medical Systems*, *44*(12), 205. https://doi.org/10.1007/s10916-020-01669-5

Chandran, D., Robbins, D. A., Chang, C. K., Shetty, H., Sanyal, J., Downs, J., Fok, M., Ball, M., Jackson, R., Stewart, R., Cohen, H., Vermeulen, J. M., Schirmbeck, F., de Haan, L., & Hayes, R. (2019). Use of Natural Language Processing to identify Obsessive Compulsive Symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder. *Scientific Reports*, *9*(1), 1–7. https://doi.org/10.1038/s41598-019-49165-2

Chen, Y., & Fu, Z. (2023). Multi-Step Ahead Forecasting of the Energy Consumed by the Residential and Commercial Sectors in the United States Based on a Hybrid CNN-BiLSTM Model. *Sustainability*, *15*, 1895. https://doi.org/10.3390/su15031895

Choi, Y. H., Yang, K. I., Yun, C. H., Kim, W. J., Heo, K., & Chu, M. K. (2021). Impact of Insomnia Symptoms on the Clinical Presentation of Depressive Symptoms: A Cross-Sectional Population Study. *Frontiers in Neurology*, *12*. https://doi.org/10.3389/fneur.2021.716097

Chung, J., & Teo, J. (2022). Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. *Applied Computational Intelligence and Soft Computing*, *2022*. https://doi.org/10.1155/2022/9970363

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60. https://doi.org/10.3115/v1/W14-3207

Cruz-Gonzalez, P., He, A. W. J., Lam, E. P. P., Ng, I. M. C., Li, M. W., Hou, R., Chan, J. N. M., Sahni, Y., Vinas Guasch, N., Miller, T., Lau, B. W. M., & Sánchez Vidaña, D. I. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis,

monitoring, and intervention applications. In *Psychological Medicine* (Vol. 55). Cambridge University Press. https://doi.org/10.1017/S0033291724003295

De Choudhury, M. (2013). Role of Social Media in Tackling Challenges in Mental Health. *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia, SAM '13*, 49–52. https://doi.org/10.1145/2509916.2509921

De Vaus, June, Hornsey, Matthew J, Kuppens, Peter, & Bastian, Brock. (2017). Exploring the East-West Divide in Prevalence of Affective Disorder: A Case for Cultural Differences in Coping With Negative Emotion. *Personality and Social Psychology Review*, *22*(3), 285–304. https://doi.org/10.1177/1088868317736222

Delfani, J., Orăsan, C., Saadany, H., Temizöz, Ö., Taylor-Stilgoe, E., Kanojia, D., Braun, S., & Schouten, B. (2024). *Google Translate Error Analysis for Mental Healthcare Information: Evaluating Accuracy, Comprehensibility, and Implications for Multilingual Healthcare Communication*. https://www.skype.com/en/features/skype-translator/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-99167-6

Dixon-Gordon, K., Chapman, A., Lovasz, N., & Walters, K. (2011). Too Upset to Think: The Interplay of Borderline Personality Features, Negative Emotions, and Social Problem Solving in the Laboratory. *Personality Disorders: Theory, Research, and Treatment*, *2*, 243–260. https://doi.org/10.1037/a0021799

Dyson, H., & Gorvin, L. (2017). How Is a Label of Borderline Personality Disorder Constructed on Twitter: A Critical Discourse Analysis. *Issues in Mental Health Nursing*, *38*(10), 780–790. https://doi.org/10.1080/01612840.2017.1354105

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68. https://doi.org/https://doi.org/10.1016/j.jrp.2017.02.005

Ezerceli, Ö., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science*, *7*(3), 2277–2307. https://doi.org/10.1007/s42001-024-00307-1

Faurholt-Jepsen, M., Frost, M., Vinberg, M., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2014). Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry Research*, *217*(1), 124–127. https://doi.org/https://doi.org/10.1016/j.psychres.2014.03.009

Franco-Martín, M. A., Muñoz-Sánchez, J. L., Sainz-de-Abajo, B., Castillo-Sánchez, G., Hamrioui, S., & de la Torre-Díez, I. (2018). A Systematic Literature Review of Technologies for Suicidal Behavior Prevention. *Journal of Medical Systems*, *42*(4). https://doi.org/10.1007/s10916-018-0926-5

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*(6), 1240–1248. https://doi.org/10.1038/s41591-022-01811-5

Giuntini, F. T., Cazzolato, M. T., dos Reis, M. de J. D., Campbell, A. T., Traina, A. J. M., & Ueyama, J. (2020). A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 4713–4729. https://doi.org/10.1007/s12652-020-01726-4

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, *7*, 1–10. https://doi.org/10.1038/srep45141

Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of Medical Informatics*, *26*(1), 214–227. https://doi.org/10.15265/IY-2017-029

Gowen, L. K. (2013). Online Mental Health Information Seeking in Young Adults with Mental Health Challenges. *Journal of Technology in Human Services*, *31*(2), 97–111. https://doi.org/10.1080/15228835.2013.765533

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (n.d.). *Learning Word Vectors for 157 Languages*. Retrieved https://fasttext.cc/

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of*

Clinical Epidemiology, *46*(12), 1417–1432. https://doi.org/10.1016/0895-4356(93)90142-N

Gunderson, J. G., & Lyons-Ruth, K. (2008). BPD's interpersonal hypersensitivity phenotype: A gene-environment- developmental model. In *Journal of Personality Disorders* (Vol. 22, Number 1, pp. 22–41). https://doi.org/10.1521/pedi.2008.22.1.22

Ha, J., Kambe, M., & Pe, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Hambleton, R. K., & Patsula, L. N. (1999). *Increasing the Validity of Adapted Tests: Myths to be Avoided and Guidelines for Improving Test Adaptation Practices*. https://api.semanticscholar.org/CorpusID:146275641

Hasan, K., & Saquer, J. (2024). *A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit*. https://doi.org/10.1109/ICMLA61862.2024.00209

Hawton, K., Saunders, K. E. A., & O'Connor, R. C. (2012). Self-harm and suicide in adolescents. *The Lancet*, *379*(9834), 2373–2382. https://doi.org/https://doi.org/10.1016/S0140-6736(12)60322-5

Herdiansyah, H., Roestam, R., Kuhon, R., & Santoso, A. S. (2023). Their post tell the truth: Detecting social media users mental health issues with sentiment analysis. *Procedia Computer Science*, *216*(2022), 691–697. https://doi.org/10.1016/j.procs.2022.12.185

Hidalgo-Mazzei, D., Mateu, A., Reinares, M., Undurraga, J., Bonnín, C. del M., Sánchez-Moreno, J., Vieta, E., & Colom, F. (2015). Self-monitoring and psychoeducation in bipolar patients with a smart-phone application (SIMPLe) project: Design, development and studies protocols. *BMC Psychiatry*, *15*(1). https://doi.org/10.1186/s12888-015-0437-6

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofmann, S. G., Anu Asnaani, M. A., & Hinton, D. E. (2010). Cultural aspects in social anxiety and social anxiety disorder. In *Depression and Anxiety* (Vol. 27, Number 12, pp. 1117–1127). https://doi.org/10.1002/da.20759

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. In *Cognitive Therapy and Research* (Vol. 36, Number 5, pp. 427–440). Springer New York LLC. https://doi.org/10.1007/s10608-012-9476-1

Huang, C. (2017). Time Spent on Social Network Sites and Psychological Well-Being: A Meta-Analysis. *Cyberpsychology, Behavior, and Social Networking*, *20*(6), 346–354. https://doi.org/10.1089/cyber.2016.0758

Hunt, M. G., Marx, R., Lipson, C., & Young, J. (2018). No More FOMO: Limiting Social Media Decreases Loneliness and Depression. *Journal of Social and Clinical Psychology*, *37*(10), 751–768. https://doi.org/10.1521/jscp.2018.37.10.751

Islam, K. M. S., Fields, J., & Madiraju, P. (2025). MultiMentalRoBERTa: A Fine-Tuned Multiclass Classifier for Mental Health Disorder. *2025 IEEE International Conference on Big Data (BigData)*, 3255–3264. https://doi.org/10.1109/BigData66926.2025.11400786

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2019). Digital biomarkers of mood disorders and symptom change. *Npj Digital Medicine*, *2*(1). https://doi.org/10.1038/s41746-019-0078-0

Jain, P., Srinivas, K. R., & Vichare, A. (2022). Depression and Suicide Analysis Using Machine Learning and NLP. *Journal of Physics: Conference Series*, *2161*(1). https://doi.org/10.1088/1742-6596/2161/1/012034

Jakhotiya, A., Jain, H., Jain, B., Chaniyara, C., Student, B., Professor, A., & ----------, M. (2022). Text Pre-Processing Techniques in Natural Language Processing: A Review. *International Research Journal of Engineering and Technology*, 878–880.

Jeong, L., Lee, M., Eyre, B., Balagopalan, A., Rudzicz, F., & Gabilondo, C. (2023). Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia. *Psychiatric Research and Clinical Practice*, *5*(3), 84–92. https://doi.org/10.1176/appi.prcp.20230003

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021a). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Transactions on Computational Social Systems*, *8*(1), 214–226. https://doi.org/10.1109/TCSS.2020.3021467

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021b). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Transactions on Computational Social Systems*, *8*(1), 214–226. https://doi.org/10.1109/TCSS.2020.3021467

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for*

Computational Linguistics, EACL 2017 - Proceedings of Conference, *2*, 427–431. https://doi.org/10.18653/v1/e17-2068

Kadkhoda, E., Khorasani, M., Pourgholamali, F., Kahani, M., & Ardani, A. R. (2022). Bipolar disorder detection over social media. *Informatics in Medicine Unlocked*, *32*(August), 101042. https://doi.org/10.1016/j.imu.2022.101042

Kandala, A., Kandala, R., Moharir, A. K., Manchanda, N., & Singh, S. (n.d.). *Cross-Lingual Mental Health Ontologies for Indian Languages: Bridging Patient Expression and Clinical Understanding through Explainable AI and Human-in-the-Loop Validation*.

Kannan, S. (2015). *Preprocessing Techniques for Text Mining*. (March).

Kelly, Y., Zilanawala, A., Booker, C., & Sacker, A. (2018). Social Media Use and Adolescent Mental Health: Findings From the UK Millennium Cohort Study. *EClinicalMedicine*, *6*, 59–68. https://doi.org/10.1016/j.eclinm.2018.12.005

Kemp, S. (2023). *Digital 2023: Global Overview Report — DataReportal – Global Digital Insights*. https://datareportal.com/reports/digital-2023-global-overview-report

Kessler, R. C., Wai, T. C., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. In *Archives of General Psychiatry* (Vol. 62, Number 6, pp. 617–627). https://doi.org/10.1001/archpsyc.62.6.617

Khan, A., Husain, Mohd. S., & Khan, A. (2018). ANALYSIS OF MENTAL STATE OF USERS USING SOCIAL MEDIA TO PREDICT DEPRESSION! A SURVEY. *International Journal of Advanced Research in Computer Science*, *9*, 100–106.

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1), 1–6. https://doi.org/10.1038/s41598-020-68764-y

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. https://doi.org/10.48550/ARXIV.1408.5882

Kirmayer, L. J. (2001). Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. *The Journal of Clinical Psychiatry*, *62 Suppl 13*, 22–28; discussion 29-30. https://api.semanticscholar.org/CorpusID:14612667

Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. In *Multimedia Tools and Applications* (Vol. 81, Number 17). Multimedia Tools and Applications. https://doi.org/10.1007/s11042-022-12648-y

Kumar, S., & Singh, T. D. (2022). Fake news detection on Hindi news dataset. *Global Transitions Proceedings*, *3*(1), 289–297. https://doi.org/10.1016/j.gltp.2022.03.014

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Jones Bell, M. (2024). The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Ment Health*, *11*, e59479. https://doi.org/10.2196/59479

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Leichsenring, F., Leibing, E., Kruse, J., New, A. S., & Leweke, F. (2011). Borderline personality disorder. *The Lancet*, *377*(9759), 74–84. https://doi.org/10.1016/S0140-6736(10)61422-5

Lejeune, A., Robaglia, B.-M., Walter, M., Berrouiguet, S., & Lemey, C. (2022). Use of Social Media Data to Diagnose and Monitor Psychotic Disorders: Systematic Review. *J Med Internet Res*, *24*(9), e36986. https://doi.org/10.2196/36986

Levanti, D., Monastero, R. N., Zamani, M., Eichstaedt, J. C., Giorgi, S., Schwartz, H. A., & Meliker, J. R. (2023). Depression and Anxiety on Twitter During the COVID-19 Stay-At-Home Period in 7 Major U.S. Cities. *AJPM Focus*, *2*(1), 100062. https://doi.org/10.1016/j.focus.2022.100062

Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, *364*(9432), 453–461. https://doi.org/https://doi.org/10.1016/S0140-6736(04)16770-6

Lin, L. Y., Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Hoffman, B. L., Giles, L. M., & Primack, B. A. (2016). ASSOCIATION between SOCIAL MEDIA USE and DEPRESSION among U.S. YOUNG ADULTS. *Depression and Anxiety*, *33*(4), 323–331. https://doi.org/10.1002/da.22466

Löchner, J., Carlbring, P., Schuller, B., Torous, J., & Sander, L. B. (2025). Digital interventions in mental health: An overview and future perspectives. *Internet Interventions*, *40*, 100824. https://doi.org/https://doi.org/10.1016/j.invent.2025.100824

Lopresti, A. L., Hood, S. D., & Drummond, P. D. (2013). A review of lifestyle factors that contribute to important pathways associated with major depression: Diet, sleep and exercise. *Journal of Affective Disorders*, *148*(1), 12–27. https://doi.org/https://doi.org/10.1016/j.jad.2013.01.014

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874

Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, *87*(May), 207–211. https://doi.org/10.1016/j.chb.2018.05.035

Madububambachu, U., Ukpebor, A., & Ihezue, U. (2024). Machine Learning Techniques to Predict Mental Health Diagnoses: A Systematic Literature Review. *Clinical Practice & Epidemiology in Mental Health*, *20*(1). https://doi.org/10.2174/0117450179315688240607052117

Mahdy, N., Magdi, D. A., Dahroug, A., & Rizka, M. A. (2020). *Comparative Study: Different Techniques to Detect Depression Using Social Media*.

Marchant, A., Hawton, K., Stewart, A., Montgomery, P., Singaravelu, V., Lloyd, K., Purdy, N., Daine, K., & John, A. (2017). A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. In *PLoS ONE* (Vol. 12, Number 8). Public Library of Science. https://doi.org/10.1371/journal.pone.0181722

Marks, M. (2019). Artificial Intelligence-based Suicide Prediction. *Yale Journal of Health Policy, Law and Ethics*, (1003774), 4.

McCutcheon, R. A., Keefe, R. S. E., & McGuire, P. K. (2023). Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. In *Molecular Psychiatry* (Vol. 28, Number 5, pp. 1902–1918). Springer Nature. https://doi.org/10.1038/s41380-023-01949-9

Merinda Lestandy, & Abdurrahim. (2023). Effect of Word2Vec Weighting with CNN-BiLSTM Model on Emotion Classification. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *12*(1), 99–107. https://doi.org/10.23887/janapati.v12i1.58571

Munir, S., & Takov, V. (2022). *Generalized Anxiety Disorder*. https://www.ncbi.nlm.nih.gov/books/NBK441870/

Murarka, A., & Raleigh, I. B. M. (2021). Classification of mental illnesses on social media using RoBERTa. *Proceedings Ofthe 12th International Workshop on Health Text Mining and Information Analysis*, 59–68.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., & Bartels, S. J. (2016). The future of mental health care: Peer-To-peer support and social media. *Epidemiology and Psychiatric Sciences*, *25*(2), 113–122. https://doi.org/10.1017/S2045796015001067

Oliva, V., Fico, G., De Prisco, M., Gonda, X., Rosa, A. R., & Vieta, E. (2025). Bipolar disorders: an update on critical aspects. *The Lancet Regional Health – Europe*, *48*. https://doi.org/10.1016/j.lanepe.2024.101135

Ovcharuk, O., Mazurets, O., Molchanova, M., Kirpich, A., Skums, P., Sobko, O., Barmak, O., Krak, I., & Yakovlev, S. (2025). *Detecting Mental Disorders in Social Media Using a Transformer-Based Ensemble of Binary Classifiers*. https://doi.org/10.64898/2025.12.16.25342390

Owen, M. J., Sawa, A., & Mortensen, P. B. (2016). Schizophrenia. *The Lancet*, *388*(10039), 86–97. https://doi.org/10.1016/S0140-6736(15)01121-6

Palmer, C. A., Bower, J. L., Cho, K. W., Clementi, M. A., Lau, S., Oosterhoff, B., & Alfano, C. A. (2023). Sleep Loss and Emotion: A Systematic Review and Meta-Analysis of Over 50 Years of Experimental Research. *Psychological Bulletin*, *150*(4), 440–463. https://doi.org/10.1037/bul0000410

Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. yi, Rosen, D., Colditz, J. B., Radovic, A., & Miller, E. (2017). Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine*, *53*(1), 1–8. https://doi.org/10.1016/j.amepre.2017.01.010

Rahman, A. B. S., Ta, H.-T., Najjar, L., Azadmanesh, A., & Gönül, A. S. (2024). *DepressionEmo: A novel dataset for multilabel classification of depression emotions*. http://arxiv.org/abs/2401.04655

Reavley, N. J., & Jorm, A. F. (2011). The quality of mental disorder information websites: A review. *Patient Education and Counseling*, *85*(2), e16–e25. https://doi.org/https://doi.org/10.1016/j.pec.2010.10.015

Rehm, J., & Shield, K. D. (2019). Global Burden of Disease and the Impact of Mental and Addictive Disorders. *Current Psychiatry Reports*, *21*(2), 1–7. https://doi.org/10.1007/s11920-019-0997-0

Rhanoui, M., & Mikram, M. (2019). *A CNN-BiLSTM Model for Document-Level Sentiment Analysis*. 832–847. https://doi.org/10.3390/make1030048

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Proceedings of the Demonstrations Session, 97–101. https://doi.org/10.18653/v1/n16-3020

Roepke, A. M., Jaffee, S. R., Riffle, O. M., McGonigal, J., Broome, R., & Maxwell, B. (2015). Randomized Controlled Trial of SuperBetter, a Smartphone-Based/Internet-Based Self-Help Tool to Reduce Depressive Symptoms. *Games for Health Journal*, *4*(3), 235–246. https://doi.org/10.1089/g4h.2014.0046

Rokom. (2021). *Kemenkes Beberkan Masalah Permasalahan Kesehatan Jiwa di Indonesia*. https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20211007/1338675/kemenkes-beberkan-masalah-permasalahan-kesehatan-jiwa-di-indonesia/

Santos, I., Nedjah, N., & Mourelle, L. de M. (2017). Sentiment analysis using convolutional neural network with fastText embeddings. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–5. https://doi.org/10.1109/LA-CCI.2017.8285683

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Sedgwick, R., Epstein, S., Dutta, R., & Ougrin, D. (2019). Social media, internet use and suicide attempts in adolescents. In *Current Opinion in Psychiatry* (Vol. 32, Number 6, pp. 534–541). Lippincott Williams and Wilkins. https://doi.org/10.1097/YCO.0000000000000547

Singh, J., Wazid, M., Singh, D. P., & Pundir, S. (2022). An embedded LSTM based scheme for depression detection and analysis. *Procedia Computer Science*, *215*, 166–175. https://doi.org/10.1016/j.procs.2022.12.019

Srivastava, V., Boggavarapu, L., Shin, A., Datta, A., Lu, Y., & Bhaumik, R. (2025). *Towards Understanding Bipolar Disorder Through Social Media and Transformer Models: Challenges and Insights*. https://doi.org/10.1101/2025.03.05.25323466

Stinson, F., Dawson, D., Goldstein, R., Chou, S., Huang, B., Smith, S., Ruan, W., Pulay, A., Saha, T., Pickering, R., & Grant, B. (2008). Prevalence, Correlates, Disability, and Comorbidity of DSM-IV Narcissistic Personality Disorder: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *The Journal of Clinical Psychiatry*, *69*, 1033–1045. https://doi.org/10.4088/JCP.v69n0701

Ströhle, A., Gensichen, J., & Domschke, K. (2018). Diagnostik und Therapie von Angsterkrankungen. *Deutsches Arzteblatt International*, *115*(37), 611–620. https://doi.org/10.3238/arztebl.2018.0611

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. In *Translational Psychiatry* (Vol. 10, Number 1). Springer Nature. https://doi.org/10.1038/s41398-020-0780-3

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. http://arxiv.org/abs/1703.01365

Sutranggono, A. N., Sarno, R., & Ghozali, I. (2024). Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media. *Journal of Information and Communication Technology*, *23*(1), 77–104. https://doi.org/10.32890/jict2024.23.1.4

Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., Malaspina, D., Owen, M. J., Schultz, S., Tsuang, M., Van Os, J., & Carpenter, W. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research*, *150*(1), 3–10. https://doi.org/https://doi.org/10.1016/j.schres.2013.05.028

Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *Npj Schizophrenia*, *7*(1), 25. https://doi.org/10.1038/s41537-021-00154-3

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Teferra, B., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for Depression Using Natural Language Processing: Literature Review. *Interactive Journal of Medical Research*, *13*, e55067. https://doi.org/10.2196/55067

Tejaswini, V., Babu, K. S., & Sahoo, B. (2022). Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3569580

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., & Gravier, G. (2021a). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, *11*(2), 85–103. https://doi.org/10.1007/s41060-021-00242-8

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., & Gravier, G. (2021b). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, *11*(2), 85–103. https://doi.org/10.1007/s41060-021-00242-8

Trifu, R. N., Nemeș, B., Herta, D. C., Bodea-Hategan, C., Talaș, D. A., & Coman, H. (2024). Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure. *Frontiers in Psychology*, *Volume 15-2024*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1355734

Turner, B. J., Dixon-Gordon, K. L., Austin, S. B., Rodriguez, M. A., Zachary Rosenthal, M., & Chapman, A. L. (2015). Non-suicidal self-injury with and without borderline personality disorder: Differences in self-injury and diagnostic comorbidity. *Psychiatry Research*, *230*(1), 28–35. https://doi.org/https://doi.org/10.1016/j.psychres.2015.07.058

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. In *Methods and data analysis for cross-cultural research.* Sage Publications, Inc.

Vannucci, A., Flannery, K. M., & Ohannessian, C. M. (2017). Social media use and anxiety in emerging adults. *Journal of Affective Disorders*, *207*, 163–166. https://doi.org/https://doi.org/10.1016/j.jad.2016.08.040

Wolters, N. E., Mobach, L., Wuthrich, V. M., Vonk, P., Van der Heijde, C. M., Wiers, R. W., Rapee, R. M., & Klein, A. M. (2023). Emotional and social loneliness and their unique links with social isolation, depression and anxiety. *Journal of Affective Disorders*, *329*, 207–217. https://doi.org/https://doi.org/10.1016/j.jad.2023.02.096

Woods, H. C., & Scott, H. (2016). #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of Adolescence*, *51*, 41–49. https://doi.org/https://doi.org/10.1016/j.adolescence.2016.05.008

Wu, Z., Wang, J., Zhang, C., Peng, D., Mellor, D., Luo, Y., & Fang, Y. (2024). Clinical distinctions in symptomatology and psychiatric comorbidities between misdiagnosed bipolar I and bipolar II disorder versus major depressive disorder. *BMC Psychiatry*, *24*(1), 352. https://doi.org/10.1186/s12888-024-05810-3

Xiaoyan, L., Raga, R. C., & Xuemei, S. (2022). GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews. *Journal of Sensors*, *2022*. https://doi.org/10.1155/2022/7212366

Yao, J., Adi Kasuma, S., & Noori, H. (2025). Exploring Ethical Considerations in Machine Translation: Addressing Bias and Cultural Sensitivity. *Journal of Interdisciplinary Studies in Education*, *14*, 99–116. https://doi.org/10.32674/gs97fy88

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*. http://arxiv.org/abs/1702.01923

Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z., & Chung, T. S. (2022). A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/7893775

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022a). Natural language processing applied to mental illness detection: a narrative review. *Npj Digital Medicine*, *5*(1), 1–13. https://doi.org/10.1038/s41746-022-00589-7

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022b). Natural language processing applied to mental illness detection: a narrative review. *Npj Digital Medicine*, *5*(1), 1–13. https://doi.org/10.1038/s41746-022-00589-7

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (n.d.). *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification*.

Zimmerman, M., & Chelminski, I. (2003). Generalized Anxiety Disorder in Patients With Major Depression: Is DSM-IV's Hierarchy Correct? *The American Journal of Psychiatry*, *160*, 504–512. https://doi.org/10.1176/appi.ajp.160.3.504

LAMPIRAN