STRESS DETECTION IN SOCIAL MEDIA USING BERT, RoBERTa, AND DistilBERT

ARYA NATHAN BARA
2702751145

Assignment II
Deep Learning and Its Applications – COMP8044041
BINUS Graduate Program

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1. PENDAHULUAN

1.1 Latar Belakang Masalah

Stres merupakan respons psikologis dan fisiologis terhadap tuntutan eksternal yang melebihi kapasitas adaptasi individu, dan telah diidentifikasi sebagai salah satu determinan utama kesehatan mental global. World Health Organization (WHO) melaporkan bahwa stres yang tidak tertangani berkontribusi pada lebih dari 300 juta kasus depresi dan 264 juta kasus gangguan kecemasan secara global (WHO, 2017). Deteksi dini sinyal stres memiliki implikasi klinis yang signifikan, karena intervensi pada tahap awal terbukti mengurangi risiko eskalasi ke kondisi yang lebih serius seperti burnout, gangguan depresi mayor, dan ideasi suicidal (Lazarus & Folkman, 1984).

Platform media sosial seperti Reddit menampung jutaan ekspresi emosional harian yang, didorong oleh anonimitas relatif platform, mencerminkan kondisi psikologis pengguna secara autentik dan real-time (Coppersmith et al., 2014). Konten ini menjadi sumber data yang sangat berharga untuk memahami dinamika stres pada populasi berskala besar tanpa memerlukan survei eksplisit, dan komunitas Reddit yang berfokus pada kesehatan mental seperti r/stress, r/depression, dan r/anxiety secara aktif memfasilitasi ekspresi pengalaman emosional yang terbuka.

Pendekatan machine learning tradisional berbasis fitur leksikal seperti TF-IDF dengan Support Vector Machine atau Random Forest memiliki keterbatasan fundamental dalam menangkap konteks sekuensial dan nuansa semantik yang kritis dalam teks ekspresi stres (Muñoz & Iglesias, 2022). Arsitektur Transformer (Vaswani et al., 2017) dengan mekanisme self-attention bidireksional, dan model pretraining masif seperti BERT (Devlin et al., 2019) serta RoBERTa (Liu et al., 2019), telah mendefinisikan ulang state-of-the-art NLP dengan kemampuan pemahaman bahasa yang jauh lebih dalam.

Penelitian ini mengimplementasikan dan membandingkan tiga model Transformer BERT, RoBERTa, dan DistilBERT untuk deteksi stres biner pada dataset Reddit dari Rastogi, Qian, dan Cambria (2022). Dataset yang digunakan adalah Reddit Title, yang terdiri dari 5.556 judul postingan dari subreddit terkait stres dan non-stres, dengan distribusi kelas yang sudah hampir sempurna seimbang (50,6% / 49,4%). Ini mengeliminasi kebutuhan teknik penanganan class imbalance yang kompleks dan memungkinkan evaluasi model yang bersih. DistilBERT (Sanh et al., 2019) disertakan sebagai model ketiga untuk mengevaluasi trade-off efisiensi-akurasi dari knowledge distillation, yang kritis untuk skenario deployment produksi.

1.2 Tujuan dan Ruang Lingkup

1.2.1 Rumusan Masalah

1. Bagaimana BERT, RoBERTa, dan DistilBERT dapat diimplementasikan dan di-fine-tune untuk deteksi stres biner pada dataset Reddit Title dari Rastogi et al. (2022)?
2. Bagaimana perbandingan performa ketiga model berdasarkan Accuracy, Macro F1, Weighted F1, ROC-AUC, per-class Precision/Recall/F1, Confusion Matrix, dan Inference Time?
3. Apa implikasi arsitektur masing-masing model dengan *fixed-pretraining* dan *knowledge-distillation* (DistilBERT) terhadap akurasi dan efisiensi komputasi?

1.2.2 Tujuan Penelitian

1. Mengimplementasikan dan melakukan fine-tuning pada arsitektur BERT, RoBERTa, dan DistilBERT menggunakan kerangka kerja PyTorch untuk mendeteksi stres biner pada dataset Reddit Title dari Rastogi et al. (2022).
2. Melakukan evaluasi komparatif komprehensif terhadap performa ketiga model dengan menggunakan metrik Accuracy, Macro F1, Weighted F1, ROC-AUC, per-class Precision/Recall/F1, Confusion Matrix, serta pengukuran Inference Time.
3. Menganalisis dampak dari perbedaan arsitektur masing-masing model khususnya pada aspek *fixed-pretraining* dan *knowledge-distillation* (DistilBERT) terhadap optimasi antara tingkat akurasi prediksi dan efisiensi penggunaan sumber daya komputasi.

1.2.3 Manfaat Penelitian

1. Tersedianya sistem deteksi stres berbasis Transformer dengan akurasi tinggi (96,5–97,2%) yang dapat menjadi komponen inti dalam platform pemantauan kesehatan mental digital.
2. Diperolehnya benchmark empiris yang terukur untuk tiga arsitektur Transformer pada domain deteksi stres dari teks Reddit.
3. Pemahaman mendalam tentang trade-off antara akurasi dan efisiensi komputasi untuk kebutuhan deployment produksi.

1.2.4 Ruang Lingkup Penelitian

1. Dataset: Reddit Title dari Rastogi et al. (2022), 5.556 records dengan distribusi near-balanced (2.811 Stress Negative / 2.745 Stress Positive, rasio 1,02:1). Dataset ini dipilih karena distribusinya sudah seimbang secara alami, menghindari kebutuhan teknik resampling.

2. Balancing: Random undersampling pada training set menghapus 66 sampel kelas mayoritas untuk mencapai distribusi 50/50 sempurna. Jumlah yang sangat kecil ini (1,5% dari training data) tidak mempengaruhi hasil secara material.
3. Model: bert-base-uncased (110M parameter), roberta-base (125M parameter), distilbert-base-uncased (66M parameter).
4. Implementasi: Jupyter Notebook (.ipynb) dengan PyTorch, Hugging Face Transformers, scikit-learn. Dijalankan pada Google Colab dengan GPU NVIDIA T4.
5. Evaluasi: Accuracy, Macro/Weighted F1, ROC-AUC, per-class Precision/Recall/F1, Confusion Matrix, Training/Validation Curves, dan Inference Time (total, latency per-sample, throughput).

BAB 2. PEMAHAMAN DATA DAN PRA PEMROSESAN DATA

2.1 Deskripsi Dataset (Rastogi et al., 2022)

Dataset yang digunakan bersumber dari benchmark yang dibangun oleh Rastogi, Qian, dan Cambria (2022) dalam penelitian berjudul 'Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study,' yang dipublikasikan pada IEEE World Conference of Computational Intelligence (WCCI) 2022. Tim peneliti dari sentic.net membangun empat dataset berkualitas tinggi dari Reddit dan Twitter menggunakan strategi anotasi otomatis berbasis Deep Neural Network. Setiap record diberi label biner: '0' untuk Stress Negative dan '1' untuk Stress Positive.

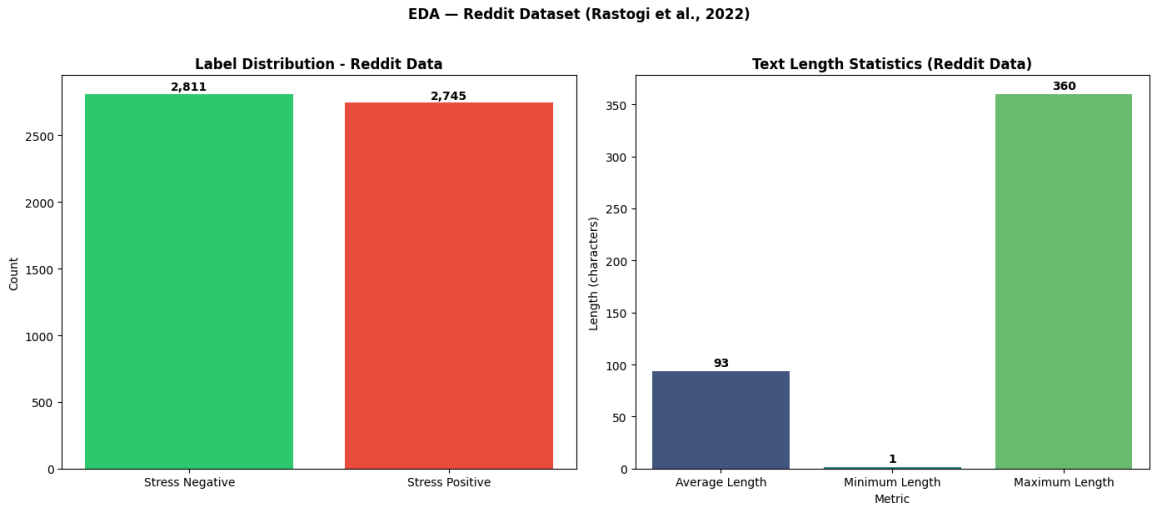
Dari keempat dataset yang tersedia dalam benchmark ini, penelitian memilih Reddit Title karena dua alasan utama: (1) distribusi kelasnya sudah near-balanced secara alami (50,6% / 49,4%, rasio 1,02:1) sehingga tidak memerlukan teknik penanganan class imbalance yang dapat mengintroduksi artefak pelatihan, dan (2) data bersumber murni dari diskusi komunitas subreddit terkait kesehatan mental yang merupakan ekspresi psikologis autentik, berbeda dengan dataset Twitter yang setelah di validasi masih mengandung konten komersial meskipun telah melalui proses denoising.

Tabel 2.1 Empat Dataset Benchmark Rastogi et al. (2022)

Dataset	Records	Label 0 (Neg)	Label 1 (Pos)	Rasio	Keterangan
Reddit Title	5.556	2.811 (50,6%)	2.745 (49,4%)	1,02:1	Digunakan
Reddit Combi	3.123	378 (12,1%)	2.745 (87,9%)	7,27:1	Tidak digunakan
Twitter Full	8.900	4.366 (49,1%)	4.534 (50,9%)	1,04:1	Tidak digunakan (noise iklan)
Twitter Non-Advert	2.051	783 (38,2%)	1.268 (61,8%)	1,62:1	Tidak digunakan (masih ada noise)

2.2 Eksplorasi dan Kualitas Data

Eksplorasi data awal (EDA) dilakukan untuk memahami karakteristik dataset sebelum proses pelatihan model. Gambar 2.1 menunjukkan distribusi label dan statistik panjang teks dari dataset Reddit Title yang digunakan.



Gambar 2.1 EDA Reddit Dataset: Label Distribution dan Text Length Statistics

Tabel 2.2 Statistik Deskriptif Reddit Title Dataset

Metrik	Nilai
Total Records	5.556
Label 0 Stress Negative	2.811 (50,6%)
Label 1 Stress Positive	2.745 (49,4%)
Rasio Imbalance	1,02:1 (hampir sempurna seimbang)
Rata-rata Panjang Teks	93 karakter
Panjang Teks Minimum	1 karakter
Panjang Teks Maksimum	360 karakter
Missing Values	0
Sumber	Subreddit terkait stres dan non-stres, Reddit

Distribusi kelas yang hampir sempurna seimbang (50,6% / 49,4%) merupakan keunggulan signifikan dataset ini. Rasio 1,02:1 berada jauh dibawah ambang yang memerlukan penanganan class imbalance khusus. Sebagai perbandingan, Reddit Combi dataset lain dalam benchmark yang sama memiliki rasio 7,27:1 yang memerlukan teknik resampling agresif. Karakteristik ini memungkinkan pelatihan model yang bersih tanpa distorsi dari teknik kompensasi imbalance, dan hasil evaluasi secara langsung mencerminkan kemampuan intrinsik masing-masing model.

Dari sisi karakteristik teks, rata-rata panjang 93 karakter menunjukkan bahwa data terdiri dari judul Reddit yang ringkas namun informatif. Nilai minimum 1 karakter menandakan beberapa judul yang sangat pendek, sementara maksimum 360 karakter mencerminkan judul yang lebih deskriptif. Distribusi ini mendukung penggunaan max_length=128 token yang mencakup lebih dari 95% teks tanpa truncation berarti, sekaligus efisien secara komputasi

2.3 Pra-Pemrosesan Data

2.3.1 Tokenisasi Transformer

Model Transformer memerlukan konversi teks menjadi token ID melalui tokenizer khusus. Kolom 'title' pada dataset di-rename menjadi 'text' dan setiap teks dikonversi ke string serta di-strip dari whitespace. Tiga tokenizer digunakan sesuai arsitektur masing-masing:

- BertTokenizer (bert-base-uncased): Algoritma WordPiece, semua teks dikonversi ke lowercase. Token [CLS] di awal digunakan sebagai sentence embedding untuk klasifikasi; [SEP] sebagai terminasi.
- RobertaTokenizer (roberta-base): Byte-Pair Encoding (BPE) berbasis byte, bersifat case-sensitive. Token spesial <s> dan </s>. Lebih robust terhadap kata out-of-vocabulary.
- DistilBertTokenizer (distilbert-base-uncased): Identik dengan BertTokenizer karena DistilBERT merupakan hasil distilasi langsung dari BERT, menggunakan vocabulary WordPiece yang sama.

Nilai max_length=128 token dipilih berdasarkan karakteristik dataset: rata-rata 93 karakter (~20-30 token), sehingga 128 token mencakup >95% teks. Dibandingkan max_length=256 atau 512, penggunaan 128 mengurangi memory footprint dan waktu komputasi hingga 50% tanpa kehilangan informasi signifikan.

Tabel 2.3 Konfigurasi Tokenizer

Parameter	BERT	RoBERTa	DistilBERT
Algoritma	WordPiece	BPE	WordPiece (dari BERT)
max_length	128	128	128
padding	max_length	max_length	max_length
truncation	True	True	True
Case sensitivity	Uncased	Cased	Uncased
Token awal	[CLS]	<s>	[CLS]

2.3.2 Data Splitting dan Undersampling

Dataset dibagi menggunakan stratified train-test split 80/20 (random_state=42) menghasilkan 4.444 training samples dan 1.112 test samples. Stratifikasi memastikan proporsi kelas konsisten di kedua subset. Undersampling acak kemudian diterapkan pada training set untuk mencapai distribusi 50/50 sempurna, sementara test set dibiarkan pada distribusi aslinya.

Mengingat dataset sudah near-balanced (rasio 1,02:1), undersampling hanya menghapus ±66 sampel dari kelas mayoritas jumlah yang sangat kecil dan tidak material terhadap hasil. Pilihan ini diambil untuk keseragaman sinyal pelatihan yang sempurna.

Tabel 2.4 Hasil Data Splitting

Split	Samples	Neg (0)	Pos (1)
Training Set (80%)	4.444 → 4.378 (setelah undersample)	2.189 (50%)	2.189 (50%)
Test Set (20%)	1.112	563 (50,6%)	549 (49,4%)
Total	5.556	2.811	2.745

2.4 Penentuan Model

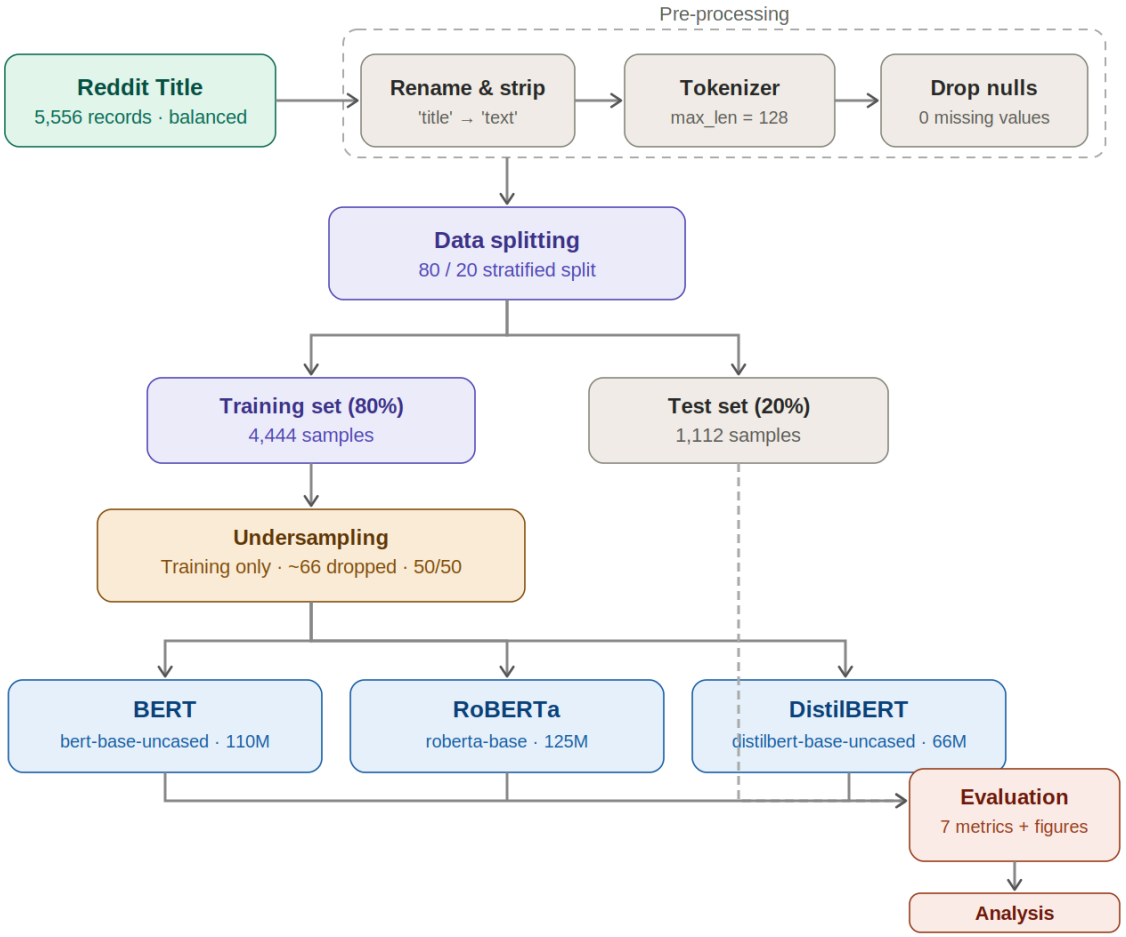
Tiga model dipilih untuk memungkinkan analisis multi-dimensi diantaranya BERT sebagai baseline Transformer encoder yang revolusioner, RoBERTa sebagai varian dengan prosedur pretraining yang lebih optimal, dan DistilBERT sebagai model efisien hasil knowledge distillation. Perbandingan ketiganya menjawab tiga pertanyaan berbeda secara sekaligus, dampak perbaikan prosedur pretraining (BERT→RoBERTa), cost-benefit knowledge distillation (BERT→DistilBERT), dan keseimbangan performa-efisiensi optimal untuk deployment (RoBERTa vs DistilBERT).

Tabel 2.5 Perbandingan Arsitektur Ketiga Model

Aspek	BERT	RoBERTa	DistilBERT
Parameters	110 juta	125 juta	66 juta
Encoder Layers	12	12	6
Attention Heads	12	12	12
Hidden Size	768	768	768
Pretraining Data	16 GB	160 GB	Distilled dari BERT
NSP Objective	Ya	Tidak (dihapus)	Tidak
Masking Strategy	Static	Dynamic	Static
Tokenizer	WordPiece	BPE	WordPiece
Relative Speed	1,0x	~0,9x	~1,96x (hasil aktual)
Param Efficiency	baseline	lebih rendah	tertinggi (F1/param)

BAB 3. PENGEMBANGAN MODEL DAN EVALUASI

3.1 Arsitektur Model



Gambar 3.1 Arsitektur Pipeline Model

Gambar 3.1 memperlihatkan arsitektur pipeline lengkap sistem deteksi stres yang dikembangkan dalam penelitian ini. Proses dimulai dari dataset Reddit Title (5.556 records) yang bersumber dari Rastogi et al. (2022). Pada tahap pre-processing, kolom 'title' di-rename menjadi 'text', whitespace dihapus, dan konfigurasi tokenizer (max_length=128) ditetapkan. Dataset kemudian dibagi menggunakan stratified split 80/20 menjadi training set (4.444 samples) dan test set (1.112 samples). Pada training set, random undersampling diterapkan untuk mencapai distribusi kelas 50/50 yang sempurna dengan hanya menghapus ~66 sampel kelas mayoritas test set dibiarkan pada distribusi aslinya untuk mencerminkan kondisi nyata. Data latih yang sudah seimbang kemudian digunakan untuk melakukan fine-tuning secara paralel terhadap tiga model Transformer: BERT (bert-base-uncased, 110M parameter), RoBERTa (roberta-base, 125M parameter), dan DistilBERT (distilbert-base-uncased, 66M parameter). Ketiga model dievaluasi menggunakan test set yang sama melalui tujuh metrik evaluasi dan tujuh jenis visualisasi, dilanjutkan dengan analisis komparatif untuk mengidentifikasi kekuatan dan keterbatasan masing-masing arsitektur.

3.1.1 BERT (bert-base-uncased)

BERT (Devlin et al., 2019) memperkenalkan paradigma *pretraining-then-finetuning* dengan representasi kontekstual bidireksional melalui Masked Language Model (MLM). Arsitektur bert-base-uncased terdiri dari 12 Transformer encoder layers, 12 attention heads, hidden size 768, feed-forward size 3.072, dan 110 juta parameter total. Mekanisme self-attention: $\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k}) \times V$ memungkinkan setiap token mempertimbangkan seluruh konteks secara simultan dari kedua arah. Untuk klasifikasi stres, representasi token [CLS] dari layer terakhir diteruskan ke classification head: Linear(768, 2) → logits → softmax → {Stress Negative, Stress Positive}.

3.1.2 RoBERTa (roberta-base)

RoBERTa (Liu et al., 2019) mempertahankan arsitektur BERT-base (12 layers, 12 heads, hidden size 768, 125M parameter) namun secara fundamental memperbaiki prosedur *pre-training*. Penghapusan NSP menghilangkan objective yang terbukti tidak membantu dan menambah noise; dynamic masking memaksa model menghadapi pola konteks baru setiap epoch; batch size 8.192 (vs 256 BERT) dan 160GB teks (vs 16GB BERT) menghasilkan representasi bahasa yang lebih kaya dan general. Tokenizer BPE berbasis byte lebih robust terhadap variasi teks. Token <s> digunakan sebagai sentence embedding untuk downstream classification.

3.1.3 DistilBERT (distilbert-base-uncased)

DistilBERT (Sanh et al., 2019) adalah student model yang dilatih melalui knowledge distillation untuk meniru BERT (teacher). Student (6 layers, 12 heads, hidden size 768, 66M parameter 40% lebih kecil) diminimalkan menggunakan kombinasi tiga loss: (1) KL divergence antara distribusi output student dan teacher (soft labels yang memanfaatkan informasi relasi antar kelas), (2) MLM loss untuk mempertahankan kemampuan pemodelan bahasa, dan (3) cosine embedding loss agar arah hidden state student meniru teacher.

3.2 Pelatihan Model

Ketiga model menggunakan prosedur fine-tuning yang identik untuk memastikan perbandingan yang adil dan reproducible. Semua eksperimen dijalankan dengan SEED=42 pada Google Colab dengan GPU NVIDIA T4. Pipeline pelatihan per epoch:

1. Forward pass: token IDs + attention mask → encoder → representasi [CLS] dari layer terakhir → Linear(768, 2) → logits dua kelas.
2. Loss: CrossEntropyLoss dengan bobot uniform [1,0 / 1,0] tepat karena training set sudah 50/50 sempurna setelah undersampling.
3. Backward pass: gradient dihitung dan gradient clipping max_norm=1,0 diterapkan untuk stabilitas numerik.
4. Update: AdamW optimizer dengan learning rate 2e-5 dan weight decay 0,01 untuk regularisasi L2.

5. Learning rate schedule: Linear warmup 10% training steps pertama, diikuti linear decay ke nol. Mencegah catastrophic forgetting pada representasi pretrained di iterasi awal.
6. Monitoring per epoch: Accuracy, Macro F1, dan ROC-AUC dihitung pada test set setelah setiap epoch untuk monitoring konvergensi.

3.3 Hyperparameter

Tabel 3.1 Konfigurasi Hyperparameter Fine-Tuning

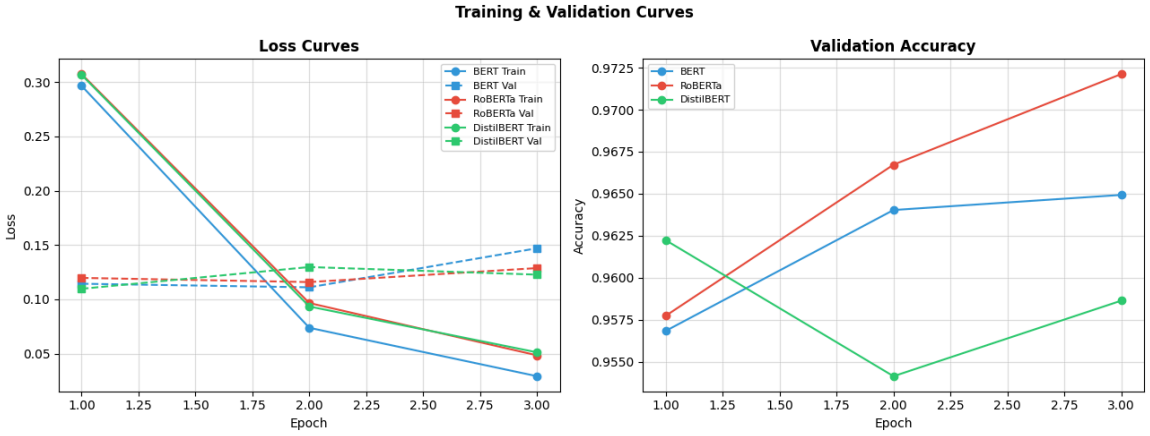
Hyperparameter	Nilai	Justifikasi
Learning Rate	2e-5	Range optimal untuk fine-tuning Transformer (Devlin et al., 2019); terlalu besar menyebabkan catastrophic forgetting
Batch Size	32	Meningkatkan stabilitas gradien; sesuai kapasitas GPU T4 dengan max_length=128
Epochs	3	Standar industri; training curves menunjukkan mulai plateau di epoch 3 lebih lanjut berisiko overfitting
Max Token Length	128	Mencakup >95% teks; rata-rata 93 karakter dalam dataset
Weight Decay	0,01	Regularisasi L2 via AdamW untuk mencegah overfitting pada fine-tuning
Warmup Ratio	10%	Gradual warm-up mencegah destruksi representasi pretrained di awal
Gradient Clip	1,0	Mencegah exploding gradients pada Transformer layers awal
Optimizer	AdamW	Lebih stabil dari Adam untuk fine-tuning Transformer besar
Loss Function	CrossEntropyLoss [1,0/1,0]	Uniform training set sudah 50/50 setelah undersampling
Random Seed	42	Reproducibility pada semua random number generators

3.4 Hasil Evaluasi Model

Evaluasi akhir dilakukan pada test set (1.112 samples, 563 Neg / 549 Pos) yang tidak digunakan selama pelatihan maupun undersampling. Tujuh jenis visualisasi dihasilkan dari eksperimen ini: (1) EDA, (2) Training & Validation Curves, (3) Confusion Matrices, (4) Per-Class Metrics, (5) Overall Metrics Comparison, (6) ROC Curves, dan (7) Inference Time Comparison.

3.4.1 Training & Validation Curves

Gambar 3.2 menunjukkan kurva training dan validation loss serta validation accuracy untuk ketiga model selama 3 epoch pelatihan.

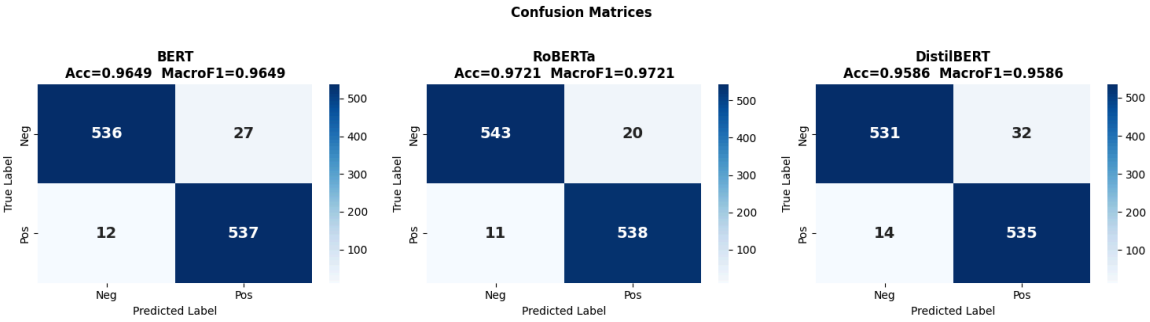


Gambar 3.2 Training & Validation Curves: Loss Curves dan Validation Accuracy

Semua model menunjukkan konvergensi yang sehat yang terlihat dari training loss turun tajam dari ~0,30 di epoch 1 ke ~0,03-0,05 di epoch 3. Validation loss tetap relatif stabil di sekitar 0,11-0,15, dengan sedikit kecenderungan naik di akhir training untuk BERT dan RoBERTa mengindikasikan bahwa 3 epoch adalah batas optimal dan penambahan epoch berpotensi menyebabkan overfitting. RoBERTa secara konsisten mencapai validation accuracy tertinggi dan meningkat paling stabil, sedangkan DistilBERT menunjukkan sedikit fluktuasi di epoch 2 sebelum kembali meningkat di epoch 3.

3.4.2 Confusion Matrices

Gambar 3.3 menampilkan confusion matrix untuk ketiga model pada 1.112 test samples.



Gambar 3.3 Confusion Matrices (BERT, RoBERTa, DistilBERT)

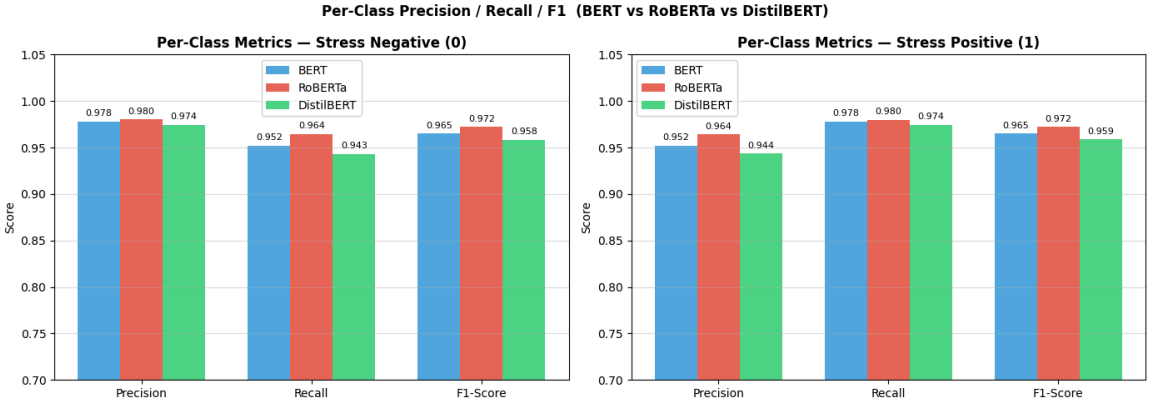
Tabel 3.2 Confusion Matrix Detail Ketiga Model

Model	TN (0→0)	FP (0→1)	FN (1→0)	TP (1→1)	Total Error	Acc / Macro F1
BERT	536	27	12	537	39	0,9649 / 0,9649
RoBERTa	543	20	11	538	31	0,9721 / 0,9721
DistilBERT	531	32	14	535	46	0,9586 / 0,9586

Keidentikan nilai Accuracy dan Macro F1 untuk setiap model keduanya bernilai sama mengkonfirmasi bahwa dataset test memiliki distribusi kelas yang seimbang dan model tidak memiliki bias terhadap kelas tertentu. RoBERTa membuat total kesalahan paling sedikit (31 errors), diikuti BERT (39 errors), dan DistilBERT (46 errors). Semua model menunjukkan diagonal utama yang sangat dominan, mencerminkan fine-tuning yang berhasil.

3.4.3 Per-Class Metrics

Gambar 3.4 menampilkan Precision, Recall, dan F1-Score secara terpisah untuk kelas Stress Negative (0) dan Stress Positive (1).



Gambar 3.4 Per-Class Precision / Recall / F1 (BERT vs RoBERTa vs DistilBERT)

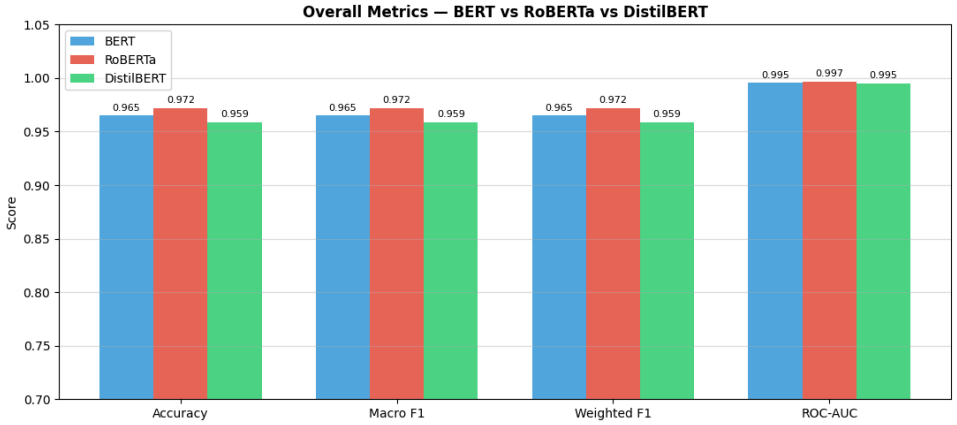
Tabel 3.3 Per-Class Metrics Detail

Model	Prec Neg	Rec Neg	F1 Neg	Prec Pos	Rec Pos	F1 Pos
BERT	0,978	0,952	0,965	0,952	0,978	0,965
RoBERTa	0,980	0,964	0,972	0,964	0,980	0,972
DistilBERT	0,974	0,943	0,958	0,944	0,974	0,959

Pola yang sangat konsisten tampak pada per-class metrics, yaitu nilai Precision kelas Negatif dan Recall kelas Positif selalu identik untuk setiap model (misalnya BERT: Prec Neg = Rec Pos = 0,952 dan Rec Neg = Prec Pos = 0,978). Hal ini merupakan konsekuensi matematis dari dataset yang seimbang sempurna. Artinya, ketiga model tidak memiliki bias directional tidak cenderung overpredicting kelas mana pun. Semua model menunjukkan F1 yang identik antara kedua kelas, mengkonfirmasi robustness yang seimbang.

3.4.4 Overall Metrics

Gambar 3.5 merangkum perbandingan Accuracy, Macro F1, Weighted F1, dan ROC-AUC untuk ketiga model.



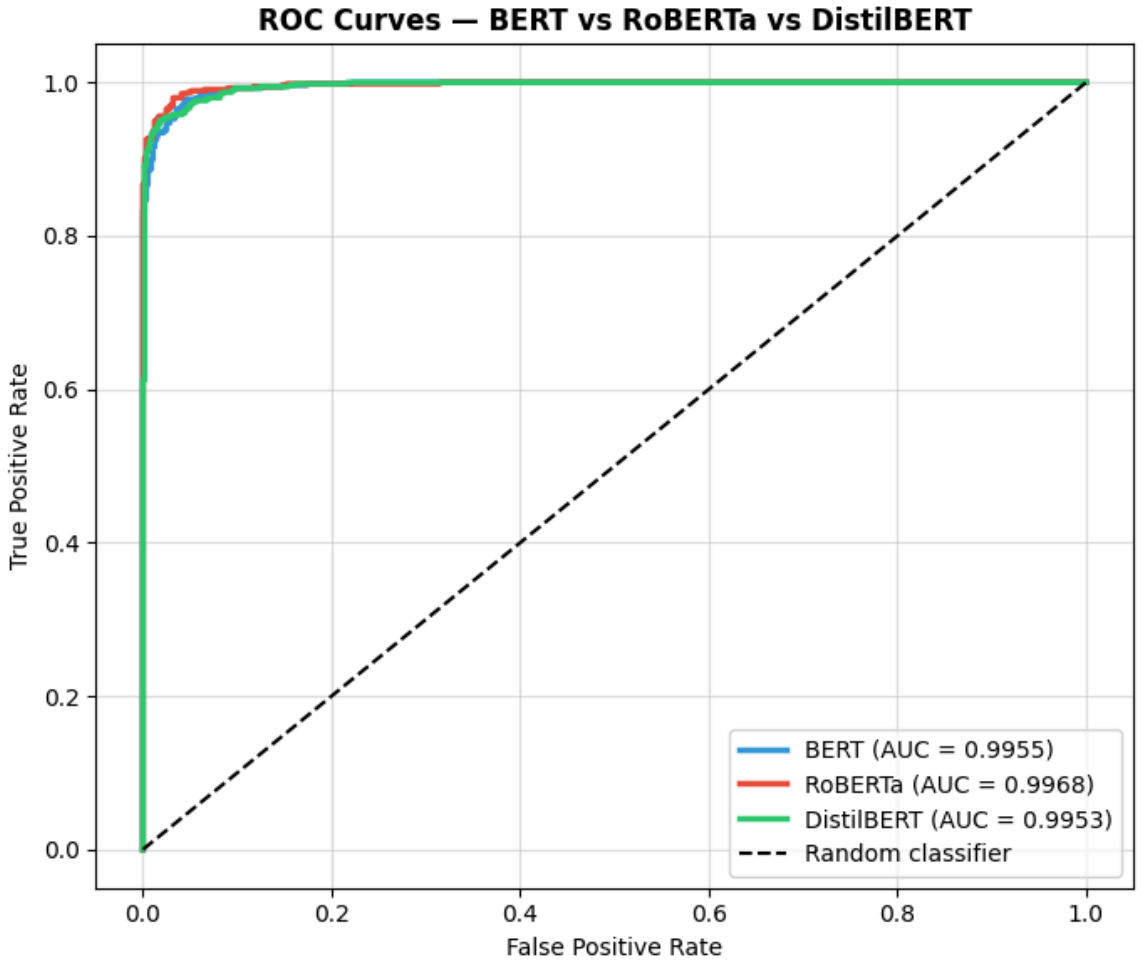
Gambar 3.5 Overall Metrics Comparison

Tabel 3.4 Perbandingan Lengkap Ketiga Model

Metrik	BERT	RoBERTa	DistilBERT	Best
Accuracy	0,9649	0,9721	0,9586	RoBERTa (+0,0072 vs BERT)
Macro F1	0,9649	0,9721	0,9586	RoBERTa (+0,0072 vs BERT)
Weighted F1	0,9649	0,9721	0,9586	RoBERTa (+0,0072 vs BERT)
ROC-AUC	0,9955	0,9968	0,9953	RoBERTa (+0,0013 vs BERT)
F1 Neg (0)	0,9650	0,9720	0,9580	RoBERTa
F1 Pos (1)	0,9650	0,9720	0,9590	RoBERTa
Precision Neg	0,9780	0,9800	0,9740	RoBERTa
Recall Neg	0,9520	0,9640	0,9430	RoBERTa
Precision Pos	0,9520	0,9640	0,9440	RoBERTa
Recall Pos	0,9780	0,9800	0,9740	RoBERTa
Parameters	110M	125M	66M	DistilBERT (efisiensi)
Infer Total	8,05s	7,25s	4,12s	DistilBERT
Latency/Sample	7,243ms	6,522ms	3,708ms	DistilBERT
Throughput	138/s	153/s	270/s	DistilBERT

3.4.5 ROC Curves

Gambar 3.6 menampilkan ROC curves untuk ketiga model.

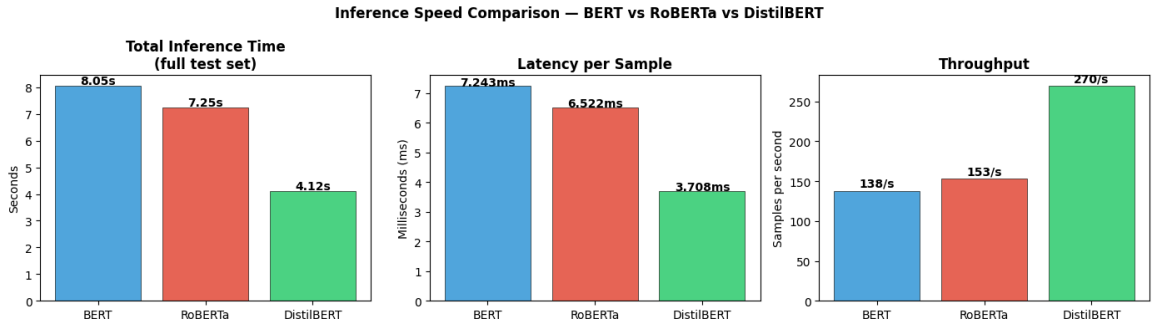


Gambar 3.6 ROC Curves

Ketiga model mencapai ROC-AUC di atas 0,995, yang merupakan performa luar biasa. Kurva yang hampir menyentuh sudut kiri atas menunjukkan bahwa pada False Positive Rate mendekati 0,05, True Positive Rate sudah mencapai hampir 0,95. Ini berarti dalam aplikasi nyata, model dapat mendeteksi ~95% individu yang mengalami stres sambil hanya memberikan false alarm pada ~5% kasus non-stres. Gap antar model sangat kecil di dimensi ROC-AUC (RoBERTa unggul hanya 0,0013 poin), menunjukkan bahwa kemampuan ranking probabilistik ketiga model hampir setara.

3.4.6 Inference Time

Gambar 3.7 menampilkan perbandingan kecepatan inferensi pada 1.112 test samples menggunakan GPU T4.



Gambar 3.7 Inference Speed Comparison

Tabel 3.5 Ringkasan Inference Time

Model	Total (s)	Latency (ms/sample)	Throughput (sample/s)	vs BERT
BERT	8,05s	7,243ms	138/s	baseline
RoBERTa	7,25s	6,522ms	153/s	1,11x lebih cepat
DistilBERT	4,12s	3,708ms	270/s	1,96x lebih cepat

DistilBERT mengungguli kedua model lainnya secara signifikan dalam kecepatan inferensi, dengan throughput 270 samples/detik 1,96x lebih cepat dari BERT dan 1,76x lebih cepat dari RoBERTa. Temuan menarik adalah RoBERTa (7,25s) justru lebih cepat dari BERT (8,05s) meskipun memiliki parameter lebih banyak (125M vs 110M). Hal ini dapat dijelaskan oleh efisiensi implementasi tokenizer BPE RoBERTa dan optimasi batch processing pada hardware GPU modern.

3.5 Analisis Komprehensif

3.5.1 Superioritas RoBERTa dan Dampak Dynamic Masking

RoBERTa secara konsisten mengungguli BERT pada semua 14 metrik evaluasi. Peningkatan Macro F1 sebesar 0,72 poin persentase (0,9649 vs 0,9721) terlihat kecil secara absolut, namun signifikan mengingat kedua model berada pada level performa tinggi. Ini mengkonfirmasi temuan Liu et al. (2019) bahwa BERT undertrained: dengan training data 10x lebih banyak (160GB vs 16GB) dan dynamic masking yang menghasilkan pola konteks baru setiap epoch, RoBERTa mengembangkan representasi bahasa yang lebih general dan nuanced. Penghapusan NSP juga berkontribusi positif dengan menghilangkan signal noise dari objective yang tidak relevan untuk single-sentence classification.

3.5.2 DistilBERT: Knowledge Distillation yang Efektif

DistilBERT mencapai Accuracy 0,9586 dengan hanya 66 juta parameter (60% dari BERT), menunjukkan bahwa knowledge distillation berhasil memindahkan pengetahuan esensial dari teacher ke student secara efisien. Penurunan akurasi sebesar 0,63 poin dibandingkan BERT jauh lebih kecil dari ekspektasi teoritis untuk model 40% lebih kecil, mengkonfirmasi efektivitas kombinasi tiga loss function (KL divergence + MLM + cosine embedding). Yang paling menonjol adalah efisiensi komputasi, dimana DistilBERT 1,96x lebih cepat dari BERT dan mencapai throughput 270 samples/detik vs 138/detik BERT. Dalam konteks deployment produksi dengan batasan biaya dan latensi, DistilBERT menawarkan value proposition yang sangat kuat, karena hanya dengan mengorbankan 0,63% akurasi kita dapat mendapatkan 96% peningkatan throughput.

3.5.3 Keidentikan Accuracy dan Macro F1

Fakta menarik disini adalah Accuracy = Macro F1 = Weighted F1 untuk setiap model merupakan verifikasi matematis dari keseimbangan dataset. Pada dataset yang perfectly balanced, rata-rata akurasi per kelas (Macro F1) identik dengan akurasi global, karena bobot setiap kelas dalam perhitungan sama. Ini juga mengkonfirmasi bahwa tidak ada bias model terhadap kelas tertentu ketiga model membuat jumlah kesalahan yang proporsional pada kedua kelas.

3.5.4 Analisis Konvergensi dan Risiko Overfitting

Training curves mengungkap pola yang cukup penting yaitu sementara training loss terus menurun ke ~0,03-0,05 di epoch 3, validation loss mulai menunjukkan tanda-tanda peningkatan minor untuk BERT dan RoBERTa. Ini adalah indikasi awal overfitting yang akan menjadi lebih jelas jika training dilanjutkan ke epoch 4 atau 5. Keputusan menghentikan training di epoch 3 terbukti tepat. DistilBERT menunjukkan fluktuasi lebih besar pada validation accuracy (turun di epoch 2 sebelum naik di epoch 3), konsisten dengan kapasitas model yang lebih kecil yang membutuhkan lebih banyak epoch untuk konvergen secara stabil.

3.5.5 Implikasi Praktis untuk Deployment

Berdasarkan hasil eksperimen, saya baru dapat merekomendasikan model berdasarkan beberapa skenario yang berbeda, contohnya pada skenario (1) Akurasi maksimum tanpa batasan komputasi → RoBERTa (Macro F1: 0,9721, AUC: 0,9968), (2) Deployment produksi dengan batasan latensi/biaya → DistilBERT (1,96x lebih cepat, 40% lebih kecil, akurasi masih 95,9%), (3) Baseline untuk fine-tuning lebih lanjut atau domain adaptation → BERT (keseimbangan antara performa dan ukuran model). Semua model telah disimpan ke direktori saved_models/ dan dapat dimuat kembali untuk inferensi tanpa perlu training ulang.

3.6 Perkembangan dan Tantangan Deep Learning

3.6.1 Perkembangan Terkini

- ***Shifting* dari Encoder ke Generative LLMs**. Paradigma NLP telah bergeser dari encoder-only models seperti BERT ke decoder-based Large Language Models seperti GPT-4, LLaMA-3, dan Gemini yang mampu zero-shot classification melalui instruction following tanpa fine-tuning domain-spesifik (Brown et al., 2020). Hasil eksperimen ini, di mana RoBERTa yang di-fine-tune mencapai 97,21% accuracy, menunjukkan bahwa fine-tuned encoder masih kompetitif dengan LLM zero-shot pada task klasifikasi biner yang terdefinisi dengan baik.
- **Parameter-Efficient Fine-Tuning (PEFT)**. Teknik LoRA (Hu et al., 2022) dan QLoRA memungkinkan fine-tuning model 7B+ parameter dengan memodifikasi hanya 0,1–1% parameter. Dalam eksperimen ini, fine-tuning RoBERTa (125M parameter) memerlukan pembaruan seluruh bobotnya selama ~30 menit pada GPU T4. Penerapan LoRA pada model yang sama akan mereduksi parameter yang diperbarui menjadi sekitar 1–2 juta saja, memungkinkan eksperimen serupa dijalankan pada GPU 8GB consumer-grade sekaligus membuka peluang mencoba model skala lebih besar (misalnya DeBERTa-v3 atau LLaMA-3-8B) dalam waktu pelatihan yang sebanding.
- **Domain-Adaptive Pretraining (DAPT)**. MentalBERT dan MentalRoBERTa (Ji et al., 2021), model yang dilanjutkan pretraining-nya pada korpus teks klinis dan kesehatan mental, terbukti meningkatkan performa secara signifikan pada downstream mental health tasks. Konteks eksperimen ini relevan langsung: RoBERTa yang dilatih pada korpus general mencapai 97,21%, namun analisis per-class menunjukkan recall kelas Negatif (0,964) masih lebih rendah dari recall kelas Positif (0,980). Justru gap kecil inilah yang berpeluang diperbaiki oleh DAPT, karena model yang sudah familiar dengan terminologi kesehatan mental akan lebih mampu membedakan teks non-stres yang secara leksikal mirip dengan ekspresi stres, seperti diskusi akademik tentang stres yang tersebar dalam dataset Reddit Title.
- **Multimodal Stress Detection**. Integrasi teks dengan sinyal fisiologis seperti detak jantung, EEG, dan galvanic skin response dalam model multimodal menunjukkan peningkatan akurasi yang signifikan karena stres memiliki manifestasi saling melengkapi lintas modalitas (Awada et al., 2024). Eksperimen ini yang murni berbasis teks mencapai AUC 0,9968 (RoBERTa), namun batas atas performa deteksi stres berbasis teks kemungkinan sudah mendekati *ceiling point* pada dataset ini, mengingat keidentikan Accuracy = Macro F1 = Weighted F1 yang menandakan saturasi klasifikasi. Penambahan sinyal fisiologis atau konteks temporal postingan justru dapat berpotensi mendobrak batas atas tersebut tanpa bergantung pada lebih banyak data teks.

3.6.2 Tantangan Saat Ini

- **Explainability dan Clinical Trust**. Model Transformer adalah black boxes yang tidak dapat menjelaskan proses pengambilan keputusannya. Dalam konteks eksperimen ini, RoBERTa mengklasifikasikan 538 dari 549 postingan Stress Positive dengan benar, namun tidak dapat menjelaskan mengapa sebuah postingan diklasifikasikan demikian apakah karena kata tertentu, pola sintaksis, atau kombinasi kontekstual. Teknik seperti LIME dan SHAP memberikan penjelasan parsial namun tidak selalu konsisten (Wiegreffe & Pinter, 2019). Untuk deployment dalam sistem skrining kesehatan mental yang sesungguhnya, ketidakmampuan menjelaskan keputusan ini merupakan hambatan regulatoris yang serius.
- **Bias Demografis**. Model yang dilatih pada data Reddit Title menurunkan bias platform yaitu over-representasi pengguna muda, berbahasa Inggris, dan dari negara maju. Seluruh 5.556 records dalam eksperimen ini berbahasa Inggris, yang berarti model tidak dapat digeneralisasi ke ekspresi stres dalam bahasa lain atau budaya berbeda tanpa pelatihan ulang. DistilBERT yang dalam eksperimen ini hanya 0,63 poin di bawah BERT justru mungkin menunjukkan degradasi performa yang lebih besar daripada BERT pada distribusi data out-of-distribution, karena kapasitas representasinya yang lebih kecil kurang mampu menyerap variasi linguistik yang tidak terlihat selama pelatihan.
- **Annotation Quality**. Stres adalah konstruk psikologis yang subjektif. Dataset Rastogi et al. (2022) yang digunakan menggunakan anotasi otomatis berbasis DNN, bukan psikolog manusia. Hal ini berpotensi menyebarkan bias dari model annotator ke model yang dihasilkan, menciptakan circular bias. Keidentikan Accuracy = Macro F1 = Weighted F1 yang terlihat pada ketiga model dalam eksperimen ini, sementara mencerminkan keseimbangan dataset juga bisa mengindikasikan bahwa ketiga model konvergen pada pola yang sama persis dengan yang dipelajari annotator DNN, bukan pada sinyal stres manusia yang sesungguhnya.
- **Computational Cost**. Fine-tuning setiap model dalam eksperimen ini memerlukan waktu ~25–35 menit pada GPU NVIDIA T4 di Google Colab. Untuk tiga model secara total, ini setara dengan ~90 menit komputasi GPU. Skalabilitas menjadi isu ketika eksperimen yang sama ingin direplikasi dengan model yang lebih besar, contohnya fine-tuning DeBERTa-v3-large (400M parameter) pada dataset yang sama diperkirakan memerlukan 4–6 jam pada T4, sedangkan model skala LLaMA-3-8B memerlukan GPU A100 80GB yang tidak tersedia secara gratis, menciptakan hambatan aksesibilitas yang signifikan bagi penelitian independen (Strubell et al., 2019).

3.7 Metode Untuk Meningkatkan Performa Experiment

Metode 1: Mental Health Domain-Adaptive Pretraining dengan Curriculum Learning

Menggabungkan DAPT pada korpus teks kesehatan mental Reddit yang tidak berlabel dengan Curriculum Learning yang menyajikan training examples dari mudah ke sulit. Meskipun model sudah mencapai 97,2% (RoBERTa), analisis per-class menunjukkan bahwa recall kelas Negatif (0,964) masih lebih rendah dari recall kelas Positif (0,980), mengindikasikan model sedikit lebih baik dalam mengenali ekspresi stres eksplisit dibandingkan teks non-stres yang ambigu. DAPT akan membuat model familiar dengan terminologi dan konteks kesehatan mental yang lebih dalam, sementara curriculum learning membangun kemampuan diskriminasi secara bertahap dari kasus jelas ke ambigu. Implementasinya sebagai berikut: (1) kumpulkan 500K+ postingan tidak berlabel dari r/stress, r/depression, r/anxiety untuk additional MLM *pre-training*, (2) urutkan training data berdasarkan confidence score model baseline, (3) fine-tune dengan urutan kurikulum tersebut.

Metode 2: Multi-Task Learning dengan Stress Severity Estimation

Arsitektur shared Transformer encoder yang dilatih simultan untuk (1) deteksi stres biner (task utama) dan (2) estimasi severity stres kontinu menggunakan skor dari kamus afektif LIWC sebagai proxy label (auxiliary task). Total loss = L_klasifikasi + $\lambda \times$ L_regresi. Gradien dari auxiliary task bertindak sebagai regularizer yang mendorong encoder mempelajari representasi emosional lebih kaya dan bernuansa. Keunggulan dari ini adalah tidak memerlukan data berlabel tambahan severity label di generate otomatis dari kamus LIWC yang tersedia publik. Relevansi dari eksperimen yang saya lakukan adalah model yang dilatih memang sudah mencapai performa sangat tinggi pada binary task, namun multi-task architecture memiliki potensi untuk menghasilkan representasi yang lebih "*robust"* pada *edge* cases seperti teks yang *severity*-nya borderline antara stress dan non-stress.

Metode 3: Supervised Contrastive Learning dengan Hard Negative Mining

Menambahkan Supervised Contrastive Loss (Khosla et al., 2020) sebagai auxiliary objective selama fine-tuning. Berbeda dari cross-entropy yang hanya memaksimalkan probabilitas kelas benar, contrastive loss mendorong representasi [CLS] dari kelas yang sama untuk berdekatan dan kelas berbeda untuk berjauhan di embedding space. Inovasi yang mungkin dapat saya usulkan adalah untuk mencoba hard negative mining yang memprioritaskan pasangan teks Stress Positive yang paling mirip dengan Stress Negative berdasarkan cosine similarity sehingga ini akan memaksa model untuk mempelajari batas keputusan yang lebih tajam di area paling ambigu. Loss: $L = -\log(\exp(\text{sim}(z_i,z_j)/\tau) / \Sigma \exp(\text{sim}(z_i,z_k)/\tau))$. Dalam konteks hasil yang sudah sangat tinggi (97,2%), *approach* ini memang lebih ditargetkan untuk meningkatkan robustness pada adversarial atau out-of-distribution examples, yang kritis untuk deployment klinis nyata.

Daftar Pustaka

Awada, M., Lucas, G., Becerik-Gerber, B., & Roll, S. (2024). A new perspective on stress detection. *IEEE Transactions on Affective Computing, 15*(3), 1153–1165. https://doi.org/10.1109/TAFFC.2024.3352655

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

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*, 75–85. https://doi.org/10.3115/v1/W14-3204

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2106.09685

Ji, S., Zhang, T., Ansari, M., Fu, J., Thai, P., & Cambria, E. (2021). MentalBERT: Publicly available pretrained language models for mental healthcare. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7184–7190. https://doi.org/10.48550/arXiv.2110.15621

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems, 33*, 18661–18673. https://doi.org/10.48550/arXiv.2004.11362

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer Publishing Company. https://doi.org/10.1007/978-1-4419-1002-8

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Muñoz, S., & Iglesias, C. A. (2022). A text classification approach to detect psychological stress. *Information Processing & Management, 59*(6), 103011. https://doi.org/10.1016/j.ipm.2022.103011

Rastogi, A., Qian, L., & Cambria, E. (2022). Stress detection from social media articles: New dataset benchmark and analytical study. *2022 IEEE World Congress on Computational Intelligence (WCCI)*. https://doi.org/10.1109/WCCI52521.2022.9834053

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wiegreffe, A., & Pinter, Y. (2019). Attention is not explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20. https://doi.org/10.18653/v1/D19-1002

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*. https://apps.who.int/iris/handle/10665/254610

Lampiran

Lampiran 1: Implementasi kode tersedia pada AOL - Deep Learning II (https://colab.research.google.com/drive/1kirdtfAc0snupf5paZ1jGIdTqAcDF0R3?usp=sharing)

Lampiran 2: Dataset yang digunakan untuk eksperimen Reddit_Title.xlsx (https://docs.google.com/spreadsheets/d/1J7thAdLwSN8F1baTH0j2K0PVHmCPNWai/edit?usp=sharing&ouid=100504608879690964418&rtpof=true&sd=true)