

EKSPLORASI METODE DETEKSI EMOSI PADA TEKS DARI TWITTER, KOMENTAR BLOG, DAN DIALOG FILM



Disusun Oleh:

N a m a : Hilqudz Dzikri
NIM : 21523017

**PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
2025**

HALAMAN PENGESAHAN DOSEN PEMBIMBING

**EKSPLORASI METODE DETEKSI EMOSI PADA TEKS DARI
TWITTER, KOMENTAR BLOG, DAN DIALOG FILM**

TUGAS AKHIR



Yogyakarta, 16 Juli 2025

Pembimbing,


(Arrie Kurniawardhani S.Si., M.Kom.)

HALAMAN PENGESAHAN DOSEN PENGUJI

EKSPLORASI METODE DETEKSI EMOSI PADA TEKS DARI TWITTER, KOMENTAR BLOG, DAN DIALOG FILM

TUGAS AKHIR

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer dari Program Studi Informatika – Program Sarjana di Fakultas Teknologi Industri Universitas Islam Indonesia
Yogyakarta, 16 Juli 2025

Tim Penguji

Arrie Kurniawardhani S.Si., M.Kom.

Anggota 1

Chanifah Indah Ratnasari, S.Kom., M.Kom.

Anggota 2

Moh. Idris, S.Kom., M.Kom.

Mengetahui,

Ketua Program Studi Informatika – Program Sarjana

Fakultas Teknologi Industri

Universitas Islam Indonesia

(Thomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Hilqudz Dzikri

NIM : 21523017

Tugas akhir dengan judul:

**EKSPLORASI METODE DETEKSI EMOSI PADA TEKS DARI
TWITTER, KOMENTAR BLOG, DAN DIALOG FILM**

Menyatakan bahwa seluruh komponen dan isi dalam tugas akhir ini adalah hasil karya saya sendiri. Apabila di kemudian hari terbukti ada beberapa bagian dari karya ini adalah bukan hasil karya sendiri, tugas akhir yang diajukan sebagai hasil karya sendiri ini siap ditarik kembali dan siap menanggung risiko dan konsekuensi apapun.

Demikian surat pernyataan ini dibuat, semoga dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 16 Juli 2025



(Hilqudz Dzikri)

HALAMAN PERSEMBAHAN

Alhamdulillah Rabbil ‘Alamin, segala puji dan syukur penulis panjatkan kehadirat Allah Swt, yang dengan rahmat dan karunia-Nya telah memberikan kekuatan, kemudahan, dan petunjuk sehingga karya ini dapat terselesaikan. Selawat serta salam semoga senantiasa tercurah kepada teladan terbaik sepanjang masa, Nabi Muhammad saw.

Skripsi ini, dengan segala kerendahan hati, penulis persembahkan kepada Orang tua tercinta, Ayahanda Yuniur Chandrawan dan Ibunda Siti Zulika Nasution. Terima kasih atas setiap tetes keringat, doa yang tak pernah putus, pengorbanan yang tak terhingga, serta cinta kasih yang menjadi sumber kekuatan terbesar. Tiada balasan yang sepadan yang bisa diberikan, selain berharap semoga karya ini menjadi secercah kebanggaan untuk Ayah dan Ibu.

Sahabat-sahabat seperjuangan yang luar biasa, Adyuta Indra Adyatma, Aldi Sri Laksono, Alexsandrea Greska, dan seluruh teman-teman lain yang tidak dapat disebutkan satu per satu. Terima kasih atas malam-malam begadang, diskusi yang tak kenal lelah, yang selalu memberikan warna, dukungan, dan canda tawa serta semangat yang saling kita bagikan dalam setiap langkah perjalanan ini. Semoga karya ini dapat menjadi bentuk bakti seorang anak kepada orang tuanya, serta wujud pertanggungjawaban untuk menyelesaikan amanah pendidikan Sarjana.

HALAMAN MOTO

“Imagine life as a game in which you are juggling some five balls in the air. You name them - work, family, health, friends and spirit - and you're keeping all of these in the air. You will soon understand that work is a rubber ball. If you drop it, it will bounce back. But the other four balls - family, health, friends and spirit - are made of glass. If you drop one of these, they will be irrevocably scuffed, marked, nicked, damaged or even shattered. They will never be the same. You must understand that and strive for balance in your life.”

- Brian Dyson

KATA PENGANTAR

Dengan menyebut nama Allah Yang Maha Pengasih lagi Maha Penyayang, segala puji dan syukur penulis panjatkan atas kehadiran-Nya, karena hanya atas rahmat dan rida-Nya, penyusunan laporan tugas akhir yang berjudul “Eksplorasi Metode Deteksi Emosi pada Teks dari Twitter, Komentar Blog, dan Dialog Film” ini dapat terselesaikan.

Penelitian ini disusun sebagai salah satu syarat untuk meraih gelar Sarjana, dengan tujuan utama untuk mengeksplorasi kemampuan berbagai pendekatan komputasi dalam memahami emosi yang tersirat dalam teks digital. Selama proses pengerjaan, penulis menyadari bahwa setiap kemudahan dan kesulitan datang atas izin-Nya. Berbagai tantangan menjadi bagian dari proses pembelajaran yang sangat berharga. Namun, berkat pertolongan Allah Swt, serta dukungan dari berbagai pihak, segala hambatan tersebut dapat dilalui. Oleh karena itu, penulis ingin menyampaikan terima kasih yang tulus kepada:

1. Bapak Prof. Fathul Wahid, S.T., M.Sc., Ph.D., selaku Rektor Universitas Islam Indonesia.
2. Bapak Prof. Dr. Ir. Hari Purnomo, M.T., IPU, ASEAN.Eng, selaku Dekan Fakultas Teknologi Industri Universitas Islam Indonesia.
3. Bapak Dr. Raden Teduh Dirgahayu, S.T., M.Sc., selaku Ketua Jurusan Informatika Fakultas Teknologi Industri Universitas Islam Indonesia.
4. Bapak DThomas Hatta Fudholi, S.T., M.Eng., Ph.D., selaku Ketua Program Studi Informatika Program Sarjana Fakultas Teknologi Industri Universitas Islam Indonesia.
5. Bapak Andhik Budi Cahyono, S.T., M.T., selaku Dosen Pembimbing Akademik, yang telah memberi informasi dan bimbingan akademik selama perkuliahan berlangsung.
6. Ibu Arrie Kurniawardhani, S.Si., M.Kom., selaku Dosen Pembimbing, yang dengan sabar telah meluangkan waktu, tenaga, dan pikiran untuk memberikan arahan, masukan, serta motivasi dari awal hingga akhir penyelesaian laporan ini.
7. Seluruh Dosen Jurusan Informatika Fakultas Teknologi Industri Universitas Islam Indonesia yang telah memberikan pengetahuan berupa ilmu yang sangat berharga selama masa perkuliahan.
8. Kedua orang tua dan seluruh keluarga tercinta, atas doa, dukungan moral, dan kasih sayang yang menjadi sumber kekuatan terbesar bagi penulis.

9. Sahabat dan rekan-rekan seperjuangan atas kebersamaan, diskusi, dan semangat yang saling menguatkan.
10. Semua pihak lain yang tidak dapat penulis sebutkan satu per satu, terima kasih atas segala bentuk bantuan yang telah diberikan.

Penulis menyadari sepenuhnya bahwa laporan ini masih jauh dari kesempurnaan. Segala kritik dan saran yang membangun akan diterima dengan lapang dada. Semoga penelitian ini dapat membawa manfaat bagi pengembangan ilmu pengetahuan dan menjadi ladang amal bagi penulis.

Yogyakarta, 16 Juli 2025



Hilqudz Dzikri

SARI

Komunikasi digital yang dominan berbasis teks seringkali kehilangan isyarat emosi nonverbal, sehingga menimbulkan kebutuhan untuk deteksi emosi secara otomatis. Namun, performa metode deteksi emosi sangat bervariasi tergantung pada karakteristik teks yang beragam, seperti *tweet* yang singkat dan informal, komentar yang naratif, dan dialog yang kontekstual. Penelitian ini bertujuan untuk mengeksplorasi kinerja pendekatan *machine learning* dan *deep learning* dalam menghadapi tiga jenis teks tersebut.

Metodologi penelitian ini mencakup penggunaan tujuh model, yaitu Naïve Bayes, SVM, dan Random Forest sebagai perwakilan *machine learning*, serta BiLSTM, BERT, RoBERTa, dan DistilBERT sebagai perwakilan *deep learning*. Kelima *dataset* publik berbahasa Inggris yang digunakan dikelompokkan berdasarkan karakteristiknya (tiga *dataset tweet*, satu *dataset* komentar, dan satu *dataset* dialog). Proses eksperimen meliputi *preprocessing* data, pembagian data, ekstraksi fitur, pelatihan model, serta evaluasi yang didukung oleh *confusion matrix* dan waktu inferensi. Eksperimen tambahan juga dilakukan untuk mengukur dampak *preprocessing* terhadap kinerja model.

Temuan utama dari penelitian ini menunjukkan bahwa tidak ada satu pendekatan pun yang unggul di semua kondisi. Pendekatan *deep learning* berbasis Transformer secara konsisten mencapai akurasi tertinggi pada teks *tweet* dan komentar, dengan *F1-score* puncak mencapai 0,96. Namun, pada teks dialog, seluruh model baik *machine learning* maupun *deep learning* menunjukkan penurunan performa yang drastis (*F1-score* tertinggi hanya 0,29), membuktikan kegagalan dalam menangkap konteks percakapan. Lebih lanjut, penelitian ini mengonfirmasi adanya *trade-off* yang jelas, yaitu model *deep learning* unggul dalam akurasi, sementara model *machine learning* unggul dalam efisiensi waktu inferensi. Temuan menarik lainnya adalah pengaruh *preprocessing* yang berlawanan, proses ini terbukti penting untuk model *machine learning*, namun justru dapat menurunkan performa model *deep learning* pada data dialog yang kaya akan konteks (*F1-score* RoBERTa naik dari 0,27 menjadi 0,45 tanpa *preprocessing*). Kesimpulannya, pemilihan metode deteksi emosi yang optimal sangat bergantung pada karakteristik data dan prioritas antara akurasi dan efisiensi komputasi.

Kata kunci: Deteksi Emosi, *Machine learning*, *Deep Learning*, Karakteristik Teks, Analisis Komparatif, *Natural Language Processing*, Transformer.

GLOSARIUM

<i>Batch Size</i>	Jumlah sampel data yang diproses oleh model dalam satu iterasi selama proses pelatihan.
<i>Case Folding</i>	Tahap <i>preprocessing</i> untuk mengubah semua huruf dalam teks menjadi huruf kecil guna menyeragamkan data.
<i>Confusion Matrix</i>	Tabel visualisasi yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan menunjukkan jumlah prediksi benar dan salah untuk setiap kelas.
<i>Deep Learning</i>	Cabang dari <i>machine learning</i> yang menggunakan jaringan saraf tiruan dengan banyak lapisan (<i>deep neural networks</i>) untuk mempelajari pola kompleks dari data.
<i>Downsampling</i>	Teknik penyeimbangan data dengan cara mengurangi jumlah sampel pada kelas mayoritas agar setara dengan jumlah sampel pada kelas minoritas.
<i>Epoch</i>	Satu siklus lengkap di mana seluruh <i>dataset</i> pelatihan telah diproses oleh model sebanyak satu kali.
Ekstraksi Fitur	Proses mengubah data mentah, seperti teks, menjadi fitur numerik yang dapat dipahami dan diolah oleh model <i>machine learning</i> .
<i>F1-score</i>	Rata-rata harmonik dari <i>precision</i> dan <i>recall</i> , digunakan sebagai metrik utama untuk mengevaluasi model pada <i>dataset</i> yang tidak seimbang.
<i>Hyperparameter</i>	Parameter konfigurasi model yang nilainya ditentukan sebelum proses pelatihan dimulai, seperti jumlah <i>epoch</i> .
<i>Machine learning</i>	Bidang ilmu komputer yang berfokus pada pengembangan algoritma yang memungkinkan sistem untuk belajar dari data.
<i>Overfitting</i>	Kondisi di mana model terlalu "menghafal" data latih sehingga kinerjanya baik pada data tersebut, tetapi buruk pada data baru.
<i>Preprocessing</i>	Serangkaian langkah pembersihan dan persiapan data mentah sebelum data tersebut digunakan untuk melatih model.
<i>Stemming</i>	Proses <i>preprocessing</i> untuk mengubah sebuah kata menjadi bentuk dasarnya dengan memotong imbuhan.
<i>Stopword</i>	Proses menghilangkan kata-kata umum yang sering muncul namun tidak memiliki makna signifikan.

TF-IDF	Metode statistik untuk mengukur bobot atau pentingnya sebuah kata dalam sebuah dokumen relatif terhadap keseluruhan kumpulan dokumen.
<i>Tokenization</i>	Proses memecah kalimat atau teks menjadi unit-unit yang lebih kecil yang disebut token (biasanya per kata).
Transformer	Arsitektur <i>deep learning</i> modern yang menggunakan mekanisme <i>self-attention</i> untuk memproses seluruh kata dalam kalimat secara bersamaan, memungkinkannya menangkap konteks secara efektif.
Waktu Inferensi	Waktu yang dibutuhkan oleh model yang telah terlatih untuk membuat prediksi pada satu sampel data baru.

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN DOSEN PEMBIMBING	ii
HALAMAN PENGESAHAN DOSEN PENGUJI	iii
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR.....	iv
HALAMAN PERSEMBAHAN	v
HALAMAN MOTO.....	vi
KATA PENGANTAR.....	vii
SARI	ix
GLOSARIUM	x
DAFTAR ISI	xii
DAFTAR TABEL	xv
DAFTAR GAMBAR.....	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian	2
1.4 Batasan Masalah	2
1.5 Manfaat Penelitian	2
BAB II KAJIAN PUSTAKA	3
2.1 Emosi dan Deteksi Emosi pada Teks	3
2.2 Pendekatan <i>Machine Learning</i>	3
2.2.1 Naive Bayes	4
2.2.2 Support Vector Machine	5
2.2.3 Random Forest	5
2.3 Pendekatan <i>Deep Learning</i>	6
2.3.1 Bidirectional Long Short-Term Memory (BiLSTM).....	7
2.3.2 Transformer.....	8
2.4 Metrik Evaluasi	11
2.4.1 Confusion Matrix	12
2.4.2 Akurasi	12
2.4.3 Presisi	12
2.4.4 <i>Recall</i>	12

2.4.5	<i>F1-score</i>	13
2.4.6	Waktu Inferensi.....	13
2.4.7	<i>Overfitting</i> dan Generalisasi Model	13
2.5	Peneltian Terdahulu	14
BAB III METODOLOGI PENELITIAN.....		17
3.1	Pengumpulan Data	17
3.1.1	<i>Dataset</i> Teks <i>Tweet</i> dari Twitter	17
3.1.2	<i>Dataset</i> Komentar	18
3.1.3	<i>Dataset</i> Dialog	19
3.1.4	Validasi Kategori Emosi Berdasarkan Kerangka Modern	19
3.2	<i>Preprocessing</i> Data	20
3.3	Pembagian Data	21
3.4	Ekstraksi Fitur	22
3.5	Pelatihan Model	22
3.6	Evaluasi Model	24
BAB IV HASIL DAN PEMBAHASAN.....		25
4.1	Hasil Pengumpulan Data.....	25
4.2	Hasil <i>Preprocessing</i> Data.....	27
4.3	Hasil Pembagian Data	29
4.3.1	<i>Dataset</i> “Emotions” oleh Elgiryewithana	30
4.3.2	<i>Dataset</i> “Emotion Dataset” oleh Parul Pandey	30
4.3.3	<i>Dataset</i> “ISEAR” oleh Faisal Santo.....	31
4.3.4	<i>Dataset</i> “Emotion Datset” oleh Abdallah Wagih.....	31
4.3.5	<i>Dataset</i> MELD.....	32
4.4	Hasil Ekstraksi Fitur.....	32
4.5	Hasil Pelatihan Model	32
4.6	Hasil Evaluasi Model	43
4.6.1	Hasil Evaluasi Model pada <i>Dataset</i> Teks <i>Tweet</i>	51
4.6.2	Hasil Evaluasi Model pada <i>Dataset</i> Teks Komentar	52
4.6.3	Hasil Evaluasi Model pada <i>Dataset</i> Teks Dialog	53
4.7	Pembahasan.....	54
4.7.1	Analisis Kelebihan dan Kekurangan Pendekatan <i>Machine Learning</i> dan <i>Deep Learning</i>	57
4.7.2	Pengaruh Karakteristik Teks terhadap Kinerja Model.....	59

4.7.3	Pengaruh <i>Preprocessing</i> terhadap Kinerja Model	60
4.7.4	Analisis Kesalahan pada <i>Dataset</i> MELD.....	67
4.7.5	Kekurangan Penelitian dan Arah Penelitian Selanjutnya.....	69
BAB V KESIMPULAN DAN SARAN		70
5.1	Kesimpulan	70
5.2	Saran.....	71
DAFTAR PUSTAKA.....		72
LAMPIRAN		77

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu	14
Tabel 3.1 Validasi Kategori Emosi per <i>Dataset</i> terhadap Model Cowen & Keltner (2017) ...	19
Tabel 3.2 Pengaturan <i>Hyperparameter</i>	23
Tabel 4.1 Contoh Data dari <i>Dataset</i> “Emotions” oleh Elgiryewithana	25
Tabel 4.2 Contoh Data dari <i>Dataset</i> “Emotion Dataset” oleh Parul Pandey	26
Tabel 4.3 Contoh Data dari <i>Dataset</i> “ISEAR” oleh Faisal Santo	26
Tabel 4.4 Contoh Data dari <i>Dataset</i> “Emotion Dataset” oleh Abdallah Wagih	26
Tabel 4.5 Contoh Data dari <i>Dataset</i> MELD	26
Tabel 4.6 Proses <i>Preprocessing</i> Bertahap pada Berbagai Contoh Teks	27
Tabel 4.7 Persebaran dan Pembagian <i>Dataset</i> “Emotions” oleh Elgiryewithana.....	30
Tabel 4.8 Persebaran Emosi <i>Dataset</i> “Emotion Dataset” oleh Parul Pandey	31
Tabel 4.9 Persebaran dan Pembagian <i>Dataset</i> “ISEAR” oleh Faisal Santo	31
Tabel 4.10 Persebaran dan Pembagian <i>Dataset</i> “Emotion Dataset” oleh Abdallah Wagih	31
Tabel 4.11 Persebaran Emosi <i>Dataset</i> MELD	32
Tabel 4.12 Contoh Hasil Ekstraksi Fitur.....	32
Tabel 4.13 Hasil Evaluasi pada <i>Dataset</i> Teks Tweet	52
Tabel 4.14 Hasil Evaluasi pada <i>Dataset</i> Teks Komentar	53
Tabel 4.15 Hasil Evaluasi pada <i>Dataset</i> Teks Dialog.....	53
Tabel 4.16 Rekapitulasi Hasil Pengujian Model.....	54
Tabel 4.17 Peringkat Model Berdasarkan Akurasi	55
Tabel 4.18 Peringkat Model Berdasarkan Waktu Inferensi.....	56
Tabel 4.19 Hasil Evaluasi pada <i>Dataset</i> “ISEAR”, <i>Dataset</i> “Emotion Dataset”, dan <i>Dataset</i> MELD tanpa <i>Preprocessing</i>	65
Tabel 4.20 Perbandingan <i>F1-score</i> (Dengan dan Tanpa <i>Preprocessing</i>)	66
Tabel 4.21 Studi Kasus Kesalahan Prediksi Model pada <i>Dataset</i> MELD Dialog 259	68

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi Konsep Klasifikasi Naive Bayes	4
Gambar 2.2 Ilustrasi Konsep Support Vector Machine	5
Gambar 2.3 Ilustrasi Konsep Random Forest	6
Gambar 2.4 Ilustrasi Perbedaan antara <i>machine learning</i> dan <i>deep learning</i>	7
Gambar 2.5 Ilustrasi Arsitektur BiLSTM	8
Gambar 2.6 Prosedur <i>pre-training</i> dan <i>fine-tuning</i> BERT	9
Gambar 2.7 Ilustrasi Optimasi Model RoBERTa dari BERT	10
Gambar 2.8 Ilustrasi Proses <i>Knowledge Distillation</i> pada DistilBERT	11
Gambar 3.1 Diagram Alur Penelitian	17
Gambar 3.2 Diagram Alur Tahapan Preprocessing Data	21
Gambar 4.1 Gambar Hasil Pelatihan Model pada <i>Dataset</i> “Emotions” oleh ElgiriyeWithana	34
Gambar 4.2 Gambar Hasil Pelatihan Model pada <i>Dataset</i> “Emotion Dataset” oleh Parul Pandey	36
Gambar 4.3 Gambar Hasil Pelatihan Model pada <i>Dataset</i> “ISEAR” oleh Faisal Santo	38
Gambar 4.4 Gambar Hasil Pelatihan Model pada <i>Dataset</i> “Emotion Dataset” oleh Abdallah Wagih	40
Gambar 4.5 Gambar Hasil Pelatihan Model pada <i>Dataset</i> “MELD”	42
Gambar 4.6 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “Emotions” oleh ElgiriyeWithana	44
Gambar 4.7 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “Emotion Dataset” oleh Parul Pandey ..	46
Gambar 4.8 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “ISEAR” oleh Faisal Santo	47
Gambar 4.9 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “Emotion Dataset” oleh Abdallah Wagih	49
Gambar 4.10 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “MELD”	51
Gambar 4.11 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “ISEAR” oleh Faisal Santo tanpa <i>Preprocessing</i>	61
Gambar 4.12 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “Emotion Dataset” oleh Abdallah Wagih tanpa <i>Preprocessing</i>	62
Gambar 4.13 <i>Confusion Matrix</i> Model pada <i>Dataset</i> “MELD” tanpa <i>Preprocessing</i>	64
Gambar 4.14 <i>Confusion Matrix</i> RoBERTa	67

BAB I PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, sebagian besar interaksi manusia berlangsung melalui teks, mulai dari pesan instan hingga komentar di media sosial. Namun, komunikasi berbasis teks tidak menyertakan sinyal nonverbal seperti nada suara atau ekspresi wajah, yang biasanya membantu kita memahami emosi lawan bicara. Akibatnya, kesalahpahaman sering terjadi karena emosi yang tersirat dalam tulisan sulit dikenali secara otomatis (Derks et al., 2008).

Teks yang digunakan dalam komunikasi digital sangat beragam jenisnya. Misalnya, *tweet* di Twitter biasanya singkat dan informal, sering kali mengandung singkatan, tagar, atau emoji untuk menyampaikan ekspresi secara cepat. Sebaliknya, komentar di blog cenderung lebih panjang dan bersifat naratif, dengan gaya bahasa yang bervariasi. Sementara itu, teks dialog memiliki tantangan tersendiri karena sering memerlukan pemahaman terhadap konteks percakapan sebelumnya. Variasi jenis teks ini membuat proses deteksi emosi menjadi lebih menantang (Hussein, 2018).

Berbagai pendekatan telah diusulkan untuk deteksi emosi pada teks, namun masing-masing memiliki keterbatasan dalam menghadapi variasi tersebut. Pendekatan *machine learning* umumnya cukup efektif pada data yang terstruktur, tetapi kesulitan memahami konteks dan ekspresi tidak langsung seperti sarkasme (Cambria et al., 2013). Di sisi lain, pendekatan *deep learning* memiliki keunggulan dalam menangkap pola pemahaman konteks tetapi butuh data yang banyak dengan proses dan waktu komputasi yang lebih besar (Socher et al., 2013).

Berdasarkan tantangan di atas, penelitian ini bertujuan untuk mengeksplorasi performa berbagai pendekatan *machine learning* dan *deep learning* dalam deteksi emosi pada teks. Fokus utamanya adalah menganalisis bagaimana kinerja berbagai pendekatan ini ketika dihadapkan pada jenis teks yang berbeda, seperti cuitan Twitter, komentar blog, dan dialog percakapan, yang telah dikumpulkan dan diuji. Dengan memahami karakteristik dari setiap jenis teks tersebut, diharapkan pendekatan yang diuji dapat mengenali emosi dalam tulisan secara lebih akurat.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, penelitian ini berusaha menjawab pertanyaan-pertanyaan berikut:

- a. Bagaimana kemampuan metode *machine learning* dan *deep learning* dalam mengklasifikasikan emosi dari teks dengan karakteristik yang beragam?
- b. Bagaimana perbandingan kinerja antara pendekatan *machine learning* dan *deep learning* ketika diuji pada dataset dengan karakteristik yang beragam?

1.3 Tujuan Penelitian

Sesuai dengan rumusan masalah di atas, penelitian ini bertujuan untuk:

- a. Mengevaluasi kemampuan pendekatan *machine learning* dan *deep learning* dalam mendeteksi emosi pada teks dengan karakteristik yang beragam.
- b. Menganalisis performa untuk melihat bagaimana karakteristik dataset yang berbeda memengaruhi kinerja setiap pendekatan yang diuji.

1.4 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus, beberapa batasan yang diterapkan dalam penelitian ini adalah sebagai berikut:

- a. Penelitian hanya menggunakan tiga variasi jenis data teks, data *tweet* dari Twitter, data komentar dari blog, dan data dialog percakapan.
- b. Penelitian hanya mencakup pendekatan beberapa *machine learning* dan *deep learning*.
- c. Seluruh data teks yang digunakan dalam penelitian terbatas pada bahasa Inggris.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan gambaran awal mengenai efektivitas berbagai pendekatan *machine learning* dan *deep learning* dalam mendeteksi emosi pada teks dengan karakteristik yang beragam. Hasil penelitian ini dapat menjadi dasar atau referensi untuk penelitian selanjutnya yang ingin mengembangkan metode deteksi emosi lebih lanjut, memperluas jenis data yang digunakan, atau mengadaptasi pendekatan serupa dalam bahasa dan konteks lain.

BAB II

KAJIAN PUSTAKA

2.1 Emosi dan Deteksi Emosi pada Teks

Emosi adalah perasaan yang muncul sebagai respons pribadi terhadap sesuatu yang kita alami atau hadapi (Kleinginna & Kleinginna, 1981). Dalam kajian psikologi modern, penelitian oleh Cowen & Keltner (2017) berhasil memetakan 27 kategori emosi yang berbeda namun saling terhubung. Kategori ini diperoleh melalui penilaian ribuan peserta terhadap 2.185 klip video pendek yang memicu beragam respons emosional. Hasilnya menunjukkan bahwa spektrum emosi manusia jauh lebih beragam dibandingkan enam emosi dasar yang sebelumnya dikemukakan oleh Ekman (1993). Contohnya, Cowen & Keltner mengidentifikasi kategori seperti *awe* (takjub), *amusement* (terhibur), *adoration* (pemujaan), hingga *boredom* (bosan), memberikan pemahaman yang lebih kaya tentang nuansa dan intensitas emosi (Cowen & Keltner, 2017). Emosi dapat diekspresikan melalui teks, ucapan, atau ekspresi wajah (Plutchik, 2001). Di era digital, komunikasi banyak berlangsung lewat teks, seperti pesan singkat atau komentar di media sosial (Kumar S. & Geetha, 2024). Meskipun tidak menyertakan nada suara atau ekspresi wajah, teks tetap bisa memuat emosi melalui pemilihan kata, tanda baca, atau simbol seperti emoji (Mahajan & Zaveri, 2021). Hal ini mendorong munculnya kebutuhan untuk mengenali dan memahami emosi hanya dari teks. Deteksi emosi pada teks bertujuan mengidentifikasi emosi-emosi tersebut secara otomatis menggunakan pendekatan komputasi. Berbeda dengan analisis sentimen yang hanya menilai apakah isi teks bersifat positif atau negatif, deteksi emosi mencoba mengenali jenis emosi spesifik yang terkandung di dalamnya. Tantangannya terletak pada gaya bahasa yang tidak langsung, seperti sarkasme, serta perlunya memahami konteks kalimat secara menyeluruh (Nandwani & Verma, 2021).

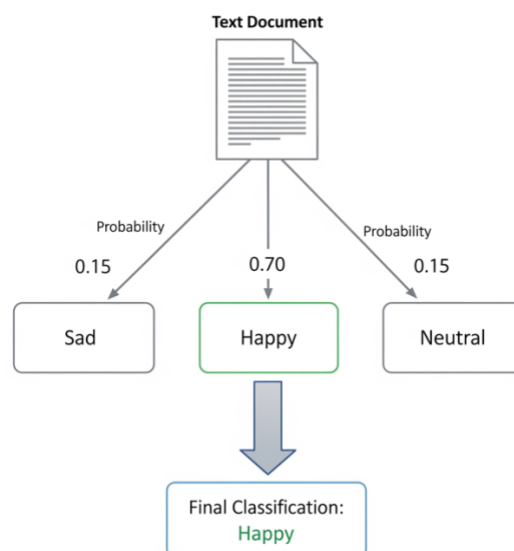
2.2 Pendekatan *Machine Learning*

Salah satu pendekatan awal yang banyak digunakan untuk deteksi emosi pada teks adalah *machine learning*. Pendekatan ini dimulai dengan mengubah teks mentah menjadi representasi numerik terlebih dahulu sebelum diolah oleh algoritma. Proses ini biasanya melibatkan pembersihan teks dari elemen-elemen yang tidak relevan (seperti tanda baca atau kata-kata umum yang disebut *stopwords*) dan kemudian mengubahnya menjadi vektor angka menggunakan teknik seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) (Sparck Jones, 1972), yang memberi bobot pada kata-kata berdasarkan seberapa penting kata tersebut

di seluruh kumpulan data. Setelah teks diubah menjadi angka, dilanjutkan untuk diolah oleh algoritma. Beberapa algoritma *machine learning* yang sering digunakan untuk melakukan klasifikasi meliputi Naïve Bayes (NB), Support Vector Machine (SVM), dan Random Forest (RF) (Aggarwal & Zhai, 2012).

2.2.1 Naïve Bayes

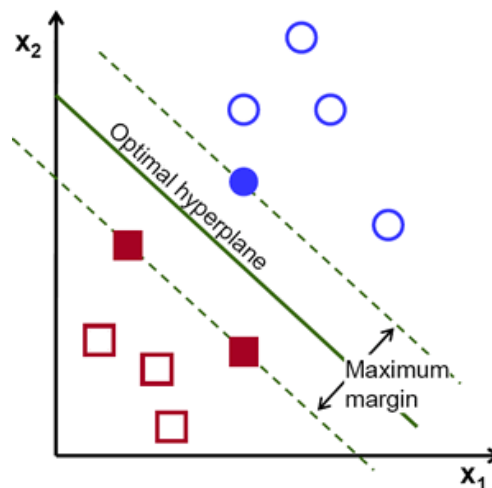
Naïve Bayes (NB) adalah salah satu algoritma yang paling sederhana dan cepat. Cara kerjanya didasarkan pada teorema probabilitas Bayes, di mana ia menghitung kemungkinan sebuah teks masuk ke dalam kategori emosi tertentu berdasarkan kata-kata yang ada di dalamnya. Algoritma ini disebut "naif" karena ia membuat asumsi sederhana bahwa setiap kata dalam teks tidak saling memengaruhi satu sama lain, sebuah asumsi yang dalam praktiknya jarang benar namun seringkali memberikan hasil yang cukup baik, terutama untuk klasifikasi teks (Manning et al., 2008). Seperti yang diilustrasikan pada Gambar 2.1, model ini bekerja dengan mempelajari probabilitas kemunculan setiap kata dalam kaitannya dengan setiap kelas dari data latih. Saat dihadapkan pada teks baru, Naïve Bayes menghitung skor probabilitas untuk setiap kelas dengan mengalikan probabilitas kata-kata di dalamnya, lalu memilih kelas dengan skor probabilitas tertinggi sebagai prediksinya. Karena kesederhanaan dan kecepatannya, Naive Bayes sering dijadikan sebagai model dasar untuk perbandingan (Wang & Manning, 2012).



Gambar 2.1 Ilustrasi Konsep Klasifikasi Naive Bayes

2.2.2 Support Vector Machine

Algoritma lain yang sering digunakan yaitu Support Vector Machine (SVM). SVM bekerja dengan cara menemukan "garis pemisah" atau *hyperplane* terbaik yang dapat memisahkan data dari kelas-kelas yang berbeda dengan *margin* atau jarak selebar mungkin seperti yang diilustrasikan pada Gambar 2.2.



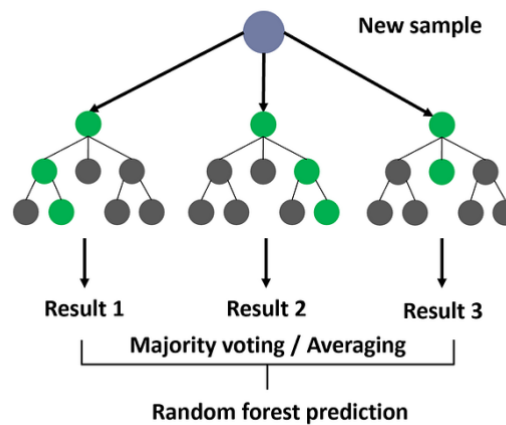
Gambar 2.2 Ilustrasi Konsep Support Vector Machine

Sumber: (Trivusi, 2022)

Keunggulan utama SVM adalah kemampuannya untuk bekerja secara efektif di ruang berdimensi tinggi, yang sangat cocok untuk data teks di mana setiap kata unik bisa dianggap sebagai satu dimensi. Dalam banyak studi perbandingan, SVM sering menunjukkan akurasi yang lebih tinggi dibandingkan Naive Bayes untuk tugas klasifikasi teks (Joachims, 1998).

2.2.3 Random Forest

Random Forest (RF) adalah pendekatan yang berbeda, yang termasuk dalam kategori *ensemble learning*, dimana Random Forest tidak membangun satu model tunggal, melainkan membangun banyak "pohon keputusan" (*decision trees*) secara acak dan kemudian menggabungkan prediksi dari semua pohon tersebut untuk membuat keputusan akhir melalui mekanisme *voting* seperti pada Gambar 2.3.



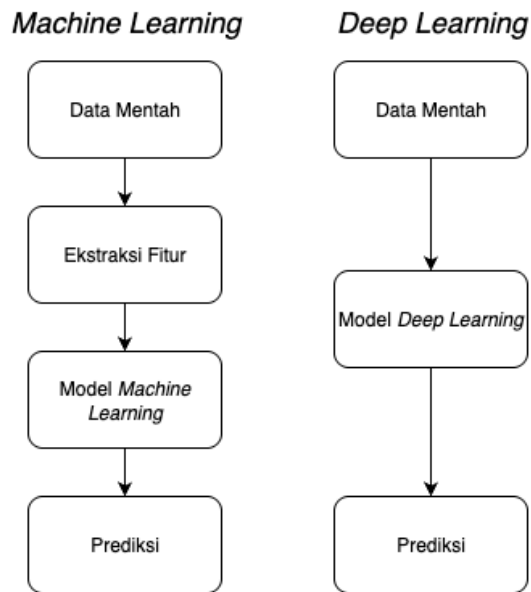
Gambar 2.3 Ilustrasi Konsep Random Forest

Sumber: (Yehoshua, 2023)

Dengan menggabungkan banyak model yang sedikit berbeda, Random Forest menjadi sangat kuat dalam mengurangi risiko *overfitting* (ketika model terlalu menghafal data latihan) dan seringkali menghasilkan akurasi yang sangat tinggi. Namun, kelemahannya adalah ia cenderung membutuhkan lebih banyak waktu dan sumber daya komputasi untuk dilatih dibandingkan dengan Naïve Bayes (Breiman, 2001).

2.3 Pendekatan *Deep Learning*

Deep learning merupakan bagian dari *machine learning* yang dibedakan oleh cara model mempelajari representasi data. Pada *machine learning*, fitur-fitur yang relevan (*feature engineering*) harus ditentukan secara manual, misalnya melalui teknik ekstraksi fitur seperti TF-IDF. Sebaliknya, *deep learning* memanfaatkan jaringan saraf tiruan berlapis-lapis (*multi-layer neural networks*) yang secara otomatis mempelajari fitur langsung dari data mentah, tanpa intervensi manusia (Bengio et al., 2014). Perbedaan antara kedua pendekatan ini diilustrasikan pada Gambar 2.4.

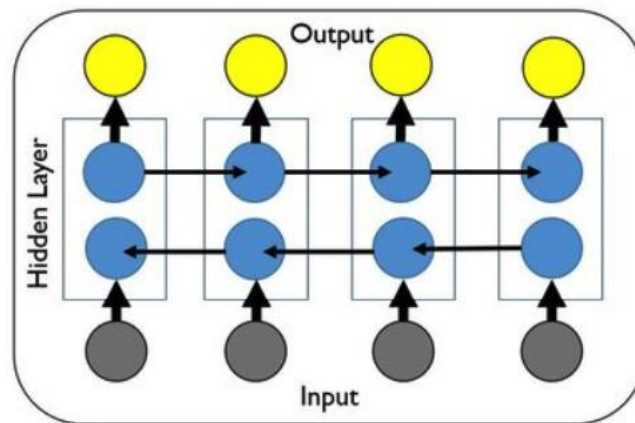


Gambar 2.4 Ilustrasi Perbedaan antara *Machine Learning* dan *Deep Learning*

Metode *machine learning* seperti SVM, Naïve Bayes, dan Random Forest terbukti efektif dalam banyak tugas klasifikasi teks, tetapi pendekatan berbasis TF-IDF hanya mempertimbangkan bobot kata tanpa menangkap urutan dan nuansa konteks. Misalnya, model-model tersebut tidak dapat membedakan “saya senang” dan “saya tidak senang” karena sama-sama melihat frekuensi kata tanpa urutan kata. Keterbatasan dalam memahami konteks inilah yang mendorong pergeseran ke arah pendekatan *deep learning*, yang mampu belajar dengan sendiri fitur-fitur penting dari bahasa dan menangkap makna yang lebih rumit (Goldberg, 2017).

2.3.1 Bidirectional Long Short-Term Memory (BiLSTM)

Salah satu pendekatan *deep learning* yang terkenal dan banyak digunakan adalah Bidirectional Long Short-Term Memory (BiLSTM). BiLSTM memproses teks secara berurutan dan memiliki "memori" untuk mengingat informasi dari kata-kata yang telah dibaca sebelumnya (Hochreiter & Schmidhuber, 1997). Sesuai namanya, BiLSTM bersifat *bidirectional*, artinya ia tidak hanya membaca kalimat dari depan ke belakang, tetapi juga dari belakang ke depan secara bersamaan (Schuster & Paliwal, 1997).



Gambar 2.5 Ilustrasi Arsitektur BiLSTM

Sumber: (Graves & Schmidhuber, 2005)

Ilustrasi arsitektur BiLSTM ditunjukkan pada Gambar 2.5, di mana data *input* masuk ke *hidden layer*, lalu diproses secara dua arah dan kemudian menghasilkan representasi yang lebih kaya sebelum menghasilkan *output*. Kemampuan untuk menangkap konteks dua arah inilah yang membuat BiLSTM dapat memahami makna sebuah kata secara lebih utuh, seperti memahami bahwa kata "tidak" dalam frasa "tidak senang" mengubah keseluruhan makna kalimat.

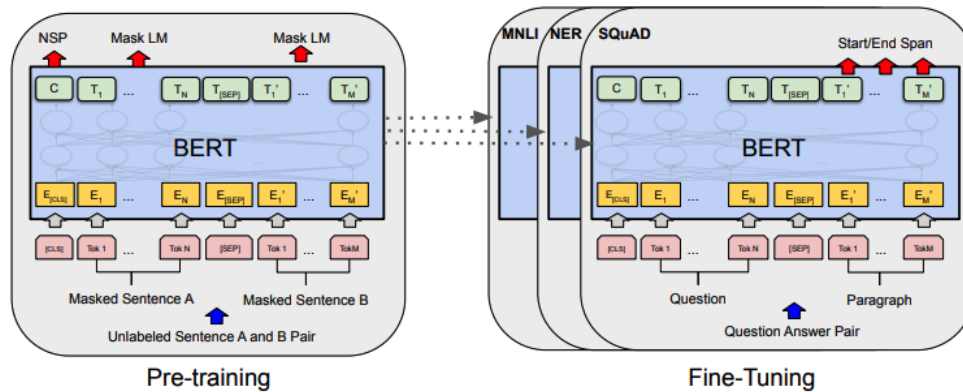
2.3.2 Transformer

Sebuah terobosan baru hadir pada tahun 2017 dengan memperkenalkan arsitektur Transformer dalam makalah *Attention Is All You Need*. Transformer tidak memproses kata secara berurutan seperti BiLSTM, tetapi menggunakan mekanisme yang disebut *self-attention*, yang memungkinkannya untuk melihat semua kata dalam kalimat secara bersamaan dan menimbang pentingnya setiap kata dalam hubungannya dengan kata-kata lain. Kemampuan untuk memproses secara paralel dan menangkap hubungan antar kata membuat model berbasis Transformer sangat kuat (Vaswani et al., 2017).

Bidirectional Encoder Representations from Transformers (BERT)

BERT adalah model yang paling terkenal dari arsitektur Transformer. Melalui proses prapelatihan (*pre-training*) pada volume data teks yang sangat besar. Proses ini menggunakan dua sumber data utama. Pertama, BookCorpus yang merupakan teks dari kumpulan 7.000 buku dengan sekitar 800 juta kata yang mencakup beragam *genre* seperti fiksi dan non-fiksi. Kedua, seluruh artikel Wikipedia berbahasa Inggris dengan sekitar 2,5 miliar kata, di mana hanya teks

utama yang diambil (mengabaikan daftar, tabel, dan header). Setelah melalui *pre-training*, BERT dapat dengan mudah disesuaikan (*fine-tuning*) untuk tugas spesifik seperti deteksi emosi dengan data terbatas.



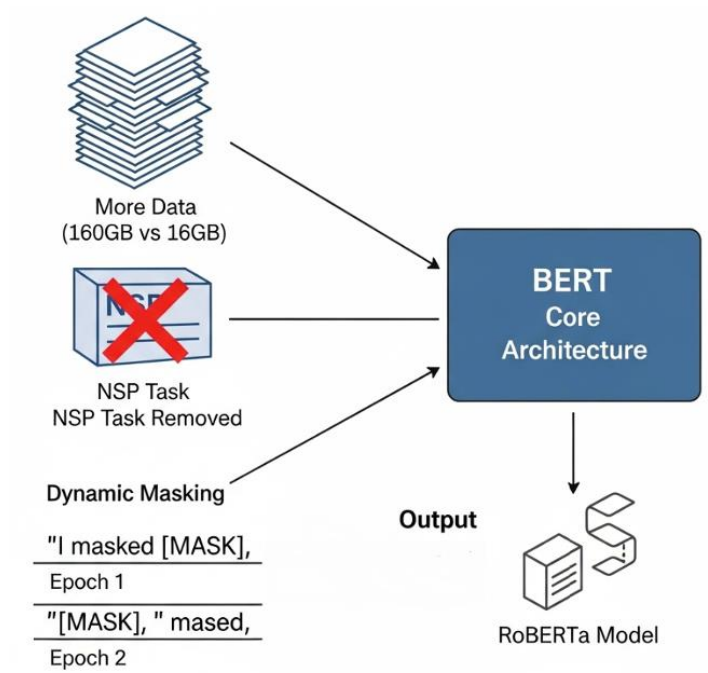
Gambar 2.6 Prosedur *pre-training* dan *fine-tuning* BERT

Sumber: (Devlin et al., 2019)

Ilustrasi prosedur *pre-training* dan *fine-tuning* BERT ditunjukkan pada Gambar 2.6. Pada tahap *pre-training*, BERT menggunakan dua strategi utama, yaitu *Masked Language Model* (MLM), menebak kata yang sengaja disembunyikan, dan *Next Sentence Prediction* (NSP), memprediksi apakah satu kalimat mengikuti kalimat lain. Sementara itu, pada tahap *fine-tuning*, struktur BERT yang sama dapat langsung digunakan untuk berbagai tugas berbeda dengan menambahkan lapisan *output* khusus sesuai kebutuhan (Devlin et al., 2019)

A Robustly Optimized BERT Approach (RoBERTa)

Sejak kemunculan BERT, tim Facebook AI menghadirkan RoBERTa (A Robustly Optimized BERT Approach) sebagai upaya untuk mengoptimalkan metode prapelatihan yang sudah ada. Dalam studi mereka menunjukkan bahwa BERT belum dimaksimalkan, sehingga RoBERTa melakukan beberapa perubahan penting, yaitu menghapus tugas *Next Sentence Prediction* (NSP) yang terbukti kurang membantu, menerapkan *dynamic masking* sehingga pola *masking* kata bervariasi setiap *epoch*, memperbesar ukuran *batch* dan durasi pelatihan, serta memperluas jumlah data menjadi lebih besar dari yang digunakan BERT asli seperti yang diilustrasikan pada Gambar 2.7.

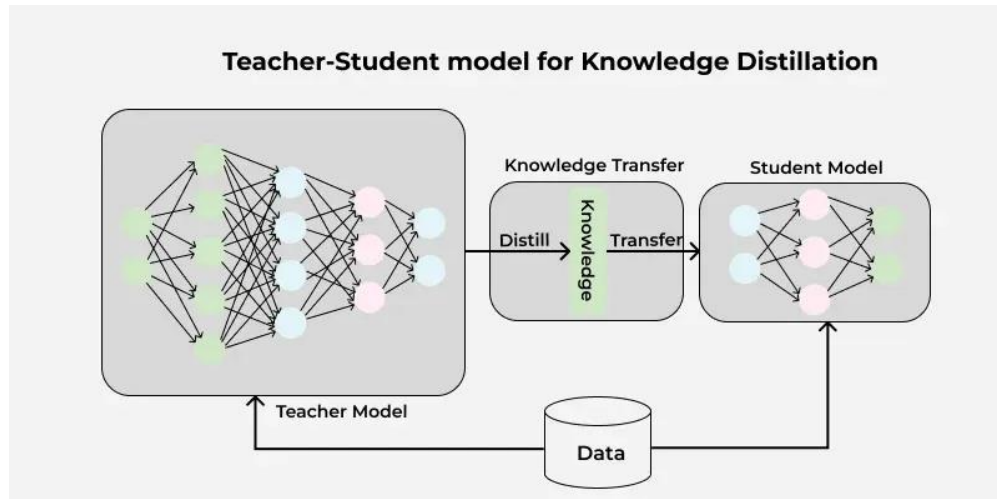


Gambar 2.7 Ilustrasi Optimasi Model RoBERTa dari BERT

Konfigurasi ini secara konsisten menghasilkan peningkatan performa pada berbagai *benchmark* seperti GLUE, RACE, dan SQuAD, menegaskan bahwa pemilihan *hyperparameter* dan data seringkali sama krusialnya dengan arsitektur model itu sendiri (Liu et al., 2019).

A Distilled Version of BERT (DistilBERT)

DistilBERT lahir sebagai versi BERT yang lebih ramping. Meskipun RoBERTa meningkatkan akurasi, ukurannya tetap besar dan memerlukan sumber daya komputasi tinggi. Sanh et al. mengadopsi teknik *Knowledge Distillation* (Bucila et al., 2006; Hinton et al., 2015).



Gambar 2.8 Ilustrasi Proses *Knowledge Distillation* pada DistilBERT

Sumber: (ayushimalm50, 2025)

Seperti yang terlihat pada Gambar 2.8, DistilBERT dilatih dengan sebuah teknik di mana pengetahuan dari model besar yang disebut *teacher* dipindahkan ke model yang lebih kecil yang disebut *student*. Dengan menggunakan BERT asli sebagai *teacher*, proses ini menghasilkan model *student* yang 40% lebih kecil dan 60% lebih cepat, sambil berhasil mempertahankan 97% performa dari *teacher*. Pendekatan ini memungkinkan penerapan pada perangkat dengan sumber daya terbatas tanpa mengorbankan akurasi secara signifikan (Sanh et al., 2019).

2.4 Metrik Evaluasi

Mulai dari pendekatan *machine learning* hingga *deep learning* kita perlu mengeksplorasi berbagai model yang telah disebutkan secara adil. Ini memerlukan serangkaian metrik evaluasi untuk mendapatkan gambaran kinerja dari masing-masing model. Metrik evaluasi yang digunakan adalah *Confusion matrix* (Fawcett, 2006). Dari matriks tersebut, kita dapat mengukur akurasi, presisi, *recall*, dan *F1-score*. Keempat metrik ini saling melengkapi, akurasi memberikan gambaran umum, presisi dan *recall* mengukur aspek kesalahan prediksi yang berbeda, dan *F1-score* memberi gambaran seimbang antara keduanya. Waktu inferensi juga diukur untuk menilai efisiensi model dalam aplikasi praktis (Manning et al., 2008). Selain itu, analisis kurva pelatihan juga dilakukan untuk mendiagnosis *overfitting* pada model *deep learning*.

2.4.1 Confusion Matrix

Confusion matrix adalah sebuah tabel yang memvisualisasikan performa model dengan menyajikan perbandingan antara hasil prediksi dengan kelas aktualnya. Tabel ini memberikan gambaran yang lebih detail daripada sekadar akurasi, karena ia menunjukkan di mana model berhasil dan di mana model membuat kesalahan. *Confusion matrix* terdiri dari empat elemen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN). TP sebagai jumlah prediksi benar untuk kelas positif, TN sebagai jumlah prediksi benar untuk kelas negatif, FP sebagai jumlah prediksi salah, di mana model memprediksi kelas positif padahal sebenarnya negatif, lalu FN sebagai jumlah prediksi salah, di mana model memprediksi kelas negatif padahal sebenarnya positif (Fawcett, 2006). Elemen-elemen ini sebagai dasar dalam menghitung metrik-metrik lain, seperti akurasi, presisi, *recall*, dan *F1-score* (Manning et al., 2008).

2.4.2 Akurasi

Akurasi secara sederhana mengukur proporsi sampel yang diprediksi dengan benar oleh model dari keseluruhan data. Secara matematis dapat dilihat pada persamaan (2.1).

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

2.4.3 Presisi

Presisi mengukur ketepatan prediksi positif yang dibuat oleh model, di mana nilai presisi yang tinggi menunjukkan tingkat *false positive* yang rendah (Manning et al., 2008). Presisi dirumuskan seperti pada persamaan (2.2).

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (2.2)$$

2.4.4 Recall

Recall atau sensitivitas menilai seberapa baik model menemukan seluruh sampel positif yang ada (Olson & Delen, 2008). *Recall* dirumuskan seperti pada persamaan (2.3).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.3)$$

2.4.5 *F1-score*

F1-score dipakai untuk mengukur keseimbangan antara presisi dan *recall* karena seringkali ada pertukaran antara keduanya (meningkatkan yang satu dapat menurunkan yang lain). Metrik ini sangat berguna untuk *dataset* tidak seimbang karena nilainya akan rendah jika salah satu dari presisi atau *recall* rendah (Manning et al., 2008). Perhitungan *F1-score* dirumuskan seperti pada persamaan (2.4).

$$F1 = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (2.4)$$

2.4.6 Waktu Inferensi

Selain metrik-metrik berbasis prediksi di atas, waktu inferensi (*Inference time*) melengkapi rangkaian metrik evaluasi dengan mengukur rata-rata waktu yang dibutuhkan model untuk memberikan prediksi pada satu contoh. Waktu inferensi umumnya menjadi tolok ukur bagi aplikasi *real-time* atau pada perangkat dengan keterbatasan sumber daya (Sanh et al., 2019).

2.4.7 *Overfitting* dan Generalisasi Model

Evaluasi pada model *deep learning* juga penting dilakukan dengan menganalisis kurva pelatihan untuk mendiagnosis *overfitting*. Model yang sehat biasanya menunjukkan kurva *training loss* dan *validation loss* yang menurun seiring bertambahnya *epoch*, dengan selisih relatif kecil antara keduanya. Sebaliknya, *overfitting* terjadi ketika *training loss* terus menurun sementara *validation loss* mulai meningkat, menandakan bahwa model terlalu menyesuaikan diri dengan data latih dan kehilangan kemampuan generalisasi. Oleh karena itu, kondisi pelatihan yang sehat dapat tercapai pada titik optimal sebelum *error* validasi tersebut mulai meningkat, di mana model menunjukkan keseimbangan antara performa pada data latih dan kemampuannya untuk generalisasi (Ying, 2019).

2.5 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

Peneliti	Masalah	Metode yang Digunakan	Kinerja Model	Dataset
Albu et al. (2022)	mengeksplorasi model ML dan DL, serta mengusulkan model <i>ensemble</i> untuk meningkatkan akurasi deteksi emosi pada data <i>tweet</i> .	NB, SVM, BERT, RoBERTa, BERTweet, BERTweet + SVM	NB 0,80, SVM akurasi 0,84, BERT akurasi 0,87, RoBERTa akurasi 0,87, BERTweet akurasi 0,89 BERTweet+SVM akurasi 0,91	WASSA-2017
Alaeddini (2024)	Perbandingan ML dan DL untuk deteksi 6 emosi dasar pada komentar Reddit.	RF, SVM, NB, LSTM, LR, XGBoost, Decision Tree	RF <i>F1-score</i> 0,56, SVM <i>F1-score</i> 0,62, NB <i>F1-score</i> 0,40, LSTM <i>F1-score</i> 0,60, LR <i>F1-score</i> 0,61, XGBoost <i>F1-score</i> 0,60, Decision Tree <i>F1-score</i> 0,52	GoEmotions
Imran (2024)	Melakukan analisis komparatif model Transformer untuk klasifikasi emosi pada teks.	BERT, RoBERTa, SentiMoji, ALBERT, DeBERTa, CodeBERT, GraphCodeBERT	Untuk dataset GitHub, BERT <i>F1-score</i> 0,591, RoBERTa <i>F1-score</i> 0,590, SentiMoji <i>F1-score</i> 0,521, ALBERT <i>F1-score</i> 0,539, DeBERTa <i>F1-score</i> 0,608, CodeBERT <i>F1-</i>	GitHub Emotion Dataset, Stack Overflow Emotion Dataset

			<i>score</i> 0,555, GraphCodeBERT <i>F1-score</i> 0,549. Untuk <i>dataset</i> Stack Overflow BERT <i>F1-score</i> 0,588, RoBERTa <i>F1-score</i> 0,599, SentiMojj <i>F1-</i> <i>score</i> 0,530, ALBERT <i>F1-</i> <i>score</i> 0,584, DeBERTa <i>F1-</i> <i>score</i> 0,607, CodeBERT <i>F1-</i> <i>score</i> 0,567, GraphCodeBERT <i>F1-score</i> 0,552.	
Cortiz (2021)	Menyelidiki dan mengeksplorasi kinerja berbagai model Transformer dalam tugas pengenalan emosi pada teks.	BERT, RoBERTa, DistilBERT, XLNet, Electra	BERT <i>F1-score</i> 0,46, RoBERTa <i>F1-score</i> 0,48, distilBERT <i>F1-</i> <i>score</i> 0,49, XLNet <i>F1-score</i> 0,48, Electra <i>F1-score</i> 0,33	GoEmotions
Farhadipour et Al. (2025)	Mengatasi tantangan deteksi emosi <i>multimodal</i> .	RoBERTa (text), Wav2Vec(audio), CNN+Transformer (Video), FacialNet(Image)	RoBERTa akurasi 0,64, Wav2Vec akurasi 0,51, CNN+Transformer akurasi 0,36, FacialNet akurasi 0,23	MELD

Dalam beberapa tahun terakhir, berbagai pendekatan telah dieksplorasi untuk tugas deteksi emosi pada teks dengan beragam jenis. Dapat dilihat dari Tabel 2.1, studi oleh Albu et al. (2022) mengeksplorasi kinerja model ML dan DL pada data *tweet* dari *dataset* WASSA-2017. Mulai dari *machine learning*, NB mencapai akurasi 0,80 dan SVM mencapai 0,84. Masuk ke *deep learning*, BERT dan RoBERTa mencapai akurasi yang sama yaitu 0,87, lalu ada BERTweet yang merupakan model BERT awal yang dilakukan spesifik *preprocessing* mencapai akurasi tertinggi sebesar 0,89. Terakhir, model gabungan yang mereka usulkan, BERTweet+SVM, mencapai akurasi 0,91. Pada jenis data komentar dari Reddit yaitu *dataset* GoEmotion, Alaeddini (2024) melakukan perbandingan antara beberapa metode. Dalam studinya, SVM menunjukkan *F1-score* tertinggi sebesar 0,62, diikuti oleh LR dengan *F1-score* 0,61 dan LSTM dengan 0,60. Selanjutnya, Imran (2024) melakukan analisis komparatif antara model-model *deep learning* berbasis Transformer pada *dataset* dari GitHub dan Stack Overflow. Dari model yang diuji, DeBERTa menunjukkan kinerja terbaik di kedua *dataset*. Pada *dataset* Github dengan *F1-score* 0,608, melampaui BERT 0,591 dan RoBERTa 0,590, pada *dataset* Stack Overflow dengan *F1-score* 0,607, melampaui RoBERTa 0,599 dan BERT 0,588. Sementara itu, Cortiz (2021) secara khusus menyelidiki berbagai varian model *deep learning* berbasis Transformer pada *dataset* GoEmotions. Penelitiannya menunjukkan bahwa DistilBERT menjadi model dengan kinerja terbaik dengan *F1-score* 0,49, diikuti oleh RoBERTa dan XLNet yang sama-sama mencapai skor 0,48. Pada data dialog percakapan, Farhadipour et al. (2025) mengatasi tantangan deteksi emosi *multimodal* pada *dataset* MELD. Dalam evaluasi *unimodal* mereka, komponen teks yang dianalisis menggunakan RoBERTa mencapai akurasi 0,64, lebih tinggi dibandingkan modalitas audio 0,51 dan video 0,36.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan langkah-langkah yang dilakukan dalam penelitian ini. Tahapan-tahapan dalam penelitian ini mencakup pengumpulan data, *preprocessing* data, pembagian data, ekstraksi fitur, pelatihan model, dan diakhiri dengan evaluasi model. Keseluruhan alur penelitian ini digambarkan pada Gambar 3.1.



Gambar 3.1 Diagram Alur Penelitian

3.1 Pengumpulan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data. Data yang digunakan merupakan data sekunder yang berasal dari berbagai sumber publik dan platform data seperti Kaggle. Penelitian ini menggunakan tiga variasi jenis data teks berbahasa Inggris (data *tweet* dari Twitter, data komentar dari blog, dan data dialog). Pemilihan *dataset* berbahasa Inggris didasarkan pada ketersediaan sumber daya bahasa Inggris jauh lebih melimpah dibandingkan bahasa lain. Studi oleh (Joshi et al., 2020) menyoroti adanya distribusi sumber daya yang sangat timpang, di mana bahasa Inggris termasuk dalam segelintir bahasa kategori "Pemenang" (*The Winners*) yang memiliki kelimpahan data. Bahasa Inggris juga memiliki dukungan model *pre-trained* yang optimal, seperti BERT, RoBERTa, dan DistilBERT, yang dilatih pada korpus besar bahasa Inggris seperti BookCorpus dan Wikipedia (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019). Dengan demikian, potensi kinerja model dapat dimaksimalkan tanpa perlu pelatihan ulang dari awal yang memerlukan sumber daya komputasi besar.

3.1.1 *Dataset* Teks *Tweet* dari Twitter

Dataset "Emotions" oleh Elgiryewithana

Dataset ini bersumber dari platform data publik Kaggle dan diunggah oleh Elgiryewithana, seorang praktisi di bidang *Data Science*. *Dataset* ini merupakan kumpulan dari sekitar 400.000 *tweet* berbahasa Inggris yang telah diberi label dengan enam emosi, marah (*anger*), takut (*fear*), senang (*joy*), cinta (*love*), sedih (*sadness*), dan terkejut (*surprise*). Setiap data terdiri dari satu teks *tweet* dan label emosi yang telah dikodekan secara numerik, *sadness*

(0), *joy* (1), *love* (2), *anger* (3), *fear* (4), dan *surprise* (5). Namun, informasi mengenai metodologi anotasi asli dan latar belakang keilmuan anotator tidak dijelaskan secara rinci dalam dokumentasi *dataset* tersebut (Elgiryewithana N, 2024).

Dataset “Emotion Dataset” oleh Parul Pandey

Dataset ini bersumber dari penelitian Saravia et al. (2018) yang berjudul *CARER: Contextualized Affect Representations for Emotion Recognition*. Proses anotasi pada *dataset* menggunakan metode *distant supervision*, di mana label emosi untuk setiap *tweet* diasumsikan berdasarkan *hashtag* yang digunakan oleh penulis *tweet* itu sendiri. Proses ini dirancang dan dilaksanakan oleh para peneliti akademis dari National Tsing Hua University (Saravia et al., 2018). Meskipun penelitian aslinya mengumpulkan data untuk delapan emosi, versi yang digunakan dalam penelitian ini adalah yang tersedia di Kaggle, yang terdiri dari 20.000 *tweet* yang telah disaring ke dalam enam emosi beserta kode numeriknya, *sadness* (0), *joy* (1), *love* (2), *anger* (3), *fear* (4), dan *surprise* (5), oleh Parul Pandey, seorang praktisi *Data Science* (Pandey P, 2021).

Dataset “ISEAR” oleh Faisal Santo

Dataset ini bersumber dari *dataset* ISEAR (Scherer & Wallbott, 1994) dengan 7 kategori emosi yaitu *joy*, *fear*, *anger*, *sadness*, *disgust*, *shame*, dan *guilt*. Proses pelabelan data dilakukan oleh Klaus R. Scherer dan Harald G. Wallbott, yang merupakan akademisi di bidang Psikologi. Versi yang digunakan dalam penelitian ini adalah yang tersedia dari Kaggle, yang telah dimodifikasi menjadi lebih pendek oleh Faisal Santo untuk penelitian sarjananya, menjadi sekitar 7.000 data dan hanya mencakup 4 emosi, yaitu *anger*, *fear*, *joy*, dan *sadness* (Santo F, 2020).

3.1.2 *Dataset* Komentar

Dataset “Emotion Dataset” oleh Abdallah Wagih

Dataset ini dikumpulkan oleh Abdallah Wagih, seorang praktisi di bidang *Artificial Intelligence* dan tersedia di platform Kaggle. *Dataset* ini berisi sekitar 6.000 data teks komentar yang merupakan kompilasi dari berbagai sumber daring. Komentar-komentar ini diklasifikasikan ke dalam tiga emosi, yaitu *anger*, *joy*, dan *fear*. Namun, informasi mengenai siapa anotator asli, latar belakang keilmuan, serta metodologi pelabelan yang digunakan tidak dijelaskan secara spesifik (Ibrahim A, 2023).

3.1.3 Dataset Dialog

Dataset MELD

MELD (Multimodal EmotionLines Dataset) merupakan pengembangan dari *dataset* EmotionLines yang diperkenalkan oleh Poria et al. (2019), sekelompok peneliti AI dan NLP dari berbagai institusi riset. Pada *dataset* EmotionLines sebelumnya, anotasi emosi dilakukan hanya berdasarkan transkrip teks. Untuk mengatasi ambiguitas yang sering muncul, pada MELD dilakukan anotasi ulang dengan metode anotasi *multimodal*. Proses ini melibatkan tiga orang anotator, yang merupakan mahasiswa pascasarjana dengan kemahiran tinggi dalam bahasa Inggris. Mereka diminta untuk memberikan label emosi dengan mempertimbangkan tidak hanya teks, tetapi juga menonton klip video dan mendengarkan audio dari setiap *utterance*. Label akhir ditentukan melalui metode *majority voting*. Contohnya, *utterance* “This guy fell asleep!” yang sebelumnya dilabeli *non-neutral*, setelah melihat ekspresi wajah yang marah dan nada suara yang tinggi dalam video, dilabeli ulang menjadi *anger* dalam MELD. *Dataset* ini terdiri dari sekitar 13.000 data percakapan yang diambil dari serial TV *Friends*, dan setiap *utterance* telah dilabeli dengan salah satu dari tujuh kategori emosi, *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, dan *surprise* (Poria et al., 2019).

3.1.4 Validasi Kategori Emosi Berdasarkan Kerangka Modern

Setelah menguraikan sumber dan karakteristik masing-masing *dataset*, penting untuk memvalidasi relevansi kategori-kategori emosi yang digunakan. Untuk tujuan ini, penelitian ini menggunakan pemetaan 27 kategori emosi dari Cowen & Keltner (2017) sebagai kerangka acuan metodologis. Kerangka ini dipilih karena merupakan salah satu pemetaan emosi empiris yang berasal dari studi skala besar terhadap ribuan respons emosional. Tabel 3.1 di bawah ini menunjukkan bagaimana kategori emosi dari setiap *dataset* divalidasi sebagai kategori yang relevan dalam kerangka kerja Cowen & Keltner.

Tabel 3.1 Validasi Kategori Emosi per *Dataset* terhadap Model Cowen & Keltner (2017)

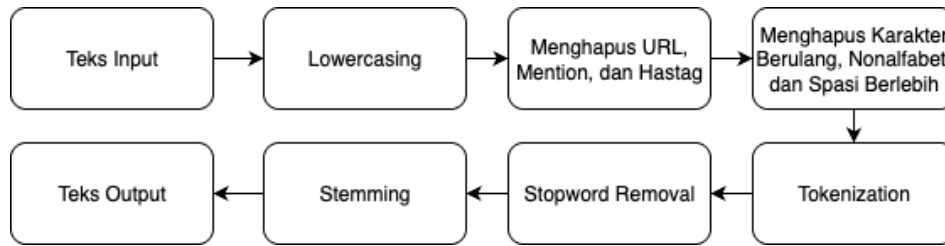
<i>Dataset</i>	Kategori Emosi yang Digunakan	Kesesuaian dengan Model Cowen & Keltner
“Emotions” oleh Elgiriye withana	<i>Love, joy, surprise, anger, sadness, fear</i>	5 dari 6 emosi memiliki kesesuaian sebagai kategori emosi. Emosi <i>love</i>

		divalidasi melalui kategori yang lebih spesifik seperti <i>adoration</i> dan <i>romance</i> .
“Emotion Dataset” oleh Parul Pandey	<i>Love, joy, surprise, anger, sadness, fear</i>	5 dari 6 emosi memiliki kesesuaian sebagai kategori emosi. Emosi <i>love</i> divalidasi melalui kategori yang lebih spesifik seperti <i>adoration</i> dan <i>romance</i> .
“ISEAR” oleh Faisal Santo	<i>Joy, anger, sadness, fear</i>	Keempat emosi yang digunakan memiliki kesesuaian sebagai kategori emosi.
“Emotion Dataset” oleh Abdallah Wagih	<i>Joy, anger, fear</i>	Ketiga emosi yang digunakan memiliki kesesuaian sebagai kategori emosi.
MELD	<i>Joy, surprise, anger, sadness, fear, disgust, neutral</i>	6 dari 7 kategori memiliki kesesuaian sebagai kategori emosi kecuali emosi <i>neutral</i> .

Beberapa emosi yang dianalisis pada penelitian ini seperti *joy*, *sadness*, *anger*, *fear*, *surprise*, dan *disgust* memiliki kesesuaian langsung sebagai kategori emosi dalam temuan Cowen & Keltner. Lebih lanjut, emosi *love*, yang juga tervalidasi keberadaannya meskipun direpresentasikan melalui kategori yang lebih spesifik seperti *adoration* (pemujaan) dan *romance* (romantisme). Sementara itu, emosi *neutral* berada di luar 27 kategori emosi tersebut karena menandakan ketiadaan muatan emosi yang kuat. Perbandingan ini secara metodologis menguatkan bahwa label-label yang digunakan dalam *dataset* penelitian ini merupakan representasi yang valid dan relevan dari spektrum emosi manusia menurut ilmu psikologi modern.

3.2 Preprocessing Data

Setelah data terkumpul, langkah selanjutnya adalah *preprocessing*, yaitu tahap untuk membersihkan dan menyiapkan data agar lebih mudah diolah oleh model *machine learning* atau *deep learning* yang akan digunakan (Aggarwal & Zhai, 2012). Alur *preprocessing* yang diterapkan secara seragam pada seluruh dataset dalam penelitian ini seperti pada Gambar 3.2.



Gambar 3.2 Diagram Alur Tahapan *Preprocessing* Data

Proses ini diawali dengan data mentah (teks *input*). Tahap pertama adalah mengubah semua huruf menjadi huruf kecil (*lowercasing*). Selanjutnya, menghapus tautan internet (URL), simbol '@' (*mention*), tagar (*hashtag*), karakter berulang, karakter nonalfabet, dan spasi berlebih. Setelah itu, dilakukan proses *tokenization*, di mana setiap kalimat dipecah menjadi per kata. Kemudian, kata-kata umum dalam bahasa Inggris yang tidak memiliki makna penting, seperti 'the', 'is', dan 'in' dihilangkan melalui tahap *stopword removal* (Manning et al., 2008). Terakhir, dilakukan proses *stemming* untuk mengubah setiap kata ke bentuk dasarnya (Porter, 1980). Misalnya, kata "merasakan", "dirasakan", dan "terasa" akan disederhanakan menjadi bentuk dasar "rasa". Hasil akhir dari rangkaian proses ini adalah teks *output*. Hal ini diharapkan dapat membantu pelatihan model tanpa terganggu oleh variasi bentuk kata atau simbol.

3.3 Pembagian Data

Setelah melalui tahap *preprocessing*, seluruh *dataset* dibagi menjadi data latih dan data uji untuk keperluan pelatihan dan evaluasi model. Untuk *dataset* yang tidak memiliki pembagian standar dari sumbernya, penelitian ini menggunakan rasio pembagian 80% untuk data latih dan 20% untuk data uji. Pemilihan rasio ini merupakan praktik umum yang didasarkan pada studi empiris yang menunjukkan bahwa alokasi 20-30% data untuk pengujian dan 70-80% sisanya untuk pelatihan cenderung memberikan hasil terbaik. Rasio 80% untuk data latih merupakan titik optimal di mana estimasi *error* model yang dihasilkan tidak hanya valid tetapi juga yang paling akurat (Gholamy et al., 2018). Khusus untuk *dataset* "Emotions" oleh Elgiryewithana, dilakukan penyeimbangan data *downsampling* karena jumlah data yang tidak seimbang sangat besar. Proses ini dilakukan dengan mengidentifikasi kelas emosi dengan jumlah paling sedikit, lalu mengambil sampel secara acak dari kelas-kelas emosi lainnya agar jumlahnya sama dengan kelas minoritas (Han et al., 2012), lalu dibatasi total data baru menjadi 25.000.

3.4 Ekstraksi Fitur

Setelah data teks dibersihkan dan dibagi, selanjutnya adalah mengubah teks menjadi dalam bentuk numerik yang dapat dipahami oleh model. Proses ini disebut dengan ekstraksi fitur atau vektorisasi. Dalam penelitian ini menggunakan metode yang berbeda untuk model *machine learning* dan model *deep learning*. Untuk model *machine learning* menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Untuk model *deep learning* proses ekstraksi fitur ini ditangani secara internal oleh *embedding layer*, yang merupakan bagian dari arsitektur model (Goldberg, 2017).

3.5 Pelatihan Model

Penelitian ini mengeksplorasi model *machine learning* dan *deep learning*. Model *machine learning* yang digunakan meliputi Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF). Model Naïve Bayes yang digunakan adalah varian MultinomialNB, yang sangat sederhana dan efektif untuk klasifikasi pada data teks (Aggarwal & Zhai, 2012). SVM dan Random Forest juga digunakan dalam penelitian ini karena sering digunakan pada penelitian terdahulu dalam studi komparatif dan biasanya menjadi dua teratas model *machine learning* terbaik (Indulkar & Patil, 2021). Untuk model *deep learning*, digunakan BiLSTM serta model-model berbasis arsitektur Transformer. BiLSTM memproses teks dari dua arah yang mampu menangkap konteks dan urutan kata, dimana ini adalah sebuah kelemahan utama dari model ML berbasis TF-IDF (Schuster & Paliwal, 1997). BiLSTM banyak digunakan dalam penelitian klasifikasi emosi dibanding dengan LSTM biasa karena memberikan performa yang lebih baik (Zhou et al., 2016). Untuk model *deep learning* dengan arsitektur Transformer, penelitian ini menggunakan model-model *pre-trained* yang disediakan oleh *library* Hugging Face Transformers (Wolf et al., 2020). Model BERT yang digunakan adalah varian bert-base-uncased yang memiliki 12 *hidden layers*, 768 *hidden size*, 12 *attention heads*, dan 110 juta *parameters* (Devlin et al., 2019). RoBERTa varian roberta-base yang memiliki 12 *hidden layers*, 768 *hidden size*, 12 *attention heads*, dan 125 juta *parameters* (Liu et al., 2019). DistilBERT varian distilbert-base-uncased yang memiliki 6 *hidden layers*, 768 *hidden size*, 12 *attention heads*, dan 66 juta *parameters* (Sanh et al., 2020). Untuk model *deep learning* dilakukan pengaturan *hyperparameter*, yang dimana *hyperparameter* merupakan parameter eksternal pada model yang ditentukan sebelum proses pelatihan, dan berpengaruh

langsung terhadap performa model (Goodfellow et al., 2016). Tabel 3.2 menampilkan pengaturan *hyperparameter* yang digunakan.

Tabel 3.2 Pengaturan *Hyperparameter*

Model	Hyperparameter	Nilai
BiLSTM	Dimensi Embedding (<i>Embedding Dim</i>)	128
	<i>Max Length</i>	128
	<i>Optimizer</i>	Adam
	<i>Learning Rate</i>	1e-3 (0,001)
	Jumlah <i>Epoch</i>	20
BERT, RoBERTa, DistilBERT	<i>Max Length</i>	128
	<i>Batch Size</i>	16
	Jumlah <i>Epoch</i>	10
	<i>Optimizer</i>	AdamW
	<i>Learning Rate</i>	2e-5 (0,00002)

Pengaturan *hyperparameter* untuk setiap model dipilih berdasarkan praktik umum dan rekomendasi dari studi-studi relevan. Untuk model berbasis Transformer, *paper* asli BERT (Devlin et al., 2019) merekomendasikan sebuah rentang *hyperparameter* untuk proses *fine-tuning*. *Learning rate* 2e-5 dipilih karena terbukti sebagai nilai yang stabil dan efektif untuk *fine-tuning* BERT (Sun et al., 2020), untuk *batch size* 16 dipilih agar lebih ringan untuk memastikan kelancaran proses pelatihan. Sementara *paper* acuan tersebut menyarankan 2-4 *epoch*, penelitian ini menggunakan 10 *epoch* dengan tujuan untuk melihat kestabilan dari model. Pengaturan *hyperparameter* yang serupa juga diterapkan pada RoBERTa dan DistilBERT, karena keduanya merupakan varian arsitektur BERT. Praktik ini didukung oleh *paper* masing-masing, di mana RoBERTa juga menggunakan rentang *batch size* dan *learning rate* yang sama untuk tugas klasifikasi GLUE (Liu et al., 2019), dan DistilBERT secara eksplisit menyatakan bahwa mereka mengadopsi praktik terbaik dari BERT dan RoBERTa (Sanh et al., 2019). Penggunaan *optimizer* AdamW juga mengikuti praktik standar untuk melatih model Transformer. Pilihan ini didasarkan pada temuan bahwa AdamW secara signifikan meningkatkan performa generalisasi dengan memperbaiki cara kerja *weight decay*, dan penggunaannya telah terbukti efektif pada arsitektur Transformer untuk tugas-tugas pemahaman bahasa alami (Loshchilov & Hutter, 2019). Untuk BiLSTM, konfigurasi *hyperparameter* yang digunakan seperti *optimizer* Adam, *learning rate* 1e-3, dan dimensi *embedding* 128 ditentukan melalui serangkaian percobaan awal. Nilai *max_length* sebesar 128 ditetapkan secara seragam untuk semua model *deep learning* karena mayoritas *dataset* yang

digunakan pada penelitian ini memiliki panjang teks relatif pendek sehingga jarang melebihi 128 token. Nilai ini juga merupakan standar yang umum digunakan untuk tugas klasifikasi teks dan dianggap sebagai jalan tengah yang optimal antara performa dan efisiensi komputasi (Devlin et al., 2019; Sun et al., 2020).

3.6 Evaluasi Model

Tahap evaluasi merupakan langkah terakhir dalam metodologi penelitian ini, yang bertujuan untuk menilai performa masing-masing model *machine learning* dan *deep learning* yang sudah dilatih. Evaluasi dilakukan menggunakan data uji yang sudah dipisahkan sebelumnya. Metode evaluasi yang digunakan dalam penelitian ini meliputi *confusion matrix*, akurasi, presisi, *recall*, *F1-score*, dan waktu inferensi. Selain metrik-metrik tersebut, dilakukan pula analisis terhadap kurva pelatihan untuk mendiagnosis ada atau tidaknya *overfitting* pada setiap model *deep learning* yang diuji.

BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan dan membahas secara mendalam seluruh hasil eksperimen yang telah dilakukan sesuai dengan metodologi pada Bab 3. Penyajian akan dimulai dari hasil pengumpulan data, *preprocessing*, pembagian data, ekstraksi fitur, pelatihan model, evaluasi model, dan diakhiri dengan pembahasan analisis. Seluruh eksperimen dijalankan menggunakan Google Colaboratory dengan spesifikasi akselerator 1x GPU T4, 15 GB VRAM, dan 12.7 GB RAM sistem, serta Kaggle Notebook yang menyediakan 2x GPU T4, total 30 GB VRAM, dan 29 GB RAM sistem. Implementasi dilakukan dalam bahasa Python dan memanfaatkan beberapa *library* utama, yaitu Scikit-learn versi 1.6.1, TensorFlow versi 2.18.0, PyTorch versi 2.6.0, dan Transformers versi 4.53.1. Seluruh kode implementasi untuk masing-masing *dataset* disediakan dalam bentuk tautan di bagian Lampiran.

4.1 Hasil Pengumpulan Data

Pada tahap awal, data dikumpulkan dari beberapa *dataset* publik yang telah banyak digunakan dalam penelitian sebelumnya. *Dataset* tersebut mencakup tiga variasi jenis teks, yaitu *tweet* dari Twitter, komentar pengguna dari blog, dan percakapan dialog. Setiap *dataset* yang ditemukan sudah berlabel emosi. Contoh data dari masing-masing *dataset* yang digunakan ditampilkan pada Tabel 4.1 sampai 4.5.

Tabel 4.1 Contoh Data dari *Dataset* “Emotions” oleh Elgiryewithana

Text	Label
i just feel really helpless and heavy hearted	4
that was what i felt when i was finally accepted in the bulgarian conservatorie after i had applied a couple of years	1
i gave up my internship with the dmrg and am feeling distraught	0

Tabel 4.2 Contoh Data dari *Dataset* “Emotion Dataset” oleh Parul Pandey

Text	Label
im feeling quite sad and sorry for myself but ill snap out of it soon	0
i feel romantic too	2
im sure its because when i am lost i feel like everyone is being hostile toward me and i hate that feeling	3

Tabel 4.3 Contoh Data dari *Dataset* “ISEAR” oleh Faisal Santo

Sentiment	Content
Anger	Police don't wanna be called pigs but they keep actin like em....honestly that's even an insult to pigs
sadness	Please bear with me, I'm not Twitter savvy ☹️ in real life I'm a Facebook person. Help me gain followers 🙏 #blog #selfhelp #depression
Fear	I'm so nervous
Joy	Love your new show @driverminnie

Tabel 4.4 Contoh Data dari *Dataset* “Emotion Dataset” oleh Abdallah Wagih

Comment	Emotion
i do not always find myself feeling thankful but over the years i ve gathered a few tricks that allow me to feel grateful in the face of moments when the last thing i want to do is say thanks	Joy
i am sitting here typing this and wondering where i belong feeling distracted feeling comfortable feeling misunderstood and hurt	Anger
i still feel its a little shaky at times and can move into the slightly odd jades hair in particular seems prone to this but generally it works well with spencers writing	Fear

Tabel 4.5 Contoh Data dari *Dataset* MELD

Dialogue ID	Utterance ID	Speaker	Utterance	Emotion	Season	Episode
1	0	Joey	But then who? The waitress I went out with last month?	Surprise	9	23
1	1	Rachel	You know? Forget it!	Sadness	9	23
1	2	Joey	No-no-no-no, no! Who, who were you talking about?	Surprise	9	23
1	3	Rachel	No, I-I-I-I don't, I actually don't know	Fear	9	23
4	0	Rachel	Hey!	Surprise	3	11
4	1	Ross	Hi!	Joy	3	11
4	2	Rachel	What are you doing here?	Surprise	3	11
4	3	Ross	Ah y'know, this building is on my paper route so I...	Neutral	3	11

4.2 Hasil *Preprocessing* Data

Setelah Pengumpulan data, Bagian ini menyajikan hasil dari serangkaian langkah *preprocessing* data yang telah dilakukan. Untuk memberikan gambaran, disajikan empat contoh data. Contoh pertama dan kedua adalah sebuah teks *tweet* yang diambil dari *dataset* "ISEAR", "Please bear with me, I'm not Twitter savvy ☹️ in real life I'm a Facebook person. Help me gain followers 🙏 #blog #selfhelp #depression." Contoh kedua, "@CorningFootball IT'S GAME DAY!!!! T MINUS 14:30 #relentless". Contoh ketiga merupakan teks komentar dari *dataset* "Emotion Dataset" oleh Abdallah Wagih, yaitu "I feel suspicious if there is no one outside like the rapture has happened or something". Terakhir, contoh dari teks dialog *dataset* MELD, "also I was the point person on my companys transition from the KL-5 to GR-6 system.". Tabel 4.6 berikut menampilkan proses *preprocessing* yang diterapkan secara bertahap pada keempat contoh teks tersebut.

Tabel 4.6 Proses *Preprocessing* Bertahap pada Berbagai Contoh Teks

No.	Langkah <i>Preprocessing</i>	<i>Dataset</i> "ISEAR" (Tweet)	<i>Dataset</i> "ISEAR" (Tweet)	<i>Dataset</i> "Emotion Dataset" (Komentar)	<i>Dataset</i> "MELD" (Dialog)

1	<i>Lowercasing</i>	please bear with me, i'm not twitter savvy ☹️ in real life i'm a facebook person. help me gain followers ☹️ #blog #selfhelp #depression.	@corningfootba ll it's game day!!!! t minus 14:30 #relentless	i feel suspicious if there is no one outside like the rapture has happened or something	also i was the point person on my companys transition from the kl-5 to gr-6 system
2	Menghapus URL, mention & Hastag	please bear with me, i'm not twitter savvy ☹️ in real life i'm a facebook person. help me gain followers ☹️ blog selfhelp depression.	corningfootball it's game day!!!! t minus 14:30 relentless	Tidak ada perubahan	Tidak ada perubahan
3	Menghapus karakter berulang	Tidak ada perubahan, karena tidak ditemukan.	corningfootball it's game day! t minus 14:30 relentless	Tidak ada perubahan	Tidak ada perubahan
4	Menghapus nonalfabet	please bear with me i m not twitter savvy in real life i m a facebook person help me gain followers blog selfhelp depression	corningfootball it s game day t minus relentless	Tidak ada perubahan	also i was the point person on my companys transition from the kl to gr system
5	Menghapus spasi berlebih	Tidak ada perubahan, karena tidak ditemukan.	corningfootball it s game day t minus relentless	Tidak ada perubahan	Tidak ada perubahan
6	Tokenisasi dan hapus <i>stopword</i>	['please', 'bear', 'twitter', 'savvy', 'real', 'life', 'facebook', 'person', 'help', 'gain',	['corningfootball', 'game', 'day', 't', 'minus', 'relentless']	['feel', 'suspicious', 'one', 'outside', 'like', 'rapture',	['also', 'point', 'person', 'companys', 'transition',

		'followers', 'blog', 'selfhelp', 'depression']		'happened', 'something']	'kl', 'gr', 'system']
7	<i>stemming</i>	['pleas', 'bear', 'twitter', 'savvi', 'real', 'life', 'facebook', 'person', 'help', 'gain', 'follow', 'blog', 'selfhelp', 'depress']	Tidak ada perubahan.	['feel', 'suspici', 'one', 'outsid', 'like', 'raptur', 'happen', 'someth']	['also', 'point', 'person', 'compani', 'transit', 'kl', 'gr', 'system']
8	Menggabungkan token	pleas bear twitter savvi real life facebook person help gain follow blog selfhelp depress	corningfootball game day t minus relentless	feel suspici one outsid like raptur happen someth	also point person compani transit kl gr system

Berdasarkan

Tabel 4.6, dapat dilihat bahwa proses *preprocessing* berhasil menghilangkan elemen-elemen yang tidak relevan seperti simbol (@, #), huruf kapital, emoji, angka, dan tanda baca. Kata-kata tidak bermakna yang termasuk dalam *stopword* (misalnya: "with", "me", "it's") juga berhasil dihapus. Selain itu, proses *stemming* berhasil menyederhanakan kata-kata ke bentuk dasarnya, misalnya kata "followers" menjadi "follow" dan "depression" menjadi "depress". Dengan demikian, data teks ini siap untuk tahap ekstraksi fitur sebelum ke pelatihan model.

4.3 Hasil Pembagian Data

Bagian ini menampilkan hasil pembagian setiap *dataset* yang digunakan dalam penelitian. Perlu diperhatikan bahwa terdapat perbedaan strategi pembagian data yang disesuaikan dengan format asli masing-masing *dataset*. Sebagian *dataset*, seperti "Emotion Dataset" oleh Parul Pandey dan MELD, telah disediakan dari sumbernya dalam format tiga bagian: data latih, data validasi, dan data uji. Penggunaan pembagian yang sudah ada ini merupakan praktik standar agar hasil penelitian dapat dibandingkan secara langsung dan konsisten dengan studi lain yang menggunakan dataset yang sama. Sementara itu, untuk

dataset yang tidak memiliki pembagian standar, penelitian ini menggunakan pembagian dua bagian: 80% data latih dan 20% data uji. Meskipun strategi pembagian bervariasi, validitas perbandingan ilmiah dalam penelitian ini tetap terjaga. Hal ini karena evaluasi akhir untuk membandingkan kinerja semua model secara seragam dilakukan pada data uji yang terpisah, yang sama sekali tidak pernah digunakan selama proses pelatihan maupun validasi. Peran data validasi adalah sebagai data terpisah yang digunakan untuk mengukur performa model selama proses pelatihan *deep learning*, memantau *overfitting*, dan membantu penyesuaian *hyperparameter*. Praktik ini merupakan standar dalam pengembangan NLP, seperti yang juga diterapkan dalam *paper* BERT (Devlin et al., 2019). Untuk *dataset* yang dibagi dengan rasio 80:20, data validasi ini secara otomatis diambil dari sebagian data latih 80% tersebut selama proses pelatihan model *deep learning*.

4.3.1 *Dataset “Emotions” oleh Elgiriye withana*

Dataset ini awalnya memiliki sekitar 400.000 data dengan persebaran label yang tidak seimbang. Oleh karena itu, dilakukan proses *downsampling* untuk menyamakan jumlah data antar kelas. Setelah itu, *dataset* dibagi menjadi data latih 80% dan uji 20%. Tabel 4.7 menampilkan persebaran data sebelum *downsampling*, setelah *downsampling*, serta hasil pembagian datanya.

Tabel 4.7 Persebaran dan Pembagian *Dataset “Emotions” oleh Elgiriye withana*

Emosi	Sebelum <i>Downsampling</i>	Setelah <i>Downsampling</i>	Latih	Uji
Sadness (0)	121187	4159	3327	832
Joy (1)	141067	4218	3375	843
Love (2)	57317	4108	3286	843
Anger (3)	47712	4073	3258	815
Fear (4)	34554	4200	3360	840
Surprise (5)	14972	4242	3394	848
Total	416809	25000	20000	5000

4.3.2 *Dataset “Emotion Dataset” oleh Parul Pandey*

Dataset ini telah disediakan dalam tiga bagian, yaitu latih, validasi dan uji. Masing-masing bagian berjumlah 16.000, 2.000, dan 2.000. Terkait persebaran emosi untuk tiap bagian ditampilkan pada Tabel 4.8.

Tabel 4.8 Persebaran Emosi *Dataset* “Emotion Dataset” oleh Parul Pandey

Emosi	Latih	Validasi	Uji
Sadness (0)	4666	550	581
Joy (1)	5362	704	695
Love (2)	1304	178	159
Anger (3)	2159	275	275
Fear (4)	1937	212	224
Surprise (5)	572	81	66
Total	16000	2000	2000

4.3.3 *Dataset* “ISEAR” oleh Faisal Santo

Dataset ini tidak memiliki pembagian data dari sumber aslinya. *Dataset* ini dibagi dengan rasio 80% data latih dan 20% data uji. Tabel 4.9 menampilkan jumlah data awal serta persebaran emosi setelah pembagian data.

Tabel 4.9 Persebaran dan Pembagian *Dataset* “ISEAR” oleh Faisal Santo

Emosi	Jumlah Asli	Latih	Uji
Joy (0)	1616	1293	323
Sadness (1)	1533	1226	307
Anger (2)	1701	1361	340
Fear (3)	2252	1801	451
Total	7102	5681	1421

4.3.4 *Dataset* “Emotion Dataset” oleh Abdallah Wagih

Dataset ini terdiri dari sekitar 6.000 data. *Dataset* ini juga dibagi dengan rasio 80:20 karena belum tersedia dari sumbernya. Tabel 4.10 menunjukkan persebaran emosi awal serta hasil pembagian data.

Tabel 4.10 Persebaran dan Pembagian *Dataset* “Emotion Dataset” oleh Abdallah Wagih

Emosi	Jumlah Data	Latih	Uji
Joy (0)	2000	1600	400
Anger (2)	2000	1600	400
Fear (3)	1937	1549	388
Total	5937	4749	1188

4.3.5 Dataset MELD

Dataset ini sudah dibagi dari sumber aslinya menjadi latih, validasi, dan uji. Dengan masing-masing jumlah per bagian yaitu latih 9.989, validasi 1.109, dan uji 2.610. Tabel 4.11 menyajikan persebaran emosi masing-masing bagian.

Tabel 4.11 Persebaran Emosi *Dataset* MELD

Emosi	Latih	Validasi	Uji
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281
Total	9989	1109	2610

4.4 Hasil Ekstraksi Fitur

Setelah pembagian data, langkah selanjutnya adalah mengubah data teks menjadi bentuk numerik melalui proses ekstraksi fitur sebelum melatih model. Berikut contoh hasil dari proses tersebut ditunjukkan pada Tabel 4.12.

Tabel 4.12 Contoh Hasil Ekstraksi Fitur

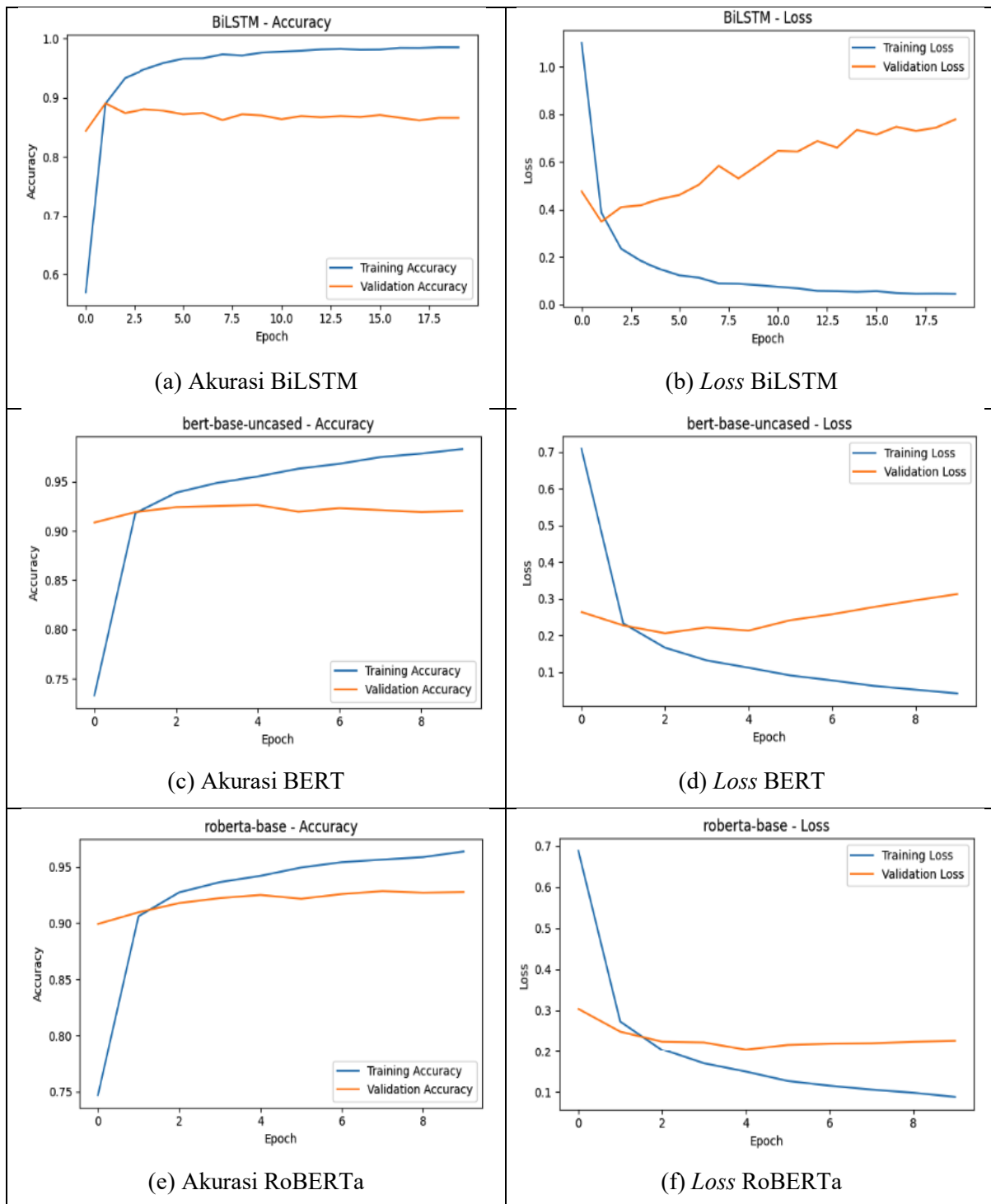
Tahap	Hasil
Teks Sebelum Ekstraksi Fitur	'sidviy uff look arsenal fan cheer everi goal unit'
Skor TF-IDF	cheer (0,3446), everi (0,3911), fan (0,4230), goal (0,4649), look (0,3237), unit (0,4780)

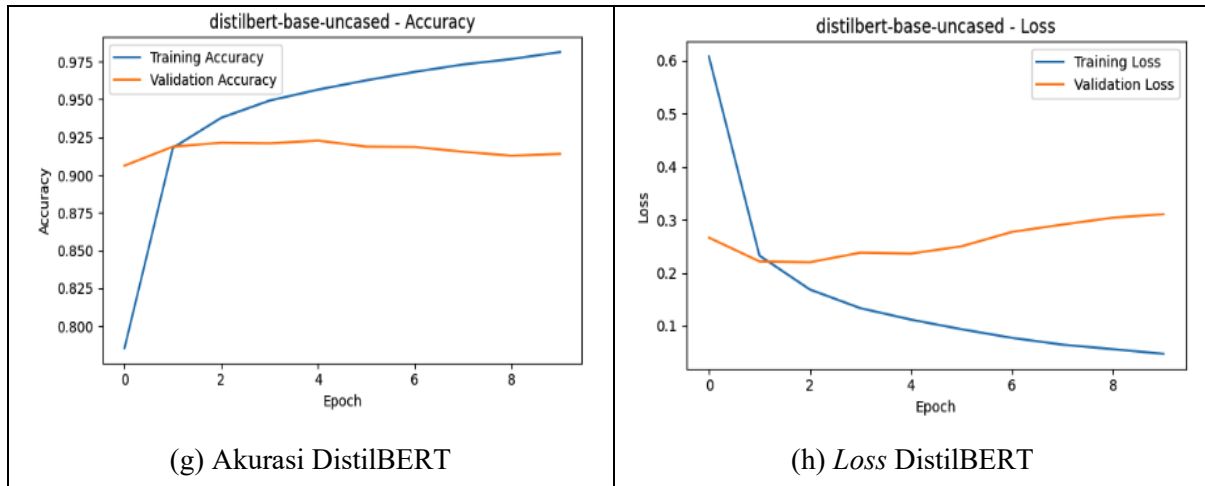
Contoh data teks yang ditampilkan pada tabel di atas diambil dari *dataset* “ISEAR” yang sudah melewati tahap *preprocessing*. Setiap skor TF-IDF kini mewakili "bobot" dari masing-masing kata.

4.5 Hasil Pelatihan Model

Hasil pelatihan model hanya terdapat pada pendekatan *deep learning* karena dilatih dalam beberapa *epoch* yang dapat dilacak selama proses pelatihan, seperti *training loss*, *validation loss*, *training accuracy*, dan *validation accuracy*. Bagian ini menyajikan hasil pelatihan dari masing-masing model *deep learning* pada setiap *dataset* yang digunakan, mulai

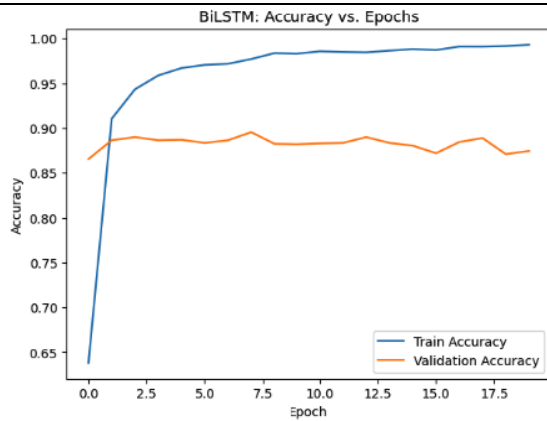
dari Gambar 4.1 hingga 4.5 dengan visualisasi grafik akurasi dan *loss*, serta analisis ada atau tidaknya *overfitting*.



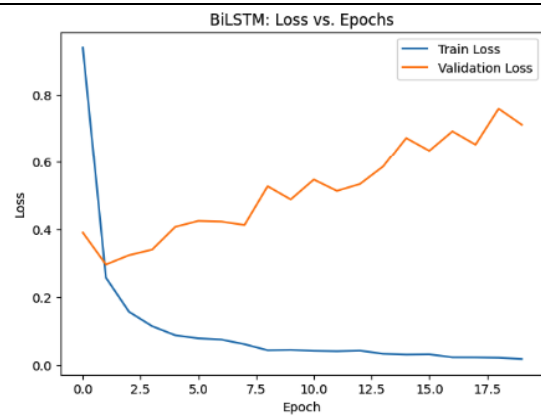


Gambar 4.1 Gambar Hasil Pelatihan Model pada *Dataset* “Emotions” oleh ElgiriyeWithana

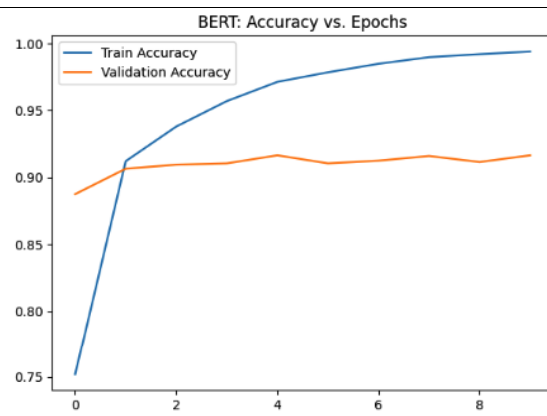
Pada *dataset* “Emotions” oleh ElgiriyeWithana, hasil pelatihan model dapat dilihat pada Gambar 4.1. Model BiLSTM (Gambar 4.1a) menunjukkan peningkatan akurasi pelatihan yang pesat, dari 36,71% pada *epoch* pertama menjadi lebih dari 98% pada *epoch* ke-20. Namun, akurasi validasi hanya naik di awal hingga puncaknya di 89,03% (*epoch* 2), lalu perlahan menurun dan stagnan di sekitar 86–87% hingga akhir pelatihan. *Validation loss* (Gambar 4.1b) juga menunjukkan pola turun tajam di awal, lalu meningkat terus-menerus dari *epoch* ke-3 ke atas, bahkan mencapai 0,77 di akhir. Pola ini adalah indikasi yang jelas bahwa model BiLSTM mengalami *overfitting*. Model BERT (Gambar 4.1c) menunjukkan pelatihan yang sehat dengan performa yang jauh lebih stabil. Akurasi pelatihan meningkat secara konsisten dari 73,30% menjadi 98,29%, sementara akurasi validasi stabil tinggi di kisaran 91–92% setelah *epoch* ke-2. Meskipun *validation loss* (Gambar 4.1d) sedikit naik setelah *epoch* ke-5 (dari 0,21 ke 0,31), performa validasi tetap cukup konsisten. Model RoBERTa (Gambar 4.1e) menunjukkan pola pelatihan sehat yang serupa dengan BERT namun sedikit lebih unggul dalam kestabilan validasi. Akurasi pelatihan meningkat dari 74,69% menjadi 96,37%, dan akurasi validasi mencapai puncak tertinggi di 92,84% pada *epoch* ke-8. *Validation loss* (Gambar 4.1f) cenderung stabil dan rendah, berkisar antara 0,20 hingga 0,30. Model DistilBERT (Gambar 4.1g) memiliki performa akurasi pelatihan meningkat dari 78,53% ke 98,14%, namun akurasi validasi cenderung menurun sedikit setelah mencapai puncak di 92,28% pada *epoch* ke-5, menjadi 91,40% di akhir pelatihan. *Validation loss* (Gambar 4.1h) juga naik perlahan dari 0,22 menjadi 0,31 menunjukkan adanya gejala *overfitting* ringan.



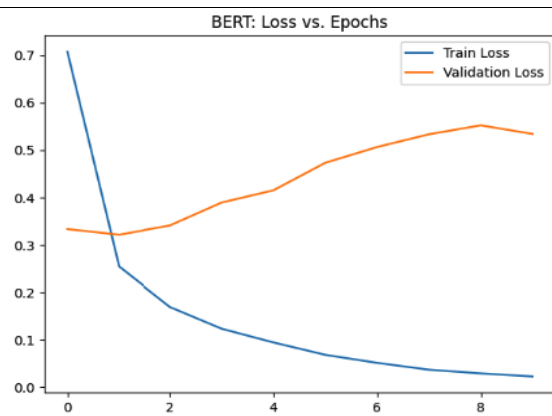
(a) Akurasi BiLSTM



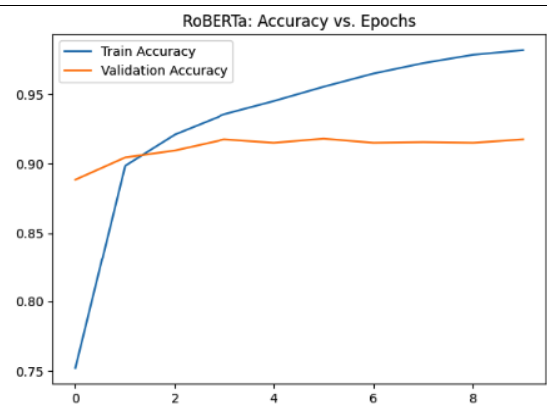
(b) Loss BiLSTM



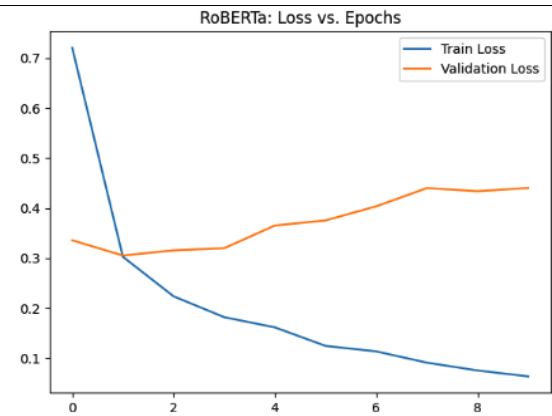
(c) Akurasi BERT



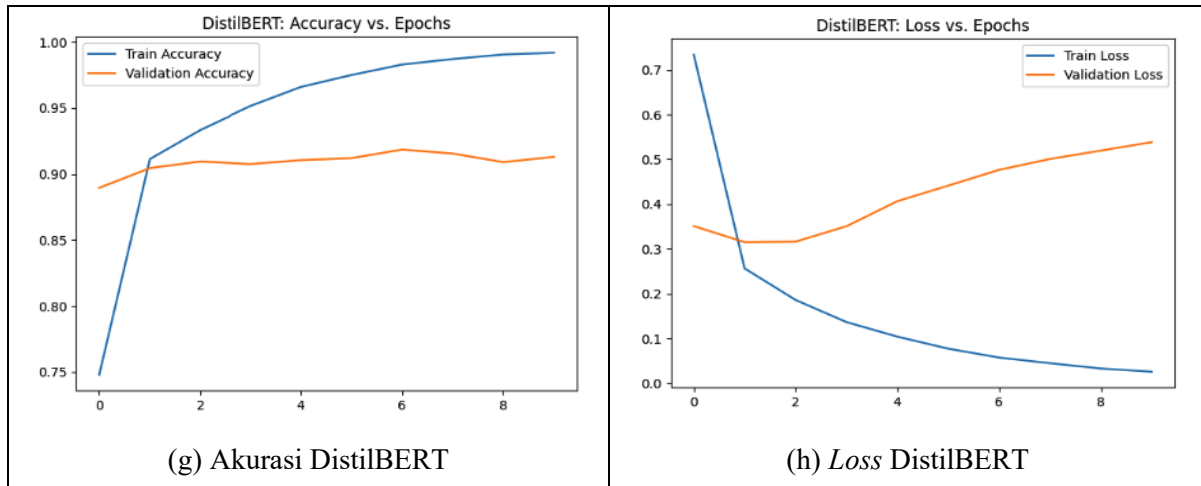
(d) Loss BERT



(e) Akurasi RoBERTa

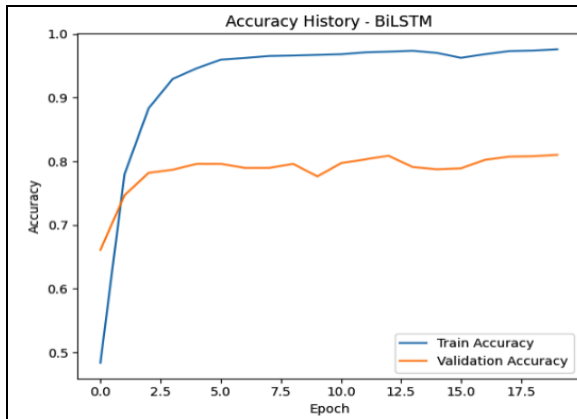


(f) Loss RoBERTa

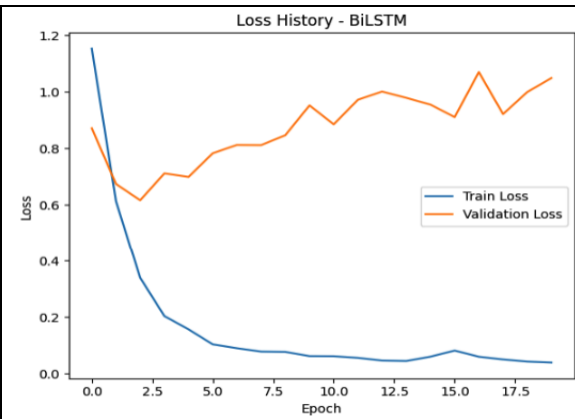


Gambar 4.2 Gambar Hasil Pelatihan Model pada *Dataset* “Emotion Dataset” oleh Parul Pandey

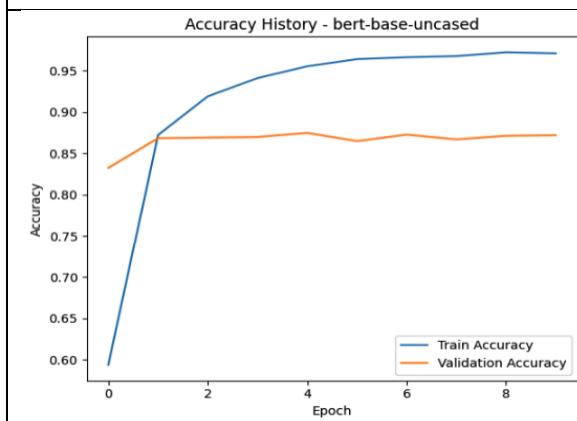
Pada *dataset* “Emotion Dataset” oleh Parul Pandey, BiLSTM (Gambar 4.2a) memiliki akurasi pelatihan meningkat pesat dari 45,30% di *epoch* pertama menjadi 99,30% pada *epoch* ke-20. Sementara itu, akurasi validasinya sempat naik dari 86,55% di *epoch* ke 1 ke puncak 89,55% di *epoch* ke 8, namun kemudian stagnan bahkan sedikit menurun. *Training loss* (Gambar 4.2b) terus turun dari 1,33 ke 0,017, namun *validation loss* justru memburuk setelah *epoch* ke-8, naik dari 0,39 menjadi 0,71 di akhir pelatihan, menunjukkan adanya gejala *overfitting*. Sebaliknya, ketiga model berbasis Transformer menunjukkan pelatihan yang sangat sehat. Model BERT (Gambar 4.2c, 4.2d) menunjukkan pelatihan yang lebih stabil. Akurasi pelatihan naik dari 75,21% di *epoch* pertama menjadi 99,40% di *epoch* ke-10. Akurasi validasi juga stabil tinggi, dimulai dari 88,75% dan mencapai 91,65% pada *epoch* ke-5 dan ke-10. *Validation loss* sempat meningkat setelah *epoch* ke-5, yaitu 0,41, namun tidak terlalu drastis. RoBERTa (Gambar 4.2e, 4.2f) menunjukkan pola serupa dengan akurasi pelatihan naik dari 75,22% menjadi 98,16%, dan akurasi validasi meningkat dari 88,85% ke 91,75% pada *epoch* terakhir. *Validation loss* sedikit naik dari 0,33 ke 0,44, namun grafik akurasi tetap stabil menunjukkan konsistensi performa antar *epoch*. DistilBERT (Gambar 4.2g, 4.2h) juga memperlihatkan pelatihan yang baik. Akurasi pelatihan meningkat dari 74,79% menjadi 99,19% dalam 10 *epoch*. Akurasi validasi naik dari 88,95% dan mencapai 91,85% di *epoch* ke-7, kemudian bertahan stabil hingga akhir. *Validation loss* juga meningkat secara perlahan dari 0,35 ke 0,53, tetapi tetap dalam batas yang wajar tanpa menimbulkan penurunan akurasi yang signifikan.



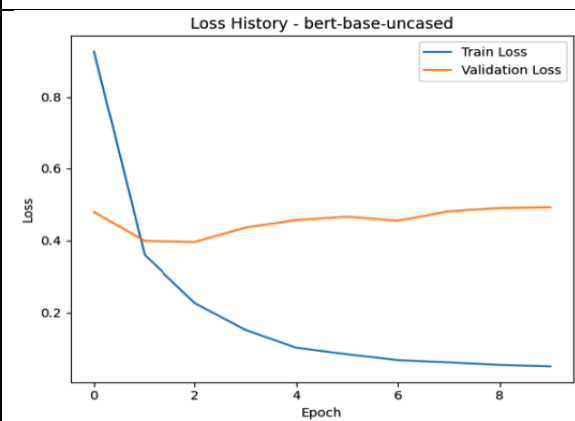
(a) Akurasi BiLSTM



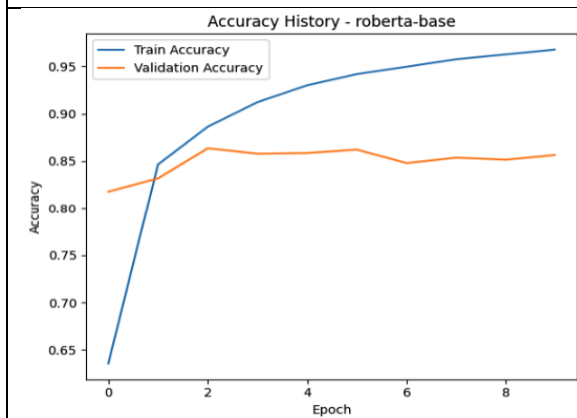
(b) Loss BiLSTM



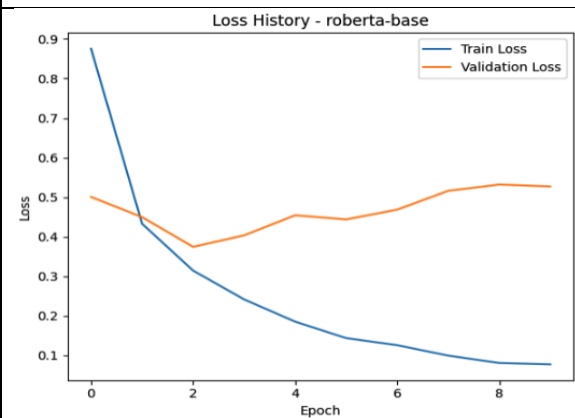
(c) Akurasi BERT



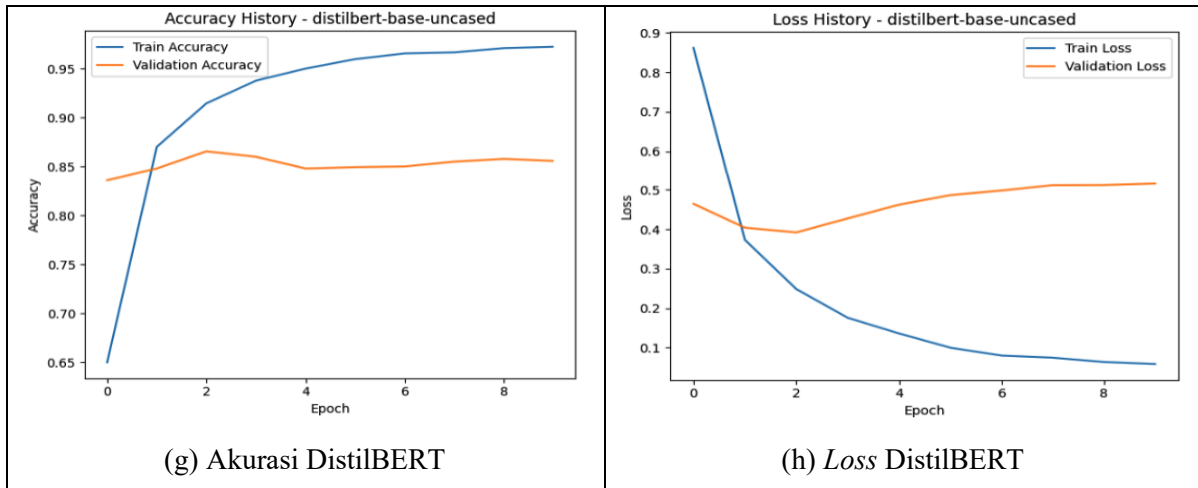
(d) Loss BERT



(e) Akurasi RoBERTa

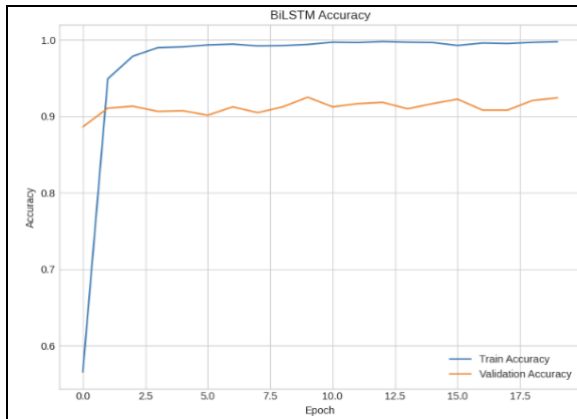


(e) Loss RoBERTa

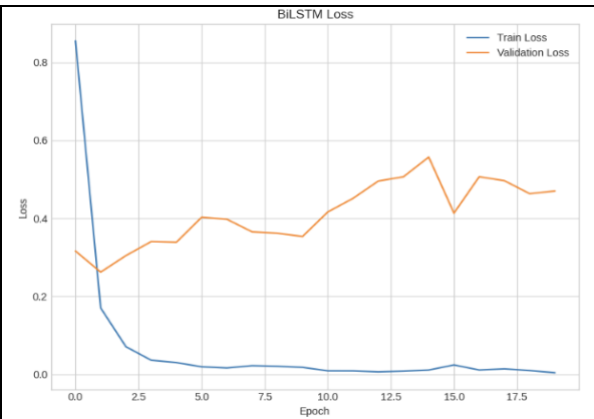
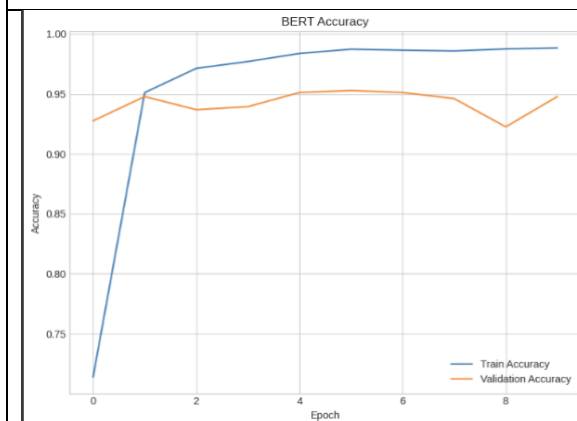


Gambar 4.3 Gambar Hasil Pelatihan Model pada *Dataset* “ISEAR” oleh Faisal Santo

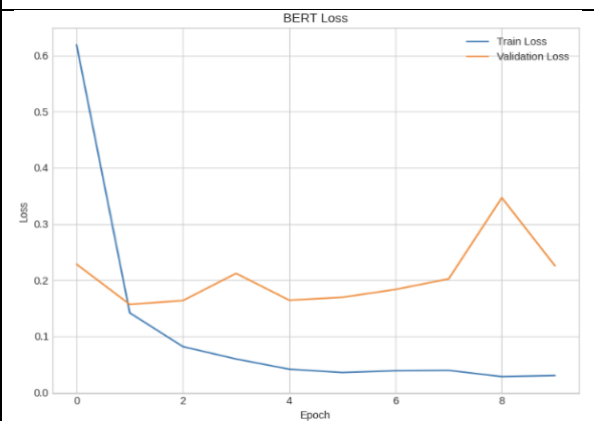
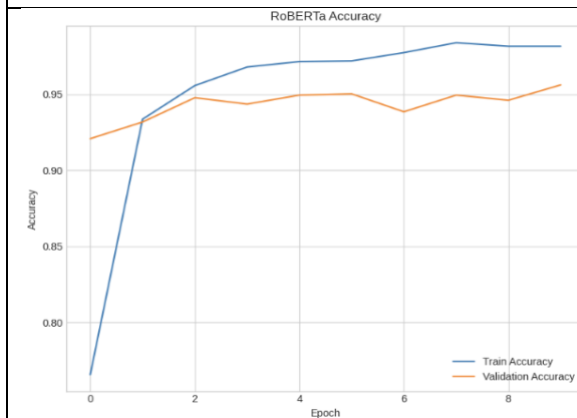
Untuk *Dataset* “ISEAR” oleh Faisal Santo, pada pelatihan model BiLSTM (Gambar 4.3a) terlihat peningkatan akurasi pelatihan dari 48,37% di *epoch* pertama hingga mencapai 97,57% pada *epoch* ke-20. Namun, akurasi validasi justru cenderung stagnan, hanya naik dari 66,08% (*epoch* 1) menjadi sekitar 81,00% di akhir pelatihan. Sementara itu, *training loss* (Gambar 4.3b) terus menurun secara signifikan, dari 1,15 menjadi 0,0387, namun *validation loss* malah meningkat dari 0,87 ke 1,04 merupakan indikator kuat adanya *overfitting*. Berbeda dengan itu, model BERT (Gambar 4.3c, 4.3d) menunjukkan pelatihan yang jauh lebih sehat. Akurasi pelatihan meningkat dari 59,37% di awal menjadi 97,08% di *epoch* ke-10, sementara akurasi validasi meningkat tajam dari 83,25% (*epoch* 1) dan stabil di kisaran 87% sejak *epoch* ke-3. *Validation loss* juga tetap relatif rendah dan konsisten, hanya naik sedikit dari 0,39 ke 0,49 menjelang akhir pelatihan. RoBERTa (Gambar 4.3e) menunjukkan peningkatan akurasi pelatihan dari 63,56% menjadi 96,76% dalam 10 *epoch*, dan akurasi validasi juga naik dari 81,77% ke puncak 86,35% di *epoch* ke-3, meskipun sempat sedikit menurun di *epoch* akhir. *Validation loss* (Gambar 4.3f) pun mengikuti tren tersebut, mencapai titik terendah 0,37 di *epoch* ke-3, lalu perlahan naik hingga sekitar 0,52. DistilBERT (Gambar 4.3g, 4.3h), akurasi pelatihan meningkat dari 64,97% menjadi 97,25%, sementara akurasi validasi naik dari 83,60% menjadi 85,78% pada *epoch* ke-9. Meskipun *validation loss* menunjukkan kenaikan bertahap dari 0,46 ke 0,51, grafik akurasi tetap cukup stabil. Model RoBERTa dan DistilBERT menunjukkan performa yang baik, namun kenaikan bertahap pada *validation loss* keduanya mengindikasikan adanya gejala *overfitting* ringan.



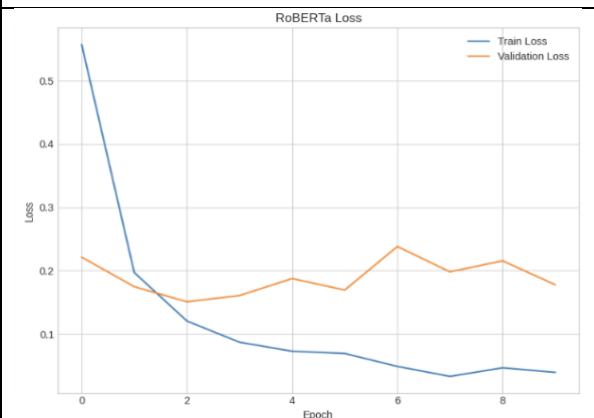
(a) Akurasi BiLSTM

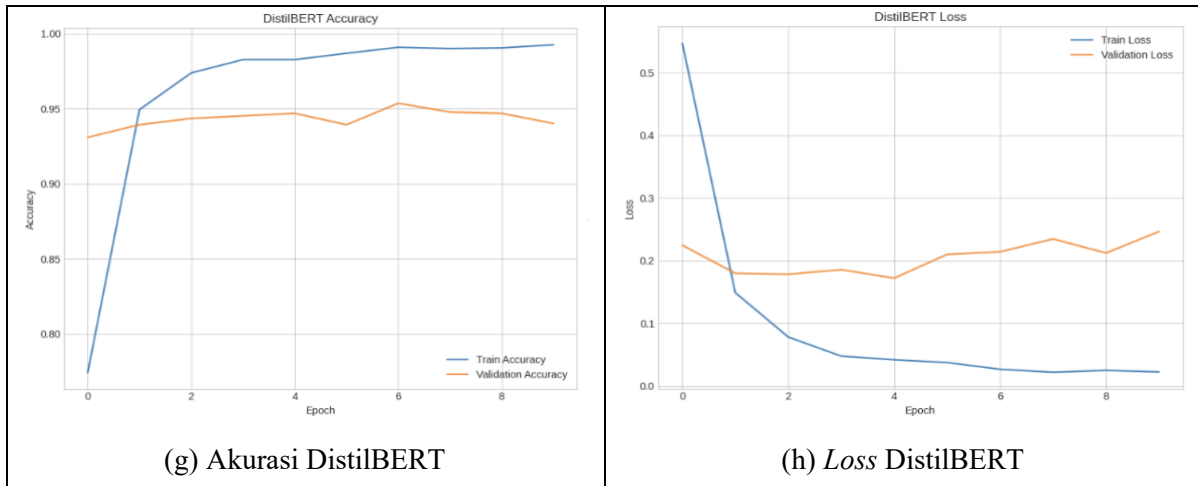
(b) *Loss* BiLSTM

(c) Akurasi BERT

(d) *Loss* BERT

(e) Akurasi RoBERTa

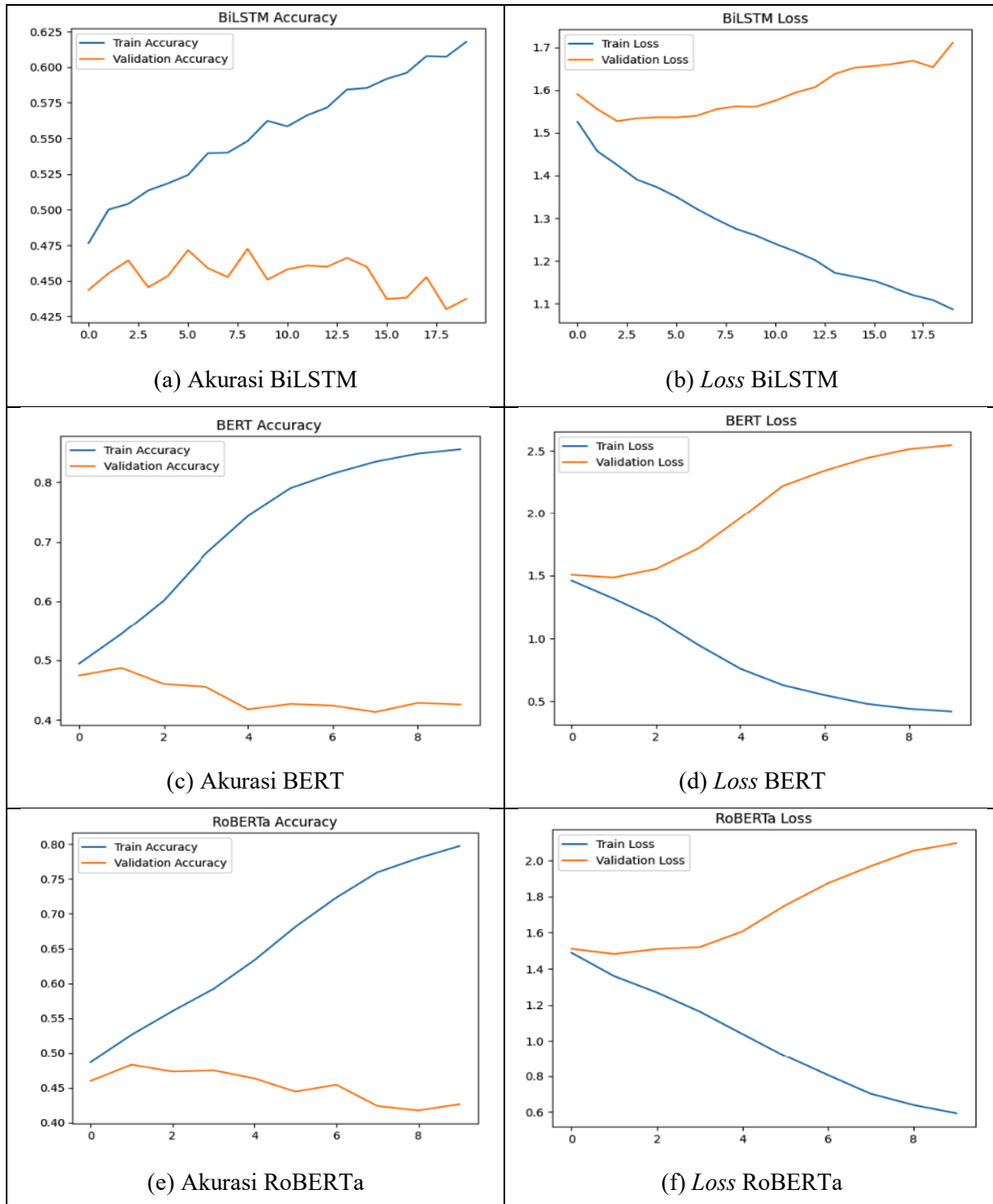
(f) *Loss* RoBERTa

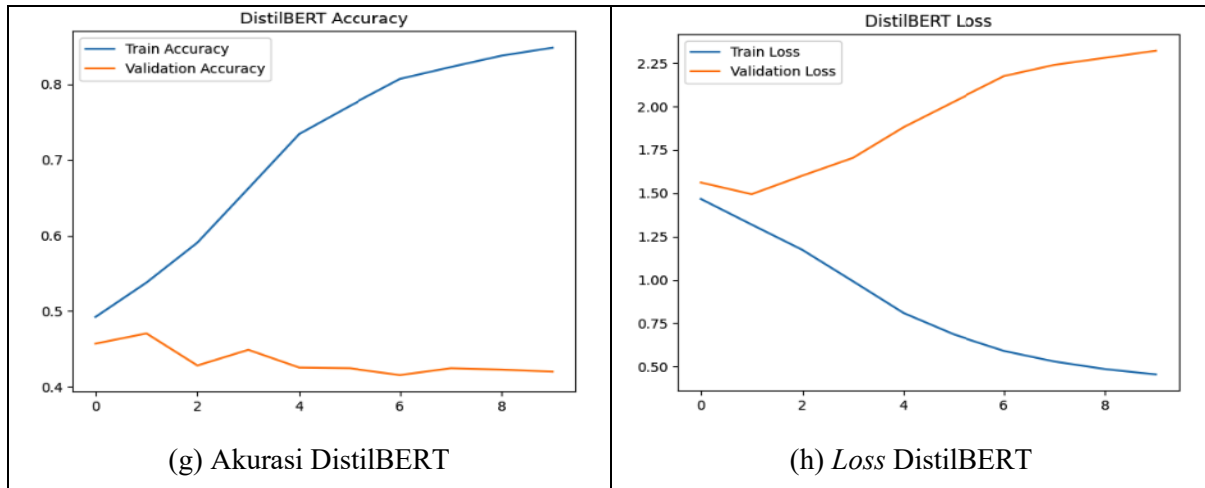


Gambar 4.4 Gambar Hasil Pelatihan Model pada *Dataset* “Emotion Dataset” oleh Abdallah Wagih

Untuk *dataset* “Emotion Dataset” oleh Abdallah Wagih, model BiLSTM (Gambar 4.4a) menunjukkan peningkatan akurasi pelatihan yang sangat cepat, dari 41,55% di *epoch* pertama menjadi 99,14% di *epoch* ke-4, dan terus meningkat hingga 99,84% di *epoch* ke-13. Namun, akurasi validasinya justru sudah tinggi sejak awal yaitu 88,64% (*epoch* 1), dan hanya naik sedikit ke sekitar 92,51% di *epoch* ke-10. Setelah itu, akurasi validasi cenderung stagnan atau fluktuatif. Sementara itu, *training loss* (Gambar 4.4b) menurun drastis dari 1,04 menjadi 0,0047 (*epoch* 20), tetapi *validation loss* yang awalnya rendah di 0,3163 (*epoch* 1) justru meningkat di tengah pelatihan hingga menyentuh 0,5571 di *epoch* ke-15. Ini menunjukkan model BiLSTM jelas mengalami *overfitting*. Di sisi lain, ketiga model berbasis Transformer menunjukkan kinerja pelatihan yang sangat baik dan sehat. Model BERT (Gambar 4.4c) memperlihatkan akurasi pelatihan meningkat dari 71,38% di awal menjadi 98,84% di *epoch* ke-10. Akurasi validasinya juga sangat tinggi sejak awal, yaitu 92,76% (*epoch* 1), dan konsisten meningkat ke puncaknya 95,29% pada *epoch* ke-6. *Validation loss* (Gambar 4.4d) pun tetap rendah di sepanjang pelatihan, berada di kisaran 0,15 hingga 0,22, meskipun sempat naik di *epoch* ke-9. RoBERTa (Gambar 4.4e) menunjukkan akurasi pelatihan dimulai dari 76,56% dan naik hingga 98,17% di *epoch* ke-10. Akurasi validasi awalnya 92,09% (*epoch* 1), lalu meningkat stabil hingga mencapai 95,62% pada *epoch* ke-10. *Validation loss* (Gambar 4.4f) paling rendah terjadi di *epoch* ke-3 dengan nilai 0,1511. Model DistilBERT (Gambar 4.4g, 4.4h) menunjukkan grafik pelatihan yang cukup mirip dengan BERT dan RoBERTa. Akurasi pelatihan naik dari 77,45% menjadi 99,26% dalam 10 *epoch*. Akurasi validasi juga tinggi sejak awal di angka 93,10% (*epoch* 1) dan naik ke 95,37% pada *epoch* ke-7. *Validation loss* cukup

stabil di sekitar 0,17 hingga 0,24. Celah yang sangat kecil antara kurva latih dan validasi pada ketiga model ini mengonfirmasi kemampuan generalisasi yang baik pada data komentar yang terstruktur.





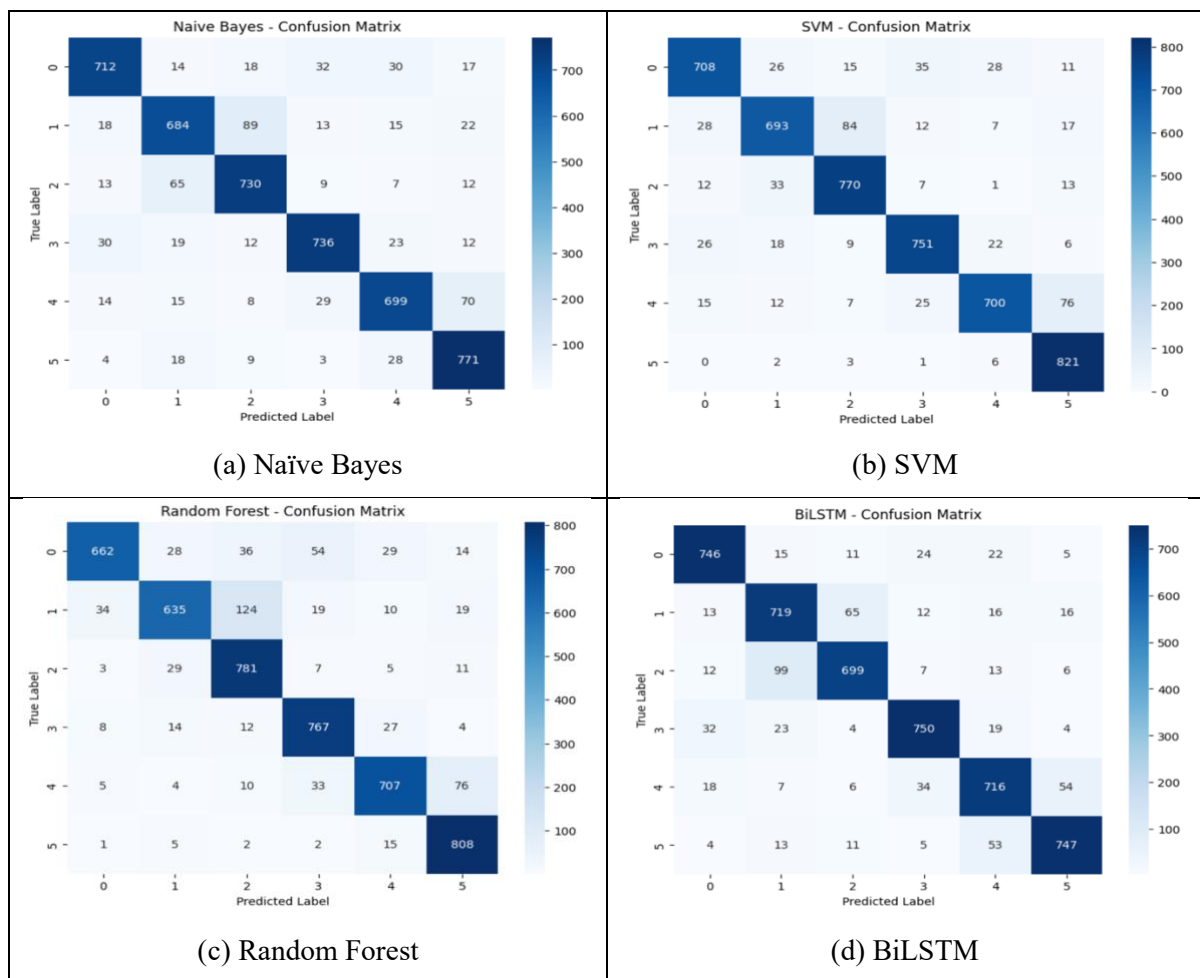
Gambar 4.5 Gambar Hasil Pelatihan Model pada *Dataset* "MELD"

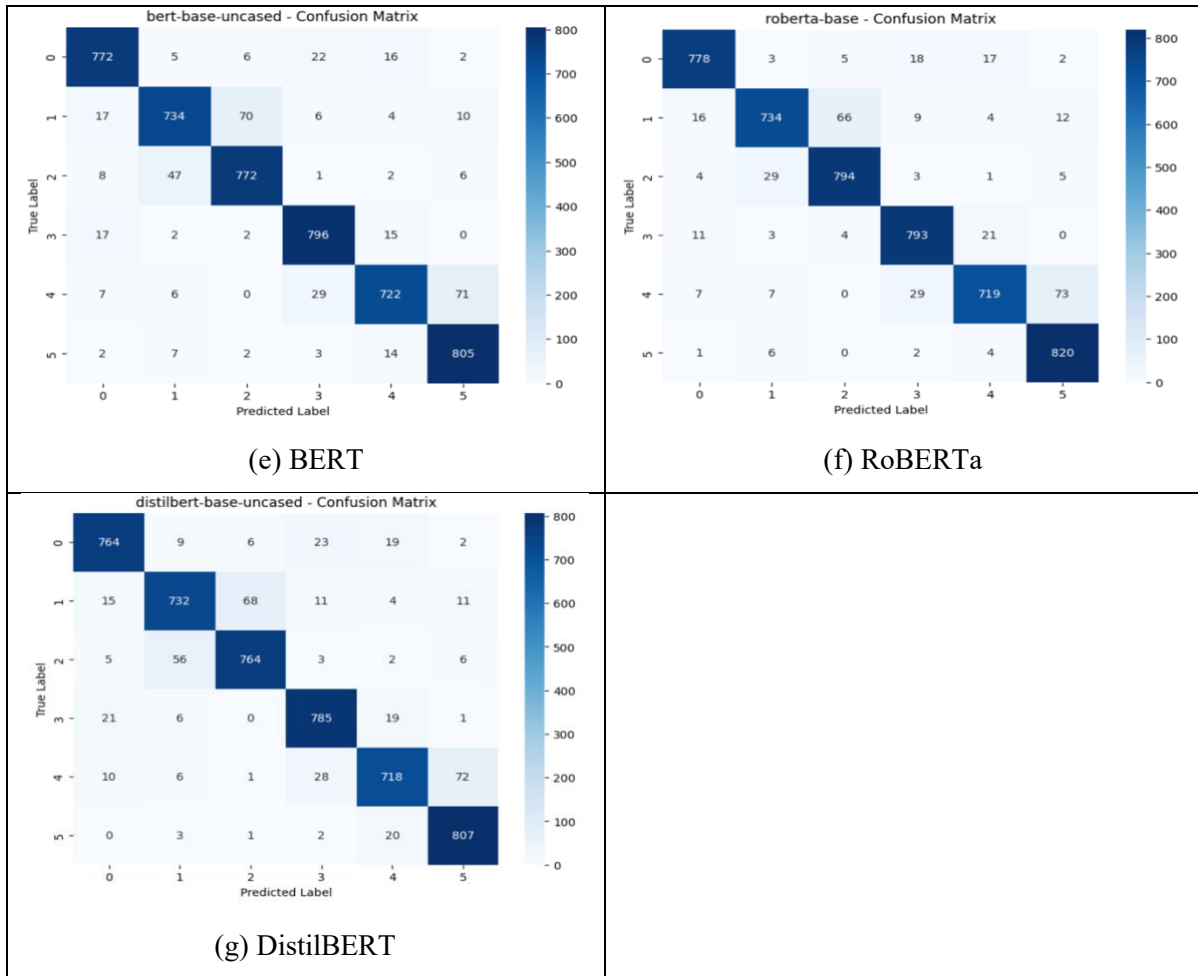
Pada *dataset* MELD, seluruh model *deep learning* menunjukkan kegagalan generalisasi yang parah dan dapat dikategorikan sebagai *overfitting*. Model BiLSTM (Gambar 4.5a, 4.5b) memperlihatkan tren peningkatan akurasi pelatihan secara perlahan dari 47,65% pada epoch pertama hingga mencapai 61,77% di akhir pelatihan (*epoch* 20). Sementara itu, akurasi validasi hanya mengalami fluktuasi kecil, berada di kisaran 43% hingga 47%, dan tidak menunjukkan peningkatan yang signifikan. *Training loss* konsisten menurun, dari 1,52 ke 1,08, namun *validation loss* justru stagnan bahkan cenderung meningkat, dari 1,59 di awal menjadi 1,71 di akhir pelatihan. Grafik ini menunjukkan bahwa model BiLSTM kesulitan dalam melakukan generalisasi terhadap data validasi, bahkan cenderung mengalami *overfitting* setelah *epoch* ke-10, ketika *validation loss* mulai naik secara konsisten meskipun *training loss* terus menurun. Model BERT (Gambar 4.5c, 4.5d) menunjukkan peningkatan performa pelatihan yang baik, dengan akurasi naik dari 49,43% di *epoch* pertama ke 85,51% di *epoch* ke-10, dan penurunan *training loss* dari 1,45 ke 0,41. Namun, akurasi validasi hanya naik sedikit dari 47,43% (*epoch* 1) ke 48,69% (*epoch* 2), lalu menurun secara konsisten hingga menyentuh 42,56% di akhir pelatihan. *Validation loss* juga memburuk, naik dari 1,50 ke 2,54. Model RoBERTa (Gambar 4.5e, 4.5f) menampilkan tren yang hampir identik dengan BERT. Akurasi pelatihan meningkat dari 48,69% (*epoch* 1) ke 79,73% (*epoch* 10), namun akurasi validasi hanya mencapai puncaknya sebesar 48,33% di *epoch* ke-2, lalu menurun secara bertahap hingga 42,65%. *Training loss* turun drastis dari 1,48 menjadi 0,59, tetapi *validation loss* naik dari 1,51 menjadi 2,09. Model DistilBERT (Gambar 4.5g, 4.5h) pun menunjukkan gejala yang serupa. Akurasi pelatihan meningkat dari 49,24% menjadi 84,78% pada akhir pelatihan, sedangkan akurasi

validasi justru turun dari 45,72% menjadi 42,02%. *Training loss* berkurang dari 1,46 menjadi 0,45, tetapi *validation loss* meningkat tajam dari 1,56 menjadi 2,32.

4.6 Hasil Evaluasi Model

Setelah *preprocessing* data, pembagian data, ekstraksi fitur, dan pelatihan, seluruh model dievaluasi dari masing-masing *dataset*. Bagian ini menyajikan *confusion matrix* dari masing-masing model per *dataset* serta hasil kuantitatif dari kinerja setiap model berdasarkan metrik evaluasi yang telah ditentukan. Hasil disajikan dalam kelompok berdasarkan jenis teks untuk memudahkan analisis.

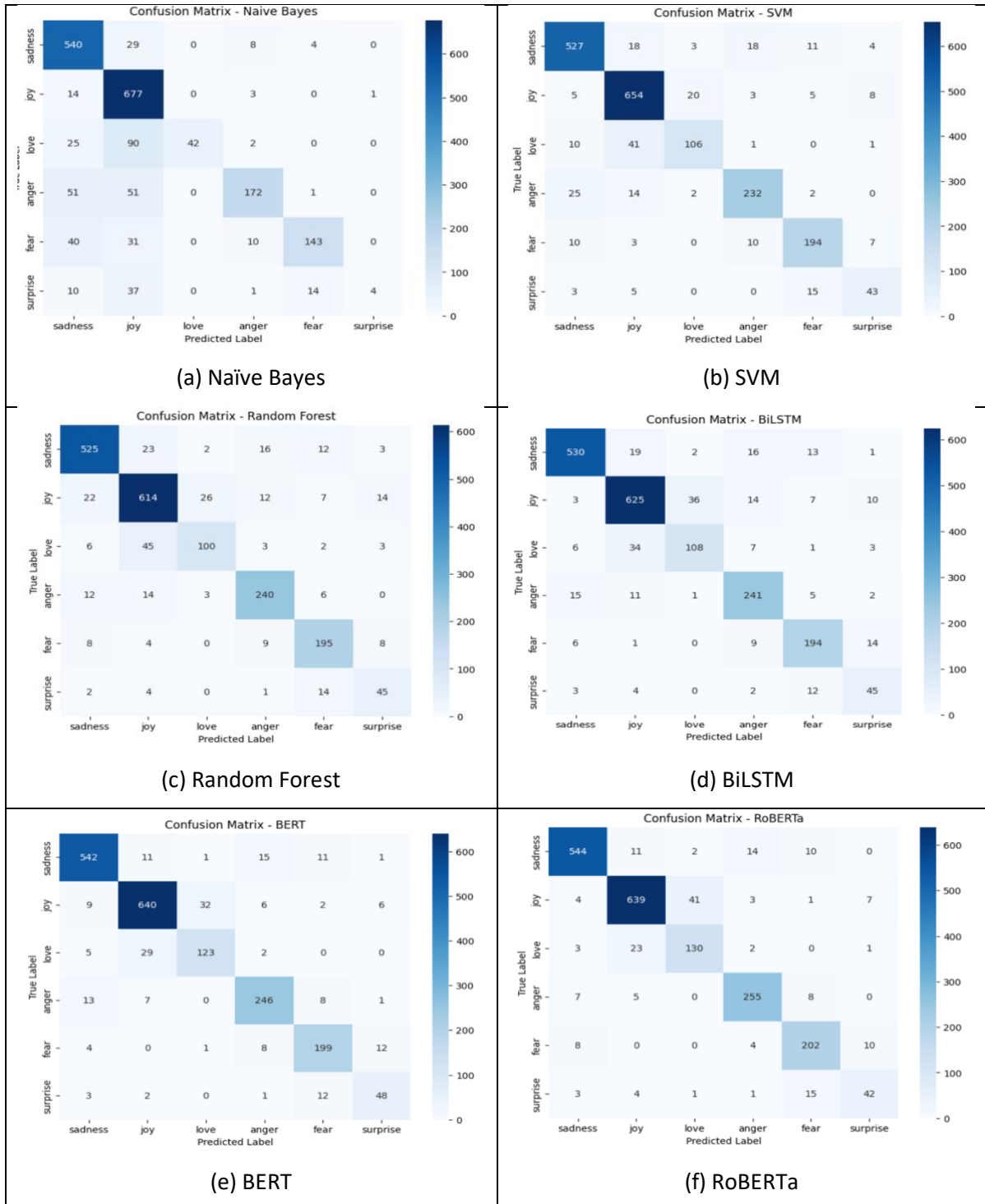


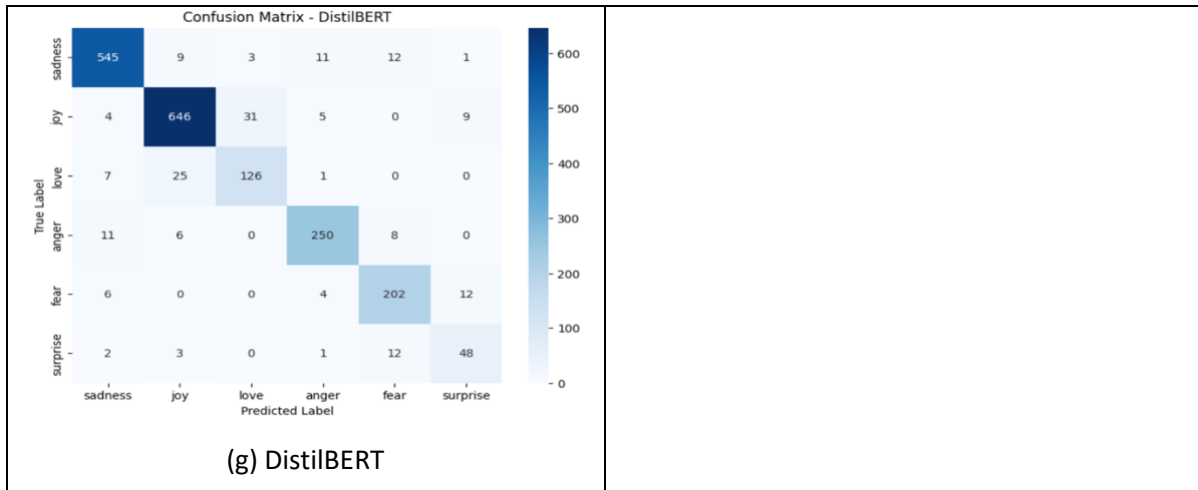


Gambar 4.6 *Confusion Matrix* Model pada *Dataset "Emotions"* oleh Elgiryewithana

Dari *confusion matrix* pada *dataset "Emotions"* oleh Elgiryewithana pada Gambar 4.6, secara umum terlihat bahwa seluruh model mampu melakukan klasifikasi dengan baik, yang ditandai oleh nilai-nilai yang tinggi pada diagonal utama. Model-model *deep learning* dengan arsitektur Transformer (BERT, RoBERTa, DistilBERT) secara visual menunjukkan matriks yang lebih bersih dengan kesalahan klasifikasi yang lebih sedikit dibandingkan dengan BiLSTM dan model *machine learning* (Naive Bayes, SVM, Random Forest). Meskipun demikian, hampir seluruh model menunjukkan pola kesalahan yang sama, yaitu sering terjadi kekeliruan antara kelas 1 dan kelas 2, serta antara kelas 4 dan kelas 5. Sebagai contoh, model RoBERTa (Gambar 4.6f) yang menunjukkan *F1-score* tertinggi, berdasarkan *confusion matrix*-nya dapat dilihat bahwa nilai pada diagonal utama sangat dominan, yang menandakan akurasi tinggi di setiap kelas. Misalnya, 820 sampel untuk kelas 5 dan 794 sampel untuk kelas 2 berhasil diprediksi dengan benar. Meskipun demikian, masih terdapat beberapa kesalahan klasifikasi seperti, 73 sampel yang seharusnya berada di kelas 4 salah diklasifikasikan oleh

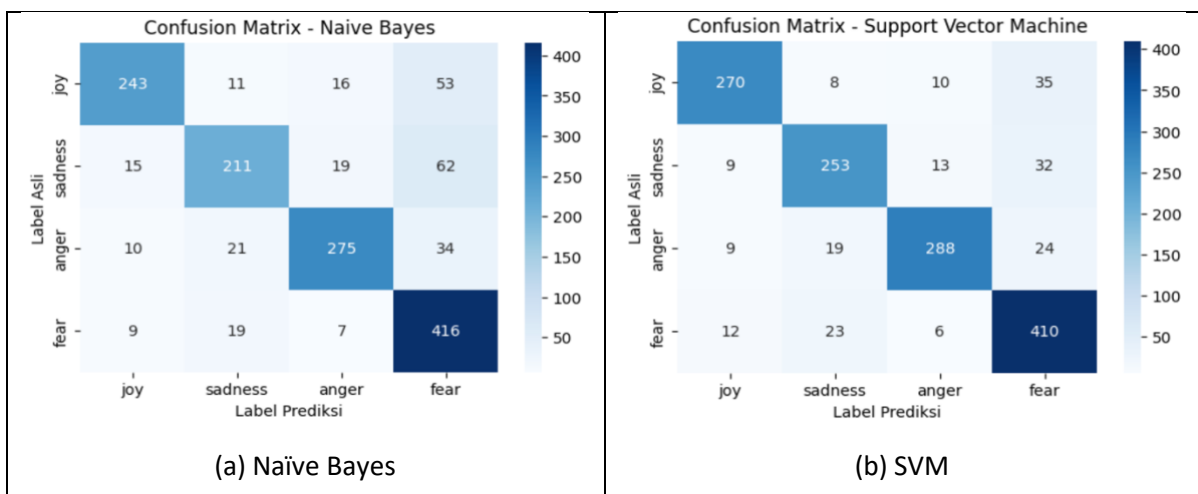
model sebagai kelas 5. Demikian pula, 66 sampel dari kelas 1 salah diklasifikasikan sebagai kelas 2.

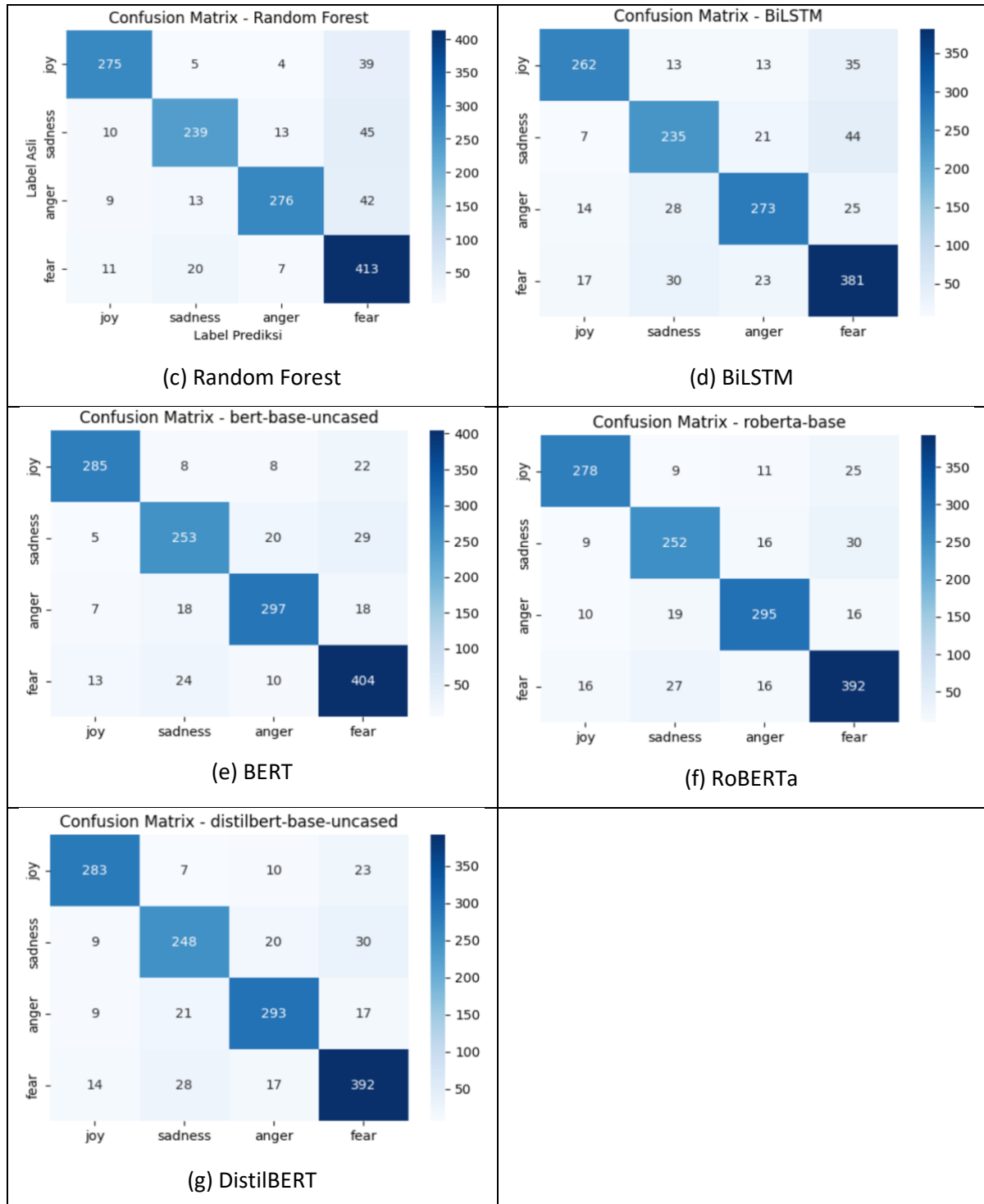




Gambar 4.7 *Confusion Matrix* Model pada *Dataset* “Emotion Dataset” oleh Parul Pandey

Untuk *Confusion Matrix* pada *dataset* “Emotion Dataset” oleh Parul Pandey pada Gambar 4.7. Pola kesalahan yang paling menonjol pada *dataset* ini adalah kebingungan antara kelas *joy* dan *love* di semua model yang dapat dilihat dari nilai prediksi benar (diagonal) yang relatif lebih rendah dibandingkan kelas lainnya. Model DistilBERT (Gambar 4.7g), yang mencapai *F1-score* tertinggi pada *dataset* ini, menunjukkan bahwa model ini salah mengklasifikasikan 31 sampel *joy* sebagai *love* dan 25 sampel *love* sebagai *joy*. Kesalahan lainnya terjadi antara *fear* dan *surprise*. Meskipun demikian, DistilBERT menunjukkan performa yang sangat baik pada kelas *sadness* dengan 545 prediksi benar dan *anger* dengan 250 prediksi benar.

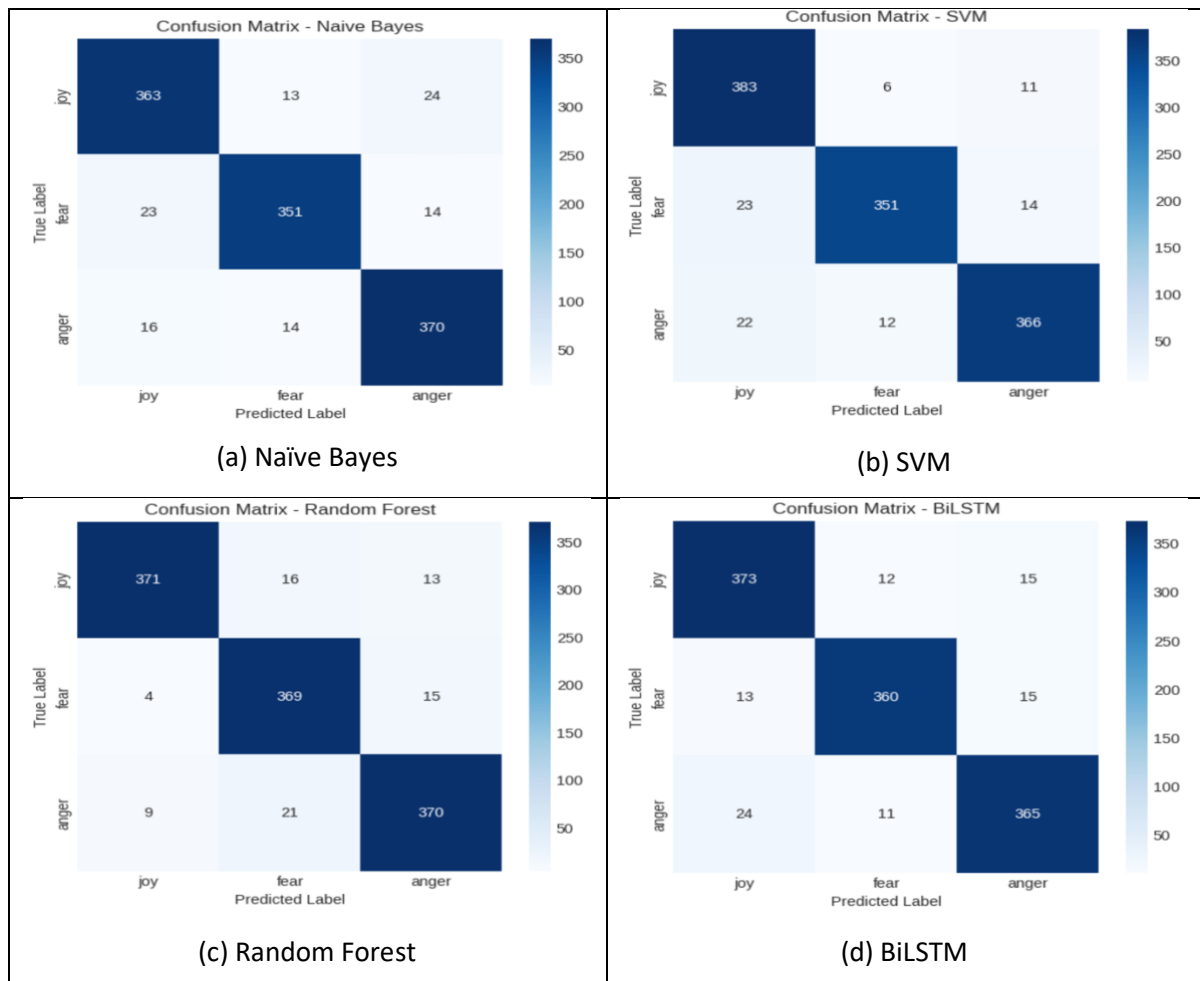


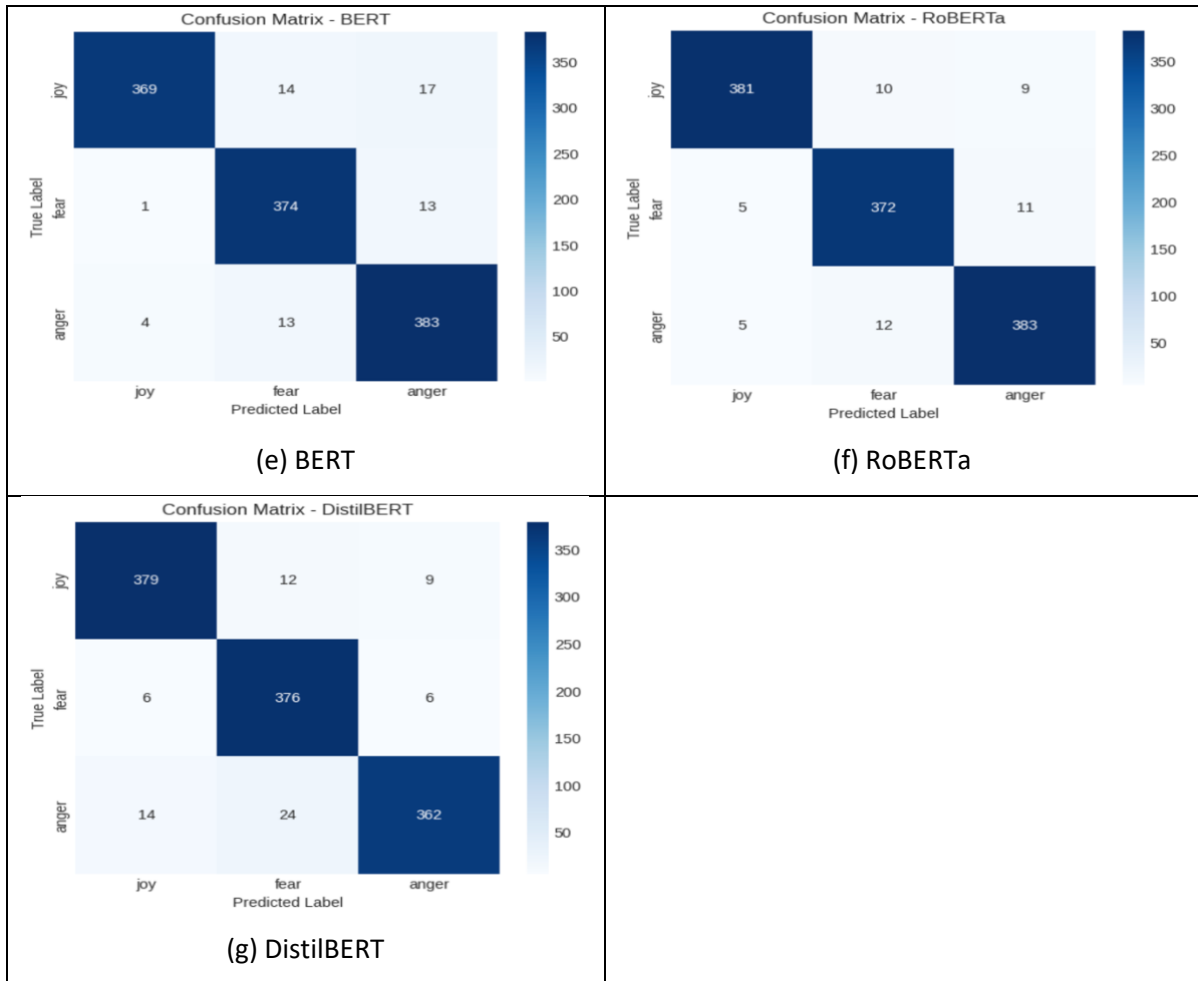


Gambar 4.8 *Confusion Matrix* Model pada *Dataset* “ISEAR” oleh Faisal Santo

Confusion Matrix pada *dataset* ISEAR Pada Gambar 4.8 menunjukkan seluruh model memiliki performa yang bagus. Secara umum, kesalahan klasifikasi yang paling sering muncul adalah antara kelas *sadness* dan *fear*. Banyak model yang keliru mengklasifikasikan *fear* sebagai *sadness* atau sebaliknya. Selain itu, kelas *joy* juga seringkali keliru diklasifikasikan

sebagai *fear* oleh model-model *machine learning*, seperti yang terlihat jelas pada model Naive Bayes dengan 53 kesalahan dan Random Forest dengan 39 kesalahan. Model BERT (Gambar 4.8e), yang meraih *F1-score* tertinggi, menunjukkan matriks yang sangat 'bersih' dengan kesalahan minimal. Terlihat pola kesalahan umum tetap ada, jumlahnya jauh lebih kecil. Misalnya, BERT hanya salah mengklasifikasikan 24 sampel *fear* sebagai *sadness* dan 29 sampel *sadness* sebagai *fear*. Performa terbaiknya terlihat pada klasifikasi kelas *anger* dengan 297 prediksi benar dan *fear* dengan 404 prediksi benar.

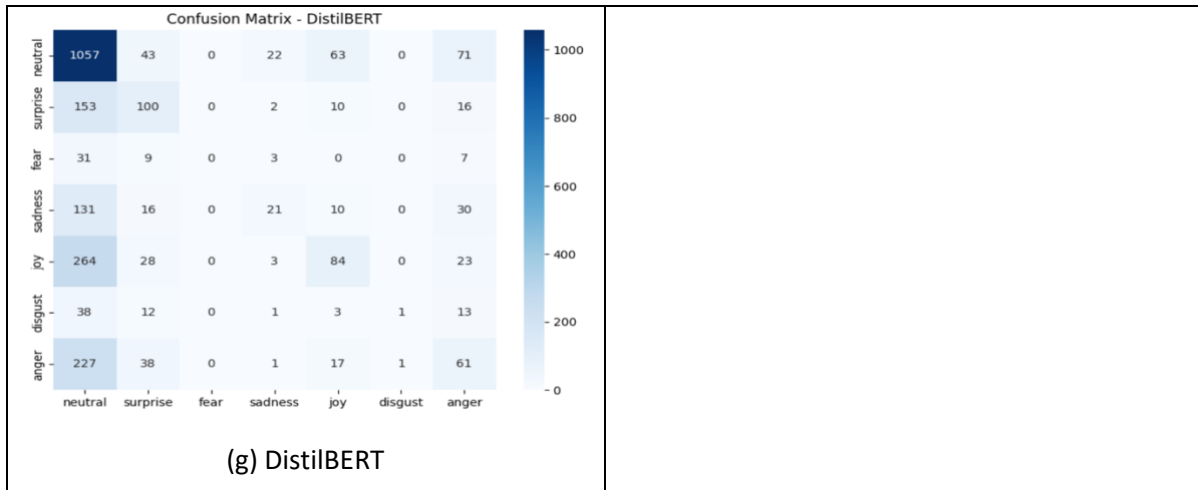




Gambar 4.9 *Confusion Matrix* Model pada *Dataset* “Emotion Dataset” oleh Abdallah Wagih

Pada *Confusion Matrix* *dataset* "Emotion Dataset" oleh Abdallah Wagih Gambar 4.9, seluruh model menunjukkan kinerja yang luar biasa tinggi. Hal ini terlihat jelas dari *confusion matrix* masing-masing model dengan nilai pada diagonal utama yang sangat dominan dan jumlah kesalahan klasifikasi di luar diagonal yang sangat minim. Model RoBERTa (Gambar 4.9f), yang menjadi model dengan *F1-score* tertinggi, dari *confusion matrix*nya terlihat bahwa model ini mampu mengklasifikasikan hampir semua sampel dengan benar, seperti 381 prediksi benar untuk *joy*, 372 untuk *fear*, dan 383 untuk *anger*. Kesalahan yang terjadi sangat kecil dan tersebar. Kesalahan terbesar yang tercatat hanya 12 sampel *anger* yang keliru diklasifikasikan sebagai *fear*.





Gambar 4.10 *Confusion Matrix* Model pada *Dataset* “MELD”

Dataset MELD, yang terdiri dari teks dialog, terbukti menjadi tantangan terbesar bagi semua model yang diuji. Dapat dilihat pada *confusion matrix* Gambar 4.10, di mana terdapat pola kesalahan yang sangat dominan. Seluruh model menunjukkan kecenderungan kuat untuk salah mengklasifikasikan hampir semua emosi sebagai *neutral*. Selain itu, banyak emosi juga keliru diklasifikasikan sebagai *joy*. Rendahnya nilai pada diagonal utama untuk kelas selain *neutral* mengonfirmasi bahwa performa model secara keseluruhan sangat rendah pada *dataset* ini.

Model BERT (Gambar 4.10e), menjadi yang terbaik dengan *F1-score* 0,29. *Confusion matrix*nya menunjukkan bahwa meskipun mampu memprediksi kelas *neutral* dengan benar sebanyak 1.045 kali, model ini juga banyak salah mengklasifikasikan emosi lain. Contohnya, 233 sampel *joy*, 230 sampel *anger*, 155 sampel *surprise*, dan 135 sampel *sadness* semuanya keliru diprediksi sebagai *neutral*. Selain itu, 130 sampel *neutral* salah diklasifikasikan sebagai *joy*.

4.6.1 Hasil Evaluasi Model pada *Dataset* Teks Tweet

Bagian ini menyajikan hasil perbandingan kinerja seluruh model pada tiga *dataset* yang berkarakteristik teks *tweet*. Mulai dari *dataset* “Emotions” oleh Elgiriye withana, *dataset* “Emotion Dataset” oleh Parul Pandey, dan *dataset* “ISEAR” oleh Faisal Santo. Tabel 4.13 merangkum hasil evaluasi untuk kelompok *dataset* ini.

Tabel 4.13 Hasil Evaluasi pada *Dataset* Teks Tweet

<i>Dataset</i>	Model	Akurasi	Presisi	<i>Recall</i>	<i>F1-score</i>	Waktu Inferensi
“Emotions” oleh Elgiryewithana	NB	0,87	0,87	0,87	0,87	0,001s
	SVM	0,89	0,89	0,89	0,89	5,250s
	RF	0,87	0,87	0,87	0,87	0,317s
	BiLSTM	0,88	0,88	0,88	0,88	2,307s
	BERT	0,92	0,92	0,92	0,92	41,602s
	RoBERTa	0,93	0,93	0,93	0,93	36,602s
	DistilBERT	0,91	0,91	0,91	0,91	20,955s
“Emotion Dataset” oleh Parul Pandey	NB	0,79	0,85	0,58	0,62	0,001s
	SVM	0,88	0,84	0,81	0,82	0,878s
	RF	0,86	0,81	0,81	0,81	0,860s
	BiLSTM	0,87	0,81	0,82	0,81	0,872s
	BERT	0,90	0,85	0,86	0,85	15,115s
	RoBERTa	0,91	0,85	0,86	0,85	13,267s
	DistilBERT	0,91	0,86	0,87	0,86	7,638s
“ISEAR” oleh Faisal Santo	NB	0,81	0,82	0,79	0,80	0,001s
	SVM	0,86	0,86	0,86	0,86	0,415s
	RF	0,85	0,85	0,85	0,85	0,084s
	BiLSTM	0,79	0,79	0,79	0,79	0,169s
	BERT	0,87	0,87	0,87	0,87	10,855s
	RoBERTa	0,86	0,86	0,86	0,86	9,622s
	DistilBERT	0,86	0,86	0,85	0,85	5,636s

Dari tabel di atas, untuk *dataset* “Emotions” oleh Elgiryewithana, model RoBERTa menunjukkan kinerja tertinggi berdasarkan *F1-score*, yaitu 0,93. Pada *dataset* “Emotion Dataset” oleh Parul Pandey, model DistilBERT mendapatkan *F1-score* tertinggi dengan capaian 0,86. Untuk *dataset* “ISEAR” oleh Faisal Santo, BERT dengan *F1-score* 0,87 mencapai tertinggi diantara model-model yang lain. Sementara itu, dari segi waktu inferensi, model Naive Bayes unggul telak dengan waktu inferensi tercepat di semua *dataset*.

4.6.2 Hasil Evaluasi Model pada *Dataset* Teks Komentar

Selanjutnya, kinerja model diuji pada *dataset* teks komentar yaitu *dataset* “Emotion Dataset” oleh Abdallah Wagih. Berbeda dengan teks *tweet*, dataset ini memiliki karakteristik teks yang cenderung lebih panjang dan bersifat naratif. Tabel 4.14 menyajikan hasil evaluasi pada *dataset* tersebut.

Tabel 4.14 Hasil Evaluasi pada *Dataset* Teks Komentar

<i>Dataset</i>	Model	Akurasi	Presisi	<i>Recall</i>	<i>F1-score</i>	Waktu Inferensi
“Emotion Dataset” oleh Abdallah Wagih	NB	0,91	0,91	0,91	0,91	0,001s
	SVM	0,93	0,93	0,93	0,93	1,048s
	RF	0,93	0,93	0,93	0,93	0,116s
	BiLSTM	0,92	0,92	0,92	0,92	0,978s
	BERT	0,95	0,95	0,95	0,95	3,937s
	RoBERTa	0,96	0,96	0,96	0,96	3,101s
	DistilBERT	0,94	0,94	0,94	0,94	1,994s

Pada *dataset* teks komentar, seluruh model yang diuji menunjukkan performa yang sangat tinggi, dengan *F1-score* terendah berada di angka 0,91. RoBERTa mencapai *F1-score* tertinggi 0,96, diikuti oleh BERT 0,95, dan DistilBERT 0,94. Dari sisi waktu inferensi, Naive Bayes tetap paling unggul.

4.6.3 Hasil Evaluasi Model pada *Dataset* Teks Dialog

Bagian terakhir menyajikan hasil evaluasi pada *dataset* teks dialog, yaitu *dataset* MELD. *Dataset* ini merepresentasikan tantangan yang berbeda karena sifatnya yang percakapan dan memerlukan pemahaman konteks. Tabel 4.15 menunjukkan hasil kinerja model pada *dataset* ini.

Tabel 4.15 Hasil Evaluasi pada *Dataset* Teks Dialog

<i>Dataset</i>	Model	Akurasi	Presisi	<i>Recall</i>	<i>F1-score</i>	Waktu Inferensi
MELD	NB	0,51	0,41	0,17	0,15	0,001s
	SVM	0,52	0,44	0,23	0,25	1,410s
	RF	0,50	0,35	0,23	0,25	0,122s
	BiLSTM	0,51	0,31	0,23	0,23	0,291s
	BERT	0,52	0,45	0,27	0,29	22,164s
	RoBERTa	0,53	0,43	0,25	0,27	19,325s
	DistilBERT	0,51	0,37	0,24	0,25	11,149s

Pada *dataset* MELD, *F1-score* tertinggi didapat oleh model BERT dengan nilai 0,29, disusul oleh RoBERTa sebesar 0,27. Selisih performa antara model *machine learning* dan *deep learning* pada *dataset* ini sangat tipis dan model dengan waktu inferensi terunggul tetap Naive Bayes.

4.7 Pembahasan

Bagian ini akan membahas temuan-temuan yang telah disajikan pada hasil evaluasi. Pembahasan akan difokuskan untuk membahas kinerja model, menganalisis pengaruh karakteristik data, efek *preprocessing*, dan analisis kesalahan (*error analysis*). Untuk perbandingan kinerja di seluruh pembahasan, *F1-score* akan digunakan sebagai metrik evaluasi utama. Dalam tugas klasifikasi teks, terutama untuk deteksi emosi, penggunaan *F1-score* sebagai metrik utama telah umum digunakan di berbagai studi (Alaeddini, 2024; Cortiz, 2021; Imran, 2024), sehingga menjadikannya tolok ukur yang kuat untuk analisis dalam penelitian ini. Sebelum masuk ke analisis mendalam, Tabel 4.16 menyajikan rekapitulasi hasil pengujian akurasi dan waktu inferensi dari seluruh model dan *dataset*.

Tabel 4.16 Rekapitulasi Hasil Pengujian Model

<i>Dataset</i>	Model	Akurasi	Waktu Inferensi
“Emotions” oleh ElgiriyeWithana	NB	0,87	0,001s
	SVM	0,89	5,250s
	RF	0,87	0,317s
	BiLSTM	0,88	2,307s
	BERT	0,92	41,602s
	RoBERTa	0,93	36,602s
	DistilBERT	0,91	20,955s
“Emotion Dataset” oleh Parul Pandey	NB	0,79	0,001s
	SVM	0,88	0,878s
	RF	0,86	0,860s
	BiLSTM	0,87	0,872s
	BERT	0,90	15,115s
	RoBERTa	0,91	13,267s
	DistilBERT	0,91	7,638s
“ISEAR” oleh Faisal Santo	NB	0,81	0,001s
	SVM	0,86	0,415s
	RF	0,85	0,084s
	BiLSTM	0,79	0,169s
	BERT	0,87	10,855s
	RoBERTa	0,86	9,622s

	DistilBERT	0,86	5,636s
“Emotion Dataset” oleh Abdallah Wagih	NB	0,91	0,001s
	SVM	0,93	1,048s
	RF	0,93	0,116s
	BiLSTM	0,92	0,978s
	BERT	0,95	3,937s
	RoBERTa	0,96	3,101s
	DistilBERT	0,94	1,994s
MELD	NB	0,51	0,001s
	SVM	0,52	1,410s
	RF	0,50	0,122s
	BiLSTM	0,51	0,291s
	BERT	0,52	22,164s
	RoBERTa	0,53	19,325s
	DistilBERT	0,51	11,149s

Berdasarkan rekapitulasi hasil pengujian pada Tabel 4.16, terlihat adanya variasi kinerja antar model baik dari sisi akurasi maupun waktu inferensi. Untuk memberikan gambaran yang lebih jelas mengenai performa antar model, Tabel 4.17 menyajikan peringkat model berdasarkan akurasi tertinggi pada masing-masing *dataset*, sedangkan Tabel 4.18 menyajikan peringkat model berdasarkan waktu inferensi.

Tabel 4.17 Peringkat Model Berdasarkan Akurasi

<i>Dataset</i>	Model	Akurasi
“Emotions” oleh ElgiriyeWithana	RoBERTa	0,93
	BERT	0,92
	DistilBERT	0,91
	SVM	0,89
	BiLSTM	0,88
	NB	0,87
	RF	0,87
“Emotion Dataset” oleh Parul Pandey	DistilBERT	0,91
	RoBERTa	0,91
	BERT	0,90

	SVM	0,88
	BiLSTM	0,87
	RF	0,86
	NB	0,79
“ISEAR” oleh Faisal Santo	BERT	0,87
	RoBERTa	0,86
	DistilBERT	0,86
	SVM	0,86
	RF	0,85
	NB	0,81
	BiLSTM	0,79
“Emotion Dataset” oleh Abdallah Wagih	RoBERTa	0,96
	BERT	0,95
	DistilBERT	0,94
	SVM	0,93
	RF	0,93
	BiLSTM	0,92
	NB	0,91
MELD	RoBERTa	0,53
	BERT	0,52
	SVM	0,52
	DistilBERT	0,51
	BiLSTM	0,51
	NB	0,51
	RF	0,50

Tabel 4.18 Peringkat Model Berdasarkan Waktu Inferensi

<i>Dataset</i>	Model	Waktu Inferensi
“Emotions” oleh ElgiriyeWithana	NB	0,001s
	RF	0,317s
	BiLSTM	2,307s
	SVM	5,250s
	DistilBERT	20,955s
	RoBERTa	36,602s

	BERT	41,602s
“Emotion Dataset” oleh Parul Pandey	NB	0,001s
	RF	0,860s
	BiLSTM	0,872s
	SVM	0,878s
	DistilBERT	7,638s
	RoBERTa	13,267s
	BERT	15,115s
“ISEAR” oleh Faisal Santo	NB	0,001s
	RF	0,084s
	BiLSTM	0,169s
	SVM	0,415s
	DistilBERT	5,636s
	RoBERTa	9,622s
	BERT	10,855s
“Emotion Dataset” oleh Abdallah Wagih	NB	0,001s
	RF	0,116s
	BiLSTM	0,978s
	SVM	1,048s
	DistilBERT	1,994s
	RoBERTa	3,101s
	BERT	3,937s
MELD	NB	0,001s
	RF	0,122s
	BiLSTM	0,291s
	SVM	1,410s
	DistilBERT	11,149s
	RoBERTa	19,325s
	BERT	22,164s

4.7.1 Analisis Kelebihan dan Kekurangan Pendekatan *Machine Learning* dan *Deep Learning*

Berdasarkan hasil evaluasi yang telah dilakukan pada berbagai jenis *dataset*, terlihat bahwa pendekatan *deep learning*, khususnya yang menggunakan arsitektur Transformer seperti BERT, RoBERTa, dan DistilBERT, menunjukkan performa yang secara konsisten lebih

unggul dibandingkan pendekatan *machine learning* seperti Naive Bayes, SVM, dan Random Forest. Keunggulan ini terutama terlihat pada metrik *F1-score* dan akurasi, yang menggambarkan kemampuan model dalam memahami makna emosional dalam teks.

Model-model berbasis Transformer memiliki keunggulan utama dalam menangkap konteks dan nuansa semantik berkat mekanisme *self-attention* yang mereka miliki. Keunggulan ini menjadi sangat menonjol ketika model dihadapkan pada data-data pendek dan tidak terstruktur seperti *tweet*. Namun, performa yang tinggi ini harus dibayar pada aspek efisiensi. Model-model Transformer membutuhkan waktu inferensi yang jauh lebih lama dibandingkan model *machine learning*. Sebagai contoh, BERT dan RoBERTa bisa membutuhkan lebih dari 20 detik untuk menyelesaikan proses inferensi satu *batch* data. Hal ini menjadikannya kurang ideal untuk aplikasi *real-time* atau untuk diterapkan pada perangkat dengan sumber daya komputasi terbatas. Meskipun DistilBERT hadir sebagai versi ringan dari BERT dan berhasil memangkas waktu inferensi hampir setengahnya, tetap saja waktu yang dibutuhkan masih jauh di atas model *machine learning* seperti Naive Bayes.

Sementara itu, model *machine learning* seperti Naive Bayes, SVM, dan Random Forest tetap menunjukkan kekuatannya dari sisi efisiensi komputasi. Waktu inferensi pada model ini sangat cepat, bahkan hanya membutuhkan waktu dalam hitungan milidetik. Karakteristik ini menjadikan model Machine Learning *machine learning* sangat cocok untuk sistem yang mengutamakan kecepatan respon dan bekerja di lingkungan dengan keterbatasan komputasi. Namun demikian, kelemahan mendasar dari model *machine learning* terletak pada ketidakmampuannya dalam menangkap konteks dan urutan kata. Representasi teks dalam model ini sangat bergantung pada teknik ekstraksi fitur seperti TF-IDF, yang tidak mempertimbangkan urutan maupun keterkaitan antar kata dalam kalimat. Akibatnya, model ini cenderung mengalami kesulitan ketika dihadapkan pada teks yang membutuhkan pemahaman konteks, seperti dialog dalam *dataset* MELD.

Model BiLSTM menempati posisi di antara kedua pendekatan tersebut. BiLSTM dirancang untuk memproses urutan kata secara dua arah, menunjukkan performa yang cukup baik pada dataset komentar dan *tweet*. Namun, BiLSTM cenderung membutuhkan waktu pelatihan dan inferensi yang lebih lama dibandingkan model *machine learning*, meskipun masih lebih cepat dibandingkan Transformer. Selain itu, model ini relatif rentan terhadap *overfitting*, terutama pada dataset yang kecil.

Secara keseluruhan, setiap pendekatan memiliki keunggulan dan keterbatasannya masing-masing. Model *machine learning* sangat cepat, tetapi memiliki keterbatasan dalam

menangkap pemahaman konteks. BiLSTM mampu memproses kata secara dua arah dengan akurasi yang cukup baik, tetapi kurang cepat. Model *deep learning* dengan arsitektur Transformer unggul dalam hal pemahaman konteks dan akurasi klasifikasi emosi, namun dengan beban komputasi yang tinggi. Oleh karena itu, pemilihan model yang tepat sangat bergantung pada kebutuhan dan konteks penerapan, termasuk pada jenis teks yang dianalisis. Hal ini akan dibahas lebih lanjut dalam bagian berikutnya yang membahas pengaruh karakteristik teks terhadap performa model.

4.7.2 Pengaruh Karakteristik Teks terhadap Kinerja Model

Setelah sebelumnya dibahas kelebihan dan kekurangan masing-masing pendekatan. Pada bagian ini akan dianalisis bagaimana jenis teks, seperti *tweet*, komentar, atau dialog mempengaruhi kinerja pendekatan model tertentu. Pada *dataset* teks *tweet*, yang bahasanya singkat dan informal, model *deep learning* berbasis Transformer secara konsisten menunjukkan kinerja terbaik. Namun, tingkat keunggulannya bervariasi. Pada *dataset* “Emotions”, RoBERTa dengan *F1-score* 0,93 unggul dari SVM 0,89, menunjukkan performa *machine learning* bisa bersaing dengan *deep learning*. Pada *dataset* “Emotion Dataset” oleh Parul Pandey, di mana DistilBERT dengan *F1-score* 0,86 mengungguli SVM sebagai model ML terbaik dengan *F1-score* 0,82, dan Naive Bayes paling kecil dengan 0,62. Sementara itu, pada *dataset* “ISEAR”, BERT dengan *F1-score* 0,87 unggul satu poin dari SVM 0,86. Ini menunjukkan bahwa meskipun Transformer unggul, model *machine learning* cukup bisa bersaing untuk teks jenis *tweet*.

Pada *dataset* teks komentar, yang cenderung lebih panjang dan naratif, seluruh model menunjukkan performa yang sangat tinggi. RoBERTa kembali menjadi yang paling unggul dengan *F1-score* 0,96, selisih kinerjanya dengan model *machine learning* terbaik seperti SVM dan Random Forest, dimana keduanya 0,93 cukup sedikit. Bahkan, model dengan performa terendah dalam kelompok ini, Naive Bayes, mampu mencapai *F1-score* 0,91. Hal ini mengindikasikan bahwa kalimat yang terstruktur pada teks komentar lebih mudah diproses dan dipahami oleh semua jenis pendekatan, baik *machine learning* maupun *deep learning*.

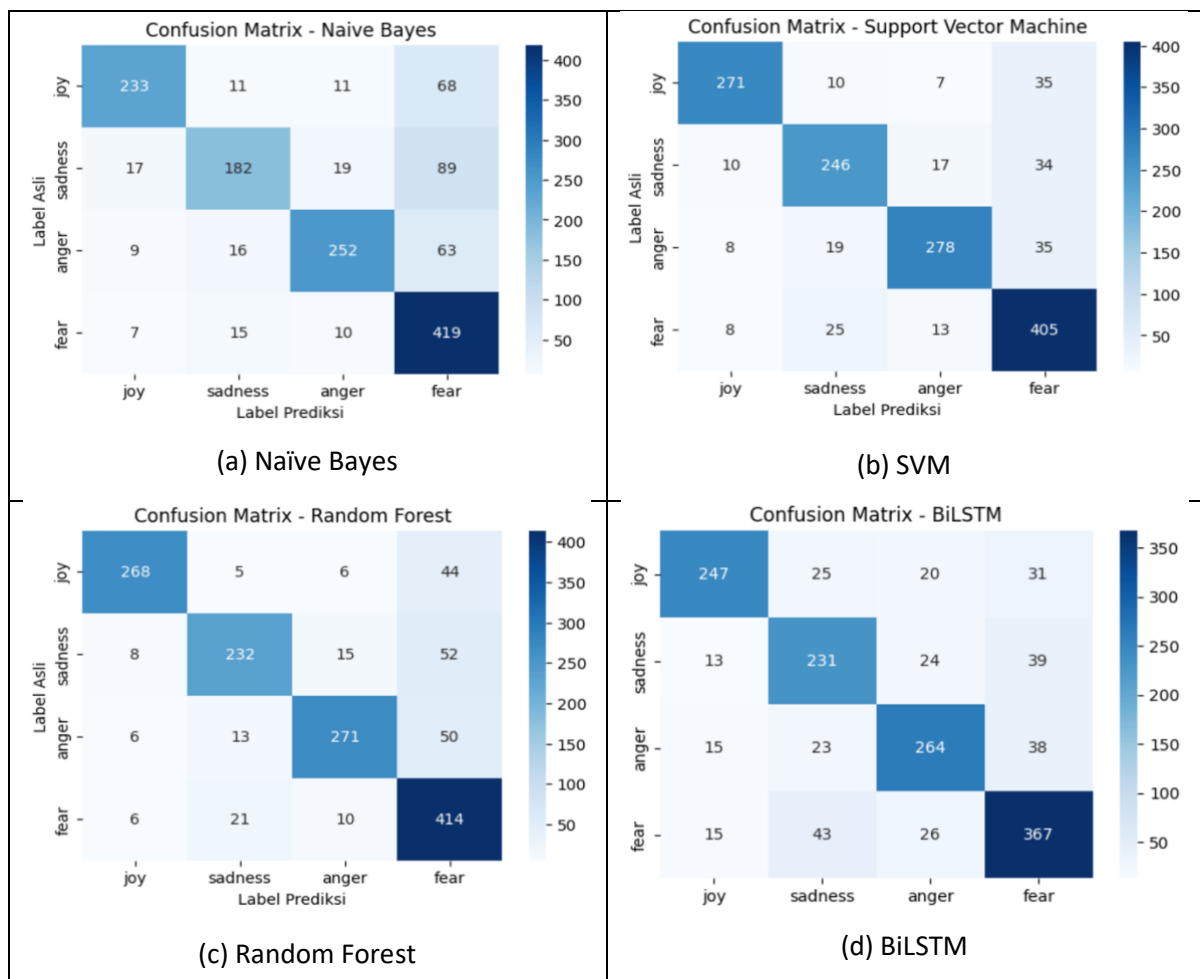
Pada *dataset* teks dialog, di mana seluruh model menunjukkan penurunan kinerja. *F1-score* tertinggi hanya mencapai 0,29, yang diperoleh oleh model BERT. Angka ini hanya sedikit lebih baik dibandingkan model *machine learning* terbaik yaitu SVM dan Random Forest, dimana keduanya 0,25, dan menunjukkan betapa sulitnya tugas ini bagi semua pendekatan. Naive Bayes di posisi terendah dengan *F1-score* 0,15, disebabkan oleh data dialog

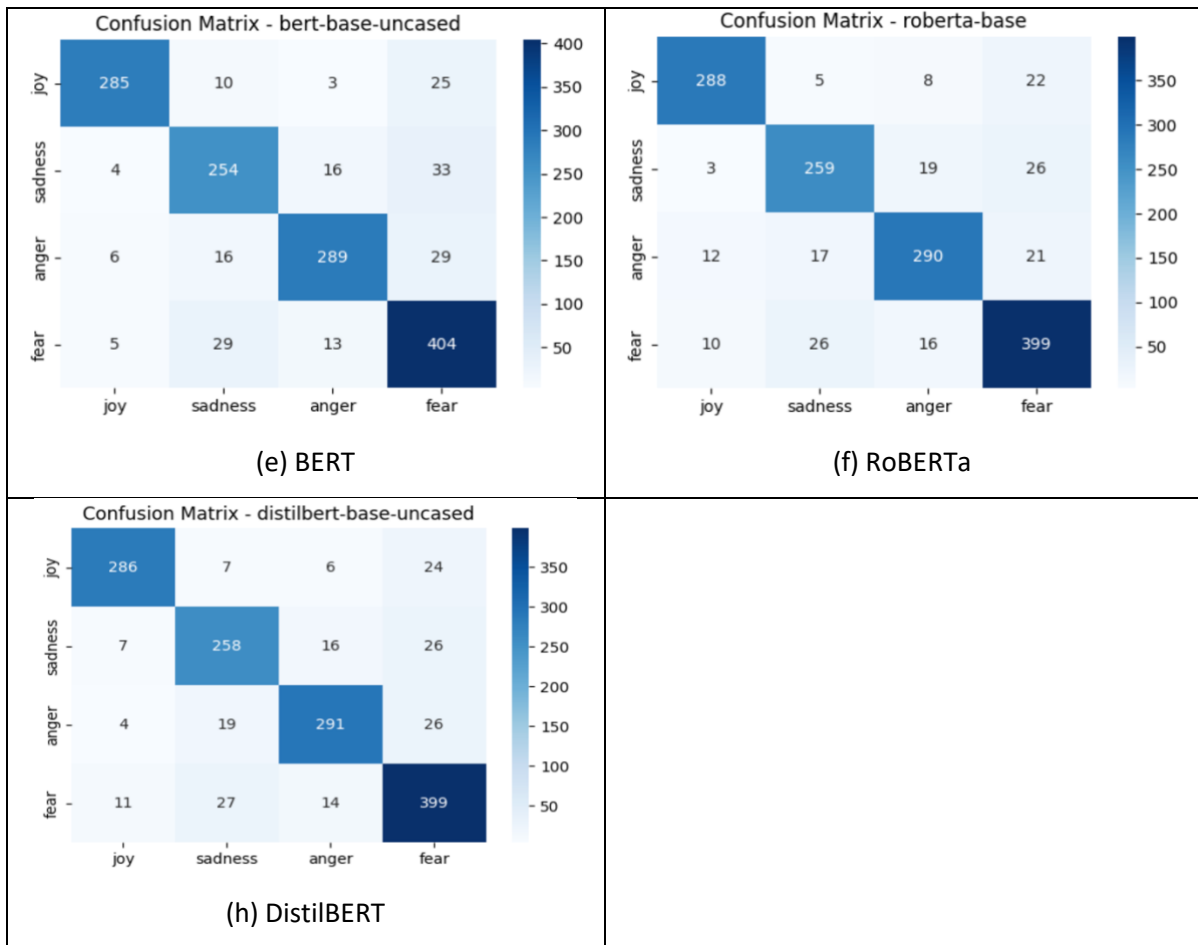
yang sangat bergantung pada konteks antar ucapan. Karena semua model dalam penelitian ini hanya menganalisis satu kalimat sendiri-sendiri, model gagal menangkap alur percakapan untuk memahami emosi yang sebenarnya.

Secara ringkas, model *deep learning* dengan arsitektur Transformer, seperti BERT, RoBERTa, dan DistilBERT menunjukkan dominasi yang bervariasi pada jenis teks *tweet* dan komentar. Namun, pada teks dialog yang kompleks semua model kesulitan. Perbedaan performa ini menegaskan bahwa tidak ada satu model pun yang dapat mengungguli semua jenis teks secara seragam.

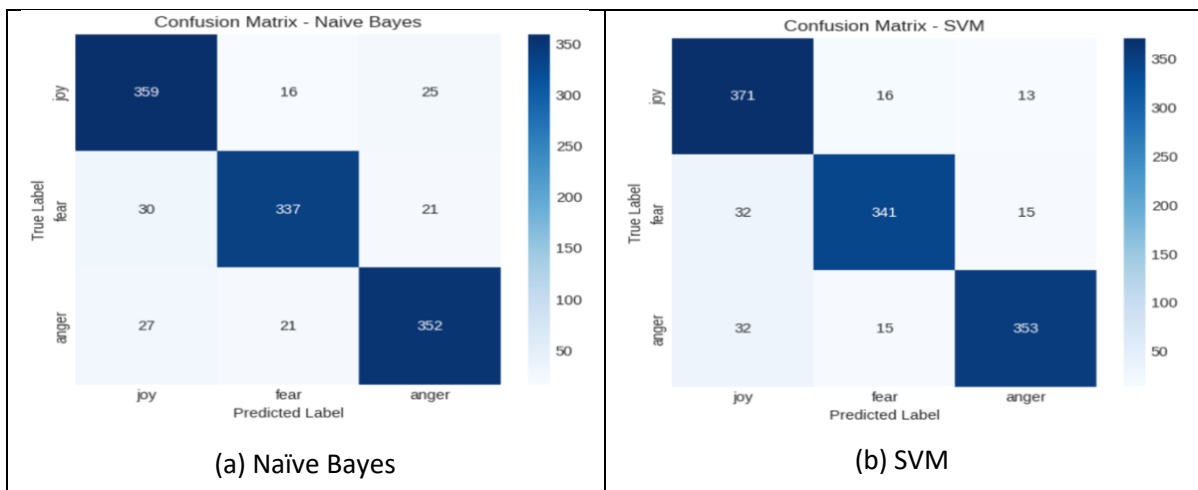
4.7.3 Pengaruh *Preprocessing* terhadap Kinerja Model

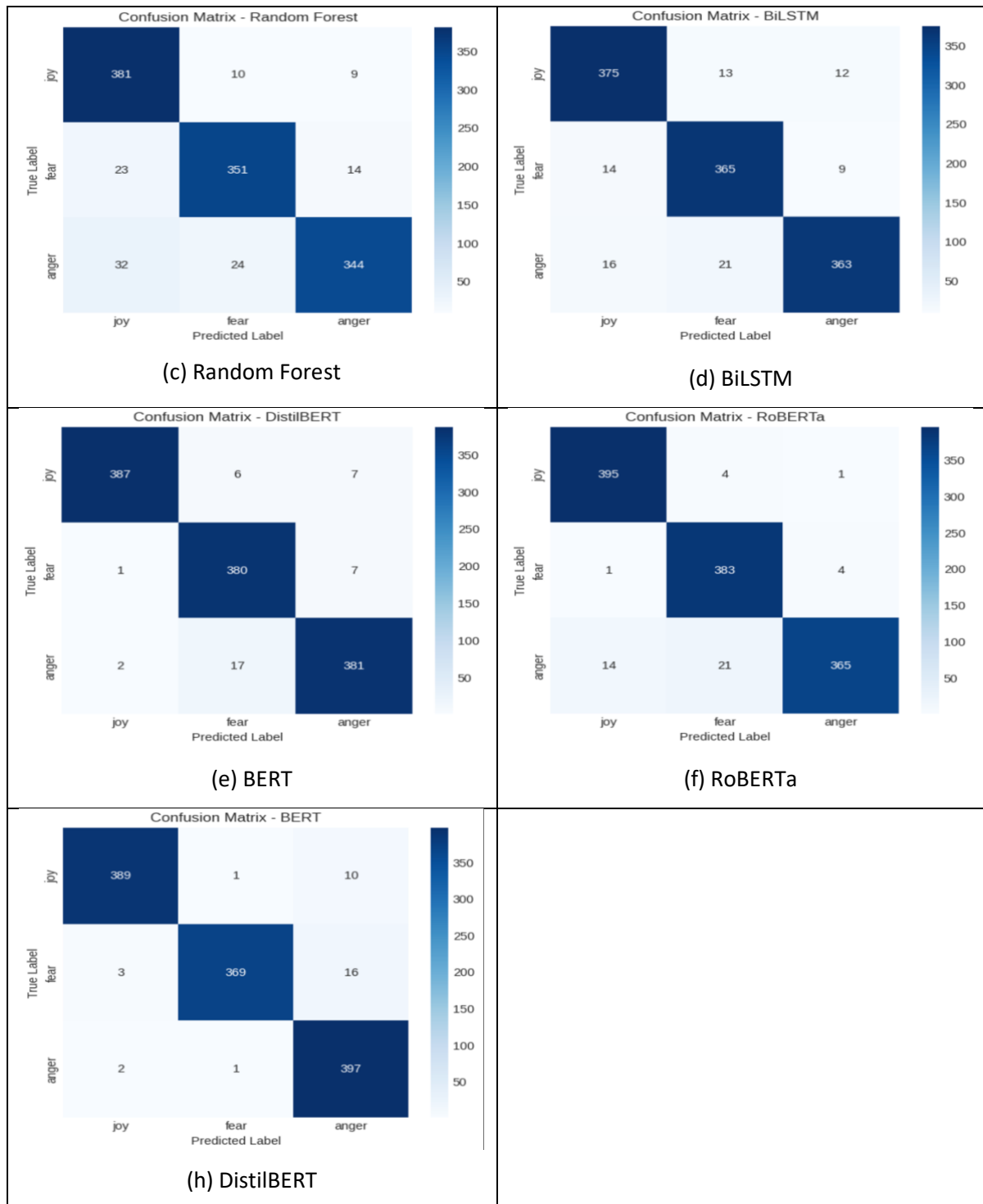
Eksperimen tambahan dilakukan untuk melihat sejauh mana pengaruh *preprocessing* terhadap kinerja model, dilakukan pada tiga *dataset* yang mewakili masing-masing jenis teks, yaitu *dataset* “ISEAR” oleh Faisal Santo, *dataset* “Emotion Dataset” oleh Abdallah Wagih dan *dataset* MELD. *Confusion matrix* dari setiap model pada kondisi tanpa *preprocessing* juga disajikan secara berurutan, dilanjutkan dengan hasil kuantitatifnya.



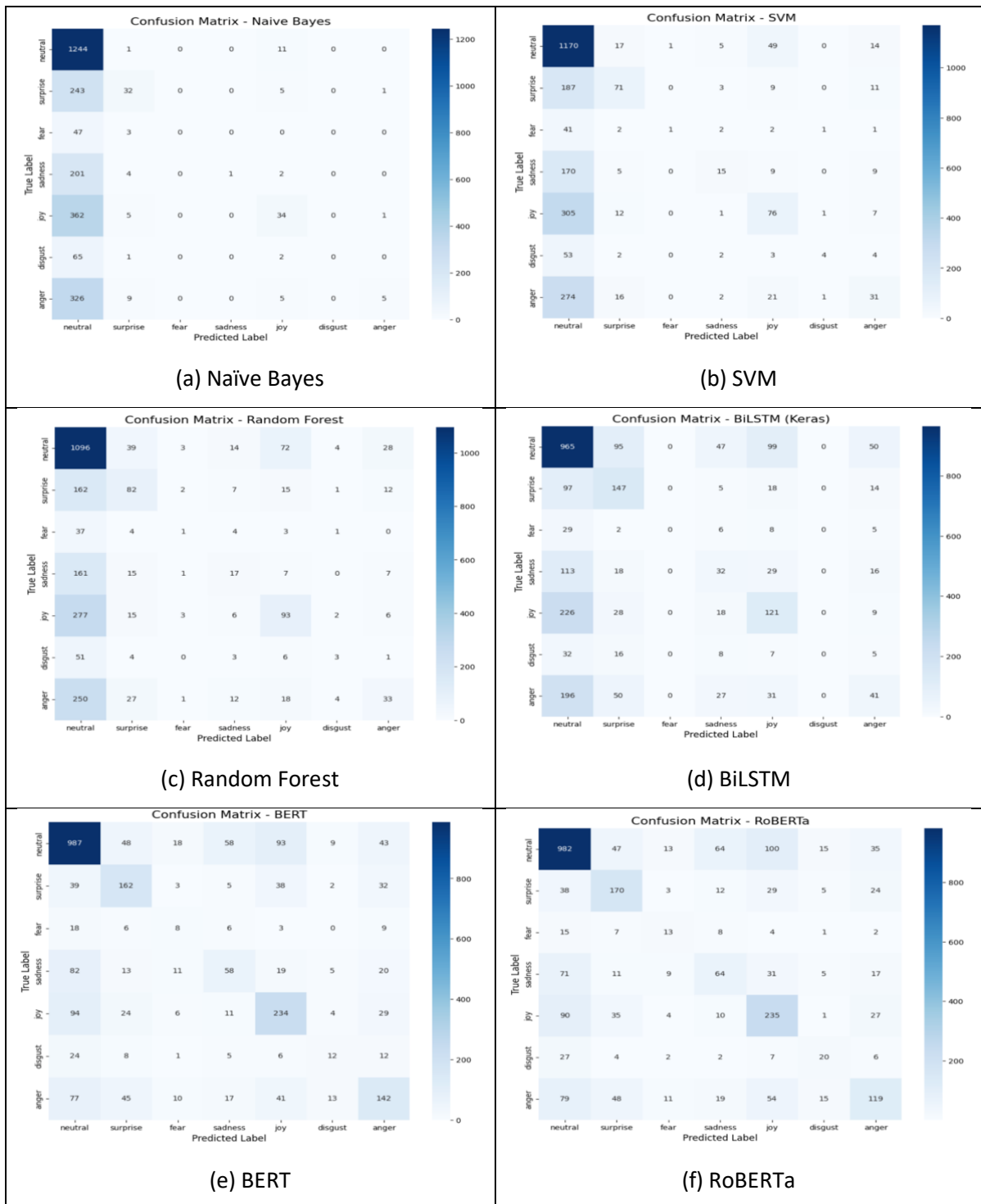


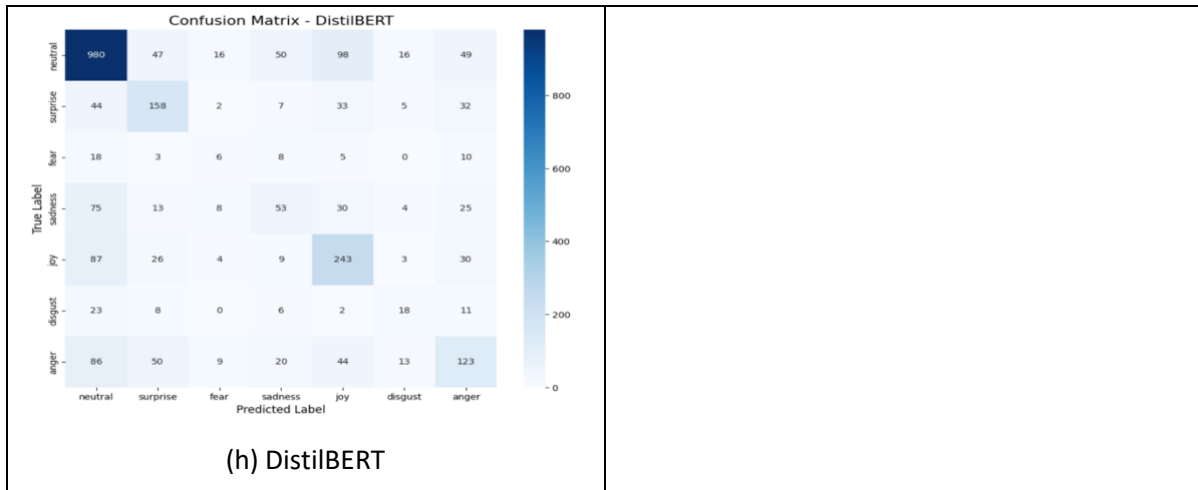
Gambar 4.11 *Confusion Matrix* Model pada Dataset “ISEAR” oleh Faisal Santo tanpa *Preprocessing*





Gambar 4.12 *Confusion Matrix* Model pada *Dataset* “Emotion Dataset” oleh Abdallah Wagih tanpa *Preprocessing*





Gambar 4.13 *Confusion Matrix* Model pada Dataset “MELD” tanpa *Preprocessing*

Dari beberapa *confusion matrix* di atas, terlihat perbedaan dampak yang jelas antara model *machine learning* dan *deep learning*. Pada dataset ISEAR (Gambar 4.11) dan dataset "Emotion Dataset" (Gambar 4.12), model *machine learning* menunjukkan peningkatan jumlah kesalahan klasifikasi tanpa *preprocessing*. Sebaliknya, model *deep learning* menunjukkan performa yang stabil. Peningkatan terlihat pada dataset MELD (Gambar 4.13). Tanpa *preprocessing*, model-model *deep learning* terutama yang dengan arsitektur Transformer, meskipun masih sangat bias terhadap kelas *neutral* dan *joy*, mulai menunjukkan kemampuan yang jauh lebih baik dalam mengidentifikasi kelas-kelas emosi lainnya, seperti yang terlihat dari nilai diagonal yang lebih tinggi untuk *surprise*, *anger*, dan *disgust* dibandingkan saat dengan *preprocessing*. Selanjutnya, untuk melihat hasil kuantitatif dari eksperimen ini, Tabel 4.19 menyajikan hasil evaluasi lengkap dari pengujian tanpa *preprocessing*.

Tabel 4.19 Hasil Evaluasi pada *Dataset* “ISEAR”, *Dataset* “Emotion Dataset”, dan *Dataset* MELD tanpa *Preprocessing*

<i>Dataset</i>	Model	Akurasi	Presisi	Recall	F1-score	Waktu Inferensi
“ISEAR” oleh Faisal Santo	NB	0,76	0,75	0,76	0,76	0,001s
	SVM	0,84	0,84	0,84	0,84	0,812s
	RF	0,83	0,84	0,83	0,84	0,063s
	BiLSTM	0,78	0,78	0,78	0,78	0,152s
	BERT	0,87	0,87	0,86	0,87	10,025s
	RoBERTa	0,87	0,87	0,87	0,87	9,136s
	DistilBERT	0,87	0,87	0,87	0,87	5,072s
“Emotion Dataset” oleh Abdallah Wagih	NB	0,88	0,88	0,88	0,88	0,001s
	SVM	0,89	0,89	0,89	0,89	2,062s
	RF	0,90	0,90	0,90	0,90	0,119s
	BiLSTM	0,92	0,92	0,92	0,92	1,027s
	BERT	0,97	0,97	0,97	0,97	4,738s
	RoBERTa	0,96	0,96	0,96	0,96	4,445s
	DistilBERT	0,96	0,96	0,96	0,96	2,374s
MELD	NB	0,50	0,48	0,17	0,15	0,001s
	SVM	0,52	0,50	0,23	0,24	0,872s
	RF	0,51	0,34	0,23	0,24	0,119s
	BiLSTM	0,50	0,27	0,27	0,26	1,713s
	BERT	0,61	0,44	0,42	0,43	21,735s
	RoBERTa	0,61	0,46	0,45	0,45	19,321s
	DistilBERT	0,61	0,43	0,42	0,42	10,960s

Tabel di atas memberikan gambaran lengkap evaluasi model. Untuk mengukur secara langsung sejauh mana *preprocessing* memberikan dampak, perlu dilakukan perbandingan. Oleh karena itu, Tabel 4.20 berikut menyajikan perbandingan berdasarkan *F1-score* untuk melihat perubahan kinerja model dengan dan tanpa *preprocessing*

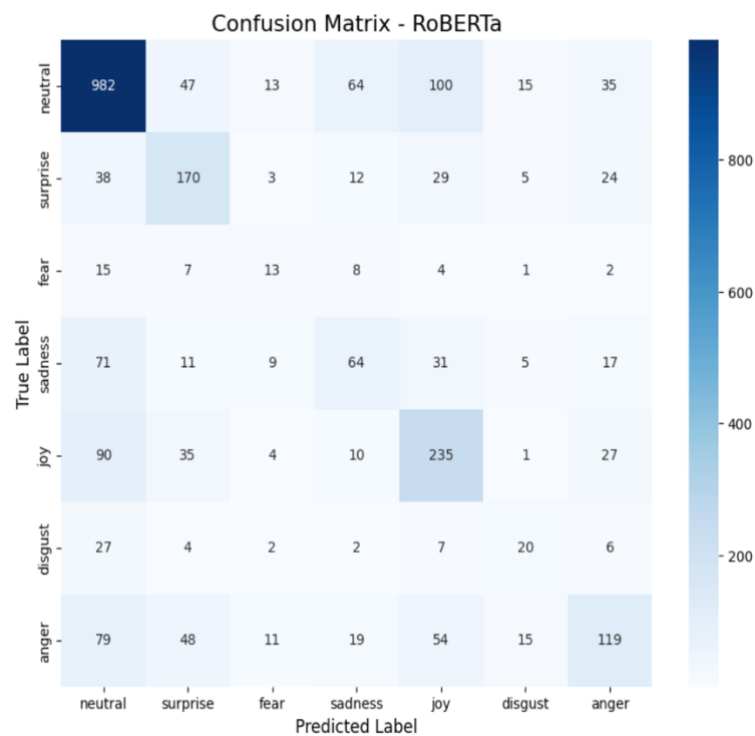
Tabel 4.20 Perbandingan *F1-score* (Dengan dan Tanpa *Preprocessing*)

<i>Dataset</i>	Model	<i>F1-score</i> (Dengan <i>Preprocessing</i>)	<i>F1-score</i> (Tanpa <i>Preprocessing</i>)
“ISEAR” oleh Faisal Santo	NB	0,80	0,76
	SVM	0,86	0,84
	RF	0,85	0,84
	BiLSTM	0,79	0,78
	BERT	0,87	0,87
	RoBERTa	0,86	0,87
	DistilBERT	0,85	0,87
“Emotion Dataset” oleh Abdallah Wagih	NB	0,91	0,88
	SVM	0,93	0,89
	RF	0,93	0,90
	BiLSTM	0,92	0,92
	BERT	0,95	0,97
	RoBERTa	0,96	0,96
	DistilBERT	0,94	0,96
MELD	NB	0,15	0,15
	SVM	0,25	0,24
	RF	0,25	0,24
	BiLSTM	0,23	0,26
	BERT	0,29	0,43
	RoBERTa	0,27	0,45
	DistilBERT	0,25	0,42

Pada *dataset* MELD, menghilangkan *preprocessing* memberikan peningkatan performa secara signifikan pada model *deep learning*, terutama arsitektur Transformer. *F1-score* BERT dari 0,29 menjadi 0,43, RoBERTa dari 0,27 naik hingga 0,45, dan DistilBERT dari 0,25 menjadi 0,42. Hal ini menunjukkan bahwa *preprocessing* mungkin justru menghilangkan konteks emosi penting dalam dialog. Pada *dataset tweet* “ISEAR” oleh Faisal Santo dan *dataset* Komentar “Emotion Dataset” oleh Abdallah Wagih, dampak menghilangkan *preprocessing* pada model *deep learning* tidak signifikan. Namun, pada model *machine learning*, *preprocessing* terbukti penting. Tanpa *preprocessing*, performa justru menurun. Sebagai contoh, pada *dataset* ISEAR, akurasi Naive Bayes menurun dari 0,81 menjadi 0,76, sedangkan pada *dataset* komentar penurunan terjadi dari 0,91 menjadi 0,88. Hal ini disebabkan karena model *machine learning* sangat bergantung pada bentuk kata bersih. Temuan ini memperkuat bahwa strategi *preprocessing* harus disesuaikan dengan model yang digunakan dan karakteristik datanya. *Preprocessing* bermanfaat bagi model *machine learning* untuk menyederhanakan representasi teks, namun bisa tidak bermanfaat pada model *deep learning* yang sudah terlatih memahami bentuk bahasa alami.

4.7.4 Analisis Kesalahan pada *Dataset* MELD

Meskipun performa pada MELD meningkat tanpa *preprocessing*, hasilnya masih jauh dari memuaskan. *F1-score* tertinggi yang berhasil dicapai hanya sebesar 0,45, jauh di bawah performa pada jenis teks lain seperti *tweet* dan komentar. Untuk memahami mengapa, analisis kesalahan dilakukan menggunakan *confusion matrix* dari hasil prediksi terbaik (RoBERTa tanpa *preprocessing*). Gambar 4.14 menampilkan *confusion matrix* model RoBERTa.



Gambar 4.14 *Confusion Matrix* RoBERTa

Berdasarkan *confusion matrix*, terlihat bahwa model menunjukkan kemampuan yang relatif baik untuk mengenali emosi-emosi yang *neutral*, *surprise*, *joy*, dan *anger*. Hal ini terlihat dari angka prediksi benar (diagonal utama) yang tinggi untuk kelas-kelas tersebut. Kesalahan prediksi paling signifikan terjadi pada kelas-kelas emosi yang jumlah datanya lebih sedikit, yaitu *disgust* dan *fear*. Hal ini menunjukkan bahwa model kesulitan membedakan emosi-emosi ini. Untuk memahami mengapa pola kesalahan ini terjadi, Studi kasus pada Dialog 259 berikut menampilkan bagaimana keterbatasan pemahaman konteks menjadi akar dari masalah ini. Tabel 4.21 menampilkan data pada Dialog 259.

Tabel 4.21 Studi Kasus Kesalahan Prediksi Model pada *Dataset* MELD Dialog 259

<i>Utterance</i>	<i>Speaker</i>	<i>Dialogue Id</i>	<i>Utterance Id</i>	Emosi Sebenarnya	Prediksi Model
Are you kidding?	Ross	259	0	surprise	anger
Okay, look.	Ross	259	1	neutral	neutral
I-I studied evolution.	Ross	259	2	joy	neutral
Remember, evolution?	Ross	259	3	joy	neutral
Monkey into man?	Ross	259	4	anger	neutral
Plus, Im a doctor, and I had a monkey.	Ross	259	5	neutral	neutral
Im Doctor Monkey!	Ross	259	6	anger	surprise
Im not arguing with that.	Chandler	259	7	neutral	neutral
All right, Ive heard enough. Ive made my decision.	Monica	259	8	neutral	anger
You are both idiots.	Monica	259	9	disgust	anger
The joke is not funny, and its offensive to women, and doctors, and monkeys!	Monica	259	10	disgust	anger
You shouldnt be arguing over who gets credit, you should be arguing over who gets blamed for inflicting this horrible joke upon the world!	Monica	259	11	anger	disgust
Now let it go!	Monica	259	12	anger	anger
The joke sucks!	Monica	259	13	anger	sadness
Its your joke.	Ross	259	14	neutral	anger
Is not.	Chandler	259	15	anger	neutral

Percakapan diawali dengan argumen Ross yang emosinya terus meningkat. Model gagal dalam menangkap ini, ia salah mengira *surprise* Ross sebagai *anger*, lalu gagal mendeteksi kebanggaan (*joy*) saat Ross berbicara tentang studinya dan menganggapnya *neutral*. Puncaknya adalah ketika Ross mengatakan "I'm Doctor Monkey!", yang emosi sebenarnya adalah *anger*, namun diprediksi sebagai *surprise*. Kegagalan beruntun ini menunjukkan bahwa model tidak mampu melacak perkembangan emosional seorang pembicara dari satu kalimat ke kalimat berikutnya.

Ketika Monica masuk untuk menengahi, kalimatnya "You are both idiots," yang dilandasi oleh rasa jijik (*disgust*) terhadap kekonyolan argumen teman-temannya, salah diinterpretasikan sebagai kemarahan (*anger*). Menariknya, pada kalimat selanjutnya saat Monica benar-benar marah, model justru memprediksinya sebagai *disgust*, menunjukkan kebingungan yang bolak-balik antara kedua emosi ini. Puncaknya, ekspresi kesal seperti "The joke sucks!" salah diklasifikasikan sebagai *sadness*, kemungkinan karena model mengasosiasikan kata "sucks" dengan sentimen negatif secara umum tanpa bisa membedakan

intensitas kemarahan. Studi kasus ini menegaskan bahwa model memproses setiap *utterance* sendiri-sendiri. Ia tidak memiliki memori percakapan yang memungkinkannya memahami bahwa sebuah kalimat adalah respons terhadap kalimat lain. Keterbatasan inilah yang menyebabkannya gagal konteks dalam teks dialog.

4.7.5 Kekurangan Penelitian dan Arah Penelitian Selanjutnya

Berdasarkan serangkaian analisis yang telah dilakukan, teridentifikasi beberapa kekurangan dalam penelitian yang sekaligus membuka peluang untuk arah penelitian di masa depan:

- a. Keterbatasan pemahaman konteks dialog semua model pada *dataset* dialog MELD. Analisis kesalahan pada subbab 4.7.4 menunjukkan bahwa model memproses setiap ucapan secara terpisah tanpa mempertimbangkan alur percakapan. Kekurangan ini membuka peluang untuk penelitian selanjutnya yang menerapkan model yang dirancang untuk memahami konteks percakapan.
- b. Penelitian ini hanya menggunakan data teks dari *dataset* MELD, padahal dataset tersebut bersifat *multimodal* (teks, audio, visual). Hal ini menjadi keterbatasan karena informasi emosional penting yang terkandung dalam intonasi suara dan ekspresi visual tidak dimanfaatkan. Saran untuk penelitian berikutnya adalah mengeksplorasi pendekatan *multimodal* untuk mendapatkan pemahaman emosi yang lebih kaya dan akurat.
- c. Bahasa yang digunakan seluruh *dataset* pada penelitian ini hanya berbahasa Inggris. Akibatnya, temuan dan model yang dihasilkan tidak dapat digeneralisasi secara langsung untuk bahasa lain. Hal ini menjadi saran bagi penelitian selanjutnya untuk melakukan eksperimen serupa pada *dataset* berbahasa selain bahasa Inggris, yang memiliki tantangannya sendiri.
- d. Teridentifikasinya *overfitting* pada Beberapa Model. Analisis pada subbab 4.5 menunjukkan bahwa beberapa model, khususnya BiLSTM, secara konsisten mengalami *overfitting* pada sebagian besar *dataset*. Penelitian ini berfokus pada identifikasi fenomena ini, namun tidak sampai menerapkan teknik mitigasi secara mendalam. Keterbatasan ini menggarisbawahi pentingnya penerapan strategi mitigasi *overfitting* dalam penelitian selanjutnya untuk mengoptimalkan performa model.

BAB V

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan pada bab sebelumnya, bab terakhir ini akan menyajikan kesimpulan dari temuan pada penelitian ini. Selain itu, akan dipaparkan pula beberapa saran yang dapat menjadi acuan untuk pengembangan penelitian di masa mendatang.

5.1 Kesimpulan

Berdasarkan analisis hasil evaluasi yang telah dilakukan, dapat ditarik dua kesimpulan utama:

- a. Kemampuan model *machine learning* dan *deep learning* dalam mengklasifikasikan emosi bergantung pada karakteristik teks yang diuji.
 1. Pada teks *tweet* yang tidak terstruktur dan kaya akan simbol, pendekatan *deep learning* berbasis Transformer (BERT, RoBERTa, DistilBERT) secara konsisten menunjukkan kemampuan klasifikasi yang paling unggul, meskipun model *machine learning* tetap mampu memberikan performa yang dapat bersaing.
 2. Pada teks komentar yang lebih naratif dan terstruktur, semua model menunjukkan performa yang sangat baik. Pendekatan *deep learning* kembali mencapai *F1-score* puncak (0,96), namun keunggulannya atas pendekatan *machine learning* terbaik (SVM dengan skor 0,93) tidak terlalu jauh, menandakan teks jenis ini lebih mudah diproses oleh semua pendekatan.
 3. pada teks dialog (MELD) yang memerlukan pemahaman konteks antar kalimat, kemampuan seluruh model baik *machine learning* maupun *deep learning* menurun (*F1-score* paling tinggi hanya 0,29), yang menunjukkan bahwa kedua pendekatan gagal menangkap dinamika emosi dalam percakapan ketika hanya menganalisis kalimat sendiri-sendiri.
- b. Perbandingan antara kedua pendekatan menunjukkan adanya *trade-off* antara akurasi, efisiensi, dan ketergantungan pada *preprocessing*.
 1. Dari segi akurasi, pendekatan *deep learning* secara konsisten lebih unggul karena kemampuannya memahami konteks semantik melalui arsitektur Transformer.
 2. Dari segi efisiensi, pendekatan *machine learning* jauh lebih cepat, dengan waktu inferensi dalam hitungan milidetik, sementara model *deep learning* berbasis Transformer membutuhkan waktu yang berkali-kali lipat lebih lama.

3. Dari segi pengaruh *preprocessing*, kedua pendekatan menunjukkan reaksi yang berlawanan. Kinerja *machine learning* menurun tanpa *preprocessing*, membuktikan ketergantungannya pada data teks yang bersih. Sebaliknya, kinerja *deep learning* pada data dialog (MELD) justru meningkat signifikan tanpa *preprocessing* (*F1-score* RoBERTa naik dari 0,27 menjadi 0,45), yang mengindikasikan bahwa proses pembersihan teks dapat menghilangkan sinyal kontekstual penting bagi model *deep learning* berbasis Transformer.

5.2 Saran

Berdasarkan kesimpulan dan keterbatasan dalam penelitian ini, berikut adalah beberapa saran untuk penelitian selanjutnya:

- a. Penelitian selanjutnya disarankan untuk mengeksplorasi pendekatan *multimodal* (teks, audio, dan visual) khususnya pada *dataset* dialog seperti MELD untuk menangkap konteks emosi secara lebih menyeluruh.
- b. Perlu dilakukan pengujian model dengan konteks percakapan agar performa model meningkat pada jenis data dialog.
- c. Pengembangan ke bahasa lain menjadi arah yang relevan mengingat sebagian besar penelitian saat ini masih fokus pada teks berbahasa Inggris.
- d. Menerapkan teknik mitigasi *overfitting* untuk mencapai hasil yang lebih baik, seperti *early stopping* untuk menghentikan pelatihan pada waktu yang optimal, serta regularisasi menggunakan *dropout* untuk meningkatkan kemampuan generalisasi model.

DAFTAR PUSTAKA

- Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Classification Algorithms. In *Mining Text Data* (pp. 163–222). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6
- Alaeddini, M. (2024). *Emotion Detection in Reddit: Comparative Study of Machine Learning and Deep Learning Techniques*.
- ayushimalm50. (2025, March 24). *DistilBERT in Natural Language Processing*. GeeksforGeeks. <https://www.geeksforgeeks.org/nlp/distilbert-in-natural-language-processing/>
- Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation Learning: A Review and New Perspectives*. <http://arxiv.org/abs/1206.5538>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bucila, C., Caruana, R., & Niculescu-Mizil, A. (2006). *Model compression* (Vol. 2006). <https://doi.org/10.1145/1150402.1150464>
- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 15–21. <https://doi.org/10.1109/MIS.2013.30>
- Cortiz, D. (2021). *Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA*. <https://arxiv.org/pdf/2104.02041>
- Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38), E7900–E7909. <https://doi.org/10.1073/pnas.1702247114>
- Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3), 766–785. <https://doi.org/https://doi.org/10.1016/j.chb.2007.04.004>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>

- Elgiriyeewithana N. (2024). *Emotions*. Kaggle.
<https://doi.org/10.34740/KAGGLE/DSV/7563141>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010>
- Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. https://scholarworks.utep.edu/cs_techrep/1209?utm_source=scholarworks.utep.edu%2Fs_techrep%2F1209&utm_medium=PDF&utm_campaign=PDFCoverPages
- Goldberg, Y. (2017a). *Neural Network Methods for Natural Language Processing*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02165-7>
- Goldberg, Y. (2017b). *Neural Network Methods for Natural Language Processing*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02165-7>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>
- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 4, 2047–2052 vol. 4. <https://doi.org/10.1109/IJCNN.2005.1556215>
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier. <https://doi.org/10.1016/C2009-0-61819-5>
- Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4), 330–338. <https://doi.org/https://doi.org/10.1016/j.jksues.2016.04.002>
- Ibrahim A. (2023). *Emotion Dataset*. Kaggle.
- Imran, M. M. (2024). Emotion Classification in Software Engineering Texts: A Comparative Analysis of Pre-Trained Transformers Language Models. *Proceedings - 2024 ACM/IEEE International Workshop on NL-Based Software Engineering, NLBSE 2024*, 73–80. <https://doi.org/10.1145/3643787.3648034>
- Indulkar, Y., & Patil, A. (2021). Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis. *2021 International Conference on Emerging Smart*

- Computing and Informatics (ESCI)*, 295–299.
<https://doi.org/10.1109/ESCI50559.2021.9396925>
- Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (pp. 137–142). Springer Berlin Heidelberg.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.acl-main.560>
- Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4), 345–379.
<https://doi.org/10.1007/BF00992553>
- Kumar S., S. A., & Geetha, A. (2024). Emotion Detection from Text using Natural Language Processing and Neural Networks. *International Journal of Intelligent Systems and Applications in Engineering*, 12(14s), 609–615.
<https://www.ijisae.org/index.php/IJISAE/article/view/4707>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.
- Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization*.
<http://arxiv.org/abs/1711.05101>
- Mahajan, R., & Zaveri, M. (2021). Harnessing Emotive Features for Emotion Recognition from Text. *International Journal of Advanced Computer Science and Applications*, 12(7).
<https://doi.org/10.14569/IJACSA.2021.0120719>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. [https://doi.org/DOI: 10.1017/CBO9780511809071](https://doi.org/DOI:10.1017/CBO9780511809071)
- Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), 81. <https://doi.org/10.1007/s13278-021-00776-6>
- Olson, D., & Delen, D. (2008). *Advanced Data Mining Techniques*. In *Springer. USA*.
<https://doi.org/10.1007/978-3-540-76917-0>
- Pandey P. (2021). *Emotion Dataset for Emotion Recognition Tasks*. Kaggle.
<https://www.kaggle.com/datasets/parulpandey/emotion-dataset>

- Plutchik, R. (2001). The Nature of Emotions. *American Scientist*, 89(4), 344.
<https://doi.org/10.1511/2001.28.344>
- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*.
<https://doi.org/10.18653/v1/p19-1050>
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.
<https://doi.org/10.1108/eb046814>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, *abs/1910.01108*.
<http://arxiv.org/abs/1910.01108>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*.
- Santo F. (2020). *ISEAR Dataset*. Kaggle.
- Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3687–3697). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1404>
- Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2), 310–328. <https://doi.org/10.1037/0022-3514.66.2.310>
- 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>
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Association for Computational Linguistics. <https://aclanthology.org/D13-1170/>
- Sparck Jones, K. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, 28(1), 11–21.
<https://doi.org/10.1108/eb026526>

- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). *How to Fine-Tune BERT for Text Classification?*
<http://arxiv.org/abs/1905.05583>
- Trivusi. (2022, July 3). *Penjelasan Lengkap Algoritma Support Vector Machine (SVM)*.
<https://www.trivusi.web.id/2022/04/algoritma-svm.html>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 90–94). Association for Computational Linguistics.
<https://aclanthology.org/P12-2018/>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. Le, Gugger, S., ... Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*.
- Yehoshua, R. (2023, March 25). *Random Forests*. Medium.
<https://medium.com/@roiyehe/random-forests-98892261dc49>
- Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2). <https://doi.org/10.1088/1742-6596/1168/2/022022>
- Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). *Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling*.

LAMPIRAN

Lampiran A. Kode Program Google Colab dan Kaggle Notebook

- <https://www.kaggle.com/code/hilqudz/emotionssss> (Dataset Emotions oleh Elgiryewithana)
- <https://www.kaggle.com/code/hilqudz/zip-dataset> (Dataset Emotion Dataset oleh Parul Pandey)
- <https://colab.research.google.com/drive/1EW7uu9XOTzgU5QhmS6TRN9U76pRaIHv?usp=sharing> (Dataset ISEAR oleh Faisal Santo)
- <https://colab.research.google.com/drive/1xnLPPszFoIRGZXVlxf1eYRsePCqE6-00?usp=sharing> (Dataset Emotion Dataset oleh Abdallah Wagih)
- <https://www.kaggle.com/code/hilqudz/melddd-ok> (Dataset MELD)