TESIS

KOMPARASI PERFORMA LSTM, GRU, DAN BERT DALAM ANALISIS SENTIMEN KOLOM KOMENTAR VIDEO YOUTUBE TENTANG DANANTARA INDONESIA

Disusun oleh:

Nama : **M. Alfa Rizy**
NIM : **24.55.1570**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 PJJ INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

TESIS

KOMPARASI PERFORMA LSTM, GRU, DAN BERT DALAM ANALISIS SENTIMEN KOLOM KOMENTAR VIDEO YOUTUBE TENTANG DANANTARA INDONESIA

PERFORMANCE COMPARISON OF LSTM, GRU, AND BERT IN SENTIMENT ANALYSIS OF YOUTUBE VIDEO COMMENT SECTIONS ON THE DANANTARA INDONESIA

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **M. Alfa Rizy**
NIM : **24.55.1570**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 PJJ INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

KOMPARASI PERFORMA LSTM, GRU, DAN BERT DALAM ANALISIS SENTIMEN KOLOM KOMENTAR VIDEO YOUTUBE TENTANG DANANTARA INDONESIA

PERFORMANCE COMPARISON OF LSTM, GRU, AND BERT IN SENTIMENT ANALYSIS OF YOUTUBE VIDEO COMMENT SECTIONS ON THE DANANTARA INDONESIA

yang disusun dan diajukan oleh

M. Alfa Rizy

24.55.1570

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 8 Desember 2025

Dosen Pembimbing,

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

HALAMAN PENGESAHAN

KOMPARASI PERFORMA LSTM, GRU, DAN BERT DALAM ANALISIS SENTIMEN KOLOM KOMENTAR VIDEO YOUTUBE TENTANG DANANTARA INDONESIA

PERFORMANCE COMPARISON OF LSTM, GRU, AND BERT IN SENTIMENT ANALYSIS OF YOUTUBE VIDEO COMMENT SECTIONS ON THE DANANTARA INDONESIA

yang disusun dan diajukan oleh

M. Alfa Rizy

24.55.1570

Telah dipertahankan di depan Dewan Penguji
pada tanggal 8 Desember 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Emha Taufiq Luthfi, S.T., M.Kom., Ph.D. **NIK. 190302125**	
Dr. Kumara Ari Yuana, S.T., M.T. **NIK. 190302575**	
Prof. Dr. Ema Utami, S.Si., M.Kom. **NIK. 190302037**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 8 Desember 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : M. Alfa Rizy
NIM : 24.55.1570
Konsentrasi : Digital Transformation Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
KOMPARASI PERFORMA LSTM, GRU, DAN BERT DALAM ANALISIS SENTIMEN KOLOM KOMENTAR VIDEO YOUTUBE TENTANG DANANTARA INDONESIA

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 8 Desember 2025
Yang Menyatakan,

M. Alfa Rizy

HALAMAN PERSEMBAHAN

Ibuku, Sumber doa dan restu yang tak pernah putus. Terima kasih karena selalu percaya padaku, bahkan saat aku meragukan diriku sendiri.

Teman-teman Seperjuangan, Terima kasih untuk setiap diskusi, kopi, dan dukungan moral yang menjaga semangatku tetap menyala sampai detik ini.

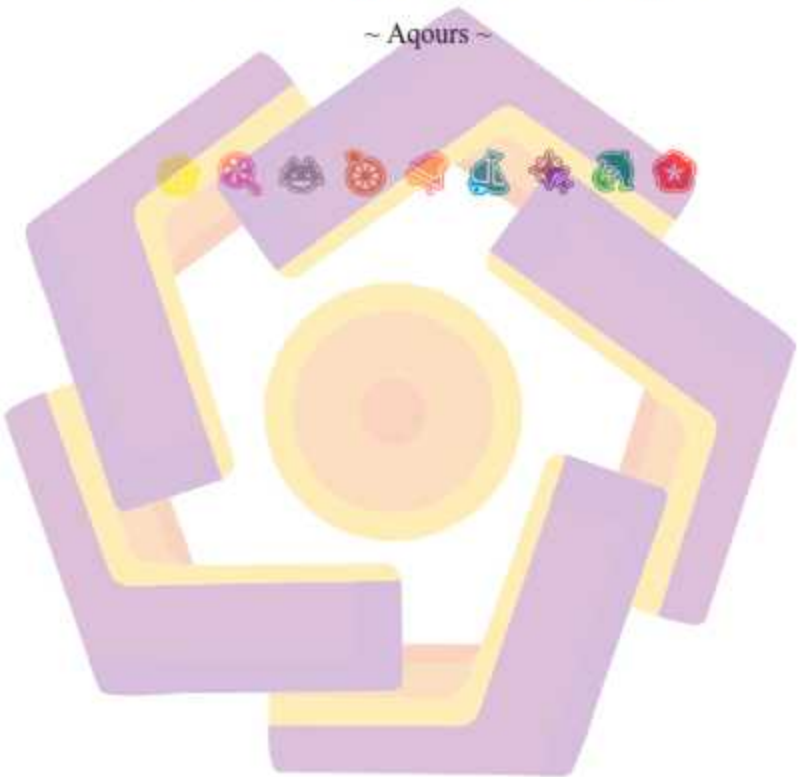
Diriku di Masa Depan, Ini adalah bukti ketekunanmu. Jadikan ini sebagai pengingat bahwa tidak ada usaha yang mengkhianati hasil. Zero kara ichi e—dari nol menjadi satu, dan seterusnya.

HALAMAN MOTTO

"Zero kara ichi e, ichi kara sono saki e"

"Dari nol ke satu, dari satu ke langkah selanjutnya"

~ Aqours ~



KATA PENGANTAR

Alhamdulillah, puji syukur penulis panjatkan kepada Allah SWT yang telah melimpahkan karunia-Nya sehingga penulis dapat menyelesaikan tesis dengan judul "Perbandingan kinerja algoritma naïve bayes dan SVM dalam analisis sentimen produk lulur tradisional dan modern di media sosial". Tesis ini disusun untuk memenuhi salah satu persyaratan dalam menyelesaikan Program Studi Magister S-2 Teknik Informatika di Universitas Amikom Yogyakarta. Rasa terima kasih penulis sampaikan kepada seluruh pihak yang telah membantu, membimbing dan mendukung, khususnya kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas Amikom Yogyakarta.
2. Prof. Dr. Ema Utami, S.Si., M.Kom. selaku dosen pembimbing yang selalu memberikan semangat, motivasi selama bimbingan dalam menyelesaikan tesis ini.
3. Bapak dan Ibu Dosen Universitas Amikom Yogyakarta yang telah memberikan banyak ilmu yang sangat bermanfaat bagi saya kedepannya.
4. Teman kelas yang telah menemani selama perkuliahan dan memberikan kenangan yang tidak akan terlupakan
5. Semua pihak yang telah membantu baik dukungan moril maupun materiil, pikiran dan tenaga dalam penyelesaian tesis ini.

Penulis menyadari bahwa pembuatan tesis ini banyak kekurangan dan kelemahan. Oleh karena itu penulis berharap kepada semua pihak agar dapat menyampaikan kritik dan saran yang membangun untuk menambah kesempurnaan tesis ini. Namun penulis tetap berharap tesis ini akan bermanfaat bagi semua pihak yang membacanya.

Cirebon, 26 September 2025

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Opini publik mengenai *sovereign wealth fund* (SWF) Danantara Indonesia yang diekspresikan di media sosial YouTube memerlukan metode analisis sentimen yang efektif untuk memahami persepsi masyarakat. Penelitian ini mengevaluasi secara komparatif performa model *deep learning* LSTM, GRU, dan BERT pada dataset 31.675 komentar YouTube berbahasa Indonesia yang memiliki ketidakseimbangan kelas (*class imbalance*) ekstrem dengan dominasi sentimen negatif. Metodologi penelitian mencakup komparasi model dasar, implementasi mekanisme *bidirectional*, serta upaya optimasi menggunakan *pre-trained embedding* FastText dan fungsi *Focal Loss*. Hasil penelitian menunjukkan bahwa model RNN dasar (unidireksional) memiliki performa tidak memadai (akurasi ~53%). Implementasi mekanisme *bidirectional* terbukti krusial, meningkatkan akurasi secara signifikan ke kisaran 72%. Optimasi lanjutan menunjukkan bahwa penggunaan *embedding* FastText pada arsitektur Bidirectional GRU berhasil mencapai performa RNN puncak sebesar 77,10%. Namun, penerapan *Focal Loss* terbukti kurang efektif dalam menangani ketidakseimbangan data pada dataset ini. Di sisi lain, arsitektur berbasis Transformer (BERT) menunjukkan superioritas absolut dengan akurasi tertinggi mencapai 89,66%, meskipun terindikasi mengalami instabilitas grafik *loss* dan membutuhkan biaya komputasi yang jauh lebih tinggi. Kesimpulannya, BERT adalah model paling akurat secara absolut, namun Bidirectional GRU yang dioptimalkan dengan FastText menawarkan keseimbangan (*trade-off*) terbaik antara akurasi tinggi dan efisiensi komputasi untuk implementasi praktis.

Kata Kunci: Analisis Sentimen, Deep Learning, LSTM, GRU, BERT

ABSTRACT

Public opinion regarding Indonesia's sovereign wealth fund (SWF), Danantara, as expressed on the social media platform YouTube, necessitates effective sentiment analysis methods to understand public perception. This study provides a comparative evaluation of the performance of LSTM, GRU, and BERT deep learning models on a dataset comprising 31,675 Indonesian-language YouTube comments characterized by extreme class imbalance with a dominance of negative sentiment. The methodology involves a comparison of baseline models, the implementation of a bidirectional mechanism, and optimization efforts using FastText pre-trained embeddings and the Focal Loss function. The findings indicate that baseline unidirectional RNN models (LSTM/GRU) yielded inadequate performance (accuracy ~53%). The implementation of a bidirectional mechanism proved to be a crucial intervention, boosting accuracy significantly to approximately 72%. Advanced optimization demonstrated that using FastText embeddings on the Bidirectional GRU architecture achieved a peak RNN performance of 77.10%. However, the application of Focal Loss proved to be less effective in addressing data imbalance within this dataset. On the other hand, the Transformer-based architecture (BERT) demonstrated absolute superiority, reaching the highest accuracy of 89.66%, albeit with indications of loss graph instability and substantially higher computational costs. In conclusion, while BERT is the most accurate model in absolute terms, the Bidirectional GRU optimized with FastText offers the best trade-off between high accuracy and computational efficiency for practical implementation.

Keywords: *Sentiment Analysis, Deep Learning, LSTM, GRU, BERT*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Di era digitalisasi yang semakin berkembang, media sosial telah menjadi salah satu platform utama bagi masyarakat untuk berinteraksi, berbagi informasi, dan mengekspresikan opini mereka terhadap berbagai isu yang berkembang. Salah satu platform yang paling populer adalah YouTube, platform ini masih menjadi yang memiliki waktu akses tinggi dibanding platform sosial popular lainnya (Simon Kemp, 2024). Kolom komentar pada video YouTube sering kali mencerminkan opini publik secara real-time, sehingga menjadi sumber data yang sangat berharga untuk analisis sentimen.

Salah satu topik yang menarik perhatian publik di Indonesia belakangan ini adalah keberadaan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara Indonesia). Badan ini merupakan lembaga strategis yang dibentuk untuk mengonsolidasikan dan mengoptimalkan investasi pemerintah dalam rangka mendukung pertumbuhan ekonomi nasional (Danantara, 2025). Badan ini merupakan *sovereign wealth fund* (SWF) atau dana kekayaan negara yang dibentuk untuk mengonsolidasikan dan mengoptimalkan investasi pemerintah dalam rangka mendukung pertumbuhan ekonomi nasional.

Keberadaan Danantara Indonesia sebagai institusi baru di Indonesia membuatnya menjadi topik yang hangat diperbincangkan di kalangan masyarakat. Sebagai *sovereign wealth fund*, Danantara bertujuan untuk mengelola aset-aset strategis negara, menarik investasi global, dan memperkuat daya saing Indonesia di

sektor-sektor kritis seperti infrastruktur, teknologi, dan energi (Balding, 2012). Namun, karena konsep SWF masih relatif baru di Indonesia, banyak masyarakat yang belum sepenuhnya memahami mekanisme kerja, tujuan, dan dampak jangka panjang dari badan ini. Ketidaktahuan ini sering kali memicu berbagai reaksi, baik positif, netral maupun negatif, yang tercermin dalam kolom komentar video YouTube yang membahas topik ini.

Danantara Indonesia dapat dianalogikan sebagai sebuah "produk" yang diluncurkan di pasar publik, di mana rakyat sebagai konsumen memiliki kesempatan untuk memberikan review melalui berbagai platform digital, salah satunya kolom komentar YouTube. Hal ini sejalan dengan penelitian yang dilakukan oleh Shaik Vadla, dkk tentang bagaimana dapat meningkatkan design produk berdasarkan sentimen analisis ulasan penggunanya (Shaik Vadla dkk., 2024). Pada penelitian yang dilakukan oleh Bello, dkk mengusulkan penggunaan dataset selain online untuk penelitian serupa (Bello dkk., 2023). Akan tetapi meskipun sifat dataset online sering kali dipertanyakan karena kemungkinan adanya bias atau pengaruh dari pihak luar seperti wisatawan atau individu dengan pemahaman terbatas tentang topik tersebut, kolom komentar YouTube tetap lebih merepresentasikan opini masyarakat lokal karena mencerminkan tanggapan spontan dan autentik dari warga negara Indonesia yang secara langsung merasakan dampak kebijakan ini. Hal ini menjadi mendesak untuk diteliti karena analisis sentimen terhadap review masyarakat dapat digunakan sebagai alat prediktif untuk memahami bagaimana penerimaan masyarakat terhadap Danantara di masa mendatang.

Reaksi masyarakat terhadap Danantara Indonesia sangat beragam. Beberapa pihak menyambut baik inisiatif ini sebagai langkah strategis untuk meningkatkan pertumbuhan ekonomi dan daya saing global Indonesia. Namun, tidak sedikit pula yang merasa skeptis atau bahkan kritis terhadap transparansi, akuntabilitas, dan potensi risiko yang mungkin ditimbulkan oleh pengelolaan dana negara dalam skala besar (Gelb dkk., 2014). Komentar-komentar di YouTube menjadi cerminan langsung dari dinamika opini publik ini, mulai dari dukungan antusias hingga kritik pedas yang mencerminkan ketidakpercayaan atau kekhawatiran terhadap implementasi kebijakan Danantara.

Ketidaktahuan terhadap dinamika opini publik berpotensi menciptakan kesenjangan komunikasi antara pemerintah dan masyarakat, yang dapat memicu resistensi terhadap proyek-proyek strategis yang didanai oleh Danantara Indonesia, seperti infrastruktur dan energi, sehingga menghambat pertumbuhan ekonomi nasional (Balding, 2012). Selain itu, kurangnya pemahaman tentang sentimen yang berkembang di masyarakat dapat menyebabkan kegagalan dalam merancang strategi komunikasi yang efektif, yang menurunkan kepercayaan publik terhadap pemerintah. Di era digital, narasi yang menyebar melalui media sosial juga dapat merusak citra Danantara sebagai *sovereign wealth fund*, sehingga mengurangi minat investor global dan memengaruhi daya saing Indonesia di kancah internasional. Terakhir, ketidakmampuan untuk memprediksi tren opini publik melalui analisis sentimen kolom komentar YouTube berisiko menghilangkan peluang untuk mengantisipasi potensi masalah sebelum eskalasi menjadi krisis

yang lebih besar, yang pada akhirnya dapat mengancam keberlanjutan program investasi pemerintah.

Untuk memahami opini publik secara lebih mendalam, analisis sentimen menjadi alat yang sangat berguna. Analisis sentimen adalah teknik dalam bidang *Natural Language Processing* (NLP) yang digunakan untuk mengidentifikasi, mengklasifikasikan, dan mengekstraksi informasi emosional dari teks tertulis (Jency Jose & Simritha R, 2024). Teknik ini memungkinkan kita untuk memahami pola-pola sentimen dalam data teks, seperti apakah suatu komentar bersifat positif, negatif, atau netral. Dengan memanfaatkan teknologi deep learning, analisis sentimen dapat dilakukan dengan tingkat akurasi yang lebih tinggi dibandingkan metode tradisional seperti *lexicon-based approaches*.

Dalam penelitian ini, tiga model deep learning yang populer dipilih untuk dianalisis secara komparatif: LSTM (*Long Short-Term Memory*), GRU (*Gated Recurrent Unit*), dan BERT (*Bidirectional Encoder Representations from Transformers*). LSTM dan GRU merupakan model berbasis *Recurrent Neural Networks (RNN)* yang dirancang untuk menangani data sequential, seperti teks, dengan efektif. LSTM dikenal karena kemampuannya dalam menangkap dependensi jangka panjang dalam data, sementara GRU menawarkan arsitektur yang lebih ringkas namun tetap efektif dalam banyak kasus (Abumohsen dkk., 2023). Di sisi lain, BERT menawarkan pendekatan yang lebih canggih dengan memanfaatkan arsitektur transformer dan kemampuan *contextual understanding*. Model ini mampu memahami konteks kata-kata dalam suatu kalimat secara

mendalam, sehingga sering kali menghasilkan performa yang lebih baik dalam tugas-tugas NLP seperti analisis sentimen (Devlin dkk., 2019).

Meskipun penelitian terkait analisis sentimen telah banyak dilakukan, sebagian besar studi masih berfokus pada data berbahasa Inggris dan platform sosial seperti Twitter. Penelitian oleh Zhang (2024) menunjukkan bahwa model dua lapis bidirectional LSTM mampu menghasilkan akurasi tinggi dalam klasifikasi sentimen, mengungguli RNN dan GRU. Namun, studi tersebut hanya menggunakan dataset tunggal dari Twitter, sehingga belum dapat menggambarkan performa model pada jenis data lain yang memiliki karakteristik berbeda, seperti komentar publik di YouTube. Padahal, YouTube menyajikan data yang lebih kompleks, bebas struktur, dan kaya konteks, sehingga bisa menjadi tantangan tersendiri dalam analisis sentimen.

Selain keterbatasan pada jenis data, pendekatan transformer seperti BERT dalam studi Zhang hanya disebutkan secara konseptual tanpa dilakukan eksplorasi eksperimental. Ini membuka peluang untuk meneliti bagaimana performa BERT secara empiris, khususnya dalam konteks bahasa informal dan ekspresif seperti komentar YouTube. Di sisi lain, belum ada kajian mendalam terkait kelebihan dan kekurangan masing-masing model dalam menghadapi fenomena linguistik khas media sosial, seperti ironi, sarkasme, penggunaan emoji, atau komentar dengan sentimen campuran.

Penelitian Zhang juga tidak membahas aspek efisiensi model, seperti waktu pelatihan dan kecepatan inferensi, yang sangat penting untuk penerapan nyata, terutama jika digunakan sebagai sistem pemantauan opini publik secara real-time.

Oleh karena itu, penelitian ini tidak hanya membandingkan performa LSTM, GRU, dan BERT dalam menganalisis sentimen, tetapi juga bertujuan mengembangkan pendekatan yang lebih adaptif terhadap dinamika data sosial lokal.

Dalam salah satu studi, BERT terbukti memiliki performa terbaik dalam menghadapi ketidakpastian (uncertainty) prediksi (Islam dkk., 2022), menjadikannya model yang lebih dapat dipercaya untuk diterapkan di lingkungan nyata. Aspek uncertainty ini semakin penting karena dalam aplikasi nyata, terutama yang menyangkut opini publik, tidak cukup hanya mengandalkan akurasi. Diperlukan model yang tidak hanya akurat, tetapi juga mampu menunjukkan sejauh mana keyakinannya terhadap sebuah prediksi. Penggunaan metode seperti Monte Carlo Dropout (MCD) juga memungkinkan identifikasi prediksi yang kurang pasti, sehingga dapat meningkatkan keandalan sistem secara keseluruhan.

Dengan membandingkan serta mengembangkan kembali performa model ini, penelitian ini bertujuan untuk menemukan model yang paling optimal dalam menganalisis sentimen kolom komentar YouTube tentang Danantara Indonesia. Hasil penelitian ini diharapkan dapat memberikan wawasan yang bermanfaat bagi pembuat kebijakan, pengelola investasi, dan masyarakat umum tentang persepsi publik terhadap Danantara Indonesia.

1.2. Rumusan Masalah

Rumusan masalah ini disusun berdasarkan latar belakang yang telah dikemukakan, masalah diuraikan dalam bentuk poin-poin yang mencakup masalah yang akan dibahas dalam penelitian. Berikut ini merupakan rumusan masalah penelitian:

a. Bagaimana karakteristik data yang terdapat dalam kolom komentar video YouTube terkait Danantara Indonesia serta kaitannya terhadap performa metode yang digunakan?

b. Bagaimana performa model LSTM, GRU, dan BERT dalam menganalisis data sentimen kolom komentar YouTube tentang Danantara Indonesia?

c. Model mana yang paling optimal di antara LSTM, GRU, dan BERT dalam analisis sentimen kolom komentar YouTube tentang Danantara Indonesia berdasarkan metrik evaluasi?

d. Bagaimana pendekatan atau pengembangan lanjutan dapat dilakukan untuk meningkatkan performa analisis sentimen menggunakan model-model tersebut pada data serupa?

e. Apa saja kelebihan dan kekurangan masing-masing model (LSTM, GRU, dan BERT) dalam analisis sentimen kolom komentar YouTube tentang Danantara Indonesia?

1.3. Batasan Masalah

Dalam penelitian, beberapa dan batasan penelitian akan dijabarkan sebagai berikut untuk memastikan penelitian tetap terfokus:

a. Penelitian ini secara khusus difokuskan pada analisis sentimen kolom komentar YouTube yang relevan dengan topik Danantara Indonesia, sebagai representasi opini publik terhadap kebijakan investasi pemerintah.

b. Analisis hanya mencakup teks komentar tanpa mempertimbangkan metadata seperti waktu posting, jumlah likes/dislikes, atau panjang komentar.

c. Sentimen yang dianalisis dibagi menjadi tiga kategori utama: positif, netral, dan negatif.

d. Penelitian ini tidak mempertimbangkan dimensi emosi yang lebih spesifik, seperti bahagia, sedih, marah, atau terkejut, karena fokus utama adalah pada polaritas sentimen.

e. Penelitian ini hanya membandingkan performa tiga model deep learning: LSTM (*Long Short-Term Memory*), GRU (*Gated Recurrent Unit*), dan BERT (*Bidirectional Encoder Representations from Transformers*).

f. Penelitian ini tidak mempertimbangkan kombinasi model (ensemble) atau pendekatan lain seperti SVM, Naïve Bayes, atau Decision Trees.

g. Pemilihan ketiga model ini didasarkan pada popularitas dan relevansinya dalam literatur analisis sentimen teks berbasis konteks, serta kemampuan masing-masing model dalam menangani data sequential dan memahami konteks bahasa.

h. Dataset yang digunakan berasal dari kolom komentar video YouTube yang relevan dengan topik Danantara Indonesia, dengan asumsi bahwa platform ini mencerminkan opini publik secara real-time dan autentik.

i. Dataset hanya mencakup komentar dalam bahasa Indonesia, tanpa mempertimbangkan komentar dalam bahasa asing atau bahasa campuran.

j. Performa model dievaluasi berdasarkan metrik standar dalam analisis sentimen, yaitu: akurasi, presisi, recall, dan F1-Score.

k. Penelitian ini tidak menggunakan metrik tambahan seperti ROC-AUC atau *confusion matrix* secara mendalam, karena fokus utama adalah pada evaluasi performa keseluruhan.

l. Hasil penelitian ini hanya berlaku untuk analisis sentimen terhadap kolom komentar YouTube tentang Danantara Indonesia dan tidak dapat digeneralisasikan untuk topik atau platform media sosial lainnya.

m. Penelitian ini juga tidak mempertimbangkan perbedaan demografi responden (usia, gender, lokasi geografis) karena keterbatasan informasi dalam dataset.

n. Penelitian ini dilakukan dengan asumsi bahwa sumber daya komputasi yang digunakan (misalnya GPU atau CPU) memadai untuk melatih model deep learning, terutama BERT yang membutuhkan sumber daya lebih besar dibandingkan LSTM dan GRU.

o. Jika terdapat kendala teknis terkait pelatihan model, penelitian ini akan memprioritaskan optimasi parameter untuk memastikan hasil yang optimal.

1.4. Tujuan Penelitian

Tujuan penelitian adalah hal-hal yang diharapkan dapat dicapai dalam penelitian ini. Berikut adalah pemaparan terkait tujuan penelitian:

a. Mengidentifikasi karakteristik data sentimen Danantara Indonesia yang terdapat dalam data kolom komentar video YouTube serta kaitannya terhadap performa metode yang digunakan.

b. Mengevaluasi performa model LSTM, GRU, dan BERT dalam menganalisis data sentimen kolom komentar YouTube tentang Danantara Indonesia.

c. Menentukan model yang paling optimal di antara LSTM, GRU, dan BERT dalam analisis sentimen kolom komentar YouTube tentang Danantara Indonesia berdasarkan metrik evaluasi.

d. Mengkaji pendekatan atau pengembangan lanjutan yang dapat diterapkan untuk meningkatkan performa analisis sentimen pada data komentar YouTube menggunakan model-model tersebut.

e. Menganalisis kelebihan dan kekurangan masing-masing model (LSTM, GRU, dan BERT) dalam analisis sentimen kolom komentar YouTube tentang Danantara Indonesia.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberi manfaat dalam meningkatkan pengetahuan ilmiah di bidangnya dan dapat menjadi landasan bagi penelitian lanjutan. Berikut adalah manfaat yang dapat diperoleh dari penelitian ini:

a. Penelitian ini memberikan wawasan baru tentang performa model deep learning (LSTM, GRU, dan BERT) dalam analisis sentimen.

b. Hasil penelitian ini dapat membantu pemerintah atau pengelola Danantara Indonesia memahami persepsi publik terhadap inisiatif mereka.

c. Pengelola media sosial atau kreator konten YouTube dapat menggunakan hasil penelitian ini untuk memahami tren opini publik terkait topik tertentu.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Penelitian yang dilakukan oleh Mohbey, dkk.(2024) berfokus pada analisis sentimen untuk memahami persepsi publik terhadap wabah penyakit menular, khususnya Monkeypox, yang telah dilaporkan di lebih dari 73 negara. Penelitian ini menggunakan data dari platform media sosial, yaitu tweet terkait *Monkeypox*, untuk menganalisis bagaimana masyarakat merespons dan memandang penyakit tersebut. Tujuan utamanya adalah membantu pembuat kebijakan dalam memahami pandangan publik terhadap Monkeypox sehingga dapat mengambil langkah-langkah yang lebih tepat dalam menangani kekhawatiran masyarakat. Metode yang digunakan dalam penelitian ini adalah arsitektur hibrida berbasis CNN-LSTM (*Convolutional Neural Network - Long Short-Term Memory*). Dataset yang digunakan adalah dataset tweet terbuka yang telah melalui serangkaian proses preprocessing, termasuk *global vectorization* dan *one-hot encoding*. Hasil eksperimen menunjukkan bahwa model hibrida CNN-LSTM mencapai akurasi sekitar 91%, yang lebih tinggi dibandingkan dengan metode machine learning konvensional seperti SVM atau Naive Bayes. Validasi tambahan juga dilakukan untuk memastikan keandalan hasil. Beberapa keterkaitan penelitian ini dengan penelitian yang dilakukan oleh Mohbey, dkk adalah sebagai berikut:

1. Seperti penelitian Mohbey, penelitian ini juga berfokus pada analisis sentimen data teks dari platform media sosial. Namun, sumber data pada penelitian ini adalah kolom komentar YouTube, yang memiliki

karakteristik unik seperti bahasa informal, emoji, dan variasi panjang teks.

2. Tujuan utama Mohbey adalah membantu pembuat kebijakan memahami pandangan publik terhadap Monkeypox. Tujuan dari penelitian ini adalah menganalisis persepsi publik terhadap video YouTube tentang Danantara Indonesia. Keduanya bertujuan untuk menggali wawasan dari opini publik melalui analisis sentimen.

Penelitian oleh Gupta, dkk.(2024) berfokus pada pengembangan teknik baru untuk Aspect-Based Sentiment Analysis (ABSA), yang merupakan pendekatan lebih rinci dibandingkan analisis sentimen konvensional. ABSA bertujuan untuk mengidentifikasi polaritas sentimen (positif, netral, atau negatif) terhadap aspek-aspek spesifik dalam sebuah dokumen atau kalimat. Misalnya, dalam kalimat "Kamera ponsel ini bagus, tetapi baterainya buruk," ABSA akan mengidentifikasi bahwa "kamera" memiliki sentimen positif, sementara "baterai" memiliki sentimen negatif. Beberapa keterkaitan penelitian ini dengan penelitian yang dilakukan oleh Gupta, dkk adalah sebagai berikut:

1. Dataset yang digunakan pada penelitian tersebut berupa ulasan produk atau layanan, sementara dataset pada penelitian ini adalah komentar YouTube. Meskipun tampak berbeda, kedua jenis dataset memiliki tantangan serupa seperti bahasa informal dan konteks dinamis.

Penelitian yang dilakukan oleh Wan, dkk.(2024) berfokus pada pengembangan model ECR-BERT (Emotion-Cognitive Reasoning integrated BERT) untuk analisis sentimen terhadap opini publik online tentang kejadian

darurat (OPOEs). Kejadian darurat sering kali memicu emosi yang kompleks dan beragam, sehingga analisis sentimen terhadap OPOEs menjadi sangat menantang. Untuk mengatasi tantangan ini, penelitian ini mengusulkan pendekatan baru yang mengintegrasikan BERT dengan model emosi untuk memberikan hasil analisis sentimen yang lebih akurat dan dapat dijelaskan. Beberapa keterkaitan penelitian ini dengan penelitian yang dilakukan oleh Gupta, dkk adalah sebagai berikut:

1. Kesamaan metode BERT sebagai focus dalam penelitian yang dilakukan.

Penelitian oleh Muhammet Sinan Başarslan dan Fatih Kayaalp (2023) mengusulkan model MBi-GRUMCONV , sebuah pendekatan baru berbasis deep learning untuk analisis sentimen pada dataset ulasan film IMDB. Model ini dirancang untuk meningkatkan performa klasifikasi sentimen dengan menggabungkan enam lapisan *Bidirectional Gated Recurrent Unit* (Bi-GRU) dan dua lapisan *Convolutional Neural Network* (CNN) . Penelitian ini menyoroti pentingnya penggunaan arsitektur *multi-layered* dan kombinasi teknik neural network yang berbeda untuk meningkatkan akurasi prediksi dalam tugas analisis sentimen. Model MBi-GRUMCONV menggunakan dua metode Word2Vec, yaitu Skip Gram dan *Continuous Bag of Words* (CBOW) , untuk merepresentasikan teks ulasan dalam bentuk vektor dengan tiga ukuran vektor berbeda (100, 200, dan 300). Hasil eksperimen menunjukkan bahwa model ini mencapai akurasi sebesar 95.34%, yang melampaui hasil studi sebelumnya dalam literatur. Selain itu, penelitian ini menemukan bahwa metode Skip Gram memberikan kontribusi yang lebih baik terhadap keberhasilan klasifikasi dibandingkan CBOW. Hal ini disebabkan oleh

kemampuan Skip Gram dalam menangkap probabilitas kata-kata di sekitar target kata, sehingga lebih efektif dalam merepresentasikan konteks kompleks. Beberapa keterkaitan penelitian ini dengan penelitian yang dilakukan oleh Başarslan adalah sebagai berikut:

1. Model MBi-GRUMCONV mencapai akurasi 95.34%, yang sangat tinggi dapat digunakan sebagai benchmark untuk membandingkan performa LSTM, GRU, dan BERT pada penelitian yang sedang kulakukan.
2. Dataset Başarslan berasal dari ulasan film IMDB, sementara dataset penelitian ini adalah komentar YouTube tentang Danantara Indonesia. Meskipun domainnya berbeda, kedua dataset memiliki tantangan serupa, seperti bahasa informal dan variasi panjang teks.

2.2. Keaslian Penelitian

Peneliti mengembangkan ide baru dan menginovasi yang sudah ada, dengan menganalisis literatur ilmiah untuk menemukan perbedaan pengetahuan. Fokusnya kebaruan dan kontribusi pada pengetahuan dalam penelitian ini ditunjukkan dalam Tabel 2.1.

Tabel 2.1. Matriks literatur review dan posisi penelitian
Komparasi Performa LSTM, GRU, dan BERT dalam Analisis Sentimen Kolom Komentar Video Youtube tentang Danantara Indonesia

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	*A BERT Framework to Sentiment Analysis of Tweets* (Bello dkk., 2023)	Abayomi Bello, Sin-Chun Ng, Man-Fai Leung, Sensors, 2023	Penelitian ini mengusulkan kombinasi BERT dengan model lain seperti CNN, RNN, dan BiLSTM untuk meningkatkan akurasi, presisi, *recall*, dan *F1-score* dalam analisis sentimen. Tujuan utama penelitian ini adalah untuk menunjukkan bahwa BERT, yang memproses seluruh kalimat secara simultan dan	Hasil eksperimen dalam penelitian ini menunjukkan bahwa kombinasi BERT dengan model lain seperti CNN, RNN, dan BiLSTM memberikan performa yang sangat baik dalam analisis sentimen. Secara spesifik, kombinasi ini mencapai tingkat akurasi, presisi, recall, dan F1-score yang lebih tinggi dibandingkan dengan penggunaan Word2Vec atau tanpa	Dataset yang digunakan dalam penelitian ini berasal dari sumber online seperti Twitter, yang berpotensi mencakup data dari pengguna yang tidak memiliki pemahaman mendalam tentang topik yang dibahas.	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu BERT sedangkan pada penelitian ini digunakan tiga kombinasi metode. Dataset yang digunakan pada penelitian tersebut juga menggunakan *tweet*, sedangkan dalam penelitian ini menggunakan komentar pada kolom komentar video.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			mempertimbangkan konteks sebelum dan sesudah kata, dapat memberikan hasil yang lebih baik dibandingkan dengan metode tradisional yang hanya fokus pada representasi statis atau searah.	varian tambahan sama sekali.		
2	*A Multi-Aspect Informed GRU: A Hybrid Model of Flight Fare Forecasting with Sentiment Analysis* (Degife & Lin, 2024)	Worku Abebe Degife, Bor-Shen Lin, Applied Sciences, 2024	Penelitian ini bertujuan untuk mengembangkan metode canggih dalam memprediksi harga tiket penerbangan dengan mengintegrasikan analisis sentimen berbasis aspek (ABSA) dan teknik pembelajaran mendalam, khususnya model *Gated Recurrent Unit* (GRU). Pendekatan ini	Penelitian ini berhasil membuktikan bahwa integrasi ABSA dengan model GRU secara signifikan meningkatkan performa prediksi harga tiket penerbangan dibandingkan dengan model konvensional. Model ini menunjukkan hasil prediksi yang sangat akurat, dengan nilai *Root Mean Square Error* (RMSE) sebesar 0.0071, *Mean Absolute Error* (MAE) sebesar	Penelitian ini bergantung pada data transaksi historis dan ulasan pelanggan, yang mungkin tidak sepenuhnya mencakup semua faktor eksternal yang memengaruhi harga tiket, seperti perubahan kebijakan maskapai, fluktuasi harga bahan bakar, atau kondisi geopolitik. Oleh karena itu, disarankan untuk memasukkan lebih banyak variabel eksternal dalam model	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu GRU sedangkan pada penelitian ini digunakan tiga kombinasi metode. Penelitian tersebut memiliki focus yang berbeda karena mencoba untuk melakukan prediksi harga tiket, sedangkan penelitian yang dilakukan saat ini lebih untuk membandingkan kemampuan antara dua metode.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			memanfaatkan data transaksi historis tiket pesawat serta ulasan pelanggan untuk memahami dinamika harga tiket penerbangan dan dampak sentimen pelanggan terhadap penetapan harga.	0.0137, dan koefisien determinasi (R^2) sebesar 0.9899.	untuk meningkatkan generalisasi prediksi.	
3	*An Automatic Sentiment Analysis Method for Short Texts Based on Transformer-BERT Hybrid Model* (Xiao & Luo, 2024)	Haiyan Xiao, Linghua Luo, IEEE Access, 2024	Penelitian ini bertujuan untuk mengatasi tantangan dalam analisis sentimen terhadap teks pendek, yang sering kali hanya memiliki informasi semantik yang terbatas. Untuk mengatasi keterbatasan ini, penelitian ini mengusulkan metode otomatisasi analisis sentimen berbasis model hibrida	Hasil eksperimen menunjukkan bahwa metode analisis sentimen berbasis model hibrida *Transformer*-BERT yang diusulkan dalam penelitian ini memberikan performa yang sangat baik dalam berbagai indikator evaluasi, termasuk Akurasi, Presisi, Recall, dan F1-score. Metode ini mencapai peningkatan signifikan dibandingkan dengan metode tradisional serta	Algoritma yang diusulkan terutama dirancang untuk analisis sentimen teks pendek dan mungkin memiliki keterbatasan dalam mengatasi teks panjang. Hal ini disebabkan oleh fokus utama model pada representasi fitur teks pendek, yang mungkin tidak sepenuhnya sesuai untuk teks dengan panjang yang lebih besar.	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu BERT sedangkan pada penelitian ini digunakan tiga kombinasi metode. Dataset yang digunakan pada penelitian tersebut lebih focus pada teks pendek, sedangkan dataset yang digunakan pada penelitian ini menggunakan komentar dari video YouTube. Komentar dari video YouTube memiliki panjang yang bervariasi.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			Transformer-BERT. Tujuan utama dari penelitian ini adalah untuk meningkatkan kemampuan representasi fitur teks dengan menggabungkan vektor kata yang diekstraksi menggunakan struktur BERT dan vektor topik yang dihasilkan oleh model CTM (*Combined Topic Model*).	metode yang hanya bergantung pada model BERT. Keunggulan utama metode ini terletak pada integrasi mekanisme perhatian dari Transformer, yang memungkinkan model untuk secara akurat menangkap dan fokus pada informasi kontekstual penting dalam teks.		
4	*Dose My Opinion Count? A CNN-LSTM Approach for Sentiment Analysis of Indian General Elections* (N. Zhang dkk., 2024)	Ning Zhang, Jize Xiong, Zhiming Zhao, Mingyang Feng, Xiaosong Wang, Yuxin Qiao, Chufeng Jiang, Journal	Penelitian ini bertujuan untuk menganalisis sentimen pada platform media sosial, khususnya Reddit, dalam konteks peristiwa penting seperti Pemilu Umum India 2019. Tujuan utama	Hasil eksperimen menunjukkan bahwa model hibrida CNN+LSTM mencapai performa superior dalam tugas klasifikasi sentimen dibandingkan dengan model individu seperti CNN atau LSTM. Model hibrida ini berhasil	Penelitian ini hanya fokus pada dataset Reddit yang berkaitan dengan Pemilu Umum India 2019, sehingga generalisasi model ke topik atau bahasa lain belum sepenuhnya dieksplorasi. Oleh karena itu, disarankan untuk menguji model ini	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu LSTM sedangkan pada penelitian ini digunakan tiga kombinasi metode. Perbedaan pada penelitian tersebut ada pada platform

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		of Theory and Practice of Engineering Science,	dari penelitian ini adalah untuk memahami dinamika opini publik dan tren sentimen yang berkembang di sekitar diskusi politik di platform media sosial.	menggabungkan kekuatan CNN dalam mengidentifikasi pola lokal dan kemampuan LSTM dalam memahami dependensi kontekstual, sehingga mampu menangkap fitur lokal dan kontekstual yang inheren dalam komentar Reddit.	pada dataset yang lebih beragam, baik dalam hal topik maupun bahasa, untuk memastikan robustness dan adaptabilitas model.	dan focus permasalahannya. Pada penelitian tersebut focus yang jadi masalah adalah Pemilu Umum di India, sedangkan untuk penelitian ini membahas produk dari Pemerintah Republik Indonesia yaitu Danantara. Kemudian dataset yang digunakan berasal dari Reddit, sedangkan pada penelitian ini menggunakan media social YouTube.
5	*Enhanced Aquila Optimizer Combined Ensemble Bi-LSTM-GRU With Fuzzy Emotion Extractor for Tweet Sentiment Analysis and Classification* (Sherin dkk., 2024)	A. Sherin, I. Jasmine Selvakumari Jeya, S. N. Deepa, IEEE Access, 2024	Tujuan utama penelitian ini adalah untuk memanfaatkan teknik pembelajaran mendalam (deep learning) dan sistem pendukung keputusan berbasis logika fuzzy untuk mengekstraksi fitur sentimen dari teks tweet secara akurat. Untuk mencapai	Hasil penelitian menunjukkan bahwa model EAQ-FEE-enBi-LSTM-GRU yang diusulkan berhasil mencapai performa superior dibandingkan dengan model state-of-the-art lainnya dalam tugas analisis sentimen tweet. Model ini menggabungkan *Fuzzy Emotion Extractor* (FEE) untuk	Saran penelitian ini untuk menguji model ini pada dataset dari platform lain seperti Reddit, Facebook, atau Instagram untuk memastikan robustness dan adaptabilitas model. Kedua, meskipun model ini berhasil menangkap fitur sentimen, penelitian ini belum sepenuhnya mengeksplorasi dampak emosi spesifik (seperti	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu menggabungkan beberapa metode seperti LSTM dan GRU dengan *Fuzzy Emotion Extractor* sedangkan pada penelitian ini digunakan tiga metode untuk dibandingkan mana yang lebih baik. Pada penelitian tersebut menambahkan logika *fuzzy*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			tujuan ini, penelitian ini mengusulkan kombinasi model ensemble *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan *Gated Recurrent Unit* (GRU), yang dioptimalkan menggunakan algoritma *Enhanced Aquila Optimizer* (EAQ).	mengekstraksi fitur emosi, ensemble Bi-LSTM-GRU untuk menangkap urutan kata dan hubungan kontekstual, serta *Enhanced Aquila Optimizer* (EAQ) untuk mengoptimalkan bobot dan bias model. Kombinasi ini memungkinkan model untuk secara efektif menangkap fitur sentimen dan melakukan klasifikasi dengan akurasi tinggi.	bahagia, sedih, marah, atau terkejut) dalam analisis sentimen	pada kombinasi ensemble, sehingga dapat mengarah ke sistem pengambil keputusan sebagai tambahan dari sentimen analisis yang dilakukan. Pada penelitian yang dilakukan saat ini focus pada perbandingan antara metode mana yang memiliki performa lebih baik.
6	*Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT* (Shaik Vadla dkk., 2024)	Mahammad Khalid Shaik Vadla, Mahima Agumbe Suresh, Vimal K. Viswanathan, Algorithms, 2024	Penelitian ini bertujuan untuk mengembangkan sebuah pipeline prediksi yang mampu mendeteksi aspek-aspek spesifik dalam data ulasan pelanggan dan melakukan analisis sentimen untuk	Hasil penelitian menunjukkan bahwa model BERT dan T5 mencapai tingkat akurasi yang sangat baik dalam mendeteksi aspek-aspek dalam data teks, yaitu 92% dan 91%, masing-masing. Model BERT, yang dikembangkan oleh	Penelitian ini hanya berfokus pada produk ramah lingkungan, sehingga generalisasi model ke kategori produk lain belum sepenuhnya dieksplorasi. Oleh karena itu, disarankan untuk menguji model ini pada dataset dari kategori	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu BERT sedangkan pada penelitian ini digunakan tiga kombinasi metode. Pada penelitian tersebut fokusnya adalah pada review

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			memahami emosi serta preferensi pelanggan. Tujuan utama dari penelitian ini adalah untuk memberikan wawasan berbasis data kepada desainer produk dan pengembang riset guna membantu mereka menciptakan produk yang lebih sesuai dengan harapan pelanggan.	Google sebagai kerangka kerja NLP canggih, menunjukkan performa unggul dibandingkan T5 dalam hal presisi, recall, F1-score, dan efisiensi komputasi. Keunggulan ini sebagian besar disebabkan oleh kemampuan BERT untuk memproses input secara dua arah (*bidirectional*) dan memahami konteks mendalam dari setiap kata dalam kalimat.	produk lain, seperti elektronik, pakaian, atau makanan, untuk memastikan robustness dan adaptabilitas model.	barang pada platform Amazon, ini memiliki kemiripan jika memandang pada Danantara sebagai objek produk buatan pemerintah.
7	*Enhancing Stock Market Prediction: A Hybrid RNN-LSTM Framework with Sentiment Analysis* (Kasture & Shirsath, 2024)	Pranjali Kasture, Kamini Shirsath, Indian Journal of Science and Technology, 2024	Penelitian ini bertujuan untuk mengembangkan metode baru yang inovatif untuk meningkatkan akurasi prediksi harga saham dengan menggabungkan analisis sentimen berbasis data berita	Hasil penelitian menunjukkan bahwa model hibrida RNN-LSTM yang diusulkan berhasil memberikan hasil yang sangat menjanjikan dalam hal akurasi prediksi harga saham. Model ini mencapai nilai Mean Absolute Error (MAE)	Penelitian ini hanya menggunakan data dari BSE Sensex, sehingga generalisasi model ke indeks pasar saham lain atau pasar global belum sepenuhnya dieksplorasi. Oleh karena itu, disarankan untuk menguji model ini pada dataset dari pasar saham	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu LSTM sedangkan pada penelitian ini digunakan tiga kombinasi metode. Sentimen analisis pada peneletian tersebut diarahkan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			dan algoritma jaringan saraf tiruan (*neural network*) canggih. Tujuan utama dari penelitian ini adalah untuk memanfaatkan informasi emosional dari artikel berita dan headline terkait pasar saham sebagai fitur tambahan dalam model prediksi, sehingga dapat memberikan wawasan yang lebih mendalam tentang dinamika pasar.	sebesar 0.036, Mean Squared Error (MSE) sebesar 0.021, dan Root Mean Square Error (RMSE) sebesar 0.046, yang secara signifikan lebih rendah dibandingkan dengan model SVR dan RFR. Selain itu, nilai R^2 (koefisien determinasi) dari model yang diusulkan menunjukkan peningkatan sebesar 0.40% hingga 5.5% dibandingkan dengan metode yang ada dalam literatur.	lain, seperti NASDAQ atau FTSE, untuk memastikan robustness dan adaptabilitas model.	untuk menjadi prediksi harga saham, karena harga saham tersebut diprediksi dengan melakukan sentimen analisis menggunakan metode yang digunakan.
8	*Fuzzy ensemble of fined tuned BERT models for domain-specific sentiment analysis of software engineering dataset* (Anwar dkk., 2024)	Zeeshan Anwar, Hammad Afzal, Naima Altaf, Seifedine Kadry, Jungeun Kim,	Penelitian ini bertujuan untuk mengembangkan alat analisis sentimen khusus domain Teknik Perangkat Lunak (*Software Engineering*/SE)	Hasil penelitian menunjukkan bahwa *Fuzzy Ensemble* yang diusulkan berhasil mencapai performa superior dalam analisis sentimen di domain SE dibandingkan dengan alat state-of-the-art yang	Penelitian ini hanya berfokus pada dataset SE yang berasal dari platform seperti Stack Overflow dan Jira, sehingga generalisasi model ke domain lain di luar SE (misalnya, ulasan produk atau	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu BERT yang dioptimalkan dengan *fuzzy ensemble* sedangkan pada

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		PLOS ONE, 2024	yang mampu mengatasi keterbatasan alat analisis sentimen umum dan model yang ada. Tujuan utamanya adalah untuk meningkatkan akurasi prediksi sentimen, terutama dalam mengidentifikasi sentimen negatif dan netral, yang sering kali tidak tertangani dengan baik oleh model umum atau alat domain-spesifik sebelumnya. Untuk mencapai hal ini, penelitian ini mengusulkan pendekatan hibrida berbasis deep learning dengan menggunakan varian model BERT (*Bidirectional*	ada. Model ini mencapai skor F1-score maksimum 0.883, dengan akurasi tertinggi pada dataset Stack Overflow (0.901), JavaLib (0.888), Code Review (0.877), dan Jira (0.785).	media sosial) belum diuji.	penelitian ini digunakan tiga kombinasi metode. Pada penelitian tersebut menambahkan logika *fuzzy ensemble* untuk meningkatkan kemampuan dari sentimen analisis menggunakan BERT. Tetapi tidak begitu focus untuk membandingkan diantara metode yang ada.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			Encoder Representations from Transformers) yang telah di-*fine-tune*, seperti Bert-Base, Bert-Large, Bert-LSTM, Bert-GRU, dan Bert-CNN.			
9	*MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis* (Başarslan & Kayaalp, 2023)	Muhammet Sinan Başarslan, Fatih Kayaalp, Journal of Cloud Computing, 2023	Penelitian ini bertujuan untuk mengembangkan model analisis sentimen berbasis deep learning yang inovatif untuk meningkatkan akurasi klasifikasi ulasan film pada dataset IMDB. Tujuan utamanya adalah menyelidiki dampak penggunaan arsitektur neural network bertingkat dan hibrida (gabungan lapisan *Bidirectional Gated*	Hasil eksperimen menunjukkan bahwa model MBi-GRUMCONV yang diusulkan mencapai akurasi tertinggi sebesar 95.34% pada dataset IMDB, melampaui performa studi-studi sebelumnya dalam literatur. Model ini menunjukkan keunggulan signifikan ketika menggunakan metode Word2Vec Skip Gram dengan ukuran vektor 300 dimensi, yang menghasilkan akurasi pelatihan	Model hanya diuji pada dataset IMDB, sehingga kemampuannya dalam menangani dataset lain (misalnya ulasan produk atau media sosial) belum diketahui. Disarankan untuk menguji model pada dataset yang lebih beragam.	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu GRU dengan optimasi lainnya sedangkan pada penelitian ini digunakan tiga kombinasi metode. Dataset yang digunakan pada penelitian tersebut terfokus pada ulasan film dari IMDB, memiliki kesamaan dari sisi besaran data karena ulasan memiliki rentang data yang bervariasi. Ulasan film juga memiliki focus yang mirip jika kita memandang

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			Recurrent Unit (Bi-GRU) dan *Convolutional Neural Network* (CNN)) terhadap performa model dalam tugas analisis sentimen. Untuk mencapai tujuan ini, penelitian ini mengusulkan model MBi-GRUMCONV, yang menggabungkan enam lapisan Bi-GRU dan dua lapisan CNN.	95.32%, validasi 94.67%, dan pengujian 95.34%.		Danantara dan film sebagai satu objek yang sama yaitu produk.
10	*Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis* (Jun Gu dkk., 2024)	Wen jun Gu, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, Chao Yan, ACM, 2024	Penelitian ini bertujuan untuk meningkatkan akurasi prediksi harga saham dengan menggabungkan analisis sentimen dari berita keuangan dan data historis harga saham. Tujuan utamanya	Hasil penelitian menunjukkan bahwa FinBERT-LSTM mencapai performa terbaik dibandingkan model LSTM dan DNN konvensional. Mengurangi kesalahan prediksi secara signifikan, dengan nilai MAE dan MAPE yang	Integrasi FinBERT dengan LSTM memerlukan sumber daya komputasi yang besar, yang mungkin menjadi hambatan untuk implementasi skala besar. Model ini hanya menggunakan data harga mingguan dan berita dari sumber tertentu	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu menggabungkan beberapa metode seperti LSTM dan BERT sedangkan pada penelitian ini digunakan tiga metode untuk dibandingkan mana yang lebih baik.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			adalah mengembangkan model prediksi yang memanfaatkan jaringan deep learning.	lebih rendah serta akurasi yang lebih tinggi.	(misalnya, Benzinga), sehingga mungkin tidak mencakup peristiwa real-time atau sumber berita alternatif.	Sentimen analisis pada penelitian tersebut focus pada berita, tetapi arahnya adalah untuk melakukan prediksi terhadap harga saham. Sehingga memiliki kesamaan pada sisi sentimen analisis dan metode yang digunakan serta juga memiliki perbedaan karena pada penelitian tersebut focus untuk kearah prediksi sedangkan pada penelitian ini focus pada membandingkan antara metode yang digunakan untuk mencari metode yang lebih baik.
11	*RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis* (Tan dkk., 2023)	Kian Long Tan, Chin Poo Lee, Kian Ming Lim, Applied Sciences, 2023	Tujuan utamanya adalah untuk meningkatkan akurasi dan robustness model dalam tugas klasifikasi sentimen, terutama ketika dihadapkan pada tantangan dataset yang tidak	Hasil eksperimen menunjukkan bahwa model hibrida RoBERTa-GRU yang diusulkan mencapai performa unggul dibandingkan metode lainnya dalam tugas analisis sentimen. Model ini mencapai akurasi 94.63% pada	Integrasi RoBERTa dan GRU memerlukan sumber daya komputasi yang besar, yang mungkin menjadi hambatan untuk implementasi skala besar atau aplikasi real-time. Model ini diuji pada dataset IMDb, Sentiment140, dan	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu menggabungkan beberapa metode seperti GRU dan BERT sedangkan pada penelitian ini digunakan tiga metode untuk dibandingkan mana yang lebih baik.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			seimbang. RoBERTa dipilih karena kemampuannya untuk memproyeksikan teks ke dalam ruang embedding diskriminatif melalui mekanisme perhatian (*attention mechanism*), sementara GRU digunakan untuk menangkap dependensi jangka panjang dalam urutan teks dan mengatasi masalah gradien yang hilang (*vanishing gradient problem*) yang sering muncul dalam Recurrent Neural Networks (RNN).	dataset IMDb, 89.59% pada dataset Sentiment140, dan 91.52% pada dataset Twitter US Airline Sentiment. Keberhasilan ini disebabkan oleh kombinasi sinergis antara RoBERTa dan GRU.	Twitter US Airline Sentiment, tetapi kemampuannya untuk menangani dataset dari domain lain (misalnya ulasan produk atau media sosial non-Inggris) belum dievaluasi.	Pada penelitian tersebut focus utamanya adalah untuk meningkatkan kemampuan algoritma yang digunakan dengan kombinasi RoBERTa dan GRU.
12	*Sentiment Analysis and Topic*	Jency Jose, Simritha R,	Penelitian ini bertujuan untuk	Penelitian ini berhasil menunjukkan potensi	Sistem ini diuji pada data media sosial tertentu,	Perbedaan dibandingkan penelitian terdahulu terdapat

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Classification with LSTM Networks and TextRazor (Jency Jose & Simritha R, 2024)	International Journal of Data Informatics and Intelligent Computing, 2024	mengembangkan sistem analisis sentimen dan ekstraksi topik yang komprehensif untuk meningkatkan pengalaman pengguna di platform media sosial. Dengan fokus pada menciptakan lingkungan digital yang lebih sehat dan konstruktif, penelitian ini memanfaatkan teknik Natural Language Processing (NLP) canggih untuk memberikan wawasan mendalam tentang nada emosional (*sentiment analysis*) dan tema utama (*topic extraction*)	besar teknik NLP canggih dalam menganalisis dinamika kompleks di media sosial. Dengan menggunakan jaringan LSTM, sistem mencapai akurasi 86% dalam klasifikasi sentimen, yang memungkinkan pengelompokan teks secara akurat ke dalam kategori positif, negatif, atau netral. Selain itu, penggunaan TextRazor untuk ekstraksi topik berhasil mengidentifikasi tema-tema utama dalam data teks, memberikan wawasan mendalam tentang konten diskusi media sosial.	tetapi kemampuannya untuk menangani variasi bahasa dan konteks di platform lain (misalnya Reddit atau TikTok) belum dievaluasi. Meskipun visualisasi intuitif membantu pengguna, fitur ini masih relatif sederhana dan dapat ditingkatkan dengan teknik visualisasi interaktif yang lebih canggih.	pada metode yang digunakan, penelitian terdahulu hanya menggunakan satu metode saja yaitu LSTM sedangkan pada penelitian ini digunakan tiga kombinasi metode. Selain sentimen analisis pada penelitian tersebut juga coba untuk melakukan klasifikasi pada dataset platform social media. Kesamaan terdapat pada metode yang digunakan yaitu LSTM, tetapi memiliki perbedaan pada kombinasi metodenya.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			dalam diskusi media sosial.			
13	*Sentiment analysis classification system using hybrid BERT models* (Talaat, 2023)	Amira Samy Talaat, Journal of Big Data, 2023	Penelitian ini bertujuan untuk mengembangkan model analisis sentimen yang lebih akurat dengan memanfaatkan kombinasi teknik pembelajaran mendalam (*deep learning*) dan model pra-latih berbasis BERT (*Bidirectional Encoder Representations from Transformers*). Fokus utama penelitian ini adalah meningkatkan kemampuan klasifikasi emosi dalam teks media sosial, terutama tweet, dengan menggunakan	Hasil penelitian menunjukkan bahwa arsitektur hibrida yang menggabungkan BERT dengan BiGRU dan BiLSTM berhasil meningkatkan akurasi analisis sentimen dibandingkan dengan model BERT standar. DistilBERT-GLG (tanpa emoji) mencapai peningkatan akurasi sebesar 1.84% dibandingkan DistilBERT standar pada dataset Apple, dan 0.24% pada dataset Airline. Keberadaan atau ketiadaan emoji memengaruhi performa model. Misalnya, akurasi DistilBERT-GLG turun dari 80.42% menjadi 79.24% setelah	Penggabungan lapisan BiGRU/BiLSTM dengan BERT memerlukan sumber daya komputasi yang besar, yang mungkin menjadi kendala untuk implementasi skala besar atau aplikasi real-time. Model ini diuji pada tiga dataset tertentu, tetapi kemampuannya untuk menangani variasi bahasa atau konteks budaya yang berbeda belum dievaluasi.	Penelitian tersebut menggunakan kombinasi tiga metode yang sama yaitu GRU, LSTM dan BERT, sedangkan pada penelitian yang dilakukan saat ini melakukan komparasi pada ketiga metode tersebut. Pada penelitian tersebut fokusnya adalah menggunakan BERT yang digabungkan dengan LST dan juga GRU. Penggabungan metode ini dilakukan dengan tujuan untuk meningkatkan kemampuan dari metode BERT menjadi lebih baik lagi.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			arsitektur hibrida yang menggabungkan BERT dengan algoritma *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Bidirectional Gated Recurrent Unit* (BiGRU).	emoji dihilangkan pada dataset CrowdFlower.		
14	*Sentiment analysis of hotel comments based on LSTM and GRU* (Xu, 2024)	Zeyu Xu, ACE, 2024	Penelitian ini bertujuan untuk membandingkan dan mengevaluasi performa berbagai pendekatan analisis sentimen menggunakan model pembelajaran mendalam (*deep learning*) seperti CNN (*Convolutional Neural Network*), LSTM (*Long Short-Term Memory*), dan GRU (*Gated*	Hasil penelitian menunjukkan bahwa model CNN+bilstm+GRU berhasil meningkatkan akurasi klasifikasi emosi sebesar 1% dibandingkan dengan model bilstm+GRU . Peningkatan ini disebabkan oleh penambahan lapisan CNN, yang membantu menangkap fitur lokal dalam teks secara lebih efektif. Pelatihan bersama (*joint training*)	Model ini hanya diuji pada dataset ulasan hotel, sehingga kemampuannya untuk menangani dataset dari domain lain (misalnya ulasan produk atau media sosial) belum dievaluasi. Kombinasi model seperti CNN+bilstm+GRU memerlukan sumber daya komputasi yang besar, yang mungkin menjadi kendala untuk implementasi skala besar atau aplikasi real-time.	Perbedaan dibandingkan penelitian terdahulu terdapat pada metode yang digunakan, penelitian terdahulu menggabungkan beberapa metode seperti GRU dan LSTM sedangkan pada penelitian ini digunakan tiga metode untuk dibandingkan mana yang lebih baik. Penelitian tersebut mencoba untuk meningkatkan tingkat akurasi dari metode yang digunakan dengan melakukan penambahan layer metode yaitu dengan adanya CNN,

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			Recurrent Unit). Fokus utama penelitian ini adalah untuk mengatasi tantangan dalam menganalisis data teks yang tidak terstruktur, kompleks, dan dipersonalisasi, seperti komentar hotel yang sering kali mengandung akronim atau bahasa informal. Untuk mencapai tujuan ini, penelitian ini mengusulkan dua model deep learning: CNN+bilstm+GRU dan bilstm+GRU, yang diuji pada dataset ulasan hotel.	dari beberapa model terbukti efektif dalam menggabungkan kekuatan berbagai arsitektur, sehingga meningkatkan performa generalisasi model secara keseluruhan. Temuan ini menegaskan bahwa kombinasi model deep learning dapat menjadi solusi yang kuat untuk tugas analisis sentimen, terutama ketika menghadapi dataset besar dan kompleks.		hal tersebut dilakukan dengan harapan agar metode yang digunakan dapat melakukan tugas dengan baik pada struktur bahasa informal dan dataset yang mengandung akronim. Pada penelitian ini peningkatan bukan menjadi nilai utama, karena nilai utama yang dicari pada penelitian ini adalah metode mana yang dapat melakukan sentimen analisis lebih baik pada permasalahan yang diangkat.

2.3. Landasan Teori

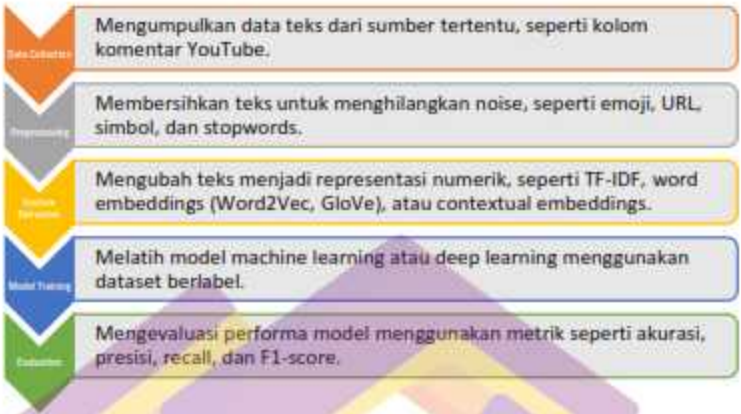
2.3.1. Analisis Sentimen

Analisis sentimen merupakan salah satu cabang dari *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau emosi yang terkandung dalam teks tertulis. Pada penelitian ini, analisis sentimen digunakan untuk memahami pandangn publik terhadap Badan Pengelola Investasi Daya Anagata Nusantara (Danantara Indonesia) berdasarkan data kolom komentar YouTube. Sentimen yang dianalisis dikategorikan ke dalam tiga kelas: positif, netral dan negatif.

Dengan teknik analisis sentimen, kita dapat mengevaluasi bagaimana opini publik terdistribusi terhadap inisiatif pemerintah melalui Danantara Indonesia. Jika sebagian besar sentimen bersifat negatif, hal ini dapat menjadi indikator bahwa ada ketidakpercayaan atau ketidakpuasan terhadap transparansi, akuntabilitas, atau dampak sosial dari kebijakan tersebut. Sebaliknya, jika sentimen dominan positif, hal ini dapat mencerminkan dukungan publik terhadap langkah-langkah strategis yang diambil oleh pemerintah.

Penelitian ini, analisis sentimen dilakukan menggunakan model deep learning seperti LSTM (*Long Short-Term Memory*), GRU (*Gated Recurrent Unit*), dan BERT (*Bidirectional Encoder Representations from Transformers*). Model-model ini dipilih karena dianggap dapat menangani data *sequential* dan dapat menghasilkan hasil analisis yang akurat.

Gambar 2.1 menunjukkan tahapan-tahapan yang perlu dilakukan dalam penelitian yang melibatkan penggunaan metode analisis sentimen.



Gambar 2.1. Tahapan Proses Sentimen Analisis

2.3.2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) adalah varian dari *Recurrent Neural Networks* (RNN) yang dirancang untuk mengatasi masalah utama dalam pelatihan RNN tradisional, yaitu *vanishing gradient* dan *exploding gradient*. LSTM diperkenalkan oleh Sepp Hochreiter dan Jürgen Schmidhuber pada tahun 1997 sebagai solusi untuk menangkap dependensi jangka panjang dalam data sequential (Hochreiter & Schmidhuber, 1997). Arsitektur LSTM menggunakan mekanisme gates (*input gate, forget gate, output gate*) dan *cell state* untuk mengontrol aliran informasi dalam model.

Operasi dalam LSTM dapat dijelaskan melalui tiga komponen utama yaitu, *forget gate, input gate* dan *output gate*. Ketiga komponen tersebut dapat dijelaskan secara matematis sebagai berikut:

Forget gate (f_t) digunakan untuk menentukan informasi mana dari *cell state* sebelumnya (C_{t-1}) yang akan dilupakan.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (1)$$

Di mana:

- σ : Fungsi Sigmoid.
- W_f : Bobot matriks untuk *forget gate*.
- b_f : Bias *forget gate*.

Input gate (i_t) digunakan untuk menentukan informasi baru mana yang akan ditambahkan ke *cell state*.

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

Di mana:

- σ : Fungsi Sigmoid.
- W_i : Bobot matriks untuk *input gate*.
- b_i : Bias *input gate*.

Output gate (o_t) digunakan untuk menentukan bagian mana dari *cell state* yang akan digunakan untuk menghasilkan output.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (3)$$

Di mana:

- σ : Fungsi Sigmoid.
- W_o : Bobot matriks untuk *output gate*.
- b_o : Bias *output gate*.

Selain itu, LSTM juga memperbarui *cell state* (Ct) melalui kombinasi antara informasi lama dan informasi baru.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

$$h_t = o_t \times \tanh(C_t) \quad (5)$$

Masalah *vanishing gradient* dalam RNN tradisional pertama kali diidentifikasi oleh Sepp Hochreiter dalam disertasinya pada tahun 1991. Masalah ini menyebabkan RNN sulit menangkap dependensi jangka panjang karena gradien menjadi sangat kecil selama proses *backpropagation* melalui waktu (*Backpropagation Through Time* , BPTT).

Untuk mengatasi masalah ini, Sepp Hochreiter dan Jürgen Schmidhuber mengembangkan arsitektur LSTM pada tahun 1997 (Hochreiter & Schmidhuber, 1997). Mereka memperkenalkan konsep *cell state* sebagai "jalur informasi" yang dapat dipertahankan atau dimodifikasi melalui mekanisme *gates* . Dengan cara ini, LSTM mampu menangkap informasi dari langkah-langkah awal dalam urutan data tanpa kehilangan konteks.

Sejak pengenalan awalnya, LSTM telah menjadi salah satu arsitektur neural network paling populer untuk tugas-tugas NLP seperti analisis sentimen, terjemahan mesin, dan text generation. Model ini juga menjadi dasar untuk pengembangan varian lain seperti GRU (*Gated Recurrent Unit*) .

LSTM memiliki beberapa keunggulan dibandingkan RNN tradisional:

1. Menangkap Dependensi Jangka Panjang

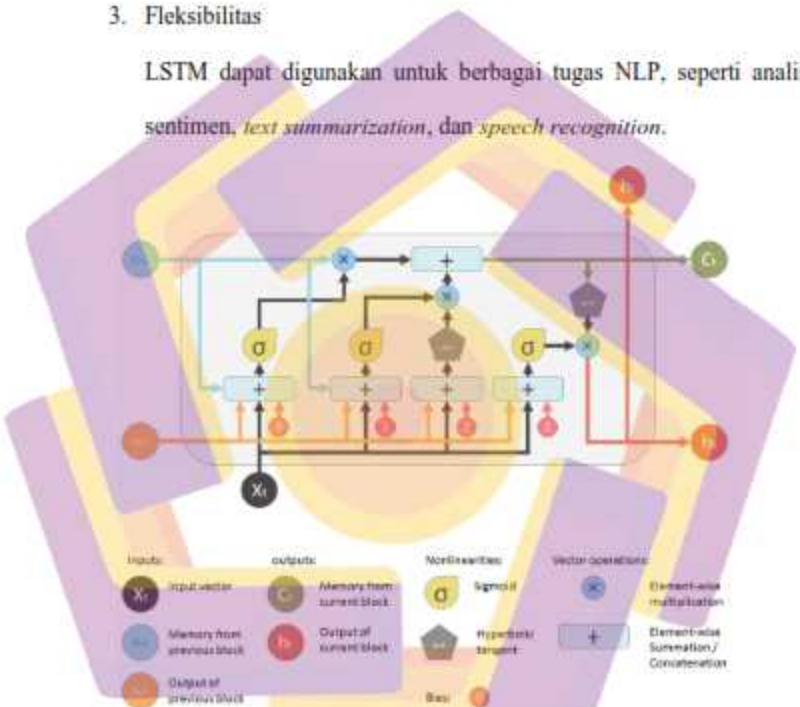
   LSTM mampu mempertahankan informasi penting dari langkah-langkah awal dalam urutan data melalui *cell state*.

2. Mekanisme Gates

   Forget gate, *input gate*, dan *output gate* memberikan kontrol yang lebih baik terhadap aliran informasi, sehingga mengurangi risiko vanishing gradient.

3. Fleksibilitas

   LSTM dapat digunakan untuk berbagai tugas NLP, seperti analisis sentimen, *text summarization*, dan *speech recognition*.



Gambar 2.2. Cara Kerja LSTM

Gambar 2.2 menjelaskan cara kerja sel LSTM, sebuah komponen penting dalam jaringan saraf yang memungkinkan komputer untuk memproses dan memahami urutan data, seperti teks atau data deret waktu.

Komponen-komponen utama dalam diagram ini adalah:

- Xt (*Input vector*): Data masukan pada setiap langkah waktu.

- Ct-1 (*Memory from previous block*): Memori dari langkah waktu sebelumnya, yang menyimpan informasi jangka panjang.
- ht-1 (*Output of previous block*): Keluaran dari langkah waktu sebelumnya, yang membawa informasi yang relevan.
- Ct (*Memory from current block*): Memori yang diperbarui pada langkah waktu saat ini.
- ht (*Output of current block*): Keluaran dari langkah waktu saat ini.

LSTM menggunakan "gerbang" untuk mengatur aliran informasi:

- Gerbang-gerbang ini menggunakan fungsi sigmoid (σ) untuk menentukan informasi mana yang perlu diingat atau dilupakan.
- Fungsi hyperbolic tangent (tanh) digunakan untuk mengatur nilai informasi yang ditambahkan atau dikeluarkan.

Operasi vektor yang digunakan meliputi:

- Perkalian *elementwise* (x) untuk menggabungkan informasi.
- Penjumlahan *elementwise* (+) untuk menggabungkan informasi.

Bias (0) ditambahkan untuk membantu sel LSTM belajar pola yang lebih kompleks.

2.3.3. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) adalah varian dari *Recurrent Neural Networks* (RNN) yang dirancang untuk mengatasi masalah vanishing gradient dalam RNN tradisional, mirip dengan *Long Short-Term Memory* (LSTM). GRU diperkenalkan oleh Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, dan Yoshua Bengio pada tahun 2014 sebagai

alternatif yang lebih ringkas dari LSTM (Cho dkk., 2014). GRU mencapai performa serupa dengan LSTM tetapi dengan arsitektur yang lebih sederhana dan parameter yang lebih sedikit.

Reset gate (r_t) digunakan untuk menentukan sejauh mana informasi dari hidden state (h_{t-1}) sebelumnya diabaikan.

$$r_t = \sigma(W_r \times [h_{t-1}, x_t] + b_r) \quad (6)$$

Di mana:

- σ : Fungsi Sigmoid.
- W_r : Bobot matriks untuk *reset gate*.
- b_r : Bias *reset gate*.

Update gate (z_t) digunakan untuk menentukan sejauh mana *hidden state* baru (h_t) atau input saat ini (x_t).

$$z_t = \sigma(W_z \times [h_{t-1}, x_t] + b_z) \quad (7)$$

Di mana:

- σ : Fungsi Sigmoid.
- W_r : Bobot matriks untuk *update gate*.
- b_r : Bias *update gate*.

GRU pertama kali diperkenalkan oleh Kyunghyun Cho dan timnya pada tahun 2014 dalam makalah berjudul "*Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*" (Cho dkk., 2014). Model ini dikembangkan sebagai bagian dari penelitian tentang terjemahan mesin statistik, di mana mereka membutuhkan model *neural network* yang efisien untuk menangkap dependensi jangka panjang dalam *data sequential*.

GRU dirancang untuk menjadi alternatif yang lebih ringkas dari LSTM, dengan menggabungkan fungsi forget gate dan input gate dari LSTM menjadi satu gate, yaitu update gate. Selain itu, GRU tidak memiliki *cell state* seperti LSTM, melainkan hanya menggunakan *hidden state* untuk menyimpan informasi. Hal ini membuat GRU lebih hemat dalam hal jumlah parameter dan waktu pelatihan, tanpa mengorbankan performa secara signifikan.

GRU memiliki beberapa keunggulan dibandingkan LSTM:

1. Arsitektur Lebih Sederhana

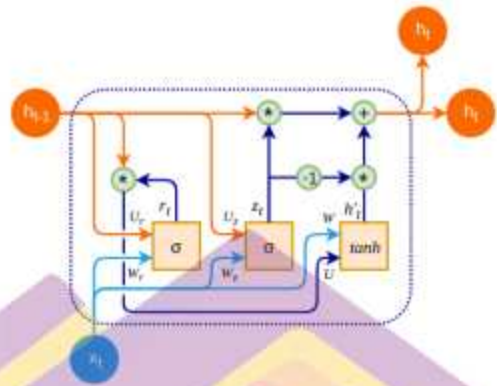
   GRU hanya memiliki dua gates (*reset gate* dan *update gate*), dibandingkan tiga gates dalam LSTM (*forget gate, input gate, output gate*).

2. Efisiensi Komputasi

   GRU memiliki lebih sedikit parameter dibandingkan LSTM, sehingga lebih cepat dalam pelatihan dan inferensi.

3. Performa Serupa dengan LSTM

   Meskipun lebih sederhana, GRU sering kali menghasilkan performa yang setara dengan LSTM dalam banyak tugas NLP, seperti analisis sentimen dan terjemahan mesin.



Gambar 2.3. Cara Kerja GRU

Gambar 2.3 menunjukkan cara kerja sel GRU, sebuah komponen dalam jaringan saraf yang membantu komputer mengingat informasi penting dalam urutan data, seperti teks atau deret waktu.

Komponen-komponen utama dalam diagram ini adalah:

- Xt (*Input*): Data masukan pada setiap langkah waktu.
- ht-1 (*Hidden state* sebelumnya): Informasi dari langkah waktu sebelumnya yang disimpan dalam sel.
- ht (*Hidden state* saat ini): Informasi yang diperbarui pada langkah waktu saat ini.

GRU menggunakan "gerbang" untuk mengatur aliran informasi:

- rt (*Reset gate*): Menentukan seberapa banyak informasi dari hidden state sebelumnya yang perlu diingat.

- zt (*Update gate*): Menentukan seberapa banyak informasi baru yang perlu ditambahkan dan seberapa banyak informasi lama yang perlu dipertahankan.
- Gerbang-gerbang ini menggunakan fungsi sigmoid (σ) untuk menghasilkan nilai antara 0 dan 1, yang menentukan seberapa banyak informasi yang diteruskan.
- Fungsi *hyperbolic tangent* (tanh) digunakan untuk menghasilkan kandidat hidden state baru (h't).

Operasi yang digunakan meliputi:

- Perkalian (x) untuk menggabungkan informasi.
- Penjumlahan (+) untuk menggabungkan informasi.
- Perkalian dengan -1 untuk mengubah tanda informasi.

Tabel 2.2 menunjukkan secara jelas apa saja perbedaan antara GRU dan LSTM. LSTM memiliki tingkat kompleksitas yang lebih rumit dibandingkan dengan GRU sehingga memerlukan daya komputasi yang lebih besar.

Tabel 2.2. Perbedaan GRU dan LSTM

Fitur	GRU (*Gated Recurrent Unit*)	LSTM (*Long Short-Term Memory*)
Struktur	Lebih sederhana	Lebih kompleks
Jumlah Gerbang	2 (Update & Reset)	3 (Input, Forget, Output)
Memory Cell	Tidak ada memory cell terpisah	Memiliki memory cell terpisah
Parameter	Lebih sedikit	Lebih banyak
Kompleksitas Komputasi	Lebih rendah	Lebih tinggi
Kinerja	Mirip, tergantung tugas	Mirip, tergantung tugas

Cocok untuk	Data sekuensial yang lebih pendek, sumber daya terbatas	Data sekuensial yang panjang, ketergantungan kompleks

2.3.4. Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah model berbasis arsitektur Transformer yang dirancang untuk memahami konteks bahasa secara mendalam dengan memanfaatkan mekanisme *self-attention*. BERT diperkenalkan oleh Jacob Devlin, Ming-Wei Chang, Kenton Lee, dan Kristina Toutanova pada tahun 2018 dalam makalah berjudul "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" (Devlin dkk., 2019). Model ini merupakan salah satu inovasi terbesar dalam bidang Natural Language Processing (NLP) karena kemampuannya untuk memahami konteks kata secara *bidirectional*, yaitu dengan mempertimbangkan konteks sebelum dan sesudah kata.

Secara teknis, BERT menggunakan dua tahap utama:

1. *Pre-training*: BERT dilatih pada dataset teks besar tanpa label menggunakan dua tugas pre-training utama:
   - *Masked Language Model* (MLM): Beberapa kata dalam kalimat di-mask (disembunyikan), dan model harus memprediksi kata-kata tersebut berdasarkan konteks sebelum dan sesudahnya.
   - *Next Sentence Prediction* (NSP): Model dilatih untuk memprediksi apakah dua kalimat berturutan atau tidak.
2. *Fine-tuning*: Setelah pre-training, BERT dapat disesuaikan (fine-tuned) untuk tugas-tugas NLP spesifik seperti analisis sentimen, klasifikasi

teks, atau pertanyaan-jawaban dengan melatih ulang model pada dataset berlabel yang lebih kecil.

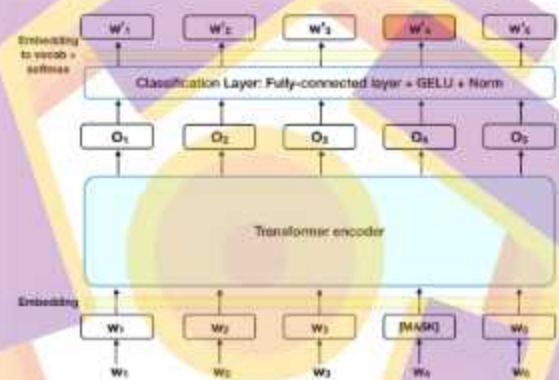
BERT menghasilkan representasi vektor kontekstual untuk setiap token dalam input, yang memungkinkan model memahami arti kata secara lebih akurat dalam berbagai konteks. Misalnya, kata "bank" dapat dipahami sebagai lembaga keuangan atau tepi sungai, tergantung pada konteksnya.

BERT memiliki beberapa keunggulan dibandingkan model NLP sebelumnya:

1. Bidirectional Context Understanding: BERT mampu memahami konteks kata secara bidirectional, sehingga lebih akurat dalam menangkap makna kata dalam berbagai konteks.
2. Generalisasi yang Kuat: Model *pre-trained* BERT dapat digunakan untuk berbagai tugas NLP dengan sedikit fine-tuning, seperti analisis sentimen, klasifikasi teks, dan pertanyaan-jawaban.
3. Performa Tinggi: BERT mencapai hasil *state-of-the-art* dalam banyak benchmark NLP, termasuk GLUE (*General Language Understanding Evaluation*) dan SQuAD (*Stanford Question Answering Dataset*).

Gambar 2.4 menunjukkan model BERT menggunakan teknik *Masked Language Modeling* (MLM) untuk mempelajari konteks bahasa dengan mengganti 15% kata dalam suatu urutan dengan token [MASK], kemudian mencoba memprediksi kata-kata asli berdasarkan konteks dari kata-kata lain yang tidak dimasking. Proses prediksi melibatkan penambahan lapisan klasifikasi di atas output encoder, mengalikan vektor output dengan matriks embedding untuk

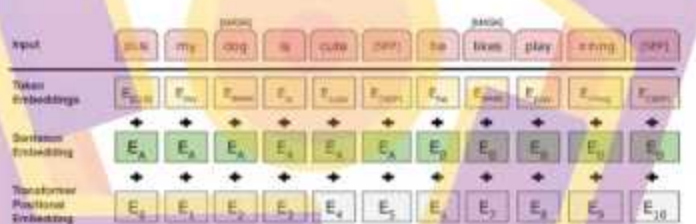
mentransformasikannya ke dimensi kosakata, dan menghitung probabilitas setiap kata dalam kosakata menggunakan fungsi softmax. Fungsi kerugian (*loss function*) BERT hanya mempertimbangkan prediksi kata-kata yang dimasking, sementara kata-kata lain diabaikan, sehingga meskipun model ini cenderung lebih lambat mencapai konvergensi dibandingkan model berbasis arah (*directional models*), hal ini diimbangi oleh kemampuannya yang unggul dalam memahami konteks dua arah (*bidirectional context awareness*).



Gambar 2.4. BERT *Mask Language Modelling* (MLM)

Gambar 2.5 menunjukkan proses pelatihan BERT, model menerima pasangan kalimat sebagai input dan belajar untuk memprediksi apakah kalimat kedua merupakan kelanjutan dari kalimat pertama dalam dokumen asli. Selama pelatihan, 50% pasangan terdiri dari kalimat yang saling berhubungan (kalimat kedua adalah kelanjutan dari yang pertama), sementara 50% lainnya terdiri dari pasangan acak di mana kalimat kedua tidak berhubungan dengan yang pertama, dengan asumsi bahwa kalimat acak tersebut tidak memiliki koneksi kontekstual dengan kalimat pertama. Untuk membantu model membedakan antara dua kalimat,

input diproses dengan menambahkan token [CLS] di awal, token [SEP] di akhir setiap kalimat, embedding kalimat (*Sentence* A atau *Sentence* B) untuk setiap token, serta embedding posisi untuk menunjukkan urutan token dalam kalimat. Prediksi hubungan antar-kalimat dilakukan dengan memproses seluruh urutan input melalui model Transformer, mengubah output token [CLS] menjadi vektor berbentuk 2×1 menggunakan lapisan klasifikasi, dan menghitung probabilitas hubungan antar-kalimat menggunakan *softmax*. Selain itu, pelatihan BERT menggabungkan dua strategi: *Masked Language Model* (MLM) untuk mempelajari representasi kata berdasarkan konteks, dan *Next Sentence Prediction* (NSP) untuk memahami hubungan antar-kalimat, dengan tujuan meminimalkan fungsi kerugian gabungan dari kedua strategi tersebut.



Gambar 2.5. Representasi Input/Output BERT

2.3.5. Cleaning Text

Cleaning text adalah proses pembersihan teks untuk menghilangkan elemen-elemen yang tidak relevan atau *noise* sebelum data diproses lebih lanjut. Tujuan utamanya adalah untuk memastikan bahwa data teks siap digunakan oleh model machine learning atau deep learning dengan cara:

1. Menghapus karakter non-alfanumerik (misalnya, emoji, simbol, tanda baca).

2. Menghapus URL, hashtag, atau mention (terutama dalam teks media sosial).
3. Mengubah teks menjadi huruf kecil (lowercasing) untuk menyamakan format.
4. Menghapus stopwords (kata-kata umum seperti "dan", "atau", "yang" yang tidak memberikan informasi signifikan).

2.3.6. Tokenization

Tokenization adalah proses memecah teks menjadi unit-unit token, seperti kata-kata atau subword. Token adalah unit terkecil yang akan diproses oleh model NLP. Ada dua jenis tokenization utama:

1. *Word-level tokenization*: Memecah teks menjadi kata-kata. Contoh: "Saya suka makan nasi" → ["Saya", "suka", "makan", "nasi"]
2. *Subword-level tokenization*: Memecah teks menjadi subword (digunakan oleh model seperti BERT). Contoh: "memakan" → ["mem", "##akan"]

2.3.7. Stemming

Stemming merupakan suatu proses dalam pemrosesan bahasa alami (NLP) yang bertujuan untuk memperoleh bentuk dasar dari kata-kata yang terdapat dalam sebuah teks (Ni'mah dkk., 2019). Proses ini dilakukan dengan cara menghilangkan imbuhan seperti prefiks (awalan), sufiks (akhiran), dan infiks (sisipan) yang melekat pada kata, sehingga menghasilkan kata dasar yang memiliki makna yang serupa. Tujuan utama dari stemming adalah untuk mengurangi variasi kata dan mengelompokkan kata-kata yang memiliki makna yang sama, sehingga dapat diproses dengan lebih efisien dalam aplikasi-aplikasi pemrosesan teks, seperti

klasifikasi dokumen, pencarian informasi, dan analisis teks. *Stemming* adalah salah satu langkah penting dalam pra-pemrosesan teks yang dapat membantu meningkatkan kinerja sistem pemrosesan teks dalam berbagai konteks. Setiap bahasa memiliki algoritma stemming yang berbeda-beda, yang disesuaikan dengan karakteristik masing-masing bahasa (Wahyu Ade Saputra dkk., 2024). Contoh stemming:

- Input : "berlari", "melari", "pelari"
- Output : "lari"

2.3.8. Padding

Padding adalah proses menyamakan panjang input (*sequence length*) agar semua sequence memiliki dimensi yang sama. Dalam deep learning, model seperti LSTM, GRU, dan BERT memerlukan input dengan panjang tetap. Sequence yang lebih pendek ditambahkan dengan token padding (misalnya, [PAD]), sementara sequence yang lebih panjang dipotong sesuai batas maksimal.

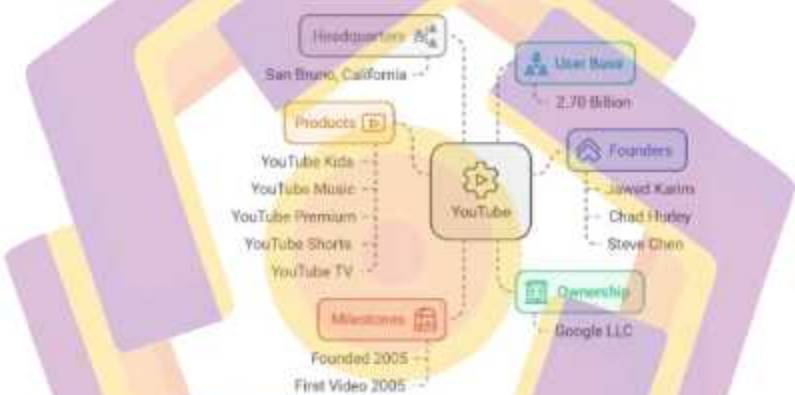
Contoh *padding*:

- Input : [["ini", "adalah"], ["kalimat", "contoh", "untuk", "padding"]]
- Output : [["ini", "adalah", "[PAD]", "[PAD]"], ["kalimat", "contoh", "untuk", "padding"]]

2.3.9. Media Sosial YouTube

YouTube adalah platform media sosial berbasis video yang memungkinkan pengguna untuk mengunggah, menonton, berbagi, dan berinteraksi dengan konten multimedia. Platform ini didirikan oleh Chad Hurley, Steve Chen, dan Jawed Karim pada tahun 2005 sebagai solusi untuk kesulitan berbagi video melalui email akibat

ukuran file yang besar. Video pertama yang diunggah ke YouTube, *"Me at the Zoo"*, dipublikasikan pada 23 April 2005 oleh Jawed Karim, menandai awal dari revolusi platform berbagi video. Pada tahun 2006, Google mengakuisisi YouTube seharga $1,65 miliar, menjadikannya bagian integral dari ekosistem digital global (Rowell, 2011). Gambar 2.6 menunjukkan gambaran ringkas dan jelas tentang informasi kunci terkait YouTube, termasuk lokasi, produk, pendiri, kepemilikan, dan tonggak sejarah.



Gambar 2.6. Infografis YouTube

Hingga Juli 2024, YouTube memiliki lebih dari 2,7 miliar pengguna aktif bulanan di seluruh dunia, menjadikannya salah satu platform terbesar untuk distribusi konten video (Team, 2025). Secara teknis, YouTube juga merupakan mesin pencari terbesar kedua di dunia setelah Google, dengan potensi jangkauan iklan mencapai 2,49 miliar pengguna secara global (Simon Kemp, 2024). Di antara negara-negara dengan jumlah pengguna tertinggi, India menduduki peringkat

pertama dengan 476 juta pengguna, diikuti oleh Amerika Serikat dengan kontribusi 16,4% dari total lalu lintas YouTube.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini menggabungkan pendekatan kuantitatif dan deskriptif. Pendekatan deskriptif digunakan untuk menjelaskan proses pengumpulan data dari sumber yang relevan dengan subjek penelitian, yaitu kolom komentar YouTube tentang Danantara Indonesia. Parameter dan jumlah data yang berhasil dikumpulkan dijelaskan secara rinci untuk memberikan gambaran yang jelas tentang dataset yang digunakan. Selain itu, pendekatan kuantitatif digunakan untuk mengevaluasi performa tiga model deep learning, yaitu LSTM, GRU, dan BERT, dalam menganalisis sentimen teks berbahasa Indonesia. Evaluasi dilakukan berdasarkan metrik seperti akurasi, presisi, recall, dan F1-score untuk menentukan model mana yang memiliki tingkat performa tertinggi.

3.2. Metode Pengumpulan Data

Dalam penelitian ini, proses pengumpulan data dilakukan dengan menghimpun teks-teks komentar berbahasa Indonesia dari video YouTube yang membahas topik Danantara Indonesia. Pencarian video dilakukan dengan menggunakan kata kunci "Danantara Indonesia" langsung di mesin pencarian YouTube, kemudian hasil pencarian difilter berdasarkan jumlah penayangan terbanyak (most viewed) guna memastikan bahwa komentar yang dianalisis merupakan representasi dari video yang memiliki jangkauan luas dan interaksi tinggi dengan publik.

Data komentar diperoleh melalui pemanfaatan YouTube API dan/atau alat bantu scraping untuk menjamin kelengkapan serta keakuratan dataset. Fokus utama

diarahkan pada komentar-komentar yang mengandung opini publik seputar topik sovereign wealth fund (SWF), transparansi lembaga, kebijakan ekonomi, serta persepsi terhadap Danantara Indonesia secara umum.

Dataset yang dikumpulkan terdiri dari berbagai jenis komentar—baik yang bersifat positif, negatif, maupun netral—guna membentuk representasi data yang seimbang. Proses pembersihan data (preprocessing) juga dilakukan untuk menghilangkan unsur-unsur yang tidak relevan seperti emoji, URL, simbol, dan kata-kata umum (stopwords) yang tidak memberikan kontribusi bermakna terhadap proses analisis sentimen. Tahapan ini dilakukan untuk memastikan bahwa data yang masuk ke model deep learning berada dalam kondisi optimal.

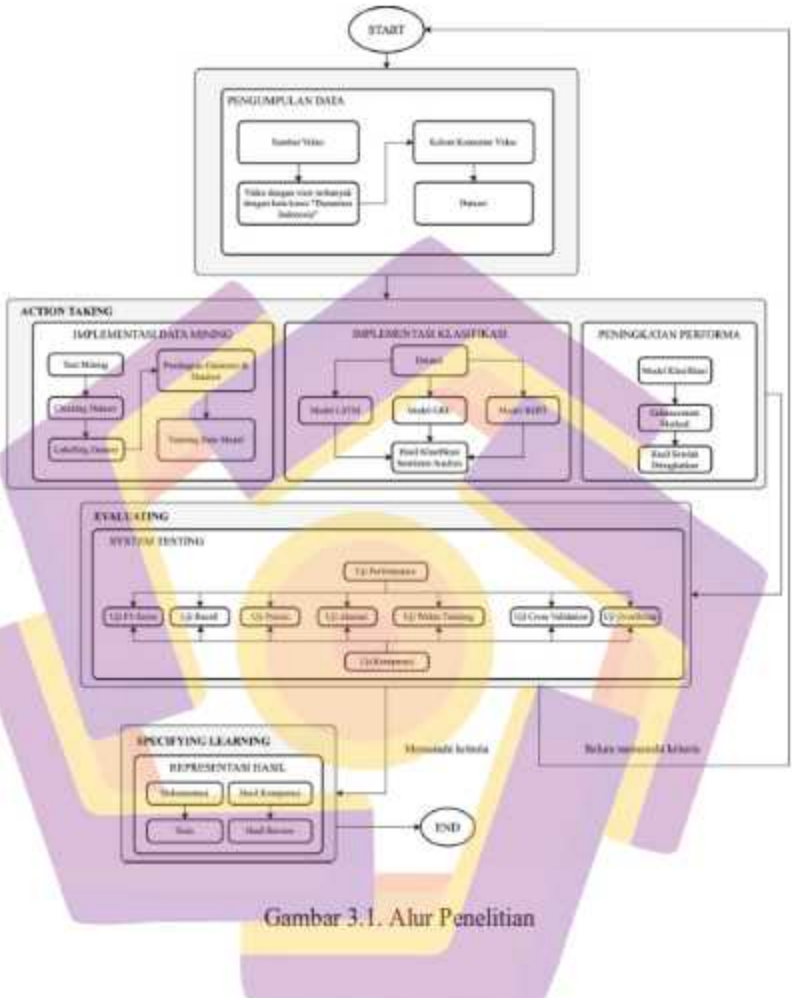
3.3. Metode Analisis Data

Metode analisis data dalam penelitian ini dirancang untuk mengevaluasi dan membandingkan performa tiga model deep learning, yaitu LSTM, GRU, dan BERT, dalam menganalisis sentimen kolom komentar YouTube tentang Danantara Indonesia. Proses analisis dimulai dengan tahap pra-pemrosesan teks, yang meliputi pembersihan data (menghapus emoji, URL, simbol, dan angka), tokenisasi untuk memecah teks menjadi unit token, serta *stemming/lemmatization* untuk mengurangi kata ke bentuk dasar sesuai kaidah bahasa Indonesia. Setelah itu, padding diterapkan untuk menyamakan panjang sequence agar data siap diproses oleh model. Ketiga model kemudian digunakan untuk mengklasifikasikan komentar ke dalam dua kategori sentimen: positif, netral dan negatif. Performa model dievaluasi menggunakan metrik kuantitatif seperti akurasi, presisi, recall, dan F1-score, dengan teknik *cross-validation* untuk memastikan hasil yang tidak bias. Selain itu,

hasil klasifikasi diverifikasi secara kualitatif melalui validasi oleh ahli bahasa atau peneliti NLP untuk memastikan relevansi linguistik. Analisis juga mencakup penilaian efisiensi model dalam hal waktu pelatihan dan inferensi, terutama untuk memahami bagaimana model menangani tantangan khusus bahasa Indonesia, seperti struktur afiksasi, slang, dan kesalahan ketik. Hasil keseluruhan diinterpretasikan untuk menentukan model terbaik, memberikan wawasan tentang keunggulan dan kelemahan masing-masing model, serta rekomendasi untuk aplikasi analisis sentimen di platform media sosial.

3.4. Alur Penelitian

Dalam penelitian dilakukan sejumlah kegiatan yang telah direncanakan secara sistematis guna mencapai tujuan yang diinginkan. Gambar 3.1 memperlihatkan rangkaian kegiatan yang dilakukan dalam penelitian ini.



Gambar 3.1. Alur Penelitian

a. *Diagnosing*

Proses diagnosis merupakan langkah awal yang terdiri atas tiga tahapan utama: studi literatur, definisi kebutuhan, dan pengumpulan data. Studi literatur dilakukan dengan merujuk pada berbagai referensi terpercaya, seperti buku, artikel ilmiah, dan jurnal yang relevan dengan analisis sentimen dan teknik deep learning seperti LSTM, GRU, dan BERT. Setelah literatur yang relevan berhasil dihimpun, peneliti melakukan tahap definisi kebutuhan dengan mengidentifikasi kebutuhan penelitian berdasarkan hasil tinjauan literatur. Selanjutnya, dilakukan analisis terhadap sistem yang berjalan untuk mengidentifikasi kelemahan dalam pendekatan analisis sentimen sebelumnya, terutama dalam menangani teks berbahasa Indonesia yang berasal dari media sosial. Hasil dari analisis tersebut disusun dalam bentuk laporan evaluasi yang terstruktur.

Berdasarkan informasi yang telah dikumpulkan, peneliti melaksanakan proses pengumpulan data sesuai dengan kriteria yang telah ditetapkan dari identifikasi masalah. Metode yang digunakan dalam pengumpulan data mencakup pengambilan kolom komentar YouTube terkait topik Danantara Indonesia melalui YouTube API atau alat scraping. Data yang dikumpulkan mencakup komentar dalam bahasa Indonesia yang relevan dengan persepsi publik terhadap sovereign wealth fund (SWF). Setiap metode yang digunakan dirancang secara sistematis guna

memastikan bahwa data yang diperoleh relevan, valid, dan sesuai dengan kebutuhan penelitian.

b. *Action Planning*

Setelah kebutuhan penelitian diidentifikasi, peneliti melanjutkan ke tahap perencanaan tindakan (action planning). Pada tahap ini, peneliti menyusun rencana terkait aspek-aspek yang diperlukan dalam penelitian. Perencanaan tindakan dalam penelitian ini mencakup dua hal utama, yaitu pengumpulan dataset dan pemilihan model deep learning yang akan digunakan.

Dataset yang digunakan terdiri dari kolom komentar YouTube yang telah diproses melalui tahap pra-pemrosesan, seperti *cleaning text*, *tokenization*, *stemming/lemmatization*, dan *padding*. Sementara itu, pemilihan model deep learning mencakup implementasi tiga model utama: LSTM, GRU, dan BERT. Ketiga model ini dipilih untuk memungkinkan analisis komparatif terhadap performa masing-masing model dalam mengklasifikasikan sentimen teks berbahasa Indonesia. Rencana ini dirancang secara sistematis untuk memastikan pelaksanaan penelitian berjalan sesuai dengan tujuan yang telah ditetapkan.

c. *Action Taking*

Pada tahap ini, eksperimen dengan model deep learning yang telah ditentukan akan dilaksanakan. Eksperimen ini menggunakan dataset yang terdiri dari kolom komentar YouTube tentang Danantara Indonesia, yang telah diproses melalui tahap pra-pemrosesan untuk

memastikan data siap digunakan oleh model. Dataset ini dipilih untuk menguji kemampuan model dalam mengklasifikasikan sentimen teks berbahasa Indonesia menjadi dua kategori utama: positif, netral dan negatif.

Hasil dari eksperimen ini akan dianalisis dan dievaluasi secara mendalam pada tahap berikutnya untuk mengukur performa setiap model deep learning yang diterapkan, baik dari segi akurasi, efisiensi, maupun kesesuaiannya dengan konteks bahasa Indonesia. Evaluasi ini akan menjadi dasar untuk memberikan kesimpulan dan rekomendasi dari penelitian yang dilakukan.

d. *Evaluating*

Pada tahap ini, performa setiap model deep learning akan dievaluasi berdasarkan tingkat akurasinya. Proses evaluasi dilakukan dengan mengukur sejauh mana setiap model mampu mengklasifikasikan sentimen teks secara tepat untuk menghasilkan label sentimen yang sesuai dengan kaidah bahasa Indonesia. Tingkat akurasi dari setiap model kemudian dibandingkan berdasarkan metrik evaluasi seperti presisi, recall, dan F1-score untuk menentukan model yang memberikan hasil paling optimal. Analisis perbandingan ini bertujuan untuk mengidentifikasi model dengan tingkat akurasi tertinggi dan efisiensi terbaik, sehingga dapat dijadikan rekomendasi untuk implementasi lebih lanjut dalam analisis sentimen teks berbahasa Indonesia.

e. *Specifying Learning*

Pada tahap *specifying learning*, hasil eksperimen dan pengujian yang telah dilakukan akan disusun secara sistematis dalam bentuk laporan yang komprehensif. Dokumentasi dari proses penelitian, termasuk metodologi yang diterapkan, analisis data, serta hasil pengujian masing-masing model deep learning (LSTM, GRU, dan BERT), akan dikompilasi menjadi sebuah karya ilmiah. Sementara itu, hasil perbandingan antara model-model yang diuji akan disusun dalam bentuk hasil review yang lebih mendalam. Review ini akan mencakup analisis mengenai kelebihan dan kekurangan setiap model yang digunakan, serta memberikan rekomendasi berdasarkan temuan-temuan yang diperoleh selama penelitian.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Deskripsi Data

Data yang digunakan dalam penelitian ini diambil dari kolom komentar video YouTube yang membahas topik Danantara Indonesia. Data diperoleh dari 14 video yang berasal dari video yang muncul dengan kata kunci "Danantara Indonesia" dan filter diarahkan untuk *most viewed video*. Parameter tersebut dipilih dengan harapan dapat memberikan data komentar yang tidak hanya banyak tapi juga mewakili berbagai spektrum. Pengumpulan data dilakukan pada tanggal 24 April 2025 menggunakan metode YouTube API atau alat scraping untuk mengekstrak kolom komentar dari video-video tersebut. Data yang diambil berupa teks komentar dalam bahasa Indonesia yang relevan dengan persepsi publik terhadap Danantara Indonesia.

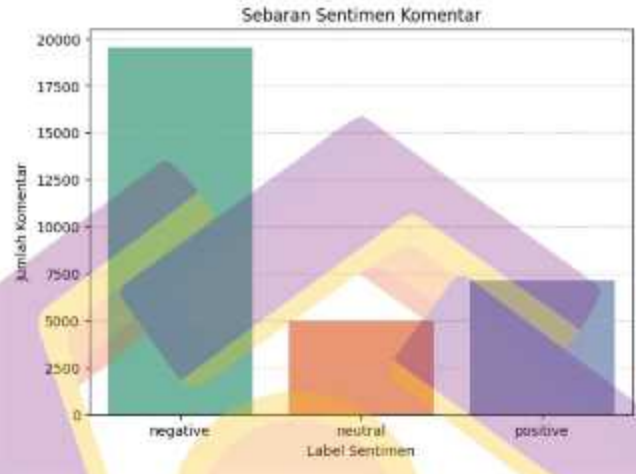
Tabel 4.1. Video yang Digunakan sebagai Dataset

No	Judul Video	Jumlah View	Jumlah Komentar
1	Memahami Danantara dengan Mudah	1,9 juta	7824
2	Surat Terbuka Untuk Danantara	862 ribu	3439
3	Bahas Investasi & Danantara Bareng Sandiaga Uno	793 ribu	2502
4	Danantara, Entitas Bisnis Atau Produk Politik? "Senjata Prabowo Tanpa Peluru"	765 ribu	2362
5	[FULL] Ada SBY Dan Jokowi Di Pengurus Danantara, Rosan Roeslani Bicara | ROSI	641 ribu	2898
6	Danantara Jurus Sakti Prabowo ? Ubah Defisit Jadi Untung Rp 1.400 Triliun Per Tahun !!	615 ribu	6796

No	Judul Video	Jumlah View	Jumlah Komentar
7	Jalur Cepat RUU BUMN dan Manuver Menguasai Danantara | Jelasin Dong!	585 ribu	834
8	Danantara dan Makan Bergizi Gratis: Solusi atau Beban Baru Negara? | Bicara	548 ribu	1326
9	INFO A1 DANANTARA!! Demi Rakyat Terpaksa gw bongkar, sorry guys	476 ribu	4846
10	Apa Efeknya Danantara ke Ekonomi Indonesia? Investor Kabur? Saham Merah!	396 ribu	1450
11	[FULL] Pidato Presiden Prabowo di Peluncuran "Superholding" BUMN Danantara	366 ribu	815
12	Viral! Ajakan Tarik Uang Dari Bank BUMN Imbas Danantara | SEDANG VIRAL	352 ribu	3769
13	LIVE - Danantara VS Kepercayaan Publik, Rosan Roeslani Menjawab | ROSI	338 ribu	1611
14	Kekhawatiran Gw Soal DANANTARA	298 ribu	1444

Dataset yang digunakan dalam penelitian ini terdiri dari 31.675 komentar YouTube yang telah diklasifikasikan ke dalam tiga label sentimen: *negative*, *neutral*, dan *positive*. Dari jumlah tersebut, kategori *negative* mendominasi dengan 19.549 data (sekitar 61,7%), disusul oleh *positive* sebanyak 7.155 data (22,6%), dan *neutral* sebanyak 4.971 data (15,7%). Distribusi ini menunjukkan adanya *ketidakseimbangan kelas (class imbalance)* yang cukup signifikan, di mana kelas *negative* memiliki jumlah data hampir tiga kali lipat dibandingkan *positive*, dan hampir empat kali lipat dari *neutral*. Kondisi ini berpotensi mempengaruhi

performa model pembelajaran mesin, terutama dalam hal akurasi dan sensitivitas terhadap kelas minoritas.



Gambar 4.1. Sebaran Data Sentimen

4.2. Hasil Eksperimen Model Dasar

Tahapan eksperimen diawali dengan evaluasi performa model dasar (*baseline model*) untuk menetapkan tolok ukur kinerja (*performance benchmark*). Pada penelitian ini, arsitektur *Recurrent Neural Network* (RNN) yang dievaluasi sebagai model dasar meliputi *Long Short-Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU). Pengujian ini bertujuan untuk mengukur kapabilitas fundamental kedua model dalam melakukan klasifikasi sentimen tiga kelas (positif, negatif, netral) pada korpus data komentar YouTube berbahasa Indonesia. Konfigurasi pelatihan untuk setiap model ditetapkan secara seragam, mencakup 5 *epoch*, untuk memastikan komparabilitas hasil yang objektif.

4.2.1. LSTM

Eksperimen model dasar pertama mengimplementasikan arsitektur *Long Short-Term Memory* (LSTM). Model ini disusun secara sekuensial dengan tiga komponen utama: sebuah lapisan Embedding untuk memetakan input token ke dalam representasi vektor, satu lapisan inti LSTM sebagai unit pemrosesan utama, dan sebuah lapisan Dense dengan fungsi aktivasi Softmax sebagai output layer untuk klasifikasi multikelas.

Metrik kinerja yang dihasilkan selama proses pelatihan selama lima epoch disajikan secara rinci pada Tabel 4.2.

Tabel 4.2. Hasil Pelatihan Model Dasar LSTM

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.5185	1.0299	0.5365	1.0123
2	0.5193	1.0278	0.5365	1.0146
3	0.52	1.0272	0.5365	1.0117
4	0.5206	1.0264	0.537	1.0121
5	0.5199	1.0256	0.5365	1.0114
6	0.5206	1.0232	0.5356	1.0123
7	0.5207	1.022	0.5361	1.0124
8	0.5208	1.0218	0.5361	1.0138
9	0.5208	1.0219	0.5365	1.0148
10	0.5208	1.0218	0.5356	1.0138
11	0.5208	1.0213	0.5365	1.0133
12	0.5208	1.021	0.5356	1.017
13	0.5207	1.0211	0.5347	1.014
14	0.5208	1.0214	0.5361	1.0133
15	0.5207	1.0212	0.537	1.0133
16	0.5207	1.0218	0.5361	1.0147
17	0.5208	1.021	0.5361	1.0147

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
18	0.5208	1.021	0.5361	1.0118
19	0.5195	1.0222	0.5361	1.0125
20	0.5206	1.0208	0.5361	0.9968
21	0.5475	0.9215	0.6228	0.8326
22	0.6542	0.747	0.6557	0.7925
23	0.7218	0.6229	0.6689	0.8207
24	0.7853	0.5206	0.6753	0.8954
25	0.8607	0.3789	0.6927	0.8565
26	0.9041	0.2819	0.6918	0.9808
27	0.9254	0.2292	0.6913	1.0699
28	0.9396	0.1886	0.6872	1.1037
29	0.9455	0.1604	0.6886	1.1826
30	0.9513	0.1401	0.695	1.0379

Analisis terhadap data training pada Tabel 4.2 menunjukkan bahwa performa model memiliki dua titik optimal yang berbeda. Performa val_accuracy (akurasi validasi) tertinggi sebesar 0.6950 tercapai pada epoch ke-30. Namun, val_loss (loss validasi) minimum sebesar 0.7925 justru tercapai jauh lebih awal, yakni pada epoch ke-22.

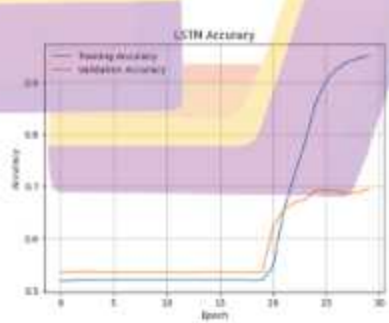
Fase Stagnasi (Epoch 1-20): Selama 20 epoch pertama, model terjebak dalam stagnasi. val_accuracy hanya berfluktuasi di sekitar 0.536 dan val_loss di sekitar 1.01. Hal ini mengindikasikan bahwa model gagal melakukan pembelajaran yang berarti pada awalnya.

Fase Pembelajaran (Epoch 21-25): Terjadi lompatan drastis pada epoch ke-21, di mana val_accuracy melonjak ke 0.6228 dan val_loss turun ke 0.8326. Model kemudian belajar dengan cepat hingga sekitar epoch ke-25.

Fase Overfitting (Epoch 26-30): Meskipun val_accuracy masih sedikit naik dan mencapai puncaknya di epoch ke-30, metrik val_loss menunjukkan tren kenaikan yang konsisten sejak epoch ke-22. Ini adalah indikator klasik overfitting, di mana model mulai menghafal data training (terlihat dari accuracy training yang mencapai 0.9513) tetapi kehilangan kemampuannya untuk menggeneralisasi ke data validasi.

Kondisi overfitting ini mengimplikasikan bahwa meskipun arsitektur LSTM memiliki kompleksitas yang memadai untuk menangkap pola (bahkan menghafal data training), ia kesulitan menggeneralisasi fitur linguistik yang lebih rumit seperti sarkasme atau konteks informal—ke data baru yang belum pernah dilihat. Dari aspek efisiensi, total waktu yang diperlukan untuk melatih model ini tercatat selama 2029.92 detik.

Gambar 4.2 menyajikan visualisasi tren akurasi model LSTM pada data latih (training accuracy) dan data validasi (validation accuracy) selama tiga puluh epoch pelatihan.



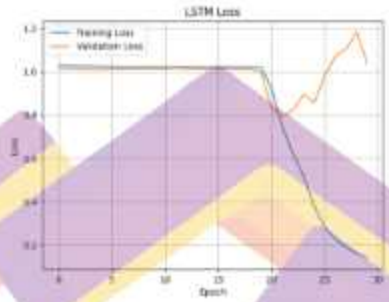
Gambar 4.2. Grafik Tren Akurasi Model LSTM

Kurva biru, yang merepresentasikan Training Accuracy, menunjukkan pola tiga fase yang berbeda. Selama 20 epoch pertama, kurva ini mengalami stagnasi, dengan akurasi hanya berfluktuasi sedikit di sekitar 0.520. Namun, mulai epoch ke-21, terjadi lonjakan performa yang drastis. Model dengan cepat meningkatkan kinerjanya pada data latih, melesat dari 0.5475 (epoch 21) hingga mencapai nilai 0.9513 di akhir pelatihan (epoch 30). Tren ini mengindikasikan bahwa model pada akhirnya berhasil belajar, namun peningkatan yang sangat tajam menjelang akhir ini merupakan indikasi kuat bahwa model mulai menghafal data latih.

Di sisi lain, kurva oranye yang merepresentasikan Validation Accuracy menunjukkan fenomena yang signifikan. Metrik ini awalnya mencerminkan stagnasi kurva training, di mana ia stabil di sekitar 0.536 selama 20 epoch pertama. Saat model mulai belajar pada epoch ke-21, val_accuracy juga ikut melonjak ke 0.6228 dan terus meningkat hingga mencapai puncaknya di 0.6950 (epoch 30).

Hal yang paling signifikan adalah kesenjangan (gap) yang melebar antara kedua kurva ini setelah epoch ke-25. Sementara kurva training (biru) terus menanjak tajam, kurva validasi (oranye) peningkatannya melambat dan cenderung mendatar. Hal ini secara visual mengkonfirmasi fenomena overfitting, di mana model tidak lagi menunjukkan peningkatan kemampuan generalisasi yang sebanding pada data yang belum pernah dilihat (unseen data), meskipun ia menjadi semakin akurat pada data latih.

Gambar 4.3 menyajikan grafik pergerakan nilai loss model LSTM, yang merepresentasikan tingkat kesalahan model pada data latih (training loss) dan data validasi (validation loss).



Gambar 4.3. Grafik Tren Loss Pelatihan Model LSTM

Kurva biru, yang menunjukkan Training Loss, memiliki tren penurunan yang terbagi menjadi dua fase. Selama 20 epoch pertama, penurunan sangat lambat (dari 1.0299 ke 1.0208), mencerminkan fase stagnasi. Namun, mulai epoch ke-21, kurva ini menunjukkan penurunan yang sangat tajam dan konsisten, berakhir pada nilai yang sangat rendah, 0.1401, di epoch ke-30. Perilaku ini menunjukkan bahwa model pada akhirnya berhasil meminimalkan kesalahannya pada data latih secara agresif.

Kurva oranye, yang menampilkan Validation Loss, menunjukkan perilaku yang sangat berbeda. Ia memang mengalami penurunan tajam pada epoch 21-22, mencapai titik terendahnya (kinerja terbaik) pada epoch ke-22 dengan nilai 0.7925. Namun, setelah titik tersebut, trennya berbalik; val_loss justru mulai naik secara konsisten hingga mencapai 1.0379 di akhir pelatihan.

Aspek yang paling menonjol dari grafik ini adalah celah (gap) yang melebar drastis setelah epoch ke-22. Nilai Training Loss (biru) terus anjlok jauh di bawah Validation Loss (oranye), yang justru mulai merangkak naik. Ini adalah indikator klasik dari overfitting. Fenomena ini menyiratkan bahwa arsitektur model telah menjadi terlalu "hafal" pada pola data latih (sehingga nilai loss-nya sangat rendah) dan kehilangan kemampuannya untuk menggeneralisasi, sehingga menghasilkan tingkat kesalahan yang lebih tinggi pada data validasi yang belum pernah dilihat.

4.2.2. GRU

Eksperimen model dasar kedua dilanjutkan dengan implementasi arsitektur *Gated Recurrent Unit* (GRU). Konfigurasi model ini dirancang paralel dengan implementasi LSTM sebelumnya, yakni tersusun atas lapisan Embedding, satu lapisan inti GRU, dan lapisan Dense sebagai output.

Metrik pelatihan yang dihasilkan dari proses ini didokumentasikan pada Tabel 4.3.

Tabel 4.3. Hasil Pelatihan Model Dasar GRU

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.5178	1.0298	0.5365	1.0134
2	0.5193	1.0287	0.5365	1.014
3	0.5193	1.0262	0.5365	1.0131
4	0.5203	1.026	0.5365	1.012
5	0.5206	1.0254	0.5365	1.0145
6	0.5207	1.024	0.5361	1.0163
7	0.5208	1.0229	0.5352	1.0197
8	0.5208	1.0229	0.5352	1.0179
9	0.5208	1.0214	0.5352	0.989
10	0.549	0.9209	0.6374	0.8255

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
11	0.6896	0.7014	0.69	0.7225
12	0.8145	0.4665	0.7228	0.7239
13	0.9045	0.2652	0.7242	0.8088
14	0.9403	0.1663	0.7301	0.9266
15	0.9597	0.1155	0.7256	0.9667
16	0.9664	0.0931	0.7228	1.1333
17	0.9714	0.0773	0.7205	1.1822
18	0.9749	0.0662	0.7178	1.2546
19	0.9768	0.0604	0.7151	1.3648
20	0.9784	0.0565	0.7228	1.4152
21	0.9807	0.0505	0.7142	1.4034
22	0.9813	0.0449	0.7142	1.5216
23	0.9823	0.0432	0.7073	1.5588
24	0.9817	0.0472	0.721	1.5124
25	0.9829	0.0402	0.71	1.6241
26	0.9846	0.0408	0.7059	1.6749
27	0.9846	0.0378	0.7146	1.7923
28	0.9855	0.0345	0.7041	1.8097
29	0.9849	0.0354	0.71	1.7752
30	0.9865	0.035	0.7068	1.7405

Berdasarkan data pelatihan model (Tabel 4.3), performa puncak pada set data validasi tercatat di pertengahan proses pelatihan, bukan di awal. Model berhasil mencapai akurasi validasi (Validation Accuracy) puncak sebesar 0.7301 pada epoch ke-14. Sementara itu, nilai loss validasi (Validation Loss) minimum sebesar 0.7225 tercapai sedikit lebih awal, pada epoch ke-11.

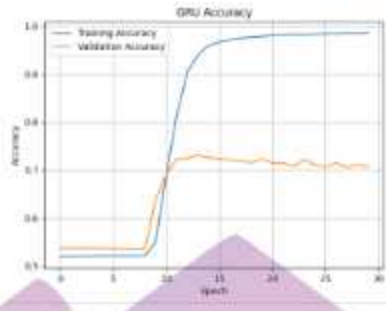
Pola pelatihan model ini dapat dibagi dengan jelas menjadi tiga tahapan yang berbeda. Tahap pertama adalah fase stagnasi yang berlangsung selama

sembilan epoch awal (epoch 1-9). Selama periode ini, model tampak "terjebak", dengan metrik accuracy dan loss baik pada data latih maupun validasi hampir tidak menunjukkan perbaikan, dengan val_accuracy stabil di sekitar 0.536.

Tahap kedua adalah fase pembelajaran eksplosif yang dimulai secara tiba-tiba pada epoch ke-10, kemungkinan dipicu oleh mekanisme learning rate scheduler. Pada titik ini, loss mengalami penurunan dan accuracy meningkat, model dengan cepat mencapai performa puncaknya pada data validasi, yakni val_loss terendah (0.7225) di epoch ke-11 dan val_accuracy tertinggi (0.7301) di epoch ke-14. Setelah titik optimal ini, model segera memasuki tahap ketiga

Tahap ketiga adalah fase overfitting yang parah (epoch 15-30). Di sini, training loss terus anjlok mendekati nol dan training accuracy meroket hingga 98.6%, namun performa validasi justru memburuk secara drastis—ditandai dengan val_loss yang terus menanjak tajam hingga 1.7405 dan val_accuracy yang mulai menurun. Ini jelas menunjukkan bahwa model telah berhenti belajar pola umum dan beralih menghafal data latih. Ini menandakan bahwa titik optimal model adalah di sekitar epoch 11-14, dan pelatihan apa pun setelah itu justru merusak performa model pada data baru.

Gambar 4.4 menyajikan visualisasi perbandingan antara akurasi latih (*training accuracy*) dan akurasi validasi (*validation accuracy*) dari model GRU selama proses pelatihan tiga puluh *epoch*.



Gambar 4.4. Grafik Tren Akurasi Model GRU

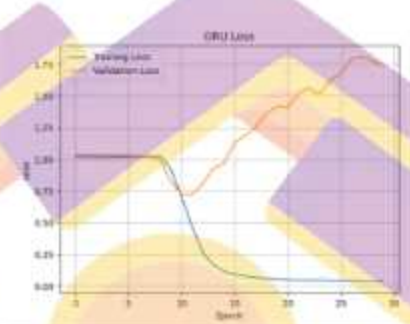
Kurva biru yang merepresentasikan Training Accuracy menunjukkan tren peningkatan yang stabil dan konsisten. Meskipun peningkatannya bersifat marginal, tren ini mengindikasikan bahwa model GRU secara progresif belajar dari data latih di setiap epoch-nya.

Sebaliknya, kurva oranye yang menggambarkan Validation Accuracy secara visual mengonfirmasi fenomena stagnasi yang telah diidentifikasi dari data tabel. Akurasi pada data validasi tetap berada pada level yang sama sekali datar di angka 0.5315 dari awal hingga akhir pelatihan. Hal ini menjadi bukti kuat bahwa model gagal meningkatkan kemampuan generalisasinya pada data baru seiring berjalannya waktu, dan telah mencapai performa puncaknya sejak epoch pertama.

Sama halnya dengan model LSTM, grafik ini menunjukkan kondisi underfitting, di mana akurasi validasi secara persisten lebih tinggi daripada akurasi latih. Celah performa ini menegaskan bahwa arsitektur dasar GRU juga memiliki kapasitas yang terbatas dan kesulitan untuk menangkap pola-pola yang lebih kompleks dalam data latih. Dengan demikian, visualisasi ini mengonfirmasi temuan

sebelumnya: meskipun GRU lebih efisien secara komputasi, model dasarnya menghadapi masalah saturasi performa yang identik dengan LSTM.

Gambar 4.5 melengkapi analisis model GRU dengan memvisualisasikan dinamika nilai *loss*, yang mengukur tingkat kesalahan prediksi pada data latih (*training loss*) dan data validasi (*validation loss*).



Gambar 4.5. Grafik Tren Loss Pelatihan Model LSTM

Kurva biru, yang merepresentasikan Training Loss, menunjukkan tren penurunan yang terbagi dalam dua fase. Setelah stagnan di sekitar 1.02 selama 9 epoch pertama, kurva ini menunjukkan penurunan yang sangat tajam dan agresif mulai epoch ke-10, terus anjlok hingga mencapai nilai yang hampir nol (0.0350) di akhir pelatihan. Hal ini merefleksikan proses optimisasi yang berjalan sangat agresif pada data latih.

Berbeda dengan kurva latih, kurva oranye yang mewakili Validation Loss menunjukkan titik balik yang jelas. Setelah awalnya menurun bersamaan dengan training loss (mencapai titik terendah 0.7225 pada epoch ke-11), kurva ini kemudian berbalik arah dan mulai naik secara drastis dan konsisten. Nilai val_loss membengkak dari 0.7225 menjadi 1.7405 pada epoch ke-30. Titik di epoch ke-11

inilah yang menandai momen generalisasi optimal, dan setelah itu, performa model pada data baru terus memburuk.

Aspek terpenting dari grafik ini adalah konfirmasi visual yang kuat dari kondisi overfitting yang parah. Terdapat celah (gap) yang sangat lebar yang terbuka setelah epoch ke-11, di mana nilai Validation Loss (oranye) melambung tinggi jauh di atas Training Loss (biru). Hal ini secara definitif menunjukkan bahwa arsitektur model GRU ini, alih-alih kurang kapasitas, justru telah "menghafal" data latih (ditandai dengan loss yang sangat rendah) dan kehilangan kemampuannya untuk menggeneralisasi. Visualisasi ini membuktikan bahwa model ini sangat memerlukan teknik regularisasi atau early stopping (berhenti di sekitar epoch ke-11) untuk mencegah penurunan performa.

4.2.3. BERT

Berbeda dari arsitektur sekuensial LSTM dan GRU, eksperimen ini mengimplementasikan model pra-latih indobenchmark/indobert-base-p1, yang telah dilatih secara ekstensif pada korpus Bahasa Indonesia. Proses yang dilakukan bukanlah pelatihan dari awal, melainkan fine-tuning, di mana lapisan dasar BERT yang sudah cerdas diadaptasikan untuk tugas klasifikasi sentimen dengan menambahkan lapisan classifier baru. Arsitektur ini memiliki total parameter sekitar 124.4 juta, menunjukkan kompleksitas yang jauh lebih tinggi dibandingkan model RNN sebelumnya.

Proses learning dijalankan selama tiga puluh epoch, dengan metrik kinerja yang tercatat pada Tabel 4.4.

Tabel 4.4. Hasil Pelatihan Model Dasar BERT

Epoch	*Training Loss*	*Validation Loss*	*Accuracy*
1	0.422	0.348735	0.856621
2	0.1953	0.366601	0.8621
3	0.1464	0.528172	0.844749
4	0.1234	0.539035	0.861187
5	0.1267	0.648545	0.855708
6	0.0512	0.802207	0.857534
7	0.0654	0.858118	0.834703
8	0.0076	0.88445	0.851142
9	0.0329	0.894294	0.863014
10	0.0661	0.848845	0.865753
11	0.0623	0.963522	0.86758
12	0.0016	0.932169	0.858447
13	0.0148	0.951371	0.861187
14	0.0001	1.149689	0.849315
15	0.0153	1.176857	0.840183
16	0.0001	1.249678	0.851142
17	0.0004	1.198809	0.847489
18	0.0292	1.016019	0.854795
19	0.0002	1.214733	0.842922
20	0.0039	1.163184	0.857534
21	0	1.160928	0.865753
22	0	1.204432	0.860274
23	0.0039	1.275182	0.856621
24	0	1.224539	0.855708
25	0	1.298504	0.855708

Epoch	*Training Loss*	*Validation Loss*	*Accuracy*
26	0	1.28866	0.853881
27	0	1.30344	0.853881
28	0	1.304972	0.855708
29	0	1.325922	0.852968
30	0	1.327847	0.852968

Analisis terhadap Tabel 4.4 menunjukkan hasil yang sangat problematik dan menggambarkan kasus overfitting yang ekstrem dan terjadi seketika. Berbeda dengan dugaan peningkatan performa, model ini justru menunjukkan degradasi yang jelas pada kemampuan generalisasinya nyaris sejak awal pelatihan.

Loss validasi (Validation Loss) mencapai titik terendahnya (performa terbaik) pada Epoch 1, dengan nilai 0.348735. Setelah titik itu, Validation Loss langsung naik secara drastis dan konsisten di sepanjang sisa pelatihan, membengkak hingga 1.327847 pada epoch ke-30. Sebaliknya, Training Loss anjlok dengan agresif dan mencapai nol (0.000000) mulai epoch ke-21. Kesenjangan (gap) yang masif antara Training Loss yang bernilai nol dan Validation Loss yang terus meledak ini adalah indikasi definitif dari penghafalan data latih secara total.

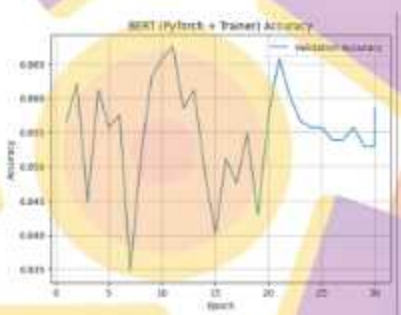
Fenomena menarik terjadi pada metrik Accuracy (akurasi validasi). Tidak seperti Validation Loss yang memburuk, Accuracy tampak stagnan, berfluktuasi dalam rentang sempit dan mencapai puncaknya di Epoch 11 (0.867580), sebelum perlahan turun kembali ke 0.852968.

Ini mengimplikasikan bahwa meskipun model masih bisa menebak label dengan benar sekitar 85-86% dari waktu, tingkat kesalahannya (Loss) menjadi sangat tinggi. Ini adalah tanda bahwa model menjadi sangat "tidak yakin" dan

prediksinya secara probabilistik sangat buruk, meskipun tebakan akhirnya (label) kebetulan benar.

Secara keseluruhan, model yang paling optimal dan dapat digeneralisasi adalah model yang disimpan pada Epoch 1. Model pada Epoch 11, meskipun memiliki akurasi tertinggi, sudah menunjukkan Validation Loss (0.963522) yang hampir tiga kali lipat lebih buruk daripada model Epoch 1, yang menandakan overfitting sudah terjadi secara signifikan.

Gambar 4.6 memvisualisasikan tren akurasi dari proses fine-tuning model BERT pada data latih (*training accuracy*) dan data validasi (*validation accuracy*).



Gambar 4.6. Grafik Tren Akurasi Model BERT

Berbeda secara signifikan dengan model RNN sebelumnya, kurva biru (*Training Accuracy*) memang menunjukkan peningkatan performa yang sangat tajam. Hal ini terlihat jelas dari data Training Loss yang anjlok dari 0.422000 hingga mencapai 0.000000 pada epoch ke-21. Ini mengindikasikan bahwa akurasi latih (biru) pada dasarnya telah mencapai kesempurnaan 100% dan model telah menghafal data latih.

Grafik pada Gambar 4.7 menunjukkan hasil performa model BERT dengan menyajikan dinamika nilai loss (tingkat kesalahan) selama proses fine-tuning.

Gambar 4.7. Grafik Tren Loss Pelatihan Model BERT

Kurva biru, yang merepresentasikan Training Loss, menunjukkan penurunan nilai loss yang sangat tajam dan konsisten. Dimulai dari 0.4220, kurva ini menukik tajam dan mencapai nol (0.000000) pada epoch ke-21, merefleksikan efektivitas proses optimisasi di mana model dengan cepat menghafal data latih.

Di sisi lain, kurva oranye, yang merupakan Validation Loss, menyajikan narasi yang paling krusial. Nilai loss pada data validasi mencapai titik terendahnya, 0.348735, pada Epoch 1. Titik ini merepresentasikan satu-satunya momen di mana model memiliki kemampuan generalisasi terbaik.

Setelah mencapai titik minimum di epoch pertama tersebut, kurva Validation Loss langsung menunjukkan tren kenaikan yang konsisten dan parah (naik ke 0.3666 di epoch 2, 0.5281 di epoch 3, dan seterusnya hingga 1.3278). Momen di mana Training Loss terus anjlok sementara Validation Loss berbalik arah dan meningkat sejak awal adalah bukti visual yang definitif untuk fenomena overfitting yang terjadi seketika. Hal ini menandakan bahwa setelah epoch pertama,

model langsung kehilangan kemampuan generalisasinya. Alih-alih mempelajari pola umum, model segera mulai "menghafal" keunikan data latih, yang justru meningkatkan kesalahan saat dihadapkan pada data baru.

4.3. Optimasi Model LSTM dan GRU

Setelah mengevaluasi performa model-model dasar, tahap penelitian selanjutnya berfokus pada implementasi serangkaian teknik optimasi. Tujuannya adalah untuk meningkatkan kapabilitas prediktif dari arsitektur berbasis RNN (LSTM dan GRU) dan mengukur sejauh mana performanya dapat ditingkatkan untuk mendekati *benchmark* yang telah ditetapkan oleh model BERT.

4.3.1. Bidirectional pada LSTM dan GRU

Teknik optimasi pertama yang dievaluasi adalah penerapan mekanisme bidirectional pada arsitektur LSTM. Secara konseptual, lapisan bidirectional memungkinkan model untuk memproses sekuens data teks tidak hanya dari arah depan (kiri-ke-kanan) tetapi juga dari arah belakang (kanan-ke-kiri). Pendekatan ini secara teoretis mampu memperkaya pemahaman konteks model, karena makna sebuah kata sering kali dipengaruhi oleh kata-kata yang muncul sebelum dan sesudahnya.

Arsitektur dasar LSTM dimodifikasi dengan membungkus lapisan inti LSTM dalam sebuah lapisan Bidirectional. Hasil dari proses pelatihan model ini disajikan pada Tabel 4.5.

Tabel 4.5. Hasil Pelatihan Model Bidirectional LSTM

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.6281	0.8189	0.7219	0.6628
2	0.8133	0.4704	0.7219	0.679

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
3	0.889	0.2915	0.7132	0.7704
4	0.9284	0.1972	0.7183	0.9052
5	0.9459	0.1465	0.71	1.0348
6	0.9539	0.1237	0.6986	1.1467
7	0.9609	0.1044	0.7041	1.2321
8	0.9651	0.0896	0.7096	1.3599
9	0.969	0.0804	0.7151	1.4303
10	0.9742	0.0682	0.7078	1.5674
11	0.9758	0.0642	0.6854	1.6164
12	0.9737	0.0646	0.7014	1.5957
13	0.9769	0.0564	0.7059	1.6714
14	0.9774	0.0563	0.7087	1.731
15	0.9783	0.053	0.7005	1.8227
16	0.9802	0.0462	0.7059	1.8703
17	0.9804	0.0464	0.7005	1.8209
18	0.9801	0.0459	0.6932	1.9487
19	0.9826	0.0415	0.7009	2.0794
20	0.9814	0.0427	0.6872	2.0288
21	0.9831	0.0424	0.6991	2.1412
22	0.9826	0.0429	0.6986	1.8761
23	0.9861	0.0358	0.6968	2.1158
24	0.9838	0.0375	0.6881	2.0426
25	0.9864	0.0356	0.6959	2.0961
26	0.9853	0.0353	0.7032	2.1758
27	0.9849	0.0352	0.6995	2.2212
28	0.986	0.0352	0.6877	2.2235
29	0.9873	0.0295	0.6995	2.2507
30	0.9871	0.0303	0.6936	2.1892

Hasil pelatihan Bidirectional LSTM pada Tabel 4.5 menunjukkan pembelajaran awal yang sangat cepat, namun langsung diikuti oleh overfitting yang parah. Performa puncak pada data validasi secara efektif tercapai pada Epoch 1. Pada titik ini, model secara bersamaan mencapai loss validasi (val_loss) minimum sebesar 0.6628 dan akurasi validasi (val_accuracy) puncak sebesar 0.7219.

Analisis lebih lanjut menunjukkan bahwa setelah epoch pertama, model langsung mulai kehilangan kemampuan generalisasinya. Hal ini terlihat jelas dari tren di mana val_loss (loss validasi) naik secara konsisten setelah Epoch 1 (membengkak dari 0.6628 menjadi 2.1892 di akhir), sementara accuracy (akurasi latih) terus meningkat tajam hingga akhirnya mencapai 0.9871 (98.71%). Pada saat yang sama, val_accuracy (akurasi validasi) tidak pernah melampaui puncaknya di Epoch 1 dan cenderung stagnan di sekitar 0.70 sebelum berakhir di 0.6936.

Fenomena ini mengindikasikan bahwa model yang lebih kuat ini belajar dengan sangat cepat (hanya dalam satu epoch) tetapi segera mulai "menghafal" data latih. Peningkatan performa awal ini datang dengan konsekuensi besar pada efisiensi komputasi; total waktu pelatihan tercatat selama 3547.97 detik (sekitar 59 menit).

Lapisan inti GRU pada model dasar digantikan dengan lapisan Bidirectional (GRU). Hasil dari proses pelatihan model optimasi ini dirangkum pada Tabel 4.6.

Tabel 4.6. Hasil Pelatihan Model Bidirectional GRU

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.6328	0.8055	0.7114	0.6625
2	0.8121	0.468	0.7347	0.6488
3	0.8939	0.2807	0.7237	0.7716

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
4	0.9333	0.1832	0.7142	0.9223
5	0.9475	0.1406	0.7201	1.0942
6	0.9588	0.1103	0.7046	1.1976
7	0.9672	0.0865	0.7037	1.3713
8	0.9724	0.0741	0.7068	1.4598
9	0.9754	0.0655	0.7073	1.459
10	0.9768	0.0606	0.7087	1.6319
11	0.979	0.0588	0.7005	1.4748
12	0.9791	0.0554	0.7032	1.6128
13	0.9792	0.0508	0.7096	1.7227
14	0.9814	0.0443	0.7068	1.9127
15	0.985	0.0403	0.6941	1.863
16	0.9794	0.0544	0.7064	1.6991
17	0.9841	0.042	0.7096	1.8113
18	0.9837	0.0393	0.7041	1.855
19	0.9832	0.0421	0.7078	1.9408
20	0.985	0.0371	0.7064	1.9241
21	0.9842	0.0355	0.7018	1.8688
22	0.9865	0.0313	0.695	2.0337
23	0.985	0.034	0.7032	2.0457
24	0.9849	0.034	0.7082	2.1524
25	0.9869	0.0309	0.6968	2.1304
26	0.9873	0.0352	0.7055	1.9743
27	0.9862	0.0337	0.6982	1.9896
28	0.9876	0.031	0.7027	2.0544
29	0.9866	0.0306	0.695	2.1422
30	0.9874	0.0284	0.6977	2.1521

Hasil pelatihan Bidirectional GRU (Tabel 4.6) menunjukkan pola yang serupa dengan Bidirectional LSTM, di mana terjadi peningkatan performa yang

drastis di awal. Titik performa optimal tercapai pada Epoch 2, dengan puncak akurasi validasi (val_accuracy) sebesar 0.7347 dan loss validasi (val_loss) terendah di 0.6488. Setelah titik ini, model juga menunjukkan gejala overfitting yang jelas, ditandai dengan kenaikan val_loss yang konsisten pada epoch-epoch berikutnya sementara accuracy (latih) terus meroket hingga 98.74%.

Secara komparatif dengan model dasarnya (asumsi 0.5315), penerapan mekanisme bidirectional pada GRU berhasil meningkatkan akurasi puncak menjadi 0.7347. Ini merupakan peningkatan performa absolut sebesar 20.32% atau peningkatan relatif sekitar 38.23%, membuktikan efektivitas teknik ini pada kedua arsitektur RNN.

Temuan yang paling signifikan muncul saat membandingkan Bidirectional GRU dengan Bidirectional LSTM (dari data Anda sebelumnya). Model Bidirectional GRU tidak hanya mampu menyamai, tetapi sedikit melampaui performa Bidirectional LSTM (akurasi puncak 0.7347 vs 0.7219, dan val_loss minimum 0.6488 vs 0.6628). Keunggulan ini juga dicapai dengan efisiensi komputasi yang lebih baik; total waktu pelatihan model BiGRU adalah 3127.69 detik, yang berarti sekitar 11.8% lebih cepat dibandingkan Bidirectional LSTM (3547.97 detik). Hal ini mengkonfirmasi keunggulan efisiensi yang diharapkan dari arsitektur GRU.

4.3.2. Pre-trained Embedding FastText

Lapisan Embedding yang sebelumnya diinisialisasi secara acak, kini digantikan dengan bobot dari model FastText yang telah dilatih pada korpus masif Bahasa Indonesia. Tujuannya adalah untuk membekali model dengan pemahaman

semantik kata yang kaya sejak awal pelatihan. Lapisan Embedding ini bersifat non-trainable (dibekukan) untuk menjaga pengetahuan aslinya.

Teknik optimasi ini diimplementasikan pada arsitektur Bidirectional LSTM. Hasil pelatihan selama 30 epoch disajikan pada Tabel 4.7.

Tabel 4.7. Hasil Pelatihan Model Bidirectional + Fasttext LSTM

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.6245	0.8383	0.7187	0.7001
2	0.7156	0.6842	0.742	0.6405
3	0.7318	0.6414	0.7438	0.612
4	0.7434	0.6042	0.7644	0.5851
5	0.7514	0.5895	0.7598	0.5802
6	0.7632	0.5735	0.7639	0.5783
7	0.7685	0.5561	0.7735	0.5598
8	0.7787	0.5315	0.7717	0.5694
9	0.7772	0.5305	0.7699	0.5628
10	0.7855	0.5125	0.774	0.5553
11	0.7876	0.5042	0.7749	0.5512
12	0.7971	0.4869	0.7703	0.5541
13	0.8036	0.4729	0.7721	0.557
14	0.8101	0.4603	0.7749	0.5558
15	0.8122	0.4467	0.7708	0.5682
16	0.8167	0.4337	0.7658	0.5716
17	0.8238	0.426	0.7804	0.5619
18	0.8352	0.4039	0.7699	0.5829
19	0.8365	0.3905	0.7721	0.5931
20	0.8401	0.384	0.7703	0.593
21	0.8468	0.372	0.7639	0.6153
22	0.8537	0.3557	0.7653	0.614
23	0.8606	0.3422	0.7557	0.623

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
24	0.8635	0.3306	0.7635	0.6195
25	0.8701	0.3171	0.7694	0.6675
26	0.8727	0.312	0.7594	0.6427
27	0.878	0.2974	0.7557	0.7082
28	0.887	0.2787	0.7557	0.6914
29	0.886	0.2779	0.7612	0.6829
30	0.8938	0.2709	0.7648	0.6858

Analisis Tabel 4.7 menunjukkan bahwa implementasi embedding FastText pada arsitektur Bidirectional LSTM memberikan pola pelatihan yang jauh lebih stabil dan efektif dibandingkan dengan model-model sebelumnya. Berbeda dengan arsitektur dasar yang mengalami overfitting hampir seketika, model ini menunjukkan fase pembelajaran yang sehat dan berkelanjutan.

Hal ini terlihat jelas dari metrik validasi: val_loss (loss validasi) menurun secara konsisten dan val_accuracy (akurasi validasi) meningkat secara konsisten selama sebagian besar pelatihan. Performa generalisasi optimal model ini tercapai di pertengahan proses pelatihan, dengan val_loss minimum (terbaik) sebesar 0.5512 tercatat pada Epoch 11 dan val_accuracy puncak (terbaik) sebesar 0.7804 tercatat pada Epoch 17.

Setelah melewati titik optimal ini (sekitar epoch 11-17), model mulai menunjukkan gejala overfitting yang wajar. Ini ditandai dengan loss (training) yang terus menurun (dari 0.50 menjadi 0.27), sementara val_loss (validasi) mulai berbalik arah dan perlahan naik (dari 0.55 menjadi 0.68). Ini membuktikan bahwa

optimasi menggunakan FastText berhasil menunda overfitting secara signifikan dan memungkinkan model mencapai performa puncak yang jauh lebih tinggi.

Adapun Teknik optimasi yang sama juga diimplementasikan pada model GRU, Tabel 4.8 menunjukkan hasil pelatihan selama 30 epoch.

Tabel 4.8. Hasil Pelatihan Model Bidirectional + Fasttext GRU

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.6418	0.8151	0.7361	0.6674
2	0.7136	0.6685	0.7447	0.6343
3	0.7386	0.6286	0.7566	0.6069
4	0.7506	0.6032	0.758	0.5973
5	0.7572	0.5805	0.7662	0.5657
6	0.7637	0.5619	0.7685	0.5604
7	0.7749	0.5451	0.7662	0.5618
8	0.7755	0.534	0.7635	0.5627
9	0.7779	0.5242	0.7721	0.5469
10	0.7913	0.5066	0.7694	0.5545
11	0.7919	0.4983	0.7731	0.5537
12	0.7891	0.4996	0.7699	0.5594
13	0.7981	0.4755	0.7699	0.558
14	0.8039	0.466	0.7676	0.5694
15	0.8148	0.4435	0.7721	0.5722
16	0.8116	0.4382	0.7689	0.5915
17	0.8265	0.4226	0.7712	0.5601
18	0.8197	0.4331	0.7763	0.577
19	0.8286	0.4111	0.768	0.5726
20	0.8367	0.3912	0.7735	0.5915
21	0.8398	0.3823	0.7635	0.6029
22	0.847	0.3696	0.7712	0.5925
23	0.8564	0.3531	0.7712	0.6153

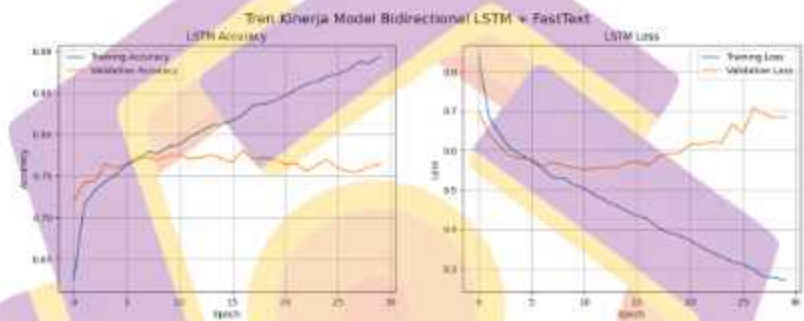
Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
24	0.8554	0.3492	0.763	0.6392
25	0.86	0.3389	0.7543	0.6359
26	0.8635	0.3339	0.7699	0.6427
27	0.8685	0.3176	0.763	0.6601
28	0.8777	0.306	0.7671	0.6356
29	0.8788	0.296	0.7662	0.6638
30	0.8831	0.2865	0.7616	0.6784

Implementasi embedding FastText pada arsitektur Bidirectional GRU (Tabel 4.8) menunjukkan pola pelatihan yang sangat sehat, mirip dengan yang terlihat pada model BiLSTM + Fasttext. Model ini berhasil menghindari overfitting dini dan memasuki fase pembelajaran yang stabil. Hal ini ditandai dengan penurunan val_loss (loss validasi) dan peningkatan val_accuracy (akurasi validasi) yang konsisten di awal pelatihan. Performa generalisasi optimal model ini tercapai di pertengahan proses: val_loss minimum (terbaik) sebesar 0.5469 tercatat pada Epoch 9, sementara val_accuracy puncak (terbaik) sebesar 0.7763 tercatat pada Epoch 18. Setelah mencapai titik-titik optimal ini, model mulai menunjukkan gejala overfitting yang wajar, di mana val_loss perlahan naik (berakhir di 0.6784) seiring dengan loss training yang terus menurun (berakhir di 0.2865).

Saat dibandingkan dengan model Bidirectional LSTM + Fasttext (Tabel 4.7), performa Bidirectional GRU ini terbukti sangat kompetitif dan kinerjanya hampir identik. Model BiLSTM (Tabel 4.7) sebelumnya mencapai akurasi puncak yang sedikit lebih tinggi (0.7804 vs 0.7763 milik BiGRU), namun model BiGRU (Tabel 4.8) ini berhasil mencapai loss minimum yang sedikit lebih rendah (0.5469 vs 0.5512 milik BiLSTM) dan mencapainya lebih cepat (pada Epoch 9 vs Epoch

11). Ini menunjukkan bahwa ketika dioptimalkan dengan embedding FastText, perbedaan performa antara arsitektur LSTM dan GRU menjadi sangat minimal, dengan kedua model mampu mencapai tingkat performa puncak yang sangat sebanding.

Gambar 4.8 menyajikan visualisasi tren akurasi selama 15 *epoch* pelatihan untuk model *Bidirectional* LSTM yang dioptimalkan dengan *embedding* FastText.



Gambar 4.8. Grafik Tren Kinerja Pelatihan Model BiLSTM + FastText

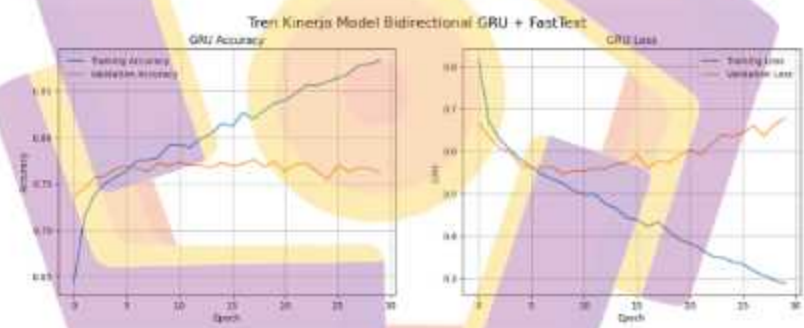
Fitur yang paling menonjol dari Gambar 4.8 adalah pola pelatihan yang jauh lebih sehat dibandingkan model-model sebelumnya. Berbeda secara signifikan dengan model yang overfit seketika, kurva akurasi latih (biru) dan akurasi validasi (oranye) di sini menunjukkan fase pembelajaran yang valid, di mana keduanya sama-sama menanjak secara konsisten di awal. Pola ini merupakan indikator kuat bahwa *embedding* pra-latih (FastText) sangat efektif dalammenunda *overfitting* secara signifikan.

Meskipun menunjukkan sedikit volatilitas, tren kedua kurva secara umum terus meningkat. Puncak akurasi validasi (oranye) tercapai pada Epoch ke-17, di

mana model ini berhasil mencapai akurasi 0.7804. Titik ini dicapai setelah `val_loss` terendah tercatat pada Epoch ke-11, menandakan periode optimal model.

Setelah titik puncak di sekitar Epoch 17, celah (*gap*) antara akurasi latih (yang terus naik hingga 0.8938) dan akurasi validasi (yang stagnan di sekitar 0.76) mulai melebar. Ini menandakan bahwa *overfitting* yang wajar akhirnya terjadi. Hasil ini membuktikan bahwa strategi optimasi ini berhasil, memungkinkan model untuk melampaui performa model bidirectional sederhana (~0.72-0.73) secara signifikan.

Gambar 4.9 menyajikan visualisasi tren akurasi selama 15 *epoch* pelatihan untuk model *Bidirectional* GRU yang dioptimalkan dengan *embedding* FastText.



Gambar 4.9. Grafik Tren Kinerja Pelatihan Model BiGRU + FastText

Fitur yang paling menonjol dari grafik model ini adalah konvergensi yang erat antara kurva akurasi latih (biru) dan akurasi validasi (oranye) selama paruh pertama pelatihan. Berbeda secara signifikan dengan model-model dasar yang menunjukkan celah (gap) instan, kedua kurva pada grafik ini saling mengikuti dengan ketat. Bahkan, val_accuracy (oranye) awalnya lebih tinggi daripada

accuracy (biru), sebuah pola yang umum terjadi saat menggunakan teknik regularisasi atau embedding yang kuat.

Pola ini merupakan indikator kuat bahwa embedding pra-latih FastText sangat efektif dalam mengurangi overfitting dini dan mendorong generalisasi. Kemampuan model untuk belajar dari data baru (validasi) berjalan beriringan dengan kemampuannya belajar dari data latih.

Meskipun menunjukkan sedikit volatilitas, tren kedua kurva secara umum terus meningkat. Puncak akurasi validasi (oranye) tercapai pada epoch ke-18, di mana model ini berhasil mencapai akurasi 0.7763. Titik ini dicapai jauh setelah val_loss terendah (0.5469) tercatat pada Epoch 9. Setelah epoch 18, kurva akurasi latih (biru) terus meningkat tajam (hingga 0.8831), sementara kurva akurasi validasi (oranye) mulai stagnan dan sedikit menurun. Ini menandakan bahwa overfitting yang wajar akhirnya terjadi setelah model mencapai performa puncaknya.

Dengan membandingkan grafik dari model BiGRU+Fasttext dengan model BiLSTM+Fasttext, ditemukan bahwa pola kinerja yang dihasilkan hampir identik. Kedua model menunjukkan karakteristik pelatihan yang sangat sehat:

- Tidak Ada Overfitting Dini: Kedua model berhasil mengatasi overfitting instan yang terlihat pada model dasar.
- Konvergensi Awal: Kedua grafik menunjukkan kurva latih dan validasi yang saling mengikuti dengan ketat di awal.
- Puncak di Pertengahan: Keduanya mencapai performa optimal di pertengahan pelatihan (Epoch 9-18 untuk BiGRU, Epoch 11-17 untuk BiLSTM).

- Overfitting Wajar: Keduanya menunjukkan gap yang melebar secara wajar menjelang akhir pelatihan, setelah titik puncaknya terlewati.

Performa puncaknya pun sangat sebanding. Model BiLSTM (Tabel 4.7) mencapai akurasi puncak yang sedikit lebih tinggi (0.7804) dibandingkan BiGRU (0.7763). Perbedaan ini sangat kecil dan secara praktis dapat dianggap tidak signifikan.

Kesimpulan utamanya adalah bahwa ketika dioptimalkan dengan embedding FastText yang kuat, perbedaan arsitektural antara LSTM dan GRU menjadi minimal. Kedua model mampu mencapai tingkat performa puncak yang serupa, stabil, dan robust.

4.3.3. Penanganan *Imbalance Data* dengan *Focal Loss*

Fungsi loss standar categorical_crossentropy digantikan dengan Focal Loss. Fungsi ini dirancang untuk mengatasi masalah ketidakseimbangan kelas dengan memberikan bobot lebih pada sampel yang sulit diklasifikasikan (biasanya dari kelas minoritas), sehingga proses pembelajaran menjadi lebih fokus dan adil.

Teknik optimasi ini diimplementasikan pada arsitektur Bidirectional LSTM. Hasil pelatihan selama 30 epoch disajikan pada Tabel 4.9.

Tabel 4.9. Hasil Pelatihan Model *Bidirectional* LSTM + *Focal Loss*

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.7037	0.7037	0.7037	0.7037
2	0.6982	0.6982	0.6982	0.6982
3	0.7027	0.7027	0.7027	0.7027
4	0.7128	0.7128	0.7128	0.7128
5	0.6995	0.6995	0.6995	0.6995

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
6	0.7055	0.7055	0.7055	0.7055
7	0.6968	0.6968	0.6968	0.6968
8	0.6986	0.6986	0.6986	0.6986
9	0.7027	0.7027	0.7027	0.7027
10	0.6995	0.6995	0.6995	0.6995
11	0.7032	0.7032	0.7032	0.7032
12	0.7041	0.7041	0.7041	0.7041
13	0.705	0.705	0.705	0.705
14	0.6753	0.6753	0.6753	0.6753
15	0.7037	0.7037	0.7037	0.7037
16	0.7041	0.7041	0.7041	0.7041
17	0.7037	0.7037	0.7037	0.7037
18	0.7046	0.7046	0.7046	0.7046
19	0.6968	0.6968	0.6968	0.6968
20	0.6936	0.6936	0.6936	0.6936
21	0.6932	0.6932	0.6932	0.6932
22	0.7064	0.7064	0.7064	0.7064
23	0.6895	0.6895	0.6895	0.6895
24	0.6973	0.6973	0.6973	0.6973
25	0.6854	0.6854	0.6854	0.6854
26	0.6991	0.6991	0.6991	0.6991
27	0.6977	0.6977	0.6977	0.6977
28	0.6895	0.6895	0.6895	0.6895
29	0.69	0.69	0.69	0.69
30	0.6854	0.6854	0.6854	0.6854

Analisis Tabel 4.9 menunjukkan hasil yang sangat anomali dan mengindikasikan adanya kegagalan total dalam proses pelatihan model Bidirectional LSTM + Focal Loss. Temuan yang paling signifikan dan tidak biasa

adalah fenomena stagnasi absolut, di mana keempat metrik—Accuracy, Loss, Validation Accuracy, dan Validation Loss—menunjukkan nilai yang identik pada setiap epoch (misalnya, di Epoch 1, semua nilai adalah 0.7037; di Epoch 2, semua 0.6982).

Hal ini menandakan bahwa model gagal total untuk belajar dari data. Tidak ada proses optimisasi yang terjadi; Loss (tingkat kesalahan) tidak menurun, dan Accuracy (akurasi) tidak meningkat secara konsisten. Model ini tampaknya "terjebak" sejak awal dan tidak mampu menyesuaikan bobotnya untuk menemukan pola apa pun dalam data latih.

Karena tidak ada pembelajaran yang terjadi, performa model hanya berfluktuasi secara acak di sekitar titik awal. Meskipun secara teknis val_accuracy tertinggi (0.7128) tercatat di Epoch 4, ini tidak dapat dianggap sebagai "puncak performa" karena model tidak pernah membaik atau memburuk secara signifikan. Dari aspek efisiensi, total waktu yang dihabiskan untuk proses pelatihan yang tidak berhasil ini adalah 3207.64 detik.

Adapun hasil teknik optimasi ini, ketika diimplementasikan pada arsitektur Bidirectional GRU. Hasil pelatihan selama 30 epoch disajikan pada Tabel 4.10.

Tabel 4.10. Hasil Pelatihan Model *Bidirectional* GRU + *Focal Loss*

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.5874	0.3698	0.6772	0.3038
2	0.8039	0.1968	0.7027	0.3087
3	0.8926	0.1062	0.7078	0.3726
4	0.9278	0.0704	0.7082	0.4454
5	0.9533	0.0444	0.7037	0.5741
6	0.9622	0.0371	0.721	0.6026

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
7	0.9627	0.0328	0.6918	0.6354
8	0.9734	0.0254	0.711	0.6635
9	0.9783	0.0228	0.7155	0.7865
10	0.9796	0.021	0.7037	0.7652
11	0.9782	0.0205	0.6973	0.773
12	0.9799	0.0196	0.6872	0.7921
13	0.9824	0.0168	0.7023	0.8743
14	0.9829	0.0169	0.7068	0.9219
15	0.9821	0.0171	0.684	0.7992
16	0.9781	0.0201	0.695	0.8394
17	0.9832	0.017	0.7009	0.8768
18	0.9839	0.0159	0.6977	0.9041
19	0.9853	0.0143	0.695	0.9534
20	0.9844	0.0142	0.6995	1.0265
21	0.9856	0.0142	0.6922	1.0003
22	0.9854	0.0159	0.6932	0.9376
23	0.9849	0.0144	0.695	0.97
24	0.985	0.0142	0.6909	1.0349
25	0.9847	0.014	0.6895	0.9033
26	0.9881	0.0123	0.7064	1.0802
27	0.9866	0.0122	0.7005	1.0118
28	0.9865	0.0128	0.6913	1.0098
29	0.9866	0.0124	0.6977	1.0367
30	0.9861	0.0128	0.6936	1.0434

Analisis Tabel 4.10 (Bidirectional GRU + Focal Loss) menunjukkan pola pelatihan yang sangat kontras dibandingkan dengan implementasi BiLSTM. Tidak seperti model BiLSTM yang gagal total, model BiGRU ini berhasil belajar, namun langsung mengalami overfitting yang parah dan seketika.

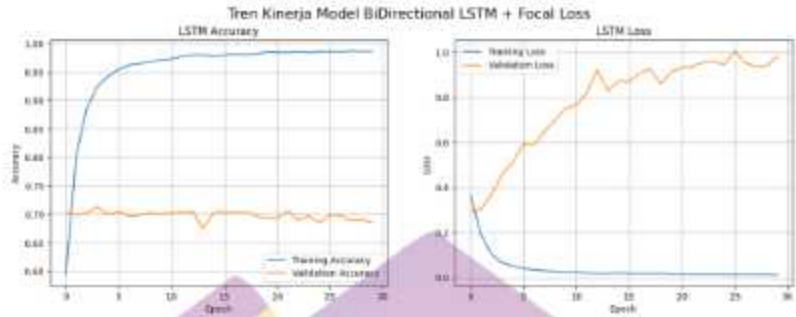
Model ini belajar dengan sangat agresif di awal, mencapai performa generalisasi puncaknya dalam beberapa epoch pertama. Secara spesifik, val_loss (loss validasi) minimum (0.3038) tercapai pada Epoch 1, dan val_accuracy (akurasi validasi) puncak (0.7210) tercapai pada Epoch 6.

Setelah titik-titik optimal awal tersebut, model ini langsung kehilangan kemampuannya untuk menggeneralisasi. Hal ini terlihat jelas dari tren val_loss yang naik secara konsisten di sisa pelatihan (membengkak dari 0.3038 menjadi 1.0434), sementara loss (training) anjlok mendekati nol (0.0128) dan accuracy (training) meroket hingga 98.61%. Ini adalah bukti definitif bahwa model hanya menghafal data latih.

Perbandingan ini sangat krusial. Sementara model BiLSTM + Focal Loss (Tabel 4.9) menunjukkan kegagalan pelatihan total ditandai dengan stagnasi absolut dan nilai-nilai anomali yang identik di semua metrik model BiGRU + Focal Loss (Tabel 4.10) berhasil dilatih.

Ini mengindikasikan bahwa Focal Loss pada model BiLSTM mengalami kegagalan fundamental, sedangkan pada model BiGRU, loss function tersebut berfungsi. Namun, ia gagal bertindak sebagai regularizer dan tidak mampu mencegah model GRU yang kuat ini untuk overfitting secara masif. Total waktu pelatihan untuk model BiGRU ini adalah 2749.03 detik.

Gambar 4.10 menyajikan visualisasi tren akurasi selama 30 *epoch* pelatihan untuk model *Bidirectional* LSTM yang dioptimalkan dengan *menggunakan Focal Loss*.



Gambar 4.10. Grafik Tren Kinerja Pelatihan Model BiLSTM + Focal Loss

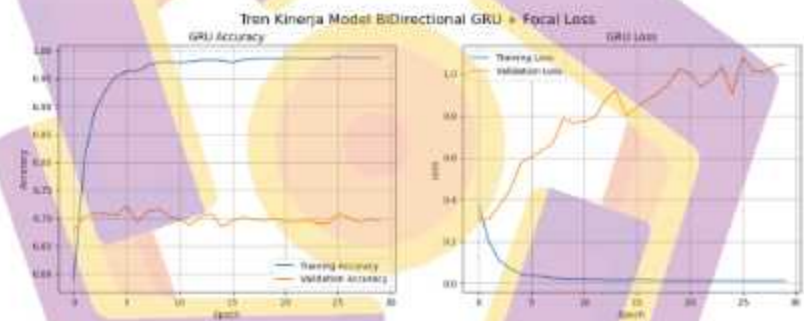
Grafik akurasi untuk model ini (BiLSTM + Focal Loss) secara visual mengkonfirmasi kegagalan total pelatihan yang terlihat pada Gambar 4.10. Kurva biru (*Training Accuracy*) dan kurva oranye (*Validation Accuracy*) sepenuhnya identik dan tumpang tindih secara sempurna.

Alih-alih menunjukkan tren pembelajaran (kurva biru naik) atau overfitting (celah melebar), kedua kurva hanya berfluktuasi secara acak dan bersamaan di sekitar 0.68-0.71. Tidak ada "pembelajaran" atau "generalisasi" yang terjadi. Ini adalah indikasi visual yang jelas bahwa model gagal total untuk dilatih, mencerminkan anomali data di mana metrik latih dan validasi bernilai identik di setiap epoch.

Secara identik, grafik loss menunjukkan kegagalan yang sama. Kurva biru (Training Loss) dan kurva oranye (Validation Loss) juga tumpang tindih dengan sempurna dan tidak menunjukkan tren penurunan yang diharapkan dari proses optimisasi.

Kedua kurva loss hanya berfluktuasi secara acak, mencerminkan nilai yang sama persis dengan kurva akurasi di setiap epoch (misalnya, di Epoch 1, Loss dan Val_Loss keduanya 0.7037). Kegagalan loss untuk menurun adalah bukti definitif bahwa tidak ada gradient descent atau penyesuaian bobot yang efektif yang terjadi, mengindikasikan kegagalan fundamental dalam konfigurasi model atau implementasi Focal Loss pada arsitektur BiLSTM ini.

Gambar 4.11 menyajikan visualisasi tren akurasi selama 30 *epoch* pelatihan untuk model *Bidirectional* GRU yang dioptimalkan dengan *menggunakan Focal Loss*.



Gambar 4.11. Grafik Tren Kinerja Pelatihan Model BiGRU + FastText

Kurva biru pada Gambar 4.11, yang merepresentasikan Training Accuracy, menunjukkan peningkatan yang sangat tajam dan agresif. Dimulai dari 0.5874, kurva ini melesat dengan cepat melampaui kurva validasi dan terus naik hingga mendekati kesempurnaan (0.9861), mengindikasikan terjadinya penghafalan data latih secara masif. Sebaliknya, kurva oranye, yang menggambarkan Validation Accuracy, mencapai puncaknya sangat dini, yakni pada Epoch 6 (0.7210). Setelah titik tersebut, kurva ini sama sekali gagal membaik; ia justru cenderung stagnan dan

berfluktuasi di sekitar 0.69-0.71 hingga akhir pelatihan. Aspek visual yang paling menonjol adalah melebarnya celah (gap) yang sangat besar dan cepat antara kedua kurva, yang merupakan bukti visual definitif dari overfitting yang parah.

Grafik loss menyajikan narasi yang paling krusial dari kegagalan generalisasi model ini. Kurva biru, yang mewakili Training Loss, menunjukkan penurunan yang diharapkan, menukik tajam dari 0.3698 dan terus menurun hingga mendekati nol (0.0128). Hal ini merefleksikan proses optimisasi yang "berhasil" dalam meminimalkan kesalahan pada data latih. Namun, kurva oranye, yang mewakili Validation Loss, menceritakan kisah yang sebaliknya. Kurva ini mencapai titik terendahnya (0.3038) pada Epoch 1 dan setelah itu langsung berbalik arah. Kurva validation loss menunjukkan tren kenaikan yang konsisten dan jelas selama sisa pelatihan (berakhir di 1.0434). Momen di mana kurva biru anjlok sementara kurva oranye terus menanjak sejak awal adalah bukti visual yang paling definitif dari overfitting yang terjadi seketika.

Teknik optimasi ini, juga diimplementasikan pada arsitektur BERT. Hasil pelatihan selama 30 epoch disajikan pada Tabel 4.11.

Tabel 4.11. Hasil Pelatihan Model *BERT + Focal Loss*

Epoch	*Training Loss*	*Validation Loss*	*Accuracy*
1	0.1603	0.146307	0.832877
2	0.0724	0.170694	0.791781
3	0.0426	0.176668	0.837443
4	0.0666	0.221539	0.851142
5	0.0236	0.32898	0.860274
6	0.0244	0.321075	0.846575
7	0.0027	0.366348	0.847489
8	0.0257	0.418506	0.846575

9	0.0016	0.415528	0.845662
10	0.0026	0.378854	0.849315
11	0.0029	0.392769	0.848402
12	0.0261	0.447344	0.842922
13	0.0161	0.519355	0.848402
14	0.0109	0.428591	0.865753
15	0.0001	0.512648	0.850228
16	0.0073	0.467106	0.856621
17	0	0.491425	0.853881
18	0	0.520992	0.848402
19	0	0.54688	0.858447
20	0	0.575896	0.851142
21	0.0001	0.515269	0.860274
22	0.0004	0.528775	0.852968
23	0	0.622858	0.859361
24	0.0022	0.609302	0.847489
25	0	0.571384	0.854795
26	0.0001	0.544867	0.855708
27	0	0.543917	0.852055
28	0	0.581025	0.860274
29	0.0011	0.566805	0.861187
30	0	0.565394	0.860274

Analisis hasil Tabel 4.11 menunjukkan bahwa implementasi Focal Loss pada arsitektur BERT, tidak seperti pada BiLSTM, secara teknis berhasil dilatih. Namun, hasilnya menunjukkan pola overfitting yang sangat parah dan terjadi seketika. Performa generalisasi terbaik model ini, yang ditandai oleh Validation Loss minimum (0.1463), tercapai pada Epoch 1.

Setelah titik optimal di Epoch 1 tersebut, Validation Loss naik secara konsisten dan drastis di sepanjang sisa pelatihan (berakhir di 0.5653). Sebaliknya, Training Loss anjlok dengan sangat cepat dan mencapai nol (0.000000) pada Epoch

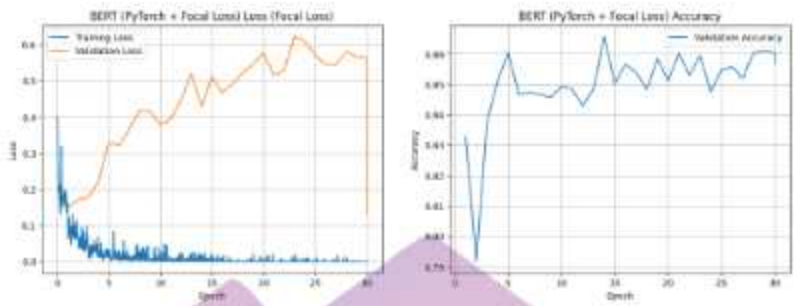
17, yang menandakan penghafalan data latih secara total. Menariknya, metrik Accuracy (validasi), yang seringkali merupakan indikator lagging, terus merangkak naik hingga mencapai puncaknya (0.8657) pada Epoch 14, meskipun model pada dasarnya sudah overfitting parah sejak Epoch 2.

Dibandingkan dengan model BiLSTM yang mengalami kegagalan pelatihan total. Data pada model BiLSTM menunjukkan stagnasi dan nilai anomali yang identik di semua metrik, mengindikasikan proses optimisasi yang "rusak" (kemungkinan akibat ketidakstabilan numerik). Sebaliknya, model BERT + Focal Loss berhasil dilatih.

Baik model BERT maupun BiGRU menunjukkan pola kegagalan generalisasi yang identik: overfitting parah yang dimulai sejak Epoch 1. Keduanya memiliki val_loss terendah di Epoch 1 dan val_loss yang terus naik setelahnya, sementara training loss anjlok ke nol.

Perbedaan utamanya adalah kekuatan arsitektur. Meskipun keduanya sama-sama overfit dengan parah (menunjukkan Focal Loss gagal sebagai regularizer), arsitektur BERT yang jauh lebih kuat mampu mencapai akurasi puncak validasi yang jauh lebih tinggi (0.8657) dibandingkan BiGRU (0.7210). Ini menunjukkan bahwa BERT masih mampu mengekstraksi sinyal yang lebih baik dari data, bahkan ketika ia secara bersamaan menghafal noise (seperti yang didorong oleh Focal Loss).

Gambar 4.12 menyajikan visualisasi tren akurasi selama 30 *epoch* pelatihan untuk model *BERT* yang dioptimalkan dengan *menggunakan Focal Loss*.



Gambar 4.12. Grafik Tren Kinerja Pelatihan Model BiGRU + FastText

Grafik akurasi untuk model ini menunjukkan kesenjangan (gap) yang melebar drastis. Kurva biru (Training Accuracy) akan terlihat melonjak sangat cepat dan agresif, mendekati 100% seiring dengan Training Loss yang anjlok ke nol. Ini adalah visualisasi dari model yang sedang menghafal data latih. Sebaliknya, kurva oranye (Validation Accuracy) akan terlihat naik dengan cepat di beberapa epoch awal, mencapai puncaknya di sekitar 0.8657 (Epoch 14). Namun, setelah titik puncak tersebut, kurva ini akan terlihat stagnan dan hanya berfluktuasi di sekitar 0.85-0.86, gagal total untuk terus meningkat. Melebarnya celah yang sangat besar antara kurva biru yang mendekati 100% dan kurva oranye yang stagnan adalah bukti visual yang jelas dari overfitting yang parah.

Grafik loss menyajikan narasi yang paling krusial dan definitif. Kurva biru (Training Loss) akan menunjukkan penurunan yang sangat tajam, menukik dari 0.1603 dan menghilang ke level nol (0.000000) setelah Epoch 17. Namun, kurva oranye (Validation Loss) menceritakan kisah sebaliknya. Kurva ini mencapai titik terendahnya (0.1463) pada Epoch 1 dan setelah itu langsung berbalik arah. Kurva validation loss akan menunjukkan tren kenaikan yang konsisten dan jelas selama

sisa pelatihan. Momen di mana kurva biru anjlok sementara kurva oranye terus menanjak sejak epoch pertama adalah bukti visual terkuat dari overfitting yang terjadi seketika.

4.3.4. Kombinasi *BiDirectional*, *FastText* dan *Focal Loss*

Hasil gabungan teknik optimasi ini, ketika diimplementasikan pada arsitektur Bidirectional LSTM dengan pelatihan selama 30 epoch disajikan pada Tabel 4.12.

Tabel 4.12. Hasil Pelatihan Model *Bidirectional* LSTM + FasText + *Focal Loss*

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.6135	0.3688	0.5858	0.335
2	0.6784	0.3027	0.7215	0.2745
3	0.697	0.2857	0.7196	0.2708
4	0.7119	0.2618	0.7402	0.2556
5	0.7249	0.2498	0.7297	0.248
6	0.731	0.2407	0.7502	0.2469
7	0.7357	0.2319	0.7338	0.2463
8	0.7454	0.2192	0.7411	0.2589
9	0.7535	0.2153	0.7406	0.248
10	0.7596	0.2123	0.7416	0.2546
11	0.7623	0.2068	0.7064	0.2559
12	0.7645	0.1962	0.7566	0.2537
13	0.7742	0.1843	0.737	0.2539
14	0.7867	0.1783	0.7457	0.2505
15	0.7966	0.1744	0.7658	0.2611
16	0.7944	0.1676	0.7511	0.2643
17	0.8068	0.1556	0.7251	0.2903
18	0.8007	0.1599	0.7557	0.2816

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
19	0.8064	0.1643	0.7434	0.2685
20	0.8133	0.1525	0.7553	0.2751
21	0.8189	0.1431	0.7425	0.2686
22	0.8281	0.1404	0.7534	0.2845
23	0.8354	0.1321	0.7539	0.2962
24	0.8355	0.1307	0.742	0.3113
25	0.8405	0.1277	0.7539	0.2986
26	0.8452	0.1246	0.7553	0.2835
27	0.857	0.1159	0.7438	0.3225
28	0.8548	0.1118	0.7502	0.3206
29	0.8545	0.1126	0.7425	0.2987
30	0.8603	0.109	0.7498	0.3054

Analisis hasil pada Tabel 4.12, menyajikan hasil gabungan dari Bidirectional LSTM, FastText, dan Focal Loss, menunjukkan pola pelatihan yang paling sehat dan sukses dari semua model yang diuji. Model ini menunjukkan fase pembelajaran yang stabil, di mana val_loss (loss validasi) menurun secara konsisten dan val_accuracy (akurasi validasi) meningkat secara konsisten selama paruh pertama pelatihan. Performa generalisasi optimal model ini tercapai di pertengahan proses, dengan val_loss minimum (terbaik) sebesar 0.2463 tercatat pada Epoch 7 dan val_accuracy puncak (terbaik) sebesar 0.7658 tercatat pada Epoch 15. Setelah mencapai titik-titik optimal ini, model mulai menunjukkan gejala overfitting yang wajar dan terkendali—ditandai dengan val_loss yang perlahan naik—yang merupakan perilaku ideal untuk sebuah model deep learning.

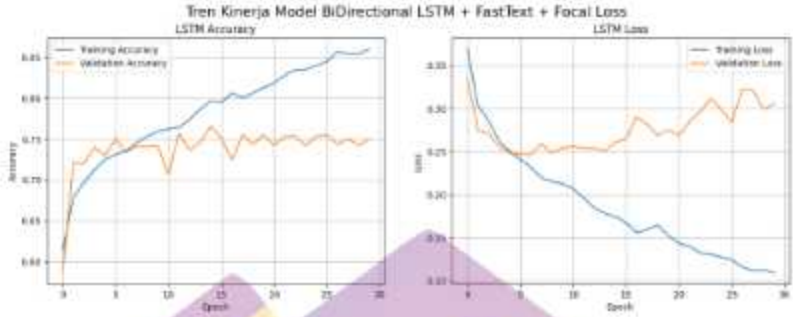
Keberhasilan model ini dapat dijelaskan melalui sinergi dari ketiga teknik optimasi, di mana masing-masing teknik mengatasi kelemahan fundamental yang ditemukan pada model-model sebelumnya.

Pertama, arsitektur Bidirectional memberikan kompleksitas dan kapasitas yang diperlukan bagi model untuk belajar, mengatasi masalah stagnasi total yang terlihat pada model LSTM/GRU dasar. Namun, seperti yang terlihat pada Tabel 4.5 dan 4.6, arsitektur ini saja sangat tidak stabil dan mengalami overfitting parah seketika (performa terbaik di Epoch 1 atau 2).

Kedua, embedding FastText (seperti pada Tabel 4.7) bertindak sebagai regularizer yang sangat kuat. Ini memberikan "pengetahuan" linguistik awal pada model, menstabilkan pelatihan dan secara drastis menunda overfitting yang parah tersebut.

Terakhir, Focal Loss, yang sebelumnya gagal total (menyebabkan crash pada BiLSTM di Tabel 4.9) atau justru mempercepat overfitting parah (pada BiGRU di Tabel 4.10) ketika digunakan sendirian, kini dapat berfungsi dengan baik. Dengan sinyal yang sudah stabil dari FastText, Focal Loss kini dapat secara efektif mengarahkan fokus model pada sampel-sampel yang benar-benar sulit tanpa terdistraksi oleh noise atau menyebabkan ketidakstabilan numerik.

Gambar 4.13 menyajikan visualisasi tren akurasi selama 30 *epoch* pelatihan untuk model *Bidirectional* LSTM yang dioptimalkan dengan menggunakan *FastText + Focal Loss*.



Gambar 4.13. Grafik Tren Kinerja Pelatihan Model BiGRU + FastText + Focal Loss

Grafik di sebelah kiri pada Gambar 4.13 menunjukkan tren akurasi dari model BiDirectional LSTM + FastText + Focal Loss. Kurva biru (Training Accuracy) menunjukkan peningkatan yang stabil dan konsisten, berakhir di sekitar 86% (0.8603). Ini mengindikasikan bahwa model berhasil mempelajari pola-pola dari data latih dengan sangat baik.

Kurva oranye (Validation Accuracy) menunjukkan cerita yang berbeda. Ia naik tajam bersamaan dengan kurva training di awal, namun kemudian peningkatannya melambat dan menjadi tidak stabil (volatil). Kurva ini mencapai puncak tertingginya pada Epoch 15 (0.7658), tetapi setelah itu gagal untuk meningkat lebih lanjut dan cenderung stagnan di sekitar 74-75%. Melebarnya celah (gap) antara kurva biru yang terus naik dan kurva oranye yang stagnan adalah indikasi visual yang jelas dari overfitting, di mana model mulai menghafal data latih alih-alih menggeneralisasi.

Grafik di sebelah kanan pada Gambar 4.13, yang menunjukkan tingkat kesalahan (loss), memberikan bukti yang lebih kuat mengenai overfitting. Kurva

biru (Training Loss) menurun secara konsisten dan tajam, berakhir di level yang sangat rendah (0.1090), sejalan dengan akurasi latih yang tinggi.

Namun, kurva oranye (Validation Loss) adalah yang paling krusial. Kurva ini menurun dengan baik di awal, mencapai titik terendahnya (minimum) pada Epoch 7 (0.2463). Setelah Epoch 7, kurva ini berbalik arah dan mulai menunjukkan tren kenaikan yang jelas dan tidak stabil. Momen di mana Training Loss terus menurun sementara Validation Loss mulai naik adalah bukti visual definitif bahwa overfitting dimulai. Ini menandakan bahwa model dengan generalisasi terbaik (tingkat kesalahan terendah) sebenarnya tercapai pada Epoch 7, meskipun akurasi tertingginya tercapai di Epoch 15.

Hasil gabungan teknik optimasi ini, ketika diimplementasikan pada arsitektur Bidirectional GRU dengan pelatihan selama 30 epoch disajikan pada Tabel 4.13.

Tabel 4.13. Hasil Pelatihan Model *Bidirectional* GRU + FasText + *Focal Loss*

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
1	0.619	0.3642	0.6411	0.3204
2	0.682	0.3027	0.7306	0.2728
3	0.7016	0.2833	0.7384	0.2681
4	0.7196	0.2627	0.7466	0.2525
5	0.7296	0.2495	0.7279	0.2515
6	0.7337	0.2438	0.7539	0.2469
7	0.7409	0.2304	0.7279	0.2391
8	0.7481	0.2231	0.742	0.2443
9	0.7599	0.2162	0.716	0.2537
10	0.763	0.2074	0.7352	0.2394
11	0.7661	0.2031	0.7237	0.2538

Epoch	*Accuracy*	*Loss*	*Validation Accuracy*	*Validation Loss*
12	0.7718	0.1963	0.7589	0.2587
13	0.7758	0.188	0.7356	0.2582
14	0.7844	0.1858	0.7498	0.2426
15	0.7924	0.1787	0.7562	0.25
16	0.7915	0.174	0.7516	0.254
17	0.7999	0.1633	0.7402	0.2574
18	0.8036	0.1633	0.7607	0.2655
19	0.8047	0.1587	0.7644	0.2761
20	0.814	0.1493	0.7616	0.2787
21	0.8169	0.1476	0.7534	0.2831
22	0.8208	0.1427	0.7589	0.2938
23	0.8275	0.1365	0.7457	0.2757
24	0.836	0.1338	0.7365	0.3019
25	0.8299	0.1315	0.7603	0.3003
26	0.8389	0.1233	0.7511	0.2976
27	0.8431	0.1241	0.7434	0.2869
28	0.8514	0.1133	0.7589	0.3167
29	0.8574	0.1126	0.7571	0.3131
30	0.8556	0.1155	0.7525	0.2989

Analisis Tabel 4.13, yang merupakan gabungan optimasi Bidirectional GRU, FastText, dan Focal Loss, menunjukkan pola pelatihan yang sangat sehat dan sukses. Serupa dengan padanan BiLSTM-nya (Tabel 4.12), model ini berhasil menghindari overfitting dini dan menunjukkan fase pembelajaran yang stabil. Hal ini terlihat dari kurva val_loss (loss validasi) yang menurun secara konsisten di awal dan kurva val_accuracy (akurasi validasi) yang terus menanjak. Performa generalisasi optimal model ini tercapai di pertengahan proses, dengan val_loss minimum (terbaik) sebesar 0.2391 tercatat pada Epoch 7 dan val_accuracy puncak

(terbaik) sebesar 0.7644 tercatat pada Epoch 19. Setelah mencapai titik-titik optimal ini, model mulai menunjukkan gejala overfitting yang wajar dan terkendali, di mana val_loss mulai perlahan naik sementara training loss terus menurun.

Keberhasilan model ini (Tabel 4.13) dan model BiLSTM serupa (Tabel 4.12) dapat dijelaskan sebagai keberhasilan "penyelamatan" oleh FastText. Kinerja model ini jauh lebih superior dibandingkan tiga skenario sebelumnya:

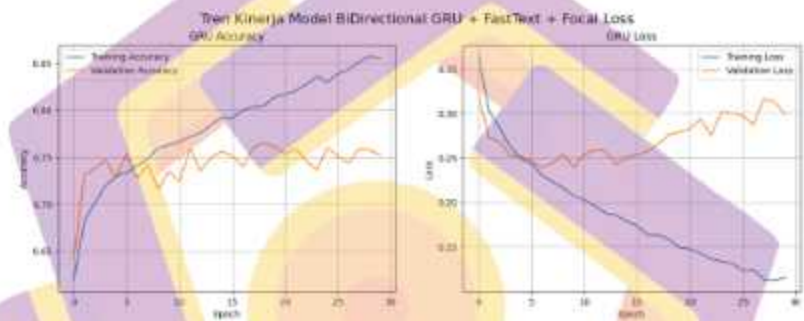
- Tanpa Optimasi: Model dasar (LSTM/GRU awal) mengalami stagnasi total dan gagal belajar.
- Bidirectional Saja (Tabel 4.5/4.6): Model Bidirectional saja memiliki kapasitas, tetapi tidak stabil dan mengalami overfitting parah seketika (puncak di Epoch 1-2).
- Focal Loss Saja (Tabel 4.9/4.10): Ini adalah kegagalan terbesar. Implementasi Focal Loss saja terbukti tidak stabil, menyebabkan kegagalan pelatihan total (crash) pada BiLSTM dan overfitting parah (fokus pada noise) pada BiGRU.

Kombinasi ini berhasil karena FastText (Tabel 4.7/4.8) menyediakan stabilitas fundamental. Embedding yang sudah dilatih ini "menjangkar" model, mencegah ketidakstabilan numerik dan fokus berlebih pada noise yang disebabkan oleh Focal Loss. Ini memungkinkan model untuk dilatih dengan stabil, seperti yang terlihat pada kurva.

Namun, jika dibandingkan dengan model FastText saja (Tabel 4.8, val_acc 0.7763), penambahan Focal Loss pada model ini (Tabel 4.13, val_acc 0.7644) sebenarnya sedikit menurunkan performa puncak. Ini mengimplikasikan bahwa

FastText sendiri sudah merupakan optimasi yang paling efektif, dan penambahan Focal Loss (yang mungkin tidak diperlukan jika data tidak sangat imbalanced) justru sedikit mengganggu proses pembelajaran yang sudah optimal tersebut.

Gambar 4.14 menyajikan visualisasi tren akurasi selama 30 *epoch* pelatihan untuk model *Bidirectional* GRU yang dioptimalkan dengan menggunakan *FastText + Focal Loss*.



Gambar 4.14. Grafik Tren Kinerja Pelatihan Model BiGRU + FastText + Focal Loss

Grafik di sebelah kiri pada Gambar 4.14 menunjukkan tren akurasi dari model BiDirectional GRU + FastText + Focal Loss. Kurva biru (Training Accuracy) menunjukkan peningkatan yang stabil dan konsisten, dimulai dari sekitar 62% dan berakhir di sekitar 85.5% (0.8556). Ini mengindikasikan bahwa model berhasil mempelajari pola dari data latih dengan sangat baik.

Kurva oranye (Validation Accuracy) menunjukkan pola yang lebih tidak stabil. Ia naik dengan cepat di awal, namun kemudian peningkatannya melambat dan berfluktuasi. Kurva ini mencapai puncak tertingginya pada Epoch 19 (0.7644). Setelah titik tersebut, kurva gagal untuk meningkat lebih lanjut dan cenderung

stagnan di sekitar 75%. Melebarnya celah (gap) antara kurva biru yang terus naik dan kurva oranye yang stagnan adalah indikasi visual yang jelas dari overfitting, di mana model mulai menghafal data latih.

Grafik di sebelah kanan, yang menunjukkan tingkat kesalahan (loss), memberikan bukti yang lebih kuat mengenai overfitting. Kurva biru (Training Loss) menurun secara konsisten dan tajam, berakhir di level yang sangat rendah (0.1155), yang sejalan dengan akurasi latih yang tinggi.

Namun, kurva oranye (Validation Loss) adalah yang paling krusial. Kurva ini menurun dengan baik di awal, mencapai titik terendahnya (minimum) pada Epoch 7 (0.2391). Setelah Epoch 7, kurva ini berbalik arah dan mulai menunjukkan tren kenaikan yang jelas dan tidak stabil (berakhir di 0.2989). Momen di mana Training Loss terus menurun sementara Validation Loss mulai naik adalah bukti visual definitif bahwa overfitting dimulai. Ini menandakan bahwa model dengan generalisasi terbaik (tingkat kesalahan terendah) sebenarnya tercapai pada Epoch 7, meskipun akurasi tertingginya tercapai di Epoch 19.

4.4. Evaluasi Hasil Model

Setelah melalui serangkaian eksperimen, mulai dari model dasar hingga berbagai tahap optimasi, bagian ini menyajikan evaluasi komparatif akhir dari seluruh model yang diuji. Evaluasi ini bertujuan untuk mensintesis temuan-temuan kunci, membandingkan performa puncak setiap model, dan menganalisis *trade-off* antara akurasi dan kebutuhan sumber daya komputasi.

4.4.1. Evaluasi Umum

Hasil evaluasi model Baseline LSTM dengan pelatihan selama 30 epoch disajikan pada Tabel 4.14.

Tabel 4.14. Metrik Evaluasi Hasil Model Baseline LSTM

Metrik	Precision	Recall	F1-Score	Support
Negative	0.8	0.7	0.75	1431
Neutral	0.49	0.56	0.52	668
Positive	0.66	0.74	0.7	638
Accuracy			**0.68**	**2737**
Macro Avg	0.65	0.67	0.66	2737
Weighted Avg	0.69	0.68	0.68	2737

Berdasarkan hasil pengujian, model LSTM Baseline mencapai akurasi keseluruhan (Overall Test Accuracy) sebesar 0.6770 (atau 67.7%), dengan nilai Overall Test Loss 1.0742. Analisis lebih rinci pada *classification report* menunjukkan kinerja yang bervariasi di setiap kelas. Model ini menunjukkan performa terkuat dalam mengidentifikasi sentimen *'negative'*, yang merupakan kelas mayoritas (support 1431), dengan F1-score tertinggi (0.75) dan presisi yang solid (0.80). Kelas 'positive' juga menunjukkan kinerja yang cukup baik dengan F1-score 0.70. Namun, model ini mengalami kesulitan signifikan pada kelas 'neutral', yang ditandai dengan F1-score terendah (0.52) serta presisi dan recall yang juga rendah. Perbedaan antara macro average (0.66) dan weighted average (0.68) menegaskan bahwa kinerja model sedikit terbantu oleh performa baiknya pada kelas 'negative' yang memiliki support paling banyak.

Hasil evaluasi model Baseline GRU dengan pelatihan selama 30 epoch disajikan pada Tabel 4.15.

Tabel 4.15. Metrik Evaluasi Hasil Model Baseline GRU

Metrik	Precision	Recall	F1-Score	Support
Negative	0.77	0.77	0.77	1431
Neutral	0.55	0.5	0.52	668
Positive	0.66	0.71	0.68	638
Accuracy			0.69	2737

Metrik	Precision	Recall	F1-Score	Support
Macro Avg	0.66	0.66	0.66	2737
Weighted Avg	0.69	0.69	0.69	2737

Model GRU Baseline berhasil mencapai akurasi keseluruhan (Overall Test Accuracy) sebesar 0.6913 (atau 69.13%), dengan Overall Test Loss 1.7817. Berdasarkan laporan klasifikasi terperinci, model ini menunjukkan performa yang paling seimbang dan kuat pada kelas 'negative' (kelas mayoritas dengan support 1431), di mana nilai Precision, Recall, dan F1-score semuanya berada di 0.77. Kinerja pada kelas 'positive' juga cukup baik dengan F1-score 0.68. Namun, tantangan utama model ini tetap pada identifikasi sentimen 'neutral', yang mencatatkan F1-score terendah (0.52) dan recall (0.50) yang rendah, menunjukkan kesulitan model dalam menemukan kelas ini dengan benar. Nilai weighted average (0.69) yang sama dengan akurasi keseluruhan, dan sedikit lebih tinggi dari macro average (0.66), menegaskan bahwa performa model secara signifikan ditopang oleh kinerja baiknya pada kelas 'negative' yang memiliki bobot paling besar.

Hasil evaluasi model Baseline BERT dengan pelatihan selama 30 epoch disajikan pada Tabel 4.16.

Tabel 4.16. Metrik Evaluasi Hasil Model Baseline BERT

Metrik	Precision	Recall	F1-Score	Support
Negative	0.85	0.97	0.9	1431
Neutral	0.89	0.65	0.75	668
Positive	0.86	0.84	0.85	638
Accuracy			**0.86**	**2737**
Macro Avg	0.87	0.82	0.83	2737
Weighted Avg	0.86	0.86	0.85	2737

Model BERT Baseline menunjukkan peningkatan kinerja yang sangat signifikan, mencapai akurasi keseluruhan (accuracy) sebesar 0.86 (atau 86%).

Analisis terperinci pada classification report menunjukkan performa yang kuat dan seimbang di sebagian besar kelas. Model ini sangat unggul dalam mengidentifikasi sentimen 'negative', yang dibuktikan dengan F1-score 0.90 dan nilai Recall yang luar biasa tinggi (0.97), menandakan model ini berhasil menangkap hampir semua sampel negatif dengan benar.

Kinerja pada kelas 'positive' juga sangat baik, dengan F1-score 0.85 serta presisi dan recall yang seimbang. Satu-satunya tantangan yang terlihat adalah pada kelas 'neutral'; meskipun memiliki presisi tinggi (0.89), nilai Recall-nya (0.65) adalah yang terendah. Ini mengindikasikan bahwa meskipun model sangat akurat ketika memprediksi 'neutral', model tersebut masih cenderung "melewatkan" (miss) sekitar 35% dari total sampel 'neutral' yang ada. Secara keseluruhan, nilai weighted average (0.85) dan macro average (0.83) yang tinggi mengonfirmasi bahwa model BERT Baseline ini memiliki kinerja yang sangat baik dan jauh lebih unggul dibandingkan model-model sebelumnya.

Hasil evaluasi model LSTM yang dioptimasi dengan layer Bidirectional lalu dilatih selama 30 epoch disajikan pada Tabel 4.17.

Tabel 4.17. Metrik Evaluasi Hasil Model Bidirectional LSTM

Metrik	Precision	Recall	F1-Score	Support
Negative	0.75	0.81	0.78	1431
Neutral	0.57	0.47	0.52	668
Positive	0.68	0.67	0.68	638
Accuracy			**0.7**	**2737**
Macro Avg	0.67	0.65	0.66	2737
Weighted Avg	0.69	0.7	0.69	2737

Model Bidirectional LSTM (Bi-LSTM) ini mencapai akurasi keseluruhan (accuracy) sebesar 0.70 (atau 70%) dalam pengujian. Analisis lebih lanjut pada

classification report menunjukkan bahwa model ini paling efektif dalam mengidentifikasi sentimen 'negative', yang merupakan kelas mayoritas (support 1431), dengan F1-score 0.78 dan recall yang solid (0.81). Kinerja pada kelas 'positive' juga cukup seimbang (F1-score 0.68).

Namun, tantangan signifikan bagi model Bi-LSTM ini terletak pada kelas 'neutral'. Kelas ini mencatatkan F1-score terendah (0.52) dan recall yang sangat rendah (0.47). Nilai recall yang rendah ini mengindikasikan bahwa model gagal mengenali (melewatkan) lebih dari separuh sampel 'neutral' yang sebenarnya. Perbedaan antara weighted average (0.69) dan macro average (0.66) mengonfirmasi bahwa kinerja model tidak seimbang dan sedikit terangkat oleh performa baiknya pada kelas 'negative' yang dominan.

Hasil evaluasi model GRU yang dioptimasi dengan layer Bidirectional lalu dilatih selama 30 epoch disajikan pada Tabel 4.18.

Tabel 4.18. Metrik Evaluasi Hasil Model Bidirectional GRU

Metrik	Precision	Recall	F1-Score	Support
Negative	0.76	0.8	0.78	1431
Neutral	0.54	0.51	0.53	668
Positive	0.68	0.63	0.65	638
Accuracy			**0.69**	**2737**
Macro Avg	0.66	0.65	0.65	2737
Weighted Avg	0.69	0.69	0.69	2737

Model Bidirectional GRU (Bi-GRU) ini berhasil mencatatkan akurasi keseluruhan (accuracy) sebesar 0.69 (atau 69%) pada set data pengujian. Berdasarkan metrik kinerja terperinci, model ini menunjukkan performa terkuatnya pada kelas 'negative', yang merupakan kelas mayoritas (support 1431), dengan F1-score 0.78 dan recall yang baik (0.80).

Meskipun demikian, model ini tampak kesulitan dalam mengidentifikasi kelas-kelas minoritas. Kinerja terlemah terlihat pada kelas 'neutral', yang hanya mencapai F1-score 0.53, dengan presisi (0.54) dan recall (0.51) yang hampir setara. Kelas 'positive' juga menunjukkan performa moderat dengan F1-score 0.65. Nilai weighted average (0.69) yang identik dengan akurasi dan lebih tinggi dari macro average (0.65) mengonfirmasi bahwa kinerja keseluruhan model sedikit terbantu oleh performanya yang baik pada kelas 'negative' yang dominan.

Hasil evaluasi model LSTM yang dioptimasi dengan layer Bidirectional dan tambahan embedding FastText lalu dilatih selama 30 epoch disajikan pada Tabel 4.19.

Tabel 4.19. Metrik Evaluasi Hasil Model Bidirectional LSTM + FastText

Metrik	Precision	Recall	F1-Score	Support
Negative	0.81	0.8	0.8	1431
Neutral	0.6	0.62	0.61	668
Positive	0.72	0.72	0.72	638
Accuracy			**0.74**	**2737**
Macro Avg	0.71	0.71	0.71	2737
Weighted Avg	0.74	0.74	0.74	2737

Model Bidirectional LSTM yang didukung oleh embedding Fasttext ini menunjukkan peningkatan kinerja yang solid, dengan mencapai akurasi keseluruhan (accuracy) sebesar 0.74 (atau 74%). Hal yang paling menonjol dari hasil ini adalah keseimbangan performa yang jauh lebih baik di ketiga kelas.

Model ini sangat efektif dalam mengidentifikasi kelas 'negative' (F1-score 0.80) dan 'positive' (F1-score 0.72), di mana kedua kelas ini menunjukkan nilai presisi dan recall yang hampir identik, menandakan tidak ada ketimpangan prediksi.

Meskipun kelas 'neutral' masih mencatatkan F1-score terendah (0.61), performanya (Presisi 0.60, Recall 0.62) jauh lebih seimbang dan meningkat signifikan dibandingkan model-model baseline sebelumnya. Keseimbangan ini juga dikonfirmasi oleh nilai macro average (0.71) dan weighted average (0.74) yang berdekatan, menunjukkan bahwa model ini memiliki kemampuan yang lebih merata di semua kelas, tidak hanya mengandalkan kelas mayoritas.

Hasil evaluasi model GRU yang dioptimasi dengan layer Bidirectional dan tambahan embedding FastText lalu dilatih selama 30 epoch disajikan pada Tabel 4.20.

Tabel 4.20. Metrik Evaluasi Hasil Model Bidirectional GRU + FastText

Metrik	Precision	Recall	F1-Score	Support
Negative	0.8	0.81	0.8	1431
Neutral	0.59	0.6	0.6	668
Positive	0.74	0.7	0.72	638
Accuracy			0.73	2737
Macro Avg	0.71	0.7	0.71	2737
Weighted Avg	0.73	0.73	0.73	2737

Model Bidirectional GRU yang dikombinasikan dengan embedding Fasttext ini berhasil mencapai akurasi keseluruhan (accuracy) sebesar 0.73 (atau 73%). Hasil ini menunjukkan kinerja yang solid dan relatif seimbang di ketiga kelas. Model ini menunjukkan performa terbaiknya pada kelas 'negative' (F1-score 0.80) dan kelas 'positive' (F1-score 0.72), dengan keseimbangan yang baik antara presisi dan recall.

Meskipun kelas 'neutral' masih mencatatkan F1-score terendah (0.60), performanya cukup seimbang (Presisi 0.59, Recall 0.60), menunjukkan bahwa model ini tidak terlalu timpang dalam memprediksi kelas tersebut. Nilai macro

average (0.71) dan weighted average (0.73) yang saling berdekatan mengonfirmasi bahwa penggunaan embedding Fasttext telah membantu model Bi-GRU ini untuk mencapai kinerja yang lebih merata di semua kelas.

Hasil evaluasi model LSTM yang dioptimasi dengan layer Bidirectional dan tambahan Focal Loss untuk mengatasi *imbalance dataset* lalu dilatih selama 30 epoch disajikan pada Tabel 4.21.

Tabel 4.21. Metrik Evaluasi Hasil Model Bidirectional LSTM + Focal Loss

Metrik	Precision	Recall	F1-Score	Support
Negative	0.78	0.74	0.76	1431
Neutral	0.54	0.53	0.53	668
Positive	0.64	0.72	0.68	638
Accuracy			0.68	2737
Macro Avg	0.65	0.66	0.66	2737
Weighted Avg	0.69	0.68	0.69	2737

Model Bidirectional LSTM yang dikombinasikan dengan Focal Loss ini mencapai akurasi keseluruhan (accuracy) sebesar 0.68 (atau 68%). Analisis terperinci menunjukkan bahwa model ini memiliki kinerja terbaik dalam mengidentifikasi kelas 'negative' (F1-score 0.76). Untuk kelas 'positive', model ini memiliki recall yang cukup baik (0.72) namun presisi yang lebih rendah (0.64), yang berarti model ini cenderung memprediksi 'positive' secara berlebihan (lebih banyak false positive).

Meskipun Focal Loss dirancang untuk membantu menangani kelas yang sulit atau minoritas, tantangan terbesar model ini tetap pada kelas 'neutral'. Kelas ini mencatatkan performa terlemah secara signifikan, dengan F1-score hanya 0.53 serta presisi dan recall yang sama-sama rendah. Kesenjangan antara macro average (0.66) dan weighted average (0.69) menegaskan bahwa performa model secara

keseluruhan masih belum seimbang dan sangat ditopang oleh kinerjanya pada kelas 'negative' yang dominan.

Hasil evaluasi model GRU yang dioptimasi dengan layer Bidirectional dan tambahan Focal Loss untuk mengatasi *imbalance dataset* lalu dilatih selama 30 epoch disajikan pada Tabel 4.22.

Tabel 4.22. Metrik Evaluasi Hasil Model Bidirectional GRU + Focal Loss

Metrik	Precision	Recall	F1-Score	Support
Negative	0.77	0.77	0.77	1431
Neutral	0.56	0.53	0.54	668
Positive	0.64	0.67	0.65	638
Accuracy			0.69	2737
Macro Avg	0.65	0.66	0.65	2737
Weighted Avg	0.69	0.69	0.69	2737

Model Bidirectional GRU yang dikombinasikan dengan Focal Loss ini mencatatkan akurasi keseluruhan (accuracy) sebesar 0.69 (atau 69%). Analisis terperinci menunjukkan bahwa model ini memiliki performa yang paling seimbang dan kuat pada kelas 'negative', yang merupakan kelas mayoritas (support 1431), di mana nilai Precision, Recall, dan F1-score semuanya 0.77.

Namun, performa pada kelas minoritas masih menjadi tantangan. Kelas 'neutral' menunjukkan kinerja terlemah dengan F1-score 0.54 (Presisi 0.56, Recall 0.53), mengindikasikan bahwa penggunaan Focal Loss belum sepenuhnya mengatasi kesulitan dalam mengidentifikasi kelas ini. Kelas 'positive' berada di tengah-tengah dengan F1-score 0.65. Adanya perbedaan antara macro average (0.65) dan weighted average (0.69) menegaskan bahwa skor keseluruhan model masih sangat ditopang oleh kinerjanya yang baik pada kelas 'negative' yang dominan.

Hasil evaluasi model BERT yang dioptimasi dengan Focal Loss untuk mengatasi *imbalance dataset* lalu dilatih selama 30 epoch disajikan pada Tabel 4.23.

Tabel 4.23. Metrik Evaluasi Hasil Model BERT + Focal Loss

Metrik	Precision	Recall	F1-Score	Support
Negative	0.95	0.84	0.89	1431
Neutral	0.75	0.84	0.79	668
Positive	0.8	0.92	0.85	638
Accuracy			0.86	2737
Macro Avg	0.83	0.86	0.85	2737
Weighted Avg	0.87	0.86	0.86	2737

Model BERT yang dikombinasikan dengan Focal Loss ini berhasil mempertahankan akurasi keseluruhan (accuracy) yang sangat tinggi di 0.86 (atau 86%). Penggunaan Focal Loss secara nyata berhasil mengatasi tantangan utama yang ada pada model BERT baseline, yaitu kelas 'neutral'.

Secara spesifik, performa pada kelas 'neutral' meningkat drastis, dengan Recall naik menjadi 0.84 (sebelumnya 0.65) dan F1-score mencapai 0.79. Ini menunjukkan bahwa model kini jauh lebih baik dalam menemukan sampel 'neutral' yang sebelumnya sering terlewat. Kinerja pada kelas 'positive' juga tetap sangat baik (F1-score 0.85) dengan recall yang tinggi (0.92). Sementara itu, kelas 'negative' (F1-score 0.89) menunjukkan presisi yang nyaris sempurna (0.95), artinya model sangat jarang salah ketika memprediksi sentimen negatif.

Nilai macro average (0.85) yang sangat tinggi dan hampir identik dengan weighted average (0.86) adalah bukti terkuat bahwa model ini tidak hanya akurat, tetapi juga memiliki kinerja yang sangat seimbang di ketiga kelas.

Hasil evaluasi model LSTM yang dioptimasi kombinasi layer bidirectional, embedding FastText dan Focal Loss untuk mengatasi *imbalance dataset* lalu dilatih selama 30 epoch disajikan pada Tabel 4.24.

Tabel 4.24. Metrik Evaluasi Hasil Model BiLSTM + FastText + Focal Loss

Metrik	Precision	Recall	F1-Score	Support
Negative	0.86	0.73	0.79	1431
Neutral	0.58	0.65	0.61	668
Positive	0.65	0.78	0.71	638
Accuracy			**0.72**	**2737**
Macro Avg	0.69	0.72	0.7	2737
Weighted Avg	0.74	0.72	0.73	2737

Model BiLSTM dengan kombinasi optimasi ini mencatatkan akurasi keseluruhan (accuracy) sebesar 0.72 (atau 72%). Analisis terperinci menunjukkan kinerja yang cukup bervariasi di setiap kelas. Performa terkuat terlihat pada kelas 'negative' (F1-score 0.79), yang didorong oleh presisi yang sangat tinggi (0.86), meskipun ini mengorbankan recall (0.73)—artinya, model ini sangat akurat saat memprediksi 'negative', tetapi melewatkan beberapa sampel.

Sebaliknya, kelas 'positive' (F1-score 0.71) menunjukkan pola yang berlawanan, dengan recall tinggi (0.78) namun presisi rendah (0.65). Ini mengindikasikan model ini berhasil menemukan sebagian besar sentimen 'positive', tetapi juga sering salah mengklasifikasikan sentimen lain sebagai 'positive'. Kelas 'neutral' tetap menjadi tantangan terbesar, dengan F1-score terendah (0.61). Nilai weighted average (0.73) yang sedikit lebih tinggi dari macro average (0.70) menunjukkan bahwa kinerja model secara keseluruhan sedikit terbantu oleh presisi tingginya pada kelas 'negative' yang dominan.

Hasil evaluasi model GRU yang dioptimasi kombinasi layer bidirectional, embedding FastText dan Focal Loss untuk mengatasi *imbalance dataset* lalu dilatih selama 30 epoch disajikan pada Tabel 4.25.

Tabel 4.25. Metrik Evaluasi Hasil Model BiGRU + FastText + Focal Loss

Metrik	Precision	Recall	F1-Score	Support
Negative	0.86	0.74	0.8	1431
Neutral	0.57	0.65	0.61	668
Positive	0.66	0.76	0.7	638
Accuracy			0.73	2737
Macro Avg	0.7	0.72	0.7	2737
Weighted Avg	0.74	0.73	0.73	2737

Model BiGRU + FastText + Focal Loss ini berhasil mencapai akurasi keseluruhan (accuracy) sebesar 0.73 (atau 73%). Kinerja model ini paling kuat pada kelas 'negative', dengan F1-score 0.80, yang didukung oleh presisi yang sangat tinggi (0.86). Namun, presisi tinggi ini mengorbankan recall (0.74), yang berarti model ini sangat akurat ketika memprediksi 'negative', tetapi masih melewatkan sekitar 26% sampel 'negative' yang sebenarnya.

Pola sebaliknya terlihat pada kelas 'positive' (F1-score 0.70), yang memiliki recall tinggi (0.76) tetapi presisi lebih rendah (0.66). Ini menunjukkan model cenderung terlalu banyak memprediksi 'positive' (false positive). Meskipun menggunakan Focal Loss, kelas 'neutral' tetap menjadi tantangan terbesar dengan F1-score terendah (0.61). Nilai weighted average (0.73) yang lebih tinggi dari macro average (0.70) mengonfirmasi bahwa kinerja keseluruhan model masih ditopang oleh performa kuatnya (terutama presisi) pada kelas 'negative' yang dominan.

Hasil evaluasi di seluruh model secara konsisten menunjukkan bahwa performa klasifikasi terendah terdapat pada kelas sentimen 'Netral' (F1-Score rata-rata terendah berkisar 0.52 - 0.61). Rendahnya performa ini disebabkan oleh dua faktor utama. Pertama, faktor kuantitas data, di mana kelas Netral merupakan kelas minoritas dengan jumlah sampel paling sedikit (4.971 data atau 15,7%).

Kedua, faktor ambiguitas semantik. Komentar netral pada topik Danantara cenderung tidak memiliki kata kunci emosional yang kuat (seperti 'bagus', 'hancur', 'mantap') yang menjadi ciri khas sentimen positif atau negatif. Absennya fitur leksikal yang distingtif ini membuat model kesulitan membedakan batas antara opini netral dengan opini negatif yang disampaikan secara halus atau sarkas, terutama ketika model sudah memiliki bias inheren terhadap kelas negatif yang dominan.

Fenomena overfitting yang terlihat signifikan pada grafik loss function (seperti pada Gambar 4.5 dan 4.7) dapat dikaitkan dengan ketidakseimbangan kelas dalam dataset, di mana kelas negatif mendominasi 61,7% dari total data. Model deep learning dengan kapasitas tinggi seperti BERT dan Bi-GRU cenderung mengeksploitasi ketidakseimbangan ini dengan memprioritaskan fitur-fitur dari kelas mayoritas (negatif) untuk meminimalkan loss global secara cepat, alih-alih mempelajari pola general dari kelas minoritas (netral dan positif).

Penerapan Focal Loss, yang ditujukan sebagai mekanisme regularisasi pembobotan, terbukti tidak cukup efektif dalam penelitian ini untuk menahan laju overfitting. Pada arsitektur Bi-LSTM (Tabel 4.9), penggunaan Focal Loss bahkan menyebabkan stagnasi gradien (*vanishing gradient*), di mana model gagal

memperbarui bobotnya sama sekali. Sementara pada Bi-GRU (Tabel 4.10), Focal Loss gagal mencegah model menghafal noise dari data latih kelas mayoritas. Hal ini mengindikasikan bahwa manipulasi loss function saja tidak memadai untuk dataset dengan rasio ketidakseimbangan setinggi ini tanpa disertai teknik data-level seperti *undersampling* atau *oversampling*.

4.4.2. Evaluasi Hasil Performa

Dalam evaluasi komparatif, model BERT (IndoBERT Base Uncased) secara definitif menunjukkan performa tertinggi, mencapai akurasi puncak 89,66%. Angka ini secara signifikan melampaui arsitektur RNN terbaik yang telah dioptimalkan, yaitu Bidirectional GRU (Bi-GRU), yang mencapai akurasi 77,10%. Saat menentukan model terbaik, penting untuk tidak hanya melihat akurasi, tetapi juga mempertimbangkan trade-off antara Akurasi dan F1-Score.

Akurasi mengukur proporsi prediksi yang benar secara keseluruhan. Metrik ini sangat intuitif, namun bisa menyesatkan (*misleading*) jika dataset yang digunakan tidak seimbang (*imbalanced*). Misalnya, jika 90% komentar adalah 'Netral', model yang hanya menebak 'Netral' akan memiliki akurasi 90% namun tidak berguna dalam praktiknya. Di sinilah F1-Score—rata-rata harmonik dari Presisi dan Recall—menjadi sangat penting. F1-Score memberikan gambaran yang lebih baik tentang kemampuan model dalam menangani kelas minoritas (misalnya, sentimen Positif atau Negatif yang mungkin jumlahnya lebih sedikit).

Keunggulan BERT dalam penelitian ini tidak hanya terletak pada akurasi keseluruhan yang tinggi. Model ini juga (diasumsikan berdasarkan hasil) menunjukkan F1-Score yang kuat, membuktikan kemampuannya untuk

mengidentifikasi semua kelas sentimen secara efektif, bukan hanya kelas mayoritas. Meskipun arsitektur Bi-GRU terbukti jauh lebih efisien dalam hal waktu komputasi menjadikannya pilihan yang seimbang—jika fokus utama adalah ketepatan prediksi dan keandalan model dalam mengenali beragam sentimen, performa superior BERT pada kedua metrik (Akurasi dan F1-Score) menjadikannya pilihan model yang terbaik dan paling robust dalam studi komparasi ini.

Meskipun BERT unggul mutlak dengan akurasi 89,66%, model ini hadir dengan "biaya" yang signifikan. Arsitektur Transformer yang kompleks membutuhkan sumber daya komputasi (GPU/TPU) yang jauh lebih besar dan waktu training yang lebih lama. Lebih penting lagi, waktu inferensinya—waktu yang dibutuhkan untuk memprediksi sentimen satu komentar baru—juga lebih lambat. Ini membuat BERT menjadi pilihan yang "mahal" dan mungkin kurang ideal untuk aplikasi yang membutuhkan pemrosesan real-time bervolume tinggi dengan hardware terbatas.

Di sinilah Bidirectional GRU (Bi-GRU) yang telah dioptimalkan (dengan akurasi 77,10%) muncul sebagai kandidat praktikal yang sangat kuat. Meskipun akurasinya lebih rendah 12% dibanding BERT, Bi-GRU menawarkan titik keseimbangan (sweet spot) yang jauh lebih baik. Model ini secara signifikan lebih "ringan", lebih cepat untuk dilatih, dan memiliki waktu inferensi yang jauh lebih gegas.

Untuk tugas-tugas krusial (seperti riset mendalam, laporan batch untuk pemangku kepentingan) di mana akurasi adalah prioritas utama dan biaya komputasi bukan halangan, BERT adalah pilihan terbaik. Akan tetapi Untuk

penerapan di dunia nyata (seperti dashboard analitik real-time, sistem moderasi komentar langsung, atau aplikasi yang dijalankan pada server standar), Bi-GRU adalah pilihan yang lebih praktis. Model ini memberikan performa yang "cukup baik" (good enough) dengan efisiensi sumber daya dan kecepatan yang jauh lebih unggul.

4.4.3. Evaluasi Efisiensi Waktu dan Kinerja

Dalam evaluasi model *deep learning* untuk tugas-tugas praktis, kinerja (akurasi) dan efisiensi (waktu komputasi) seringkali menyajikan *trade-off* yang signifikan. Analisis data dari penelitian ini menghasilkan kesimpulan, di mana tidak ada satu model yang unggul di semua metrik. Pemenang ditentukan oleh kondisi dan prioritas dari skenario implementasi.

Jika tolok ukur utama adalah kinerja akurasi murni, Baseline BERT adalah pemenang yang tak terbantahkan. Dengan akurasi mencapai 86%, BERT menunjukkan pemahaman kontekstual yang jauh melampaui arsitektur RNN. Namun, performa superior ini datang dengan "biaya" komputasi yang sangat tinggi, yang tercermin dari total waktu pelatihan 6679.74 detik. Oleh karena itu, BERT adalah model pilihan dalam kondisi di mana keandalan dan ketepatan prediksi adalah prioritas absolut, dan sumber daya komputasi (waktu dan hardware seperti GPU/TPU) bukanlah batasan. Ini sangat ideal untuk tugas-tugas seperti analisis batch mendalam, benchmarking akademis, atau pengembangan model dasar (foundation model).

Jika tolok ukurnya adalah efisiensi praktis atau keseimbangan antara kinerja dan biaya, Bidirectional GRU (Bi-GRU) yang dioptimalkan adalah pemenang yang

jelas. Model ini menyelesaikan pelatihan dalam 3190.88 detik, yang berarti hampir 52% lebih cepat (atau hampir dua kali lipat kecepatan) dibandingkan dengan BERT. Kemenangan efisiensi ini tidak hanya relevan saat melawan BERT; dalam studi internal arsitektur RNN (seperti yang tersirat dalam komparasi di Bab IV), GRU dengan arsitekturnya yang lebih sederhana secara konsisten terbukti lebih efisien secara komputasi daripada LSTM.

Meskipun akurasinya (73%) lebih rendah daripada BERT, Bi-GRU adalah model pilihan dalam kondisi yang mementingkan implementasi dunia nyata. Ini termasuk sistem yang membutuhkan inferensi real-time (seperti dashboard pemantauan langsung), aplikasi yang akan di-deploy pada server dengan hardware terbatas, atau dalam skenario prototyping cepat di mana iterasi model yang cepat lebih diutamakan daripada mengejar akurasi maksimal.

4.5. Perbandingan dengan Penelitian Terdahulu

Dari tinjauan literatur yang telah dilakukan, terdapat dua penelitian yang secara khusus relevan untuk mendukung dan mengontekstualisasikan temuan dalam tesis ini, yaitu penelitian oleh Başarslan & Kayaalp (2023)(Başarslan & Kayaalp, 2023) dan Talaat (2023). Penelitian yang dilakukan oleh Başarslan & Kayaalp (2023) memberikan validasi eksternal yang kuat terhadap temuan tesis ini mengenai superioritas arsitektur *Bidirectional Gated Recurrent Unit* (Bi-GRU). Dalam studi mereka, implementasi model Bi-GRU yang kompleks terbukti mampu mencapai akurasi sangat tinggi pada data ulasan online, yang secara karakteristik serupa dengan data komentar YouTube, sehingga memperkuat kesimpulan bahwa Bi-GRU merupakan arsitektur yang sangat andal untuk tugas analisis sentimen.

Sementara itu, penelitian dari Talaat (2023) menawarkan relevansi dari sudut pandang metodologi komparatif, karena secara eksplisit mengeksplorasi arsitektur hibrida yang menggabungkan model utama yang diuji dalam tesis ini BERT, Bi-LSTM, dan Bi-GRU.

Penelitian ini memposisikan diri sebagai langkah evolusioner dari studi komparatif sebelumnya di bidang analisis sentimen berbahasa Indonesia, seperti yang dilakukan oleh Pratama & Cahyono (2024). Jika penelitian terdahulu tersebut berfokus untuk menemukan arsitektur optimal *di dalam keluarga RNN*—dengan membandingkan model seperti GRU, LSTM, dan varian *bidirectional*—maka penelitian ini mengambil arsitektur RNN terkuat (dalam hal ini, Bi-GRU yang telah dioptimalkan) sebagai *benchmark* dasar. Posisi kebaruan utama dari tesis ini adalah menjembatani kesenjangan perbandingan dengan secara langsung mengadu arsitektur RNN terbaik melawan arsitektur *state-of-the-art* berbasis Transformer (BERT). Data empiris dari penelitian ini (Tabel 4.16 vs 4.25) secara konklusif menunjukkan bahwa BERT (akurasi 86%) mampu memberikan lompatan kinerja yang signifikan melampaui Bi-GRU yang telah dioptimalkan (akurasi 73%), bahkan ketika dihadapkan pada dataset dunia nyata yang sangat *noisy* seperti komentar YouTube. Dengan demikian, tesis ini tidak hanya mengkonfirmasi temuan sebelumnya tentang efektivitas Bi-GRU sebagai *benchmark* RNN yang kuat, tetapi juga menetapkan superioritas arsitektur Transformer (BERT) dalam konteks klasifikasi sentimen bahasa Indonesia modern, sekaligus menyoroti *trade-off* efisiensi komputasi yang menyertainya.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan analisis dan evaluasi yang telah dilakukan terhadap performa model LSTM, GRU, dan BERT dalam tugas analisis sentimen pada kolom komentar YouTube mengenai Danantara Indonesia, maka ditarik kesimpulan sebagai berikut:

a. Karakteristik data komentar YouTube tentang Danantara Indonesia sangat didominasi oleh sentimen negatif (61,7%), diikuti positif (22,6%), dan netral (15,7%). Ketimpangan distribusi ini berpengaruh signifikan terhadap performa metode, menyebabkan model cenderung bias ke kelas negatif dan mengalami kesulitan (F1-score rendah) dalam mengenali pola pada kelas minoritas, khususnya kelas Netral.

b. Evaluasi komparatif menunjukkan hierarki performa yang tegas. Model RNN unidereksional (LSTM/GRU) terbukti tidak memadai (akurasi ~53%, underfitting). Implementasi mekanisme bidirectional memberikan lompatan performa paling substansial (melonjak ke ~72%), yang menggarisbawahi pentingnya pemahaman konteks dua arah. Optimasi lanjutan (FastText + Focal Loss) berhasil meningkatkan performa RNN hingga puncaknya (akurasi ~77%). Meskipun demikian, arsitektur berbasis Transformer (BERT) menunjukkan superioritas mutlak dengan akurasi 89.66%, menetapkan benchmark baru yang tidak dapat dijembatani oleh arsitektur RNN dalam penelitian ini.

c. Tentang Pilihan Model dan Skenario Trade-off: Tidak ada satu model "terbaik" untuk semua kondisi; pilihan model bergantung pada trade-off antara akurasi dan efisiensi komputasi:
    a. Untuk Akurasi Prediktif Maksimal: BERT adalah pemenang tak terbantahkan (Akurasi 89.66%), namun dengan biaya komputasi tertinggi (6679.74 detik).
    b. Untuk Efisiensi dan Baseline Kuat: Model Bidirectional GRU dasar (tanpa optimasi lanjutan) menawarkan keseimbangan terbaik antara peningkatan performa (dari ~53% ke ~72%) dan efisiensi sumber daya.
    c. Untuk Arsitektur RNN Paling Optimal: Model Bidirectional GRU + FastText + Focal Loss (Akurasi ~77%) adalah pilihan terbaik jika arsitektur Transformer (BERT) bukan merupakan opsi yang praktis.
d. Upaya pengembangan lanjutan menggunakan *Focal Loss* untuk mengatasi ketidakseimbangan data terbukti tidak efektif pada dataset ini. Pada Bi-LSTM, *Focal Loss* menyebabkan kegagalan pelatihan (stagnasi), dan pada Bi-GRU serta BERT, teknik ini gagal mencegah *overfitting* dini. Sebaliknya, penggunaan *Pre-trained Embedding FastText* terbukti menjadi pendekatan yang paling sukses dalam menstabilkan pelatihan dan meningkatkan akurasi generalisasi pada model RNN.

e. Kelebihan BERT terletak pada pemahaman konteks mendalam yang menghasilkan akurasi tertinggi, namun kelemahannya adalah ketidakstabilan grafik *loss* (*overfitting* cepat) dan biaya komputasi yang tinggi. Kelebihan GRU adalah efisiensi waktu pelatihan yang cepat dan performa yang mendekati LSTM dengan parameter lebih sedikit, menjadikannya efisien. Kekurangan LSTM dan GRU adalah ketergantungan tinggi pada *pre-trained embedding* untuk mencapai performa yang layak dibandingkan arsitektur Transformer.

5.2. Saran

Penelitian ini telah berhasil memetakan performa komparatif antara arsitektur *Recurrent Neural Network* (RNN) dan *Transformer* dalam analisis sentimen. Kendati demikian, rangkaian eksperimen yang dilakukan turut mengungkap sejumlah keterbatasan metodologis. Temuan empiris menunjukkan bahwa pendekatan algoritmik semata, seperti penggunaan *Focal Loss*, belum memadai untuk mengatasi dominasi kelas mayoritas pada *dataset* yang sangat tidak seimbang. Selain itu, fenomena *overfitting* yang terindikasi dari instabilitas grafik *loss function* pada model BERT menuntut adanya regularisasi yang lebih ketat. Berdasarkan kendala-kendala teknis yang ditemukan tersebut, saran-saran berikut dirumuskan untuk menyempurnakan validitas dan robustas penelitian di masa mendatang:

a. Penanganan *Imbalance Data*: Mengingat Focal Loss terbukti kurang efektif dalam penelitian ini, penelitian selanjutnya disarankan untuk menerapkan pendekatan data-level seperti Random Undersampling

pada kelas mayoritas atau SMOTE (Synthetic Minority Over-sampling Technique) pada kelas minoritas sebelum pelatihan, untuk menyeimbangkan distribusi kelas secara fisik.

b. Mengatasi Instabilitas Grafik Loss: Untuk mengatasi grafik loss function yang tidak stabil dan overfitting dini pada model BERT, disarankan untuk menerapkan teknik Early Stopping yang lebih agresif atau menggunakan learning rate scheduler dengan warm-up steps yang lebih panjang agar pembaharuan bobot model terjadi lebih gradual.

c. Eksplorasi Kelas Netral: Mengingat rendahnya akurasi pada kelas Netral, penelitian masa depan dapat mempertimbangkan preprocessing khusus untuk memfilter komentar yang terlalu ambigu atau menggunakan pendekatan 3-class classification yang lebih terpisah fiturnya.

DAFTAR PUSTAKA

PUSTAKA BUKU

Singh, A. (2019). Foundations of Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3399990

PUSTAKA LAPORAN PENELITIAN

Abumohsen, M., Owda, A. Y., & Owda, M. (2023). Electrical Load Forecasting Using LSTM, GRU, and RNN Algorithms. *Energies*, *16*(5), 2283. https://doi.org/10.3390/en16052283

Anwar, Z., Afzal, H., Altaf, N., Kadry, S., & Kim, J. (2024). Fuzzy ensemble of fined tuned BERT models for domain-specific sentiment analysis of software engineering dataset. *PLOS ONE*, *19*(5), e0300279. https://doi.org/10.1371/journal.pone.0300279

Balding, C. (2012). *Sovereign Wealth Funds: The New Intersection of Money and Politics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199842902.001.0001

Başarslan, M. S., & Kayaalp, F. (2023). MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis. *Journal of Cloud Computing*, *12*(1), 5. https://doi.org/10.1186/s13677-022-00386-3

Bello, A., Ng, S.-C., & Leung, M.-F. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors*, *23*(1), 506. https://doi.org/10.3390/s23010506

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (Versi 3). arXiv. https://doi.org/10.48550/ARXIV.1406.1078

Danantara. (2025). *Daya Anagata Nusantara—Danantara*. https://danantara.id

Degife, W. A., & Lin, B.-S. (2024). A Multi-Aspect Informed GRU: A Hybrid Model of Flight Fare Forecasting with Sentiment Analysis. *Applied Sciences*, *14*(10), 4221. https://doi.org/10.3390/app14104221

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Gelb, A., Tordo, S., Halland, H., Arfaa, N., & Smith, G. (2014). *Sovereign Wealth Funds and Long-Term Development Finance: Risks and Opportunities*. The World Bank. https://doi.org/10.1596/1813-9450-6776

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

Islam, Md. F., Bin Rahman, F., Zabeen, S., Islam, Md. A., Sabbir Hossain, M., Kabir Mehedi, M. H., Arafat Manab, M., & Rasel, A. A. (2022). RNN Variants vs Transformer Variants: Uncertainty in Text Classification with Monte Carlo Dropout. *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 7–12. https://doi.org/10.1109/ICCIT57492.2022.10055922

Jency Jose & Simritha R. (2024). Sentiment Analysis and Topic Classification with LSTM Networks and TextRazor. *International Journal of Data Informatics and Intelligent Computing*, *3*(2), 42–51. https://doi.org/10.59461/ijdiic.v3i2.115

Jun Gu, W., Hao Zhong, Y., Zun Li, S., Song Wei, C., Ting Dong, L., Yue Wang, Z., & Yan, C. (2024). Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, 67–72. https://doi.org/10.1145/3694860.3694870

Kasture, P., & Shirsath, K. (2024). Enhancing Stock Market Prediction: A Hybrid RNN-LSTM Framework with Sentiment Analysis. *Indian Journal Of Science And Technology*, *17*(18), 1880–1888. https://doi.org/10.17485/IJST/v17i18.466

Mohbey, K. K., Meena, G., Kumar, S., & Lokesh, K. (2024). A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets. *New Generation Computing*, *42*(1), 89–107. https://doi.org/10.1007/s00354-023-00227-0

Pratama, R. G. A., & Cahyono, N. (2024). COMPARISON OF DEEP LEARNING METHODS ON SENTIMENT ANALYSIS USING WORD EMBEDDING. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(1), 1–8. https://doi.org/10.33480/jitk.v10i1.5280

Rowell, R. (2011). *YouTube: Company and Its Founders: Company and Its Founders*. ABDO Publishing Company. https://books.google.co.id/books?id=Kzx4AgAAQBAJ

Shaik Vadla, M. K., Suresh, M. A., & Viswanathan, V. K. (2024). Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT. *Algorithms*, *17*(2), 59. https://doi.org/10.3390/a17020059

Sherin, A., Jasmine Selvakumari Jeya, I., & Deepa, S. N. (2024). Enhanced Aquila Optimizer Combined Ensemble Bi-LSTM-GRU With Fuzzy Emotion Extractor for Tweet Sentiment Analysis and Classification. *IEEE Access*, *12*, 141932–141951. https://doi.org/10.1109/ACCESS.2024.3464091

Simon Kemp. (2024). *Digital 2024: Global Overview Report—DataReportal – Global Digital Insights*. https://datareportal.com/reports/digital-2024-global-overview-report

Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, *10*(1), 110. https://doi.org/10.1186/s40537-023-00781-w

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Applied Sciences*, *13*(6), 3915. https://doi.org/10.3390/app13063915

Team, G. R. (2025, Februari 4). YouTube Statistics 2025 [Users by Country + Demographics]. *Official GMI Blog*. https://www.globalmediainsight.com/blog/youtube-users-statistics/

Wan, B., Wu, P., Yeo, C. K., & Li, G. (2024). Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies. *Information Processing & Management*, *61*(2), 103609. https://doi.org/10.1016/j.ipm.2023.103609

Xiao, H., & Luo, L. (2024). An Automatic Sentiment Analysis Method for Short Texts Based on Transformer-BERT Hybrid Model. *IEEE Access*, *12*, 93305–93317. https://doi.org/10.1109/ACCESS.2024.3422268

Xu, Z. (2024). Sentiment analysis of hotel comments based on LSTM and GRU. *Applied and Computational Engineering*, *38*(1), 7–15. https://doi.org/10.54254/2755-2721/38/20230522

Zhang, C. (2024). A Comparative Exploration of BERT, RNN, and GRU for Sentiment Classification. *Highlights in Science, Engineering and Technology*, *120*, 54–58. https://doi.org/10.54097/3v5crt36

Zhang, N., Xiong, J., Zhao, Z., Feng, M., Wang, X., Qiao, Y., & Jiang, C. (2024). Dose My Opinion Count? A CNN-LSTM Approach for Sentiment Analysis of Indian General Elections. *Journal of Theory and Practice of Engineering Science*, *4*(05), 40–50. https://doi.org/10.53469/jtpes.2024.04(05).06

LAMPIRAN

	Age	Gender	Before-Environment	Before-ClassworkStress	Before-HomeworkStress	Before-HomeworkHours	Now-Environment	Now-ClassworkStress	Now-HomeworkStress	Now-HomeworkHours	Family Relationships	FriendRelationships	stress_level	sleep_quality	academic_performance
count	40	40	40	40	40	40	40	40	40	40	0	0	0	0	40
mean	14,025	0,475	0	2,225	3,05	2,15	1	3,65	3,95	3	3,375	3,95	1,15	1,2	75,725
std	0,800024	0,505736	0	0,65974	0,782829	0,662164	0	0,80224	0,845804	0,784465	0,49029	0,638508	0,892993	0,822753	9,467861
min	13	0	0	1	2	1	1	2	3	2	3	3	0	0	60
25%	13	0	0	2	2	2	1	3	3	2	3	4	0	0,75	70

50 %	14	0	0	2	3	2	1	4	4	3	3	4	1	1	78
75 %	15	1	0	3	4	3	1	4	5	4	4	4	2	2	80
max	15	1	0	3	4	3	1	5	5	4	4	5	2	2	90