KERJA PRAKTIK - EF234603

Disinformasi Fitnah Ujaran Kebencian - CPT & SFT

Kementerian Komunikasi dan Digital RI (KOMDIGI) & Artificial Intelligence Talent Factory (AITF)
Jl. Medan Merdeka Barat No. 9, RT.2/RW.3, Gambir, Kecamatan Gambir, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10110, Indonesia

Periode: Februari 2026 – Juni 2026

Oleh:

Daffa Rinali (5025231209)
Joe Hanzen Goenawan (5054231003)

Pembimbing Departemen
Aldinata Rizky Revanda, S.Kom., M.Kom.
Pembimbing Lapangan
Ranti Ramadhiana

DEPARTEMEN TEKNIK INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2026

KERJA PRAKTIK – EF234603

Disinformasi Fitnah Ujaran Kebencian - CPT & SFT

Kementerian Komunikasi dan Digital RI (KOMDIGI) & Artificial Intelligence Talent Factory (AITF)
Jl. Medan Merdeka Barat No. 9, RT.2/RW.3, Gambir, Kecamatan Gambir, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10110, Indonesia
Periode: Februari 2026 - Juni 2026

Oleh:

Daffa Rinali (5025231209)
Joe Hanzen Goenawan (5054231003)

Pembimbing Departemen
Aldinata Rizky Revanda, S.Kom., M.Kom.
Pembimbing Lapangan
Ranti Ramadhiana

DEPARTEMEN TEKNIK INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2026

[Halaman ini sengaja dikosongkan]

DAFTAR ISI

[Halaman ini sengaja dikosongkan]

DAFTAR GAMBAR

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

[Halaman ini sengaja dikosongkan]

LEMBAR PENGESAHAN

KERJA PRAKTIK

Disinformasi Fitnah Ujaran Kebencian - CPT & SFT

Oleh:

Daffa Rinali	(5025231209)
Joe Hanzen Goenawan	(5054231003)

Disetujui oleh Pembimbing Kerja Praktik:

1. Aldinata Rizky Revanda, S.Kom., M.Kom.
   NIP. 199806262024061002

(Pembimbing Departemen)

2. Ranti Ramadhiana
   NIP. 199403062022032010

(Pembimbing Lapangan)

[Halaman ini sengaja dikosongkan]

Disinformasi Fitnah Ujaran Kebencian - CPT & SFT

ABSTRAK

Kerja Praktik ini berfokus pada pengembangan dan adaptasi Large Language Model (LLM) yang terspesialisasi untuk tugas deteksi konten negatif berbasis teks, meliputi Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) pada media sosial berbahasa Indonesia. Pengembangan model dilakukan melalui dua tahapan utama, yaitu Continued Pre-Training (CPT) dan Supervised Fine-Tuning (SFT). Tahap CPT menggunakan korpus spesifik domain DFK sebanyak 374.930 baris data (227,4 juta token) untuk memperkaya pemahaman model dasar. Tahap SFT dilakukan untuk melatih model agar mampu mengklasifikasikan konten dan menghasilkan penalaran (reasoning) yang relevan. Eksperimen dilakukan dengan menggunakan arsitektur dasar Ministral-3-8B-Base-2512. Hasil pengujian menunjukkan bahwa tahap CPT berhasil menurunkan nilai perplexity pada domain target dari 7.7082 menjadi 5.8837 (peningkatan adaptasi 23,67%). Pada tahap SFT, model terbaik mencapai nilai akurasi dan F1-score sebesar 0,99 pada data uji, serta nilai BERTScore-F1 hingga 0,82 yang menunjukkan kualitas keselarasan penalaran dengan jawaban acuan. Model final ini kemudian di-deploy dalam bentuk layanan Application Programming Interface (API) menggunakan platform komputasi serverless, yang diintegrasikan ke dalam purwarupa Minimum Viable Product (MVP) untuk mendukung pengawasan ruang digital oleh Kementerian Komunikasi dan Digital RI.

Kata Kunci : Continued Pre-Training, Disinformasi, Fitnah, Large Language Model, Supervised Fine-Tuning, Ujaran Kebencian.

[Halaman ini sengaja dikosongkan]

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga Tim 1 DFK dapat menyelesaikan laporan Kerja Praktik berjudul "Disinformasi Fitnah Ujaran Kebencian - CPT & SFT" ini dengan baik dan tepat waktu.

Laporan ini disusun sebagai pertanggungjawaban tertulis atas proyek pengembangan Large Language Model (LLM) untuk deteksi konten negatif, hasil kolaborasi dengan Kementerian Komunikasi dan Digital RI (KOMDIGI) dan Artificial Intelligence Talent Factory (AITF).

Penyelesaian proyek ini tidak terlepas dari dukungan dan bimbingan berbagai pihak. Oleh karena itu, Tim 1 DFK mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua penulis, atas doa dan dukungan yang tak ternilai.
2. Bapak Aldinata Rizky Revanda, S.Kom., M.Kom., selaku Dosen Pembimbing Departemen atas arahan, evaluasi, dan wawasannya.
3. Ibu Ranti Ramadhiana selaku Pembimbing Lapangan dari KOMDIGI dan AITF atas kesempatan dan panduan yang diberikan.
4. Seluruh anggota Tim 1 DFK atas kolaborasi, dedikasi, dan kerja kerasnya selama masa pengembangan model.
5. Semua pihak yang telah mendukung kelancaran pelaksanaan Kerja Praktik ini.

Surabaya, 4 Juli 2026

[Halaman ini sengaja dikosongkan]

BAB I
PENDAHULUAN

1.1. Latar Belakang

Volume konten yang diproduksi pada media sosial berbahasa Indonesia telah mencapai skala yang tidak lagi mungkin ditangani secara manual oleh moderator manusia. Direktorat Jenderal Pengawasan Ruang Digital, Kementerian Komunikasi dan Digital (Komdigi), menghadapi tantangan yang semakin kompleks dalam mengawasi konten digital yang mengandung ujaran kebencian, fitnah, dan disinformasi, serta media sintetis seperti deepfake, khususnya yang menyasar tokoh pemerintahan, tokoh publik, maupun individu dengan tingkat eksposur tinggi, dan berdampak signifikan terhadap opini serta ketertiban publik. Beberapa permasalahan utama yang melatarbelakangi proyek ini adalah sebagai berikut:

1. Proses moderasi konten Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) sebagian besar masih dilakukan secara manual sehingga lambat dan sulit mengimbangi laju penyebaran konten.
2. Pendekatan berbasis kata kunci statis cepat usang menghadapi variasi bahasa tidak baku, slang, dan singkatan yang terus berkembang di media sosial.
3. Deteksi DFK menuntut pemahaman konteks, klaim, dan nada bahasa, bukan sekadar pencocokan kata kunci, sehingga membutuhkan model yang mampu bernalar (reasoning).
4. Setiap keputusan klasifikasi memerlukan alasan yang transparan agar dapat dipertanggungjawabkan dan

meningkatkan kepercayaan publik terhadap tata kelola ruang digital.

Untuk menjawab tantangan tersebut, proyek ini berfokus pada pengembangan dan adaptasi Large Language Model (LLM) yang terspesialisasi untuk mendeteksi dan mengklasifikasikan narasi DFK pada media sosial berbahasa Indonesia. Pengembangan dilakukan melalui dua tahapan utama, yaitu Continued Pre-Training (CPT) untuk memperkaya pengetahuan model dasar terhadap domain DFK berbahasa Indonesia dan Supervised Fine-Tuning (SFT) untuk melatih model mengikuti instruksi serta menghasilkan reasoning yang mendukung hasil klasifikasi. Luaran proyek diharapkan berupa model inferensi teks yang dapat diintegrasikan ke dalam sistem Agentic AI maupun purwarupa Minimum Viable Product (MVP).

Ruang lingkup proyek dibatasi pada konten berbasis teks berbahasa Indonesia, termasuk variasi bahasa tidak baku, slang, dan singkatan; metodologi teknis mencakup penyusunan dataset spesifik domain DFK yang dilanjutkan tahapan CPT dan SFT; pelabelan data diselaraskan dengan sumber tepercaya seperti pers rilis resmi kementerian/lembaga, portal berita kredibel, dan situs cek fakta; serta luaran akhir berupa model hasil fine-tuning yang diunggah ke Hugging Face dan di-deploy sebagai layanan API siap pakai.

1.2. Tujuan

Tujuan dari pelaksanaan proyek ini adalah sebagai berikut:

1. Melakukan deteksi konten DFK berbasis teks untuk mencegah penyebarannya semakin meluas atau viral.
2. Mengklasifikasikan konten dengan penyelarasan terhadap sumber tepercaya, seperti pers rilis kementerian/lembaga, portal berita, dan situs cek fakta.
3. Meningkatkan kepercayaan publik terhadap tata kelola ruang digital melalui klasifikasi yang akurat dan disertai alasan yang transparan.

1.3. Manfaat

Secara bisnis, proyek ini menyelesaikan dua tantangan utama dalam fungsi Perlindungan Ruang Digital di Komdigi. Pertama, penyaringan konten Disinformasi, Fitnah, dan Ujaran Kebencian secara cepat dan otomatis, sehingga proses pemeriksaan konten DFK yang sebelumnya dilakukan secara manual dapat dijalankan secara otomatis oleh model AI yang terintegrasi. Kedua, pemberian alasan klasifikasi yang transparan, yaitu model tidak hanya mengelompokkan konten DFK tetapi juga memberikan alasan tertulis mengapa konten tersebut dianggap melanggar. Dengan demikian, sistem ini diharapkan memangkas waktu kerja manual moderator, meningkatkan akurasi penindakan, serta memperkuat akuntabilitas tata kelola ruang digital.

1.4. Rumusan Masalah

Berdasarkan uraian pada latar belakang dan tujuan, rumusan masalah yang menjadi fokus proyek ini adalah sebagai berikut:

1. Bagaimana menyusun dataset spesifik domain DFK berbahasa Indonesia yang layak untuk melatih model deteksi konten negatif?
2. Bagaimana mengadaptasi Large Language Model melalui Continued Pre-Training (CPT) dan Supervised Fine-Tuning (SFT) agar mampu mendeteksi dan mengklasifikasikan konten DFK secara akurat dan interpretatif?
3. Bagaimana mengevaluasi performa model hasil fine-tuning dalam melakukan klasifikasi konten DFK?

1.5. Lokasi dan Waktu Kerja Praktik

Kegiatan Kerja Praktik ini dilaksanakan bersama mitra Kementerian Komunikasi dan Digital (Komdigi) melalui Pusat Pengembangan Talenta Digital, Badan Pengembangan Sumber Daya Manusia Komunikasi dan Digital (BPSDM Komdigi), yang berkantor pusat di Jalan Medan Merdeka Barat No. 9, Jakarta Pusat, 10110. Proyek dijalankan dalam kerangka program Artificial Intelligence Talent Factory (AITF), hasil kolaborasi Komdigi dengan Institut Teknologi Sepuluh Nopember (ITS).

Sehubungan dengan pola kerja kolaboratif berbasis sprint, sebagian besar pelaksanaan dilakukan secara daring (remote) dengan agenda pertemuan rutin bersama mitra, disertai beberapa sesi kerja luring bersama tim. Pelaksanaan

Kerja Praktik berlangsung pada periode Februari sampai dengan Juni 2026.

1.6. Metodologi Kerja Praktik

Pengembangan model DFK dilaksanakan secara iteratif dengan tahapan yang mencakup perumusan kebutuhan, studi literatur, analisis dan perancangan sistem, implementasi, pengujian dan evaluasi, serta penyusunan laporan pada akhir kegiatan.

1.6.1. Perumusan Kebutuhan

Tahap ini mengidentifikasi kebutuhan mitra Komdigi dalam kerangka program AITF, mencakup identifikasi permasalahan moderasi konten DFK berbahasa Indonesia serta perumusan tujuan proyek, sebagaimana diuraikan pada latar belakang dan tujuan.

1.6.2. Studi Literatur

Studi literatur dilakukan untuk mempelajari konsep dan teknologi yang digunakan, meliputi Large Language Models, studi komparatif base model (Qwen, Gemma, dan Mistral), tahapan Continued Pre-Training (CPT) dan Supervised Fine-Tuning (SFT), serta metrik evaluasi seperti perplexity dan metrik klasifikasi teks.

1.6.3. Analisis dan Perancangan Sistem

Pada tahap ini dilakukan analisis pengguna dan penyusunan user story, perancangan dataset domain DFK, serta perancangan pipeline CPT dan SFT.

1.6.4. Implementasi Sistem

Implementasi mencakup data preprocessing, pelaksanaan Continued Pre-Training dan Supervised Fine-Tuning, integrasi sistem, hingga deployment

model sebagai layanan API dan pengunggahan ke platform Hugging Face.

1.6.5. Pengujian dan Evaluasi

Pengujian dilakukan terhadap model hasil CPT dan SFT menggunakan skenario pengujian yang telah ditetapkan, dengan evaluasi berbasis perplexity untuk memantau adaptasi domain serta metrik klasifikasi (accuracy, precision, recall, dan F1-score) untuk menilai performa klasifikasi konten DFK.

1.6.6. Penyusunan Laporan

Tahap akhir berupa penyusunan laporan kerja praktik yang mendokumentasikan seluruh proses pelaksanaan, mulai dari analisis, perancangan, implementasi, hingga pengujian dan evaluasi.

1.7. Sistematika Laporan

Laporan kerja praktik ini disusun dalam tujuh bab dengan sistematika penulisan sebagai berikut:

1.7.1. Bab I Pendahuluan

Bab ini berisi latar belakang, tujuan, manfaat, rumusan masalah, lokasi dan waktu kerja praktik, metodologi kerja praktik, serta sistematika laporan.

1.7.2. Bab II Profil Perusahaan

Bab ini memaparkan profil Kementerian Komunikasi dan Digital (Komdigi) selaku mitra kerja praktik, tugas dan fungsinya, serta lokasi instansi.

1.7.3. Bab III Tinjauan Pustaka

Bab ini menguraikan dasar teori yang digunakan, meliputi Large Language Models, studi komparatif base models, dan metrik evaluasi.

1.7.4. Bab IV Analisis dan Perancangan Sistem

Bab ini menjelaskan analisis pengguna, user story, perancangan dataset, serta perancangan pipeline CPT dan SFT.

1.7.5. Bab V Implementasi Sistem

Bab ini memaparkan lingkungan pengembangan, implementasi, integrasi sistem, dan deployment model.

1.7.6. Bab VI Pengujian dan Evaluasi

Bab ini menyajikan skenario pengujian, hasil pengujian, dan evaluasi sistem.

1.7.7. Bab VII Kesimpulan dan Saran

Bab ini berisi kesimpulan dari pelaksanaan proyek serta saran untuk pengembangan selanjutnya.

[Halaman ini sengaja dikosongkan]

BAB II
PROFIL PERUSAHAAN

2.1. Profil Kementerian Komunikasi dan Digital

Kementerian Komunikasi dan Digital (Komdigi) merupakan lembaga pemerintah Republik Indonesia yang memegang otoritas penuh dalam pengelolaan, pengaturan, dan pengawasan ruang digital nasional. Dalam fungsinya, Komdigi memiliki tanggung jawab krusial untuk memastikan ekosistem siber Indonesia aman, sehat, dan bebas dari muatan negatif, khususnya konten disinformasi, fitnah, dan ujaran kebencian yang dapat mengganggu ketertiban serta keharmonisan masyarakat. Proyek ini dikerjakan dalam kerangka program AI Talent Factory (AITF) hasil kolaborasi Komdigi dengan Institut Teknologi Sepuluh Nopember (ITS). Tim kami berfokus pada pengembangan model Large Language Model (LLM) melalui proses Continue Pre-Training (CPT) dan Supervised Fine-Tuning (SFT) untuk mendeteksi serta mengklasifikasikan konten disinformasi, fitnah, dan ujaran kebencian (DFK) pada ruang digital berbahasa Indonesia.

2.2. Tugas dan Fungsi

Kementerian Komunikasi dan Digital mempunyai tugas menyelenggarakan urusan pemerintahan di bidang komunikasi dan informasi untuk membantu Presiden dalam menyelenggarakan pemerintahan negara. Dalam melaksanakan tugas tersebut, Kementerian menyelenggarakan fungsi perumusan, penetapan, dan pelaksanaan kebijakan di bidang infrastruktur digital,

teknologi pemerintah digital, ekosistem digital, pengawasan ruang digital, perlindungan data pribadi, serta komunikasi publik dan media.

Struktur organisasi Kementerian Komunikasi dan Digital terdiri atas Sekretariat Jenderal, Direktorat Jenderal Infrastruktur Digital, Direktorat Jenderal Teknologi Pemerintah Digital, Direktorat Jenderal Ekosistem Digital, Direktorat Jenderal Pengawasan Ruang Digital, Direktorat Jenderal Komunikasi Publik dan Media, Inspektorat Jenderal, serta Badan Pengembangan Sumber Daya Manusia Komunikasi dan Digital.

Dalam konteks proyek ini, peran utama Komdigi terutama bersumber pada Direktorat Jenderal Pengawasan Ruang Digital, yang memegang tanggung jawab untuk mengawasi aktivitas digital dan melindungi masyarakat dari konten berbahaya seperti disinformasi, fitnah, dan ujaran kebencian di ekosistem media sosial dan platform digital lainnya.

2.3. Lokasi

Kantor pusat Kementerian Komunikasi dan Digital (Komdigi) berlokasi di Jalan Medan Merdeka Barat No. 9, Jakarta Pusat, 10110.

BAB III
TINJAUAN PUSTAKA

3.1. Large Language Models

Large Language Models (LLM) merupakan model bahasa berbasis *deep learning* yang dilatih menggunakan korpus teks berskala besar untuk memahami pola bahasa, konteks kalimat, serta hubungan semantik antarkata. Sebagian besar LLM modern menggunakan arsitektur *Transformer* yang memanfaatkan mekanisme *attention* untuk memproses hubungan antar-token secara paralel (Vaswani *et al*., 2017).

LLM umumnya dilatih melalui proses *pre-training*, yaitu pelatihan awal pada data teks besar agar model mampu mempelajari distribusi bahasa. Pada model generatif, proses ini dilakukan dengan memprediksi token berikutnya berdasarkan konteks sebelumnya, sehingga model dapat menghasilkan teks yang koheren dan sesuai konteks (Radford *et al*., 2019).

Dalam pemrosesan bahasa alami, LLM dapat digunakan untuk berbagai tugas, seperti klasifikasi teks, tanya jawab, peringkasan, dan generasi teks. Keunggulan LLM terletak pada kemampuannya memahami konteks secara lebih luas dibandingkan pendekatan berbasis kata kunci, sehingga sesuai untuk tugas yang membutuhkan pemahaman makna dan hubungan antarpernyataan (Brown *et al*., 2020).

Pada proyek ini, LLM digunakan sebagai dasar sistem deteksi Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) berbasis teks berbahasa Indonesia. Tugas deteksi DFK membutuhkan model yang tidak hanya mampu mengenali label, tetapi juga memahami konteks, klaim, nada bahasa, serta menghasilkan alasan atau *reasoning* terhadap hasil klasifikasi (Devlin *et al*., 2019).

Meskipun *base model* telah memiliki kemampuan bahasa umum, performanya belum tentu optimal pada domain khusus seperti DFK. Oleh karena itu, dilakukan adaptasi model melalui *Continual Pre-Training* (CPT) untuk memperkaya pemahaman domain dan *Supervised Fine-Tuning* (SFT) untuk melatih model mengikuti instruksi serta menghasilkan keluaran sesuai tugas (Gururangan *et al*., 2020).

a. *Continue Pre Training*

Continued Pre-Training (CPT) merupakan sebuah tahapan dalam siklus pengembangan *Large Language Model* (LLM) di mana sebuah model dasar yang telah dilatih sebelumnya, dilatih kembali menggunakan sekumpulan data teks baru yang tidak berlabel. Secara umum, tahapan ini bertujuan untuk menggeser distribusi pengetahuan model agar lebih selaras dengan domain, bahasa, atau konteks target yang baru (Ke et al., 2023).

Melalui CPT, model secara khusus dihadapkan pada kosakata baru, jargon, hingga gaya bahasa yang khas pada suatu domain sebelum model tersebut dilatih untuk melakukan tugas akhir. Gururangan et al. (2020) dalam studinya yang menjadi landasan konsep ini membuktikan bahwa melanjutkan proses prapelatihan pada korpus data yang spesifik domain (*domain-adaptive pretraining*) sangat efektif dan krusial untuk meningkatkan pemahaman serta performa model bahasa pada domain tersebut.

b. *Supervised Finetuning*

Supervised Fine-Tuning (SFT) merupakan tahap lanjutan dalam siklus pengembangan *Large Language Model* (LLM), di mana model yang telah melalui proses pre-training atau *Continued Pre-Training* dilatih kembali menggunakan pasangan data instruksi dan respons yang berlabel (labeled instruction-response pairs). Berbeda dengan CPT yang menggunakan data tidak berlabel untuk memperkaya

pengetahuan domain, SFT bertujuan melatih model agar mampu mengikuti instruksi serta menghasilkan keluaran yang sesuai dengan format dan tugas yang diinginkan (Ouyang et al., 2022).

Melalui SFT, model diarahkan untuk menghasilkan respons yang selaras dengan kebutuhan tugas spesifik, seperti klasifikasi dan penalaran (reasoning). Wei et al. (2022) menunjukkan bahwa fine-tuning model pada kumpulan dataset instruksi yang beragam (instruction tuning) secara signifikan meningkatkan kemampuan generalisasi model terhadap tugas-tugas baru yang belum pernah dilihat sebelumnya. Pada proyek ini, tahap SFT digunakan untuk melatih model agar mampu mengklasifikasikan serta memberikan reasoning atas konten Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) berdasarkan dataset instruksi yang telah disusun.

3.2. Studi Komparatif Base Models

Dalam pengembangan sistem berbasis *Large Language Model* (LLM), pemilihan model dasar (*base model*) yang tepat sangat memengaruhi performa, efisiensi komputasi, dan kapabilitas sistem akhir. Saat ini, terdapat berbagai pilihan model berafiliasi *open-weights* terkemuka yang memiliki karakteristik dan keunggulan masing-masing. Berikut adalah perbandingan dari tiga model dasar populer, yaitu Qwen, Gemma, dan Mistral.

a. Qwen

Qwen adalah keluarga model bahasa besar yang dikembangkan oleh Alibaba Cloud. Model ini dibangun menggunakan korpus data prapelatihan berskala masif yang mencakup teks umum, kode pemrograman, dan matematika. Secara umum, keunggulan utama Qwen terletak pada

kapabilitas multibahasanya (*multilingual*) yang sangat kuat serta kemampuannya dalam memproses panjang konteks (*context length*) yang besar secara stabil. Hal ini menjadikan Qwen sebagai model dasar yang sangat fleksibel dan adaptif ketika dihadapkan pada tugas-tugas lintas bahasa di luar bahasa Inggris (Bai et al., 2023).

b. Gemma

Gemma merupakan seri model *open-weights* ringan yang dikembangkan oleh Google, dibangun menggunakan fondasi riset dan teknologi yang sama dengan model andalan mereka, Gemini. Ciri khas utama dari model Gemma adalah efisiensinya; model ini dirancang untuk memberikan performa penalaran (*reasoning*) dan pemahaman bahasa setingkat model raksasa, namun dalam ukuran parameter yang jauh lebih kecil dan ringan. Selain itu, literatur pengembangan Gemma juga sangat menekankan pada aspek keamanan (*safety*) serta keselarasan respons, menjadikannya pilihan ideal untuk aplikasi yang mengutamakan keandalan (Gemma Team, 2024).

c. Mistral

Mistral dikenal luas di komunitas pemrosesan bahasa karena keseimbangan yang sangat baik antara efisiensi komputasi dan tingginya performa yang dihasilkan. Meskipun sering kali dirilis dengan ukuran parameter yang moderat (seperti varian 7B), Mistral secara konsisten mampu mengungguli model-model lain yang berukuran jauh lebih besar. Keunggulan umum ini dicapai melalui efisiensi arsitektur yang memungkinkan model memproses teks secara cepat tanpa membebani memori perangkat, sehingga sangat cocok untuk pengembangan sistem pada sumber daya komputasi yang terbatas (Jiang et al., 2023).

3.3. Metrik Evaluasi

a. Perplexity

Perplexity (PPL) adalah metrik evaluasi intrinsik yang paling umum digunakan untuk mengukur kinerja *Language Model* (LM). Secara matematis, *perplexity* merupakan eksponensial dari *cross-entropy loss* (Jurafsky & Martin, 2023), yang mendefinisikan invers probabilitas dari urutan kata yang dinormalisasi oleh jumlah kata:

$$PP(W) = P(w_1, w_2, \ldots, w_N)^{-\frac{1}{N}}$$

PPL mengukur tingkat ketidakpastian (*surprise*) sebuah model ketika memprediksi sampel data teks baru; semakin rendah nilai *perplexity*, semakin baik model dalam memprediksi distribusi kata pada teks tersebut. Dalam konteks pemelajaran lanjutan (*Continual Pre-training*), metrik ini secara khusus digunakan untuk memantau adaptasi model terhadap domain spesifik (seperti teks Disinformasi, Fitnah, dan Kebencian) serta memastikan model tidak mengalami *catastrophic forgetting* dengan mengukur stabilitas PPL pada *dataset* pengetahuan umum.

b. Metrik Klasifikasi Teks (*Confusion Matrix*)

Metrik klasifikasi teks digunakan untuk mengukur kemampuan model dalam memprediksi label secara tepat. Pada proyek ini, metrik tersebut digunakan untuk mengevaluasi klasifikasi teks ke dalam kategori Disinformasi, Fitnah, Ujaran Kebencian, Fakta, dan Bukan DFK. Evaluasi ini umumnya didasarkan pada *confusion matrix*, yaitu tabel perbandingan antara label aktual dan label prediksi (Sokolova & Lapalme, 2009).

Confusion matrix terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Keempat komponen tersebut digunakan sebagai dasar perhitungan *accuracy*, *precision*, *recall*, dan *F1-score* dalam menilai performa klasifikasi model (Powers, 2011).

Accuracy menunjukkan proporsi prediksi benar terhadap seluruh data uji. Metrik ini memberikan gambaran umum performa model, tetapi dapat kurang representatif apabila distribusi kelas tidak seimbang. Rumus *accuracy* adalah sebagai berikut (Sokolova & Lapalme, 2009).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision mengukur ketepatan model ketika memprediksi suatu kelas. Dalam konteks DFK, *precision* penting agar model tidak terlalu sering salah menandai teks non-DFK sebagai konten bermasalah. Rumus *precision* adalah sebagai berikut (Powers, 2011).

$$Precision = TP / (TP + FP)$$

Recall mengukur kemampuan model dalam menemukan seluruh data yang seharusnya termasuk ke dalam suatu kelas. Pada deteksi DFK, recall penting agar model mampu menangkap sebanyak mungkin konten bermasalah yang benar-benar ada. Rumus recall adalah sebagai berikut (Powers, 2011).

$$Recall = TP / (TP + FN)$$

F1-score merupakan rata-rata harmonik antara precision dan recall. Metrik ini digunakan untuk melihat keseimbangan antara ketepatan prediksi dan kemampuan model menemukan data relevan, terutama ketika kesalahan false positive dan false negative sama-sama perlu diperhatikan. Rumus F1-score adalah sebagai berikut (Sokolova & Lapalme, 2009).

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Pada klasifikasi multikelas, precision, recall, dan F1-score dapat dihitung menggunakan macro average, micro average, atau weighted average. Penggunaan beberapa metrik tersebut membuat evaluasi model lebih menyeluruh dibandingkan hanya menggunakan accuracy (Sokolova & Lapalme, 2009).

c. Metrik Evaluasi Generasi Teks (Text Generation Metrics)

Untuk mengevaluasi kualitas generasi teks tersebut, digunakan dua metrik utama, yaitu ROUGE-L dan BERTScore. Kedua metrik ini dipilih karena sama-sama cocok untuk menilai kesesuaian keluaran model terhadap teks referensi, tetapi dari sudut pandang yang sedikit berbeda.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) digunakan untuk mengukur kemiripan urutan antara teks hasil generasi model dan teks referensi. Berbeda dengan pendekatan n-gram biasa, ROUGE-L memanfaatkan konsep Longest Common Subsequence (LCS), sehingga lebih toleran terhadap variasi kecil dalam susunan kata selama inti urutan informasinya masih sama. Pada konteks proyek ini, ROUGE-L bermanfaat untuk mengukur apakah model mampu menghasilkan

reasoning yang secara struktur masih selaras dengan jawaban acuan.

Sementara itu, BERTScore digunakan untuk menilai kemiripan semantik antara teks hasil generasi model dan teks referensi dengan memanfaatkan representasi kontekstual dari model transformer. Metrik ini tidak hanya melihat kecocokan kata secara literal, tetapi juga menilai apakah makna dari kalimat yang dihasilkan tetap dekat dengan makna referensi. Hal ini sangat penting untuk kasus reasoning, karena dua jawaban bisa saja memakai susunan kata yang berbeda, tapi tetap menyampaikan maksud yang sama. Dengan begitu, BERTScore dapat memberi gambaran yang lebih adil terhadap kualitas keluaran model dalam tugas generatif.

BAB IV
ANALISIS DAN PERANCANGAN SISTEM

4.1. Analisis Pengguna

Pengguna dari luaran proyek yang dikembangkan oleh tim ini dibagi menjadi dua kategori berdasarkan tingkat pemanfaatannya:

a) Direct User: Tim Pengembang Aplikasi (Tim 2 DFK) berperan sebagai pihak yang mengintegrasikan layanan API model LLM ke dalam aplikasi (Minimum Viable Product/MVP) melalui pemanggilan endpoint.

b) End-User: Kementerian Komunikasi dan Digital (KOMDIGI) khususnya tim Perlindungan Ruang Digital berperan sebagai pihak yang menggunakan fungsi deteksi teks DFK (Disinformasi, Fitnah, dan Ujaran Kebencian) secara tidak langsung melalui aplikasi MVP yang telah terintegrasi dengan API model kami.

4.2. User Story

User story dalam proyek ini dirumuskan dari sudut pandang KOMDIGI sebagai institusi pengguna:

a) Sebagai pengembang aplikasi (Tim 2 DFK), saya ingin mengakses fungsi deteksi konten DFK melalui layanan API siap pakai sehingga saya dapat mengintegrasikan kemampuan kecerdasan buatan tersebut ke dalam aplikasi MVP pesanan KOMDIGI tanpa perlu melatih model dari awal.

b) Sebagai pihak pengawas ruang digital di KOMDIGI, saya ingin sistem melalui API model yang terpasang dapat mengklasifikasikan konten berbasis teks dari media sosial secara otomatis dan memberikan alasan (reasoning) sehingga saya mendapatkan draf rujukan yang kuat untuk membantu proses penindakan konten yang mengandung unsur disinformasi, fitnah, dan ujaran kebencian.

4.3. Perancangan Dataset

4.3.1. Dataset CPT

Dataset *Continual Pre-Training* (CPT) dirancang sebagai korpus teks tidak berlabel untuk mengadaptasi *base model* terhadap domain Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) berbahasa Indonesia. Dataset ini berbentuk teks mentah, bukan instruksi atau label klasifikasi, sehingga model dapat mempelajari pola bahasa, kosakata, konteks, serta karakteristik narasi yang sering muncul pada domain DFK.

Pengumpulan dataset dilakukan melalui dua pendekatan, yaitu *scraping* data dan pemanfaatan dataset *open-source*. *Scraping* digunakan untuk memperoleh teks yang relevan dengan isu DFK, seperti politik dan pemilu, hukum dan kriminal, kesehatan publik, agama dan sosial sensitif, konflik dan keamanan, identitas kelompok, serta isu sosial lainnya. Sementara itu, dataset *open-source* dari Kaggle, Hugging Face, dan GitHub digunakan untuk melengkapi kebutuhan data DFK maupun data *general* berbahasa Indonesia.

Dataset CPT terdiri atas dua kelompok utama, yaitu data DFK dan data *general*. Data DFK digunakan untuk memperkuat pemahaman model terhadap narasi sensitif dan bermasalah, sedangkan data *general* digunakan untuk menjaga kemampuan bahasa umum model agar tidak terlalu sempit pada satu domain. Komposisi dataset ditargetkan terdiri atas 80% data DFK dan 20% data *general*, dengan

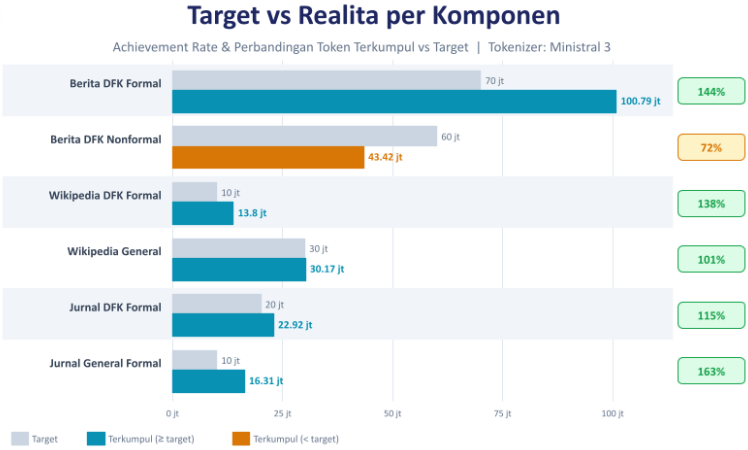
mempertimbangkan variasi bahasa formal dan nonformal.

Secara keseluruhan, dataset CPT yang terkumpul berjumlah 374.930 baris dengan total sekitar 227,4 juta token menggunakan *tokenizer* Ministral 3. Jumlah tersebut melampaui target awal 200 juta token dengan *achievement rate* sebesar 113,7%. Komposisi aktual dataset telah memenuhi target domain, yaitu 80% data DFK dan 20% data *general*. Dari sisi gaya bahasa, dataset terdiri atas 81% teks formal dan 19% teks nonformal.

Tabel 4.1 Komponen Dataset CPT

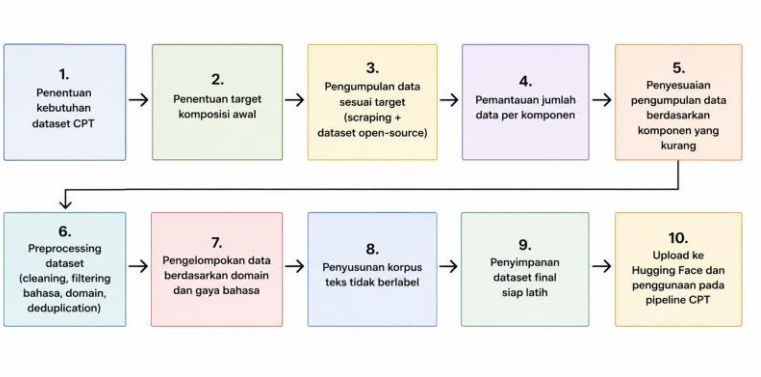
Komponen	Target	Aktual
Total token	200 juta	227,4 juta
Achievement rate	100%	113,7%
Domain DFK	80%	80%
Domain general	20%	20%
Bahasa formal	70%	81%
Bahasa nonformal	30%	19%

Visualisasi berikut menunjukkan perbandingan target dan realisasi token pada setiap komponen dataset CPT. Sebagian besar komponen telah memenuhi atau melampaui target, seperti Berita DFK Formal, Wikipedia DFK Formal, Wikipedia *General*, Jurnal DFK Formal, dan Jurnal *General* Formal. Namun, komponen Berita DFK Nonformal belum mencapai target, sehingga proporsi bahasa nonformal aktual masih lebih rendah dari rancangan awal.



Gambar 4.1 Komponen Dataset CPT

Alur perancangan dataset CPT secara umum adalah sebagai berikut.



Gambar 4.2 Flowchart Perancangan Dataset CPT

Dataset final kemudian digunakan pada *pipeline* CPT dengan pembagian data *train* dan *test* sebesar 9:1. Data *train* digunakan sebagai korpus pelatihan, sedangkan data *test*

digunakan untuk mengevaluasi kemampuan model melalui metrik *perplexity* pada domain DFK.

4.3.2 Dataset SFT

Dataset Supervised Fine-Tuning (SFT) dirancang sebagai korpus teks berinstruksi untuk melatih model dalam memahami dan merespons domain Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) berbahasa Indonesia. Berbeda dengan korpus teks mentah, dataset SFT ini disusun dalam format pasangan instruksi (prompt) dan jawaban (response) yang disertai label klasifikasi serta argumen penjelas. Pendekatan ini diterapkan agar model tidak hanya mengenali karakteristik narasi DFK, tetapi juga mampu memberikan luaran analisis yang selaras dengan instruksi spesifik yang diberikan.

Untuk keperluan fine-tuning, data teks yang digunakan terdiri dari komentar yang diklasifikasikan kedalam lima label: “Disinformasi”, “Fakta”, “Fitnah”, “Bukan DFK” dan “Ujaran Kebencian”. setiap data komentar dipasangkan dengan artikel referensi terkait. Dengan output yang disediakan meliputi label label klasifikasi serta penjelasan mengenai keselarasan atau penyimpangan informasi antara komentar dan artikel pembanding tersebut. Maksud dari artikel pembanding disini adalah artikel yang diambil dari RAG (*Retrieval-Augmented Generation*). Dengan demikian, tujuan dari format dataset ini adalah agar model dapat mempelajari pola perubahan informasi, tingkat distorsi, ataupun keselarasan klaim antara komentar dengan artikel pembandingnya.

Oleh karena itu, data test SFT tidak bisa hanya mengandalkan komentar, melainkan komentar yang digunakan harus memiliki korelasi langsung dengan sebuah artikel referensi. Strategi yang digunakan untuk mengatasi masalah tersebut adalah dengan membuat komentar secara sintetis. Dengan memanfaatkan artikel yang sudah di scraping untuk kebutuhan CPT sebagai referensinya, pembuatan komentar sintetis tersebut dilakukan dengan menggunakan model "qwen/qwen-2.5-72b-instruct" via OpenRouter. Guna menjaga konsistensi konteks, pembuatan komentar sintetis beserta anotasi luarannya (output) dilakukan dalam satu sesi inferensi yang sama.

Pada data komentar yang diklasifikasikan ke dalam label Disinformasi, artikel referensi yang digunakan sepenuhnya berbasis pada data klarifikasi resmi yang dirilis oleh Tim Internal Kementerian Komunikasi dan Digital (Komdigi). Penyertaan dokumen dari pihak otoritas ini bertujuan untuk memastikan bahwa model mempelajari narasi pembanding berdasarkan informasi penyeimbang (fact-checking) yang valid dan terpercaya. Sedangkan, khusus untuk data komentar yang diklasifikasikan sebagai ujaran kebencian (hate speech), argumentasi penjelasan model diarahkan untuk merujuk pada bentuk pelanggaran spesifik terhadap Panduan Komunitas. Panduan ini disusun ke dalam lima kategori pelanggaran utama guna memberikan batasan objektif bagi model dalam mengidentifikasi karakteristik ujaran kebencian. Rincian mengenai indikator pelanggaran,

definisi, serta kata kunci untuk masing-masing kategori disajikan pada tabel dibawah.

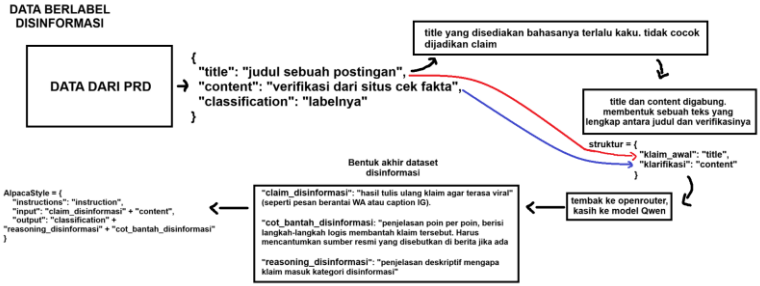
Tabel 4.2 Panduan Komunitas Ujaran Kebencian

ID	Kategori & Indikator Pelanggaran	Definisi/Isi Panduan	Kata kunci (*Keywords*)
HS_DEHUMAN	Penggunaan metafora yang menyamakan kelompok masyarakat dengan ancaman kesehatan atau kebersihan lingkungan	Segala bentuk komunikasi yang merendahkan martabat manusia dengan cara menyamakan, mengibaratkan, atau melabeli individu maupun kelompok tertentu sebagai hewan, hama, serangga, penyakit, parasit, atau benda mati yang berkonotasi kotor dan menjijikkan. Tindakan ini dilarang karena bertujuan menghilangkan sisi kemanusiaan suatu golongan sehingga memicu pembenaran atas kekerasan atau pengucilan.	Hama, parasit, virus, kuman, sampah masyarakat, kecoa, anjing, babi, tikus, kotoran, dibasmi, disemprot, dibersihkan, penyakit masyarakat, beban bumi
HS_STEREOTYPE	Penggunaan kata 'semua', 'selalu', atau penyebutan nama suku/ras yang diikuti dengan klaim perilaku negatif, watak jahat, atau	Tindakan memberikan label negatif, atribut buruk, atau tuduhan kriminal secara pukul rata (generalisasi) kepada seluruh anggota kelompok berdasarkan latar belakang suku, ras,	Suku itu licik, Ras ini pelit, memang wataknya jahat, keturunan pemberontak, genetik kriminal, tidak bisa dipercaya, sok ramah padahal busuk, orang

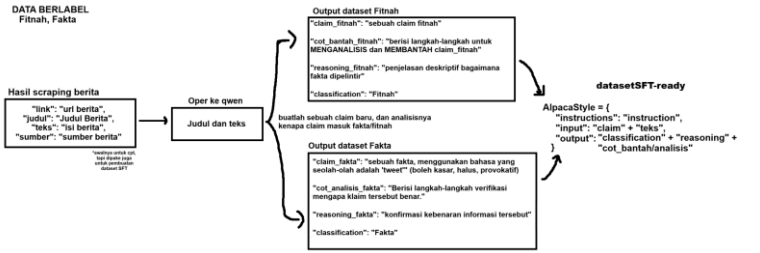
ID	Kategori & Indikator Pelanggaran	Definisi/Isi Panduan	Kata kunci (*Keywords*)
	kecenderungan kriminal kolektif.	etnis, atau agama tertentu. Pelanggaran terjadi ketika perilaku individu digunakan untuk menghakimi seluruh identitas golongan. Hal ini mencakup tuduhan sifat bawaan (genetik) yang buruk terhadap etnis tertentu.	daerah itu kasar, kaum itu memang tukang buat onar.
HS_EXCLUSION	Adanya kalimat perintah atau ajakan untuk memindahkan, mengusir, atau melarang keberadaan suatu kelompok di suatu wilayah geografis atau lingkungan sosial tertentu.	Pernyataan yang mengandung ajakan, hasutan, desakan, atau dukungan untuk melakukan pemisahan wilayah secara paksa (segregasi), pengusiran, boikot tempat tinggal, atau pembatasan hak hidup suatu golongan dari suatu lingkungan masyarakat atau negara. Tindakan ini dianggap sebagai ancaman serius terhadap integrasi sosial dan keselamatan kelompok tertentu.	Usir mereka, kembalikan ke asalnya, jangan kasih tempat tinggal, bersihkan wilayah kita, jangan mau bertetangga, boikot keberadaannya, ganyang, pindahkan paksa, jangan kasih ijin tinggal, bersihkan dari tanah kita.
HS_GENDER	Penggunaan istilah yang merendahkan harga diri	Segala bentuk penghinaan, perendahan, atau	Laki-laki banci, perempuan samsak, laki-laki semua sama

ID	Kategori & Indikator Pelanggaran	Definisi/Isi Panduan	Kata kunci (*Keywords*)
	berdasarkan peran gender, penyebutan label seksis, atau narasi yang memposisikan satu gender sebagai kelompok yang inferior/lemah.	serangan yang menargetkan identitas gender tertentu (laki-laki atau perempuan). Hal ini mencakup penggunaan kata-kata ejekan yang merendahkan maskulinitas atau feminitas, serta generalisasi negatif terhadap perilaku salah satu gender sebagai bentuk serangan kolektif.	saja, lemah, tidak berguna, objek, merendahkan martabat gender, serangan berbasis jenis kelamin.
HS_POLITICAL_SARA	Mengaitkan identitas SARA secara langsung dengan label pengkhianat negara, gerakan radikal, atau sejarah organisasi terlarang dengan tujuan menciptakan stigma berbahaya.	Tindakan memberikan label atau tuduhan keterlibatan dalam ideologi politik terlarang, kelompok teroris, atau organisasi terlarang kepada suatu kelompok etnis atau agama tanpa bukti hukum yang sah. Tuduhan ini sering digunakan untuk memicu kebencian publik dan persekusi terhadap kelompok tersebut.	Keturunan PKI, antek asing, kelompok teroris, pengkhianat negara, radikal, ekstremis, penyusup ideologi, menghancurkan negara dari dalam.

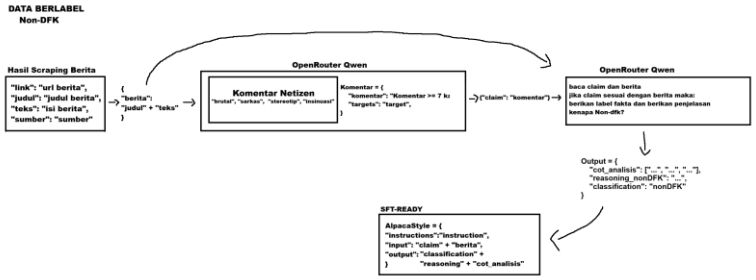
Berikut adalah flowchart pembuatan dataset SFT untuk setiap label:



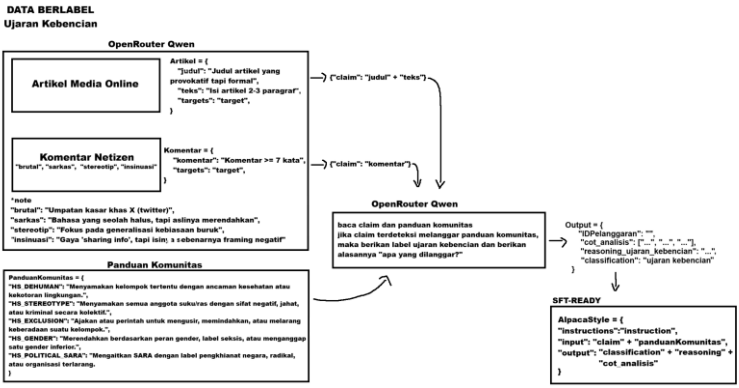
Gambar 4.3 Flowchart Pembuatan Dataset SFT Disinformasi



Gambar 4.4 Flowchart Pembuatan Dataset SFT Fitnah & Fakta



Gambar 4.5 Flowchart Pembuatan Dataset SFT Bukan DFK



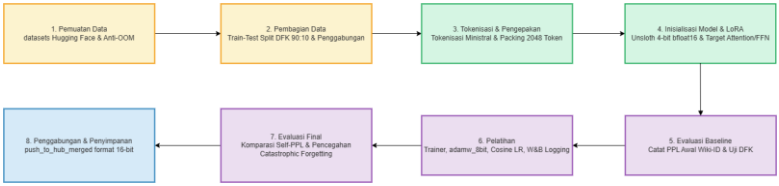
Gambar 4.6 Flowchart Pembuatan Dataset SFT Ujaran Kebencian

Distribusi data komentar perlabel, adalah sebagai berikut:

- Disinformasi : 5269 (26.60%)
- Ujaran Kebencian : 5266 (26.58%)
- Fitnah : 5172 (26.11%)
- Fakta : 5156 (26.03%)
- Bukan DFK : 5125 (25.88%)

4.4. Perancangan Pipeline CPT & SFT

a. *Flowchart* CPT



Gambar 4.7 Flowchart Pipeline CPT

Pemuatan Data: Proses awal penarikan data teks mentah menggunakan pustaka datasets Hugging Face yang dilengkapi dengan strategi Anti-OOM (Out of Memory) untuk

memastikan manajemen memori (RAM/VRAM) tetap aman saat memproses data skala besar.

Pembagian Data: Dataset khusus domain target (DFK) dipisah menggunakan metode Train-Test Split dengan rasio 90:10 (90% untuk pelatihan dan 10% untuk pengujian), yang kemudian disusun kembali dalam satu alur terintegrasi sebelum masuk ke tahap tokenisasi.

Tokenisasi & Packing: Teks mentah diubah menjadi urutan angka (token) menggunakan tokenizer Ministral, lalu digabungkan secara efisien melalui teknik Packing ke dalam panjang sekuens tetap 2048 token untuk meminimalkan ruang kosong (padding) dan mengoptimalkan kecepatan komputasi.

Inisialisasi Model & LoRA: Memuat model dasar lewat framework Unsloth dengan kuantisasi 4-bit format bfloat16 untuk menghemat VRAM, sekaligus mengaktifkan PEFT menggunakan metode LoRA (Low-Rank Adaptation) dengan target pembaruan parameter pada modul Attention dan Feed-Forward Network (FFN).

Evaluasi Baseline: Pengujian performa awal model sebelum diberikan pelatihan baru dengan mengukur dan mencatat nilai Perplexity (PPL) pada dataset umum berbahasa Indonesia (Wiki-ID) serta dataset domain spesifik target (Uji DFK) sebagai titik acuan komparasi.

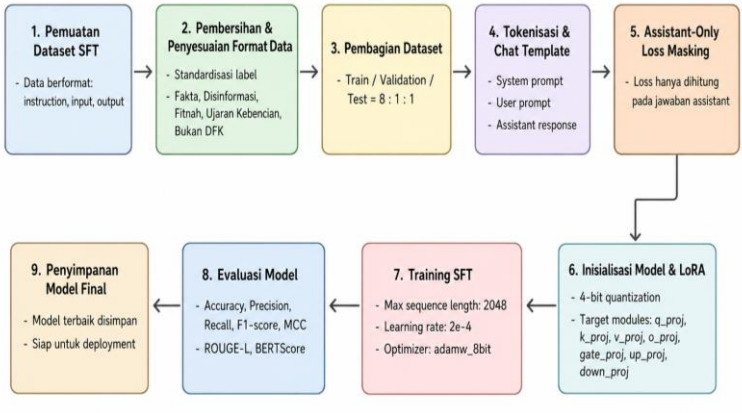
Training: Tahap pembelajaran inti menggunakan objek Trainer yang menjalankan algoritma optimasi adamw_8bit dan strategi penurunan tingkat pembelajaran Cosine LR (Cosine Learning Rate Scheduler), di mana seluruh metrik performa dipantau secara langsung melalui W&B (Weights & Biases) Logging.

Evaluasi Final: Pengujian akhir setelah proses CPT selesai untuk membandingkan nilai Self-PPL terhadap

baseline, dengan fokus utama pada Pencegahan Catastrophic Forgetting agar model berhasil menyerap keahlian domain baru (DFK) tanpa merusak kemampuan bahasa umum yang sudah dimiliki (Wiki-ID).

Penggabungan & Penyimpanan: Bobot adaptor LoRA yang telah terlatih digabungkan kembali secara permanen dengan model dasar (base model) menggunakan fungsi push_to_hub_merged, lalu diekspor ke dalam format 16-bit penuh dan diunggah ke Hugging Face Hub agar siap digunakan untuk kebutuhan produksi atau tahap post-training berikutnya.

b. *Flowchart* SFT



Gambar 4.8 Flowchart Pipeline SFT

1. Pemuatan Dataset: Proses awal untuk memuat data yang sudah terstruktur dalam format pasangan perintah dan jawaban spesifik, yang terdiri dari kolom *instruction, input,* dan *output*.

2. Pembersihan & Penyesuaian Format Data: Tahap prapemrosesan data untuk melakukan standardisasi label kategorikal ke dalam kelas-kelas spesifik yang akan dipelajari model, meliputi kategori *Fakta, Disinformasi, Fitnah, Ujaran Kebencian*, dan *Bukan DFK*.
3. Pembagian Dataset: Dataset yang telah bersih dipisahkan ke dalam tiga subset data independen menggunakan rasio proporsi 8:1:1, yaitu 80% untuk data pelatihan (*Train*), 10% untuk validasi internal model (*Validation*), dan 10% untuk pengujian performa objektif (*Test*).
4. Tokenisasi & *Chat Template*: Proses mengubah teks mentah menjadi representasi token angka yang disusun ke dalam struktur percakapan standar (*chat template*), yang secara berurutan memetakan peran instruksi sistem (*System prompt*), input pengguna (*User prompt*), dan tanggapan model (*Assistant response*).
5. *Assistant-Only Loss Masking*: Penerapan teknik penyaringan fungsi kerugian (*loss function*) di mana perhitungan nilai loss dimanipulasi secara ketat hanya pada bagian token jawaban *assistant* saja, sedangkan token instruksi (*prompt*) dari system dan user diabaikan (*masked*) agar model fokus belajar meniru jawaban target.
6. Inisialisasi Model & LoRA: Memuat model dasar menggunakan teknik kuantisasi 4-bit demi menghemat alokasi memori VRAM, sekaligus mengonfigurasi PEFT berbasis LoRA (Low-Rank Adaptation) dengan target adaptasi penuh pada seluruh modul linear utama

model (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj).

7. *Training* SFT: Tahap inti pelatihan parameter model menggunakan batasan panjang sekuens maksimal 2048 token, tingkat pembelajaran awal tetap sebesar 2e-4, dan algoritma optimasi berbasis adamw_8bit untuk menjaga stabilitas komputasi yang efisien.
8. Evaluasi Model: Pengujian performa model pasca-pelatihan dengan mengukur metrik klasifikasi standar meliputi *Accuracy, Precision, Recall, F1-score*, dan, serta metrik evaluasi generasi teks seperti *ROUGE-L* dan *BERTScore*.
9. Penyimpanan Model Final: Proses akhir untuk menyimpan bobot model dengan performa terbaik berdasarkan hasil evaluasi, yang kemudian disiapkan agar siap masuk ke tahap produksi (*deployment*).

[Halaman ini sengaja dikosongkan]

BAB V
IMPLEMENTASI SISTEM

5.1. Lingkungan Pengembangan

Implementasi dan pelatihan model dilakukan menggunakan server komputasi berbasis awan (*cloud computing*) pada platform **Google Colaboratory (Colab)** dengan unit pemrosesan grafis berspesifikasi tinggi guna menangani beban pemrosesan *Large Language Model*. Perangkat keras utama yang dialokasikan pada lingkungan virtual ini adalah GPU NVIDIA RTX PRO 6000 Blackwell Server Edition dengan kapasitas *Video RAM* (VRAM) sebesar 96 GB. Lingkungan pengembangan (*environment*) berjalan di atas sistem operasi Linux virtual bawaan platform dengan *Compute Unified Device Architecture* (CUDA) versi 12.8 dan kerangka kerja PyTorch versi 2.10.0.

Untuk mendukung efisiensi waktu dan memori pada lingkungan komputasi tersebut, sistem memanfaatkan kerangka kerja Unsloth sebagai utilitas utama percepatan *fine-tuning*. Selain itu, ekosistem perangkat lunak terintegrasi penuh dengan pustaka Hugging Face, mencakup transformers, peft (untuk metode efisiensi parameter), trl, dan datasets. Pemantauan utilisasi perangkat keras dan metrik pelatihan divisualisasikan secara langsung melalui platform *Weights & Biases* (W&B).

5.2. Implementasi

5.2.1. *Data Preprocessing* CPT

Data preprocessing CPT dilakukan untuk menyiapkan dataset mentah menjadi korpus teks siap

latih pada proses *Continual Pre-Training*. Dataset yang digunakan berasal dari hasil *scraping* dan dataset *open-source*, kemudian diproses menjadi korpus berbahasa Indonesia yang mencakup domain DFK dan *general*. Secara keseluruhan, dataset CPT berjumlah 374.930 baris dengan sekitar 227,4 juta token menggunakan *tokenizer* Ministral 3.

Tahapan *preprocessing* dataset CPT terdiri atas beberapa proses berikut.

1. *Data loading*

Dataset dimuat dari berbagai sumber data yang telah dikumpulkan, baik hasil *scraping* maupun dataset *open-source*. Tahap ini bertujuan untuk menyatukan data sebelum diproses lebih lanjut.

2. *Cleaning*

Cleaning dilakukan untuk membersihkan teks dari elemen yang tidak relevan, seperti teks kosong, karakter tidak diperlukan, simbol berlebih, atau konten yang dapat mengganggu kualitas korpus.

3. *Filtering* bahasa

Filtering bahasa dilakukan untuk memastikan bahwa data yang digunakan merupakan teks berbahasa Indonesia. Tahap ini penting karena model dikembangkan untuk memahami konteks DFK dalam bahasa Indonesia.

4. *Filtering* domain

Filtering domain dilakukan untuk memisahkan data ke dalam dua kelompok utama, yaitu data DFK dan data *general*. Data DFK mencakup topik seperti politik, hukum, kesehatan publik, agama, konflik, identitas kelompok, ekonomi-sosial, bencana, isu internasional

sensitif, dan identitas etnik. Sementara itu, data *general* mencakup topik umum seperti hiburan, kuliner, wisata, olahraga, pendidikan, geografi, sains, dan artikel non-DFK.

5. *Deduplication*

Deduplication dilakukan untuk mengurangi data yang berulang. Tahap ini bertujuan agar model tidak terlalu sering melihat teks yang sama selama proses pelatihan.

6. Pengelompokan data

Data yang telah bersih dikelompokkan berdasarkan domain dan gaya bahasa, yaitu DFK, *general*, formal, dan nonformal. Pengelompokan ini dilakukan agar komposisi dataset tetap sesuai dengan rancangan awal.

7. *Tokenization*

Tokenization dilakukan menggunakan dua *tokenizer*, yaitu Qwen *tokenizer* dan Ministral *tokenizer*. Penggunaan dua *tokenizer* diperlukan karena setiap model memiliki mekanisme pemecahan token yang berbeda, sehingga jumlah token dapat berbeda meskipun berasal dari teks yang sama.

8. Penambahan *eos_token* dan *packing*

Setiap teks ditambahkan *eos_token* sebagai penanda akhir sekuens. Setelah itu, token diproses menggunakan *packing* untuk membentuk blok dengan panjang maksimal 2048 token. Proses ini bertujuan untuk mengurangi *padding* kosong dan membuat penggunaan memori lebih efisien selama pelatihan. Konfigurasi *max length* 2048 juga digunakan pada skenario CPT.

5.2.2. Data Preprocessing SFT

A. Data Formatting & Tokenization

- Sumber Data: Dataset dimuat dari Hugging Face (Brodip/DatasetSFT_DFK_TRL).
- Format: Menggunakan Jinja Template bawaan Unsloth tanpa token <think>.

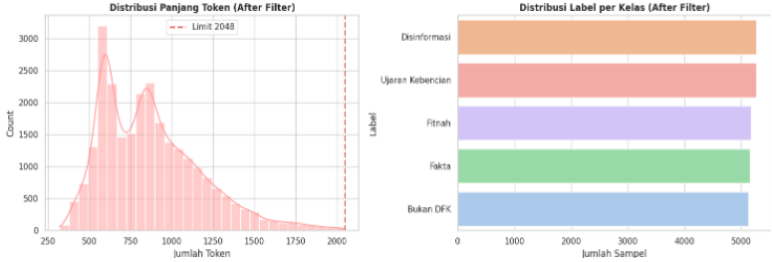
Tabel 5.1 Chat Template Jinja SFT

```
{#- Begin of sequence token. -#}
{{- bos_token }}
{#- Handle system prompt jika ada. -#}
{%- if messages[0]['role'] == 'system' %}
   {{- '[SYSTEM_PROMPT]' + messages[0]['content'] +
'[/SYSTEM_PROMPT]' }}
   {%- set loop_messages = messages[1:] %}
{%- else %}
   {%- set loop_messages = messages %}
{%- endif %}
{#- Handle iterasi untuk User dan Assistant -#}
{%- for message in loop_messages %}
   {%- if message['role'] == 'user' %}
      {{- '[INST]' + message['content'] + '[/INST]' }}
   {%- elif message['role'] == 'assistant' %}
      {{- message['content'] + eos_token }}
   {%- endif %}
{%- endfor %}
```

- Tokenisasi Awal: Mengonversi teks hasil render template percakapan menjadi urutan

nomor identifikasi (Token IDs) menggunakan tokenizer bawaan model.

- Validasi & Klasifikasi Kelas: Menerapkan deteksi berbasis pola teks (regex/string matching) pada akhir kalimat respons untuk memetakan data ke dalam empat kelas target: Netral, Disinformasi, Fitnah, atau Ujaran Kebencian. Data yang memiliki label diluar ke empat kelas target akan di drop.
- Filtering Data: Panjang token dari setiap baris data yang lolos validasi dihitung untuk memastikan tidak ada data yang melebihi batas pengaman arsitektur model (max_seq_length = 4096).
- Statistik Data Akhir: Hasil penyaringan menunjukkan dataset bersih memiliki sebaran panjang token yang sangat aman dengan nilai rata-rata (mean) sebesar 673,35 token dan panjang maksimal (max) hanya menyentuh 3.593 token. Dengan demikian, seluruh data dipastikan aman dan tidak akan mengalami pemotongan konteks (truncation) saat pelatihan.



Gambar 5.1 Distribusi Data SFT

- Volume Data Bersih: Memperoleh total 25.988 baris data siap latih setelah melalui tahap pembersihan dan penyaringan panjang token. Rasio Distribusi (80:10:10):
  - Data Latih (Training): 20.790 baris (80%)
  - Data Uji (Testing): 2.599 baris (10%)
  - Data Validasi (Validation): 2.599 baris (10%)

B. *Batching* dan *Data Collating*

Proses pengelompokan data (*batching*) dilakukan dengan dikelola secara dinamis berbasis mekanisme *On-The-Fly Data Collator*:

- Penerapan *CompletionOnlyLM:* Untuk efisiensi komputasi loss, digunakan kelas DataCollatorForCompletionOnlyLM dari library trl. Struktur pencetakan loss dikonfigurasi agar model hanya belajar memprediksi teks jawaban/respons setelah token instruksi penutup khusus [/INST], sementara area System Prompt dan User Instruction di-masking secara otomatis menggunakan nilai target token ID -100 (sehingga diabaikan dalam perhitungan gradien).
- Padding Dinamis: Batching mengelompokkan data per per_device_train_batch_size dengan gradient accumulation steps sebanyak 16. Penyelarasan panjang urutan dilakukan secara dinamis pada setiap batch (hingga batas aman max_seq_length = 4096) untuk menghemat alokasi VRAM GPU.

5.2.3 Continual Pre-training (CPT)

Proses Continued Pre-Training (CPT) pada proyek ini disusun dalam sebuah pipeline terintegrasi menggunakan framework Unsloth dan Hugging Face Transformers. Pipeline ini dirancang untuk memproses dataset teks mentah (*raw text*) berskala masif, melakukan tokenisasi dan pemadatan (*packing*), hingga melatih model secara efisien menggunakan teknik Parameter-Efficient Fine-Tuning (PEFT/LoRA) pada satu GPU.

A. Training

Proses pelatihan model dieksekusi dengan mengintegrasikan optimasi hardware tingkat tinggi dan pelacakan metrik secara real-time: Inisialisasi Trainer: Proses training dijalankan menggunakan objek Trainer standar dari Hugging Face yang telah di-patch oleh framework Unsloth agar proses backpropagation dan komputasi 16-bit LoRA berjalan hingga 2x lebih cepat.

Optimasi Memori & Hardware: Pelatihan dilakukan pada arsitektur GPU tunggal kelas enterprise (NVIDIA RTX PRO 6000) memanfaatkan tipe data Bfloat16. Fitur gradient checkpointing kustom berbasis Unsloth juga diaktifkan untuk melepaskan beban memori GPU secara cerdas (smart offloading).

Tracking & Logging: Metrik performa berupa penurunan Training Loss dan Validation Loss di-streaming secara berkala ke platform

Weights & Biases (W&B) untuk pemantauan kurva konvergensi model.

B. Output

Penyatuan dan Penyimpanan Model: Bobot LoRA adapter yang telah dilatih diekstraksi dan langsung digabungkan (*merging*) secara permanen dengan arsitektur base model menggunakan format merged_16bit pada checkpoint terbaik, lalu di-push ke Hugging Face Hub.

Export Tokenizer: Objek tokenizer diekspor secara bersamaan beserta arsitektur model final untuk memastikan konsistensi pemformatan ukuran kosa kata (*vocab size*) saat pengujian dan tahap selanjutnya.

C. Skenario dan Parameter Evaluasi

Untuk mengukur keberhasilan adaptasi domain dan memantau risiko penurunan kemampuan berbahasa umum (*catastrophic forgetting*), proses evaluasi dilakukan menggunakan dataset uji (test set) yang belum pernah dilihat oleh model selama pelatihan. Evaluasi dijalankan menggunakan perhitungan distribusi loss intrinsik (tanpa generation decoding) dengan konfigurasi sebagai berikut:

- Max Sequence Length: 2048 token (hasil packing)
- Data Collator: Default Data Collator (*unsupervised next-token prediction*)
- Padding Strategy: Group texts / Packing tanpa padding kosong

- Prompt Template Format: Murni menggunakan struktur teks mentah berurutan tanpa instruksi tanya-jawab, yang hanya dibatasi oleh token penutup sebagai berikut: [TEKS_MENTAH_ARTIKEL][TEKS_MENTAH_ARTIKEL_SELANJUTNYA]
- Metrik: Perplexity (PPL) pada domain DFK

5.2.4 Supervised Fine-Tuning (SFT)

Proses Supervised Fine-Tuning (SFT) teks pada proyek ini disusun dalam sebuah pipeline terintegrasi menggunakan framework Unsloth dan Hugging Face Transformers. Pipeline ini dirancang untuk memproses dataset teks murni, melakukan tokenisasi, hingga melatih model secara efisien menggunakan teknik Parameter-Efficient Fine-Tuning (PEFT/LoRA) pada satu GPU.

A. Training

Proses pelatihan model dieksekusi dengan mengintegrasikan optimasi hardware tingkat tinggi dan pelacakan metrik secara real-time:

- Inisialisasi Trainer: Proses training dijalankan menggunakan objek SFTTrainer dari Hugging Face yang telah dioptimalkan oleh framework Unsloth agar proses backpropagation dan komputasi 16-bit LoRA berjalan hingga 2x lebih cepat.
- Optimasi Memori & Hardware: Pelatihan dilakukan pada arsitektur GPU tunggal berkapasitas besar memanfaatkan tipe data

Bfloat16. Fitur gradient checkpointing bawaan Unsloth juga diaktifkan untuk melepaskan beban memori GPU secara cerdas (smartly offload gradients).

- Tracking & Logging: Metrik performa berupa penurunan Training Loss dan Validation Loss di-streaming secara berkala ke platform Weights & Biases (W&B) untuk pemantauan kurva konvergensi model.

B. Output

- Penyimpanan LoRA Adapters: Model yang telah dilatih tidak langsung digabungkan ke struktur base model berbobot penuh (16-bit), melainkan disimpan terlebih dahulu bobot parameter LoRA adapters secara lokal ke direktori Google Drive dan di push ke Hugging Face menggunakan metode model.save_pretrained().
- Export Tokenizer: Objek tokenizer yang membawa konfigurasi Chat Template Jinja murni tanpa komponen <think> diekspor secara bersamaan via tokenizer.save_pretrained() untuk memastikan konsistensi pemformatan teks saat pengujian.

C. Skenario dan Parameter Evaluasi:

Untuk menjaga keadilan (*fairness*) dan konsistensi saat membandingkan performa antara model *baseline* dan model hasil *Continued Pre-training*, proses evaluasi dan pengujian inferensi dilakukan menggunakan dataset uji (*test set*) yang belum pernah dilihat oleh model selama pelatihan.

Inferensi dijalankan menggunakan teknik *batching* dengan konfigurasi *generation decoding* sebagai berikut:

- Max New Tokens: 384
- Batch Size: 1
- Temperature: 0.1
- Do Sample: False
- Padding Strategy: Left Padding
- Prompt Template Format: Menggunakan struktur sekat prompt instruksi murni berbasis *no-think template* sebagai berikut:
- [SYSTEM_PROMPT]Anda adalah...[/SYSTEM_PROMPT][INST]Komentar: [Teks Komentar][/INST]
- Metrik: Accuracy, Precision, Recall, F1-Score (Macro & Weighted), BERTScore, ROUGEL

5.3. Integrasi Sistem

Integrasi sistem pada proyek ini dilakukan dengan menjadikan model hasil pelatihan sebagai layanan inferensi berbasis API yang dapat digunakan oleh sistem lain. Dalam arsitektur proyek, model yang dikembangkan oleh tim ini tidak langsung dipakai oleh pengguna akhir, melainkan diintegrasikan terlebih dahulu ke dalam aplikasi MVP yang dikerjakan oleh tim pengembang aplikasi. Dengan skema ini, model berperan sebagai komponen backend cerdas yang bertugas menerima teks masukan, memprosesnya, lalu mengembalikan hasil klasifikasi dan *reasoning* ke aplikasi.

Model final yang dipilih untuk integrasi adalah Ministral-CPT, karena hasil eksperimen menunjukkan bahwa

model ini memberikan performa terbaik pada fase SFT. Sebelum diintegrasikan, adaptor LoRA dari model tersebut diekstraksi dan dikemas ke format PEFT standar agar lebih mudah digunakan pada tahap penyajian model. Hasil model kemudian dihubungkan ke sistem aplikasi melalui endpoint HTTP berbasis REST API.

Mekanisme integrasinya cukup sederhana. Aplikasi eksternal mengirimkan permintaan ke endpoint inferensi dengan membawa teks konten yang ingin dianalisis. Selanjutnya sistem model melakukan pemrosesan, menjalankan inferensi, lalu mengembalikan hasil dalam format JSON yang berisi label kelas DFK dan penjelasan klasifikasi. Format ini memudahkan integrasi karena respons model dapat langsung dibaca dan ditampilkan oleh aplikasi lain tanpa perlu pengolahan tambahan yang rumit.

Dengan pendekatan tersebut, integrasi antara model dan sistem aplikasi menjadi lebih modular. Tim model dapat fokus pada peningkatan kualitas inferensi, sedangkan tim aplikasi bisa fokus pada antarmuka pengguna, alur kerja moderasi, dan visualisasi hasil. Pemisahan ini membuat pengembangan sistem menjadi lebih rapi dan mudah dipelihara ke depannya.

5.4. Deployment Model

Setelah model terbaik diperoleh dari rangkaian eksperimen, tahap selanjutnya adalah *deployment* agar model dapat diakses sebagai layanan inferensi yang siap digunakan. Pada proyek ini, model yang dipilih untuk *deployment* adalah Ministral-CPT, yaitu model yang telah melalui fase CPT dan SFT. Model ini dipilih karena memberikan performa terbaik dibanding model lain yang diuji, terutama dalam

menghasilkan klasifikasi yang konsisten beserta *reasoning* yang relevan.

Proses *deployment* dilakukan menggunakan platform Modal, yaitu layanan komputasi *serverless* yang mendukung eksekusi model berbasis GPU secara fleksibel. Pemilihan Modal didasarkan pada kemampuannya untuk menjalankan model besar tanpa perlu mengelola infrastruktur server secara manual. Dengan pendekatan ini, model dapat diaktifkan saat ada permintaan inferensi dan tidak harus terus berjalan sepanjang waktu, sehingga penggunaan sumber daya menjadi lebih efisien.

Pada tahap implementasi *deployment*, model dimuat ke lingkungan eksekusi bersama dependensi yang dibutuhkan, lalu diekspos melalui endpoint API berbasis FastAPI. Endpoint ini menerima masukan berupa teks dari pengguna atau sistem aplikasi, kemudian memprosesnya menggunakan model yang telah di-*deploy*, dan akhirnya mengembalikan hasil klasifikasi dalam format JSON. Hasil keluaran mencakup kelas DFK yang diprediksi beserta *reasoning* yang menjelaskan alasan klasifikasi tersebut.

Agar proses *deployment* lebih praktis, bobot adaptor hasil LoRA terlebih dahulu dipersiapkan untuk digunakan bersama model dasar pada lingkungan inferensi. Dokumentasi dan implementasi teknis terkait *deployment* ini tersedia pada repositori GitHub proyek, sehingga proses reproduksi dan pengembangan lanjutan dapat dilakukan dengan lebih mudah. Dengan demikian, model yang sebelumnya hanya berjalan pada tahap eksperimen kini sudah dapat dimanfaatkan sebagai layanan inferensi nyata untuk mendukung aplikasi moderasi konten berbasis AI.

[Halaman ini sengaja dikosongkan]

BAB VI
PENGUJIAN DAN EVALUASI

6.1. Skenario Pengujian

6.1.1. Continue Pre-Training

Pada tahap ini dilakukan pengujian perbandingan terhadap beberapa base model LLM *open-source* untuk menentukan model terbaik yang akan digunakan pada fase *Continue Pre-Training* dengan dataset skala besar. Pengujian dilakukan menggunakan dataset spesifik berdomain Disinformasi, Fitnah, dan Ujaran Kebencian (DFK) sebanyak 44.000 baris (~20 juta token) dengan proporsi pembagian data latih dan data uji sebesar 9:1. Proses pelatihan dan evaluasi model dieksekusi dengan spesifikasi konfigurasi parameter pelatihan sebagai berikut:

Tabel 6.1 Konfigurasi Parameter Benchmark Base Model

Max Sequence Length	2048
Epoch	3
Learning Rate	5e-5
Optimizer	adamw_8bit
Batch Size	8
LoRA R	16
LoRA Alpha	32
Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

Setelah pengujian perbandingan selesai dilakukan dan base model terbaik telah ditentukan,

proses dilanjutkan ke tahap pelatihan penuh menggunakan keseluruhan dataset (full dataset) berdomain DFK untuk fase Continued Pre-Training (CPT). Proporsi dataset yang digunakan yaitu 9:1, yang mana data train terdiri dari 70% domain DFK dan 20% domain umum sedangkan data test menggunakan 10% domain DFK.

Proses pelatihan full dataset ini dieksekusi menggunakan konfigurasi parameter optimal yang telah diuji pada tahap benchmark, dengan rincian parameter sebagai berikut:

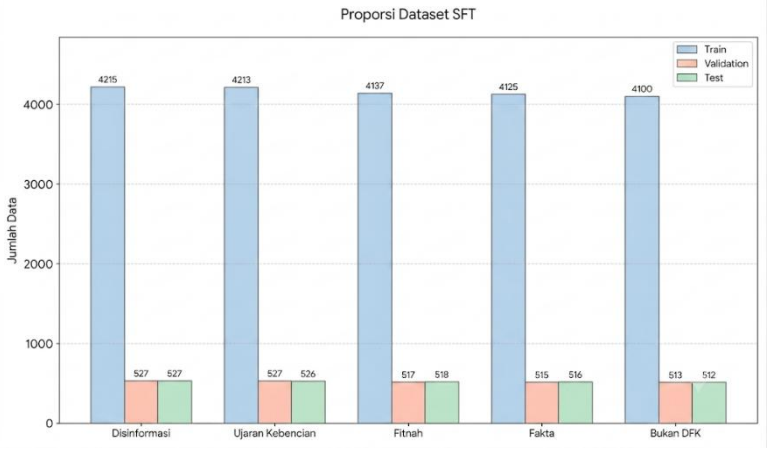
Tabel 6.2 Konfigurasi Paramater CPT Full Dataset

Max Sequence Length	2048
Epoch	3
Learning Rate	5e-5
Optimizer	adamw_8bit
Batch Size	16
LoRA R	16
LoRA Alpha	32
Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

6.1.2. Supervised Finetuning

Pada tahap ini dilakukan proses *Supervised Fine-Tuning* (SFT) sebagai bentuk pengujian perbandingan antara base model dan model hasil *Continued Pre-Training* (CPT). Kedua model tersebut dilatih menggunakan proporsi pembagian data serta konfigurasi parameter yang sama untuk mengevaluasi secara objektif sejauh mana fase

CPT berkontribusi terhadap peningkatan performa model dalam mengklasifikasikan konten DFK teks berbahasa Indonesia sekaligus memberikan alasan.



Gambar 6.1 Proporsi Dataset SFT

Eksperimen SFT ini menggunakan dataset format TRL dengan proporsi pembagian data train, test, validation sebesar 8:1:1 dengan total keseluruhan data sebanyak 25.988 dan konfigurasi parameter sebagai berikut:

Parameter	Nilai
Max Sequence Length	2048 token
Gradient Accumulation Steps	16
Batch Size	2
Learning Rate	2e-4
Number of Epochs	1
LR Scheduler	Linear
Optimizer	AdamW 8-bit
Eval Steps	Setiap 200 steps
Weight Decay	0.01
Max Grad Norm	1.0
Warmup Step	20

Tabel 6.3 Konfigurasi Parameter SFT

Tabel 6.4 Konfigurasi Hyperparameter Lora SFT

Parameter	Nilai
r (Rank)	16
lora_alpha	32
lora_dropout	0
bias	none
use_rslora	False
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, down_proj, up_proj
Gradient Checkpointing	unsloth

6.2. Hasil Pengujian

6.2.1 Continue Pre Training

Pengujian performa pada tahap *Continual Pre-training* (CPT) diawali dengan melakukan eksperimen komparatif terhadap beberapa arsitektur *base model* untuk mengidentifikasi model dengan kinerja dan efisiensi sumber daya paling optimal. Evaluasi awal ini mengukur *Training Loss*, *Evaluation Loss*, *Perplexity* (PPL) secara umum, serta

konsumsi komputasi (*Runtime* dan *VRAM Peak*). Hasil komparasi dari keenam model yang diuji dapat dilihat pada Tabel berikut.

Tabel 6.5 Hasil Benchmark Model Base

Model	Train Loss	Eval Loss	PPL	Runtime	VRAM Peak
Qwen3.5-9B-Base	2.07	2.02	7.57	~12.8 Jam	31.88 GB
Qwen3.5-4B-Base	2.22	2.17	8.74	~11.99 Jam	20.99 GB
Gemma-3-4b-pt	1.90	1.84	6.31	~9.08 Jam	13.40 GB
Gemma-4-E4B	1.85	1.78	5.92	~5.93 Jam	32.56 GB
Ministral-3-3B-Base-2512	1.83	1.79	5.98	~4.33 Jam	13.40 GB
Ministral-3-8B-Base-2512	1.68	1.64	5.15	~7.5 Jam	20.95 GB

Berdasarkan Tabel tersebut, arsitektur Ministral-3-8B-Base-2512 menunjukkan performa paling superior dengan mencatatkan *Eval Loss* terendah (1.64) dan skor *Perplexity* evaluasi pelatihan terbaik di angka 5.15. Meskipun ukuran parameternya lebih besar (8 miliar parameter), model ini terbukti sangat efisien

berkat optimalisasi kuantisasi, hanya mengonsumsi *VRAM Peak* sebesar 20.95 GB dengan waktu pelatihan sekitar 7.5 jam. Sebagai perbandingan, keluarga model Qwen membutuhkan waktu pelatihan terlama (hingga 12.8 jam) dan menghasilkan *Perplexity* yang cukup tinggi (7.57 - 8.74). Di sisi lain, meskipun Ministral-3-3B dan Gemma-4-E4B menawarkan performa yang kompetitif dengan *runtime* yang lebih singkat, Ministral-3-8B tetap dipilih sebagai model final karena memberikan representasi bahasa yang paling akurat (PPL terendah) dengan rasio efisiensi yang rasional.

Setelah Ministral-3-8B-Base-2512 ditetapkan sebagai model utama, pengujian mendalam dilakukan melalui komparasi metrik *Perplexity* antara *baseline* (sebelum pelatihan CPT) dan evaluasi final (setelah pelatihan). Pengukuran ini memanfaatkan *subset* pengujian khusus: sampel murni domain spesifik (Self-PPL) untuk mengukur tingkat serapan ilmu baru.

Proses pembaruan parameter bobot (Continual Pre-training) pada Ministral-3-8B-Base-2512 dieksekusi secara intensif menggunakan total akumulasi data bersih sebesar 374.930 baris. Volume korpus ini mencakup ekspansi hingga menyentuh 227,4 juta token, sukses melampaui target rancangan awal sebesar 200 juta token dengan tingkat pencapaian (*achievement rate*) mencapai 113,7%. Pelatihan ini berjalan di atas infrastruktur komputasi GPU NVIDIA RTX PRO 6000 Blackwell Server Edition dengan kapasitas VRAM 96 GB, menggunakan konfigurasi 3 *epochs* dan ukuran *batch* 16. Selama proses pembaruan berlangsung, model menunjukkan stabilitas komputasi

yang sangat baik, mencatat *validation loss* yang berkonvergensi di angka 1.788.

Setelah tahap pelatihan CPT selesai, evaluasi *Self-Perplexity* dilakukan menggunakan data uji (*test set*) murni dari domain DFK (Disinformasi, Fitnah, dan Ujaran Kebencian). Hasil komparasi menunjukkan penurunan tingkat ketidakpastian prediksi yang sangat masif. Pada model baseline (sebelum CPT), nilai *Perplexity* pada domain target berada di angka 7.7082. Namun, setelah diinkubasi menjadi model **KomdigiITS-8B-DFK-CPT**, angka tersebut berhasil ditekan runtuh ke titik final 5.8837.

Tabel 6.6 Hasil Evaluasi Model pada Domain Target (DFK)

Metrik Evaluasi	Ministral-3-8B-Base-2512	KomdigiITS-8B-DFK-CPT
Self-Perplexity	7.7082	5.8837

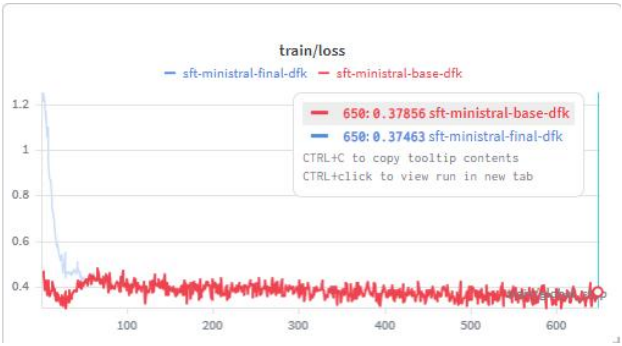
Penurunan dramatis skor *Perplexity* tersebut merepresentasikan lonjakan kapabilitas adaptasi domain spesifik (*Self Improvement*) sebesar 23,67%. Hasil empiris ini membuktikan bahwa pipeline pengumpulan dan *preprocessing* dataset CPT sukses menginjeksikan pemahaman kosakata baru, narasi sensitif, variasi *slang*, serta ambiguitas linguistik yang inheren pada bahasa informal di media sosial Indonesia. Model tidak lagi mengalami disorientasi ketika dihadapkan pada teks kotor siber.

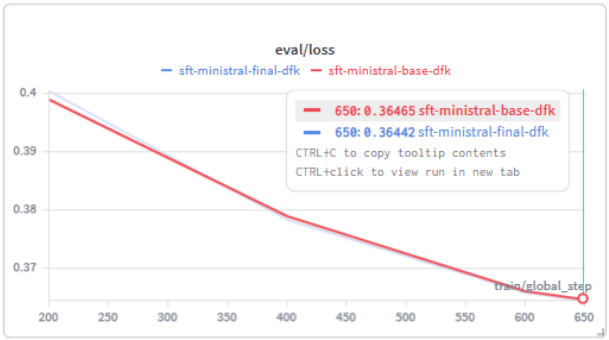
Sebagai langkah validasi tambahan untuk mencegah fenomena hilang ingatan (*Catastrophic*

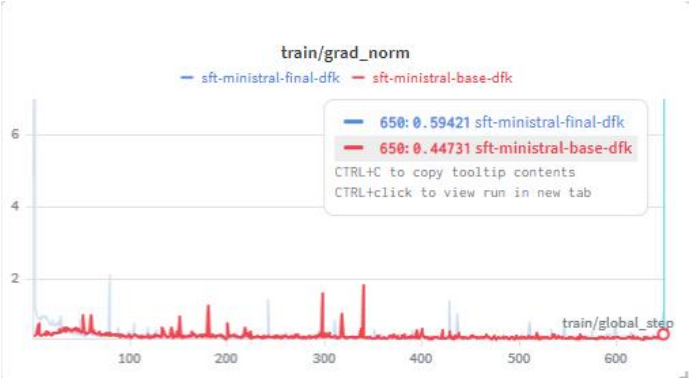
Forgetting), evaluasi berkala juga dilakukan menggunakan sampel kontrol dari Wikipedia Bahasa Indonesia (Wiki-ID). Kurva *Perplexity* pada sampel kontrol tersebut terbukti tetap stabil. Hal ini menegaskan bahwa arsitektur KomdigiITS-8B-DFK-CPT mampu menyerap representasi domain siber negatif secara komprehensif tanpa merusak kapabilitas pemahaman struktur tata bahasa fundamental dan pengetahuan umum yang telah dimilikinya.

Secara keseluruhan, metrik evaluasi intrinsik ini menunjukkan bahwa KomdigiITS-8B-DFK-CPT memiliki sensitivitas domain yang jauh lebih tinggi dan stabilitas pelatihan yang optimal. Pencapaian ini secara mutlak menjadikannya sebagai fondasi yang matang dan siap untuk diekspansi ke tahap *Supervised Fine-Tuning* (SFT).

6.2.2. Supervised Finetuning

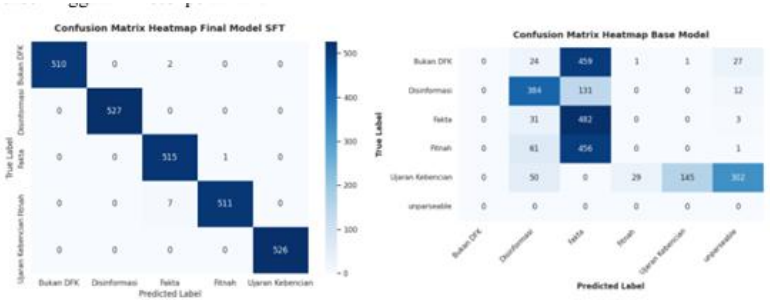




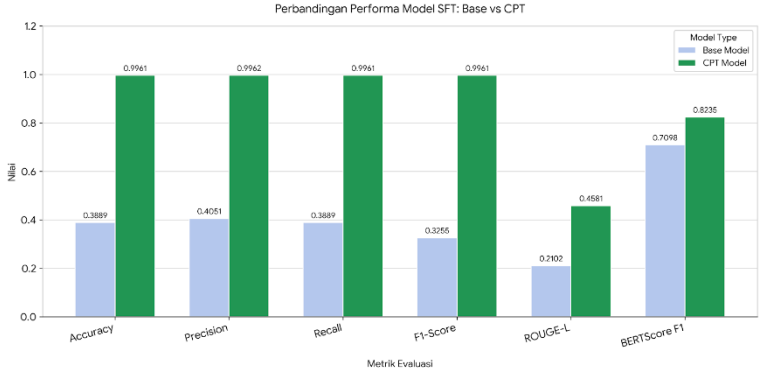


Gambar 6. 2 Kurva Loss SFT

Kurva menunjukkan pergerakan yang sangat berhimpitan secara konsisten dari step 200 hingga step 650. Hal ini mengindikasikan bahwa kemampuan generalisasi kedua model berada pada level yang sangat setara. Pada fase awal pelatihan (step < 50), model CPT menunjukkan lonjakan nilai loss yang cukup tinggi (~1.2) sebelum akhirnya menukik tajam secara drastis. Perilaku ini umum dijumpai ketika model yang baru saja menyelesaikan CPT harus melakukan penyesuaian format terhadap instruksi SFT yang baru. Setelah melewati step 100, model CPT berhasil menstabilkan diri dan secara konsisten mempertahankan posisi loss di bawah model base hingga akhir sesi pelatihan.



Gambar 6. 3 Hasil Confusion Matrix



Gambar 6.4 Perbandingan Performa SFT: Base vs CPT

Tabel 6.7 Hasil Perbandingan SFT

Metrik Evaluasi	**Ministral-3-8B-Base-2512**	**Model CPT**
Eval/Loss	0.365	0.364
Train/Loss	0.379	0.375
Train/Grad Norm	0.447	0.594
Accuracy	0.39	0.99
F1-Score	0.41	0.99
Precisison	0.39	0.99
Recall	0.32	0.99
BERTScore -F1	0.21	0.82
ROUGE-L	0.71	0.21

Berdasarkan hasil evaluasi pada test set, berikut adalah poin-poin observasi penting:

- Model KomdigiITS-8B-DFK-CPT dipilih sebagai model terbaik: Berdasarkan keseluruhan hasil evaluasi, model KomdigiITS-8B-DFK-CPT dipilih

sebagai model akhir karena menunjukkan performa yang lebih baik pada sebagian besar metrik evaluasi, khususnya Accuracy, F1-Score, BERTScore-F1, serta memiliki nilai Eval Loss yang lebih rendah dibandingkan model dasar.

- Kualitas representasi semantik meningkat pada model CPT: Nilai BERTScore-F1 model CPT jauh lebih tinggi dibandingkan model Base. Hasil ini menunjukkan bahwa respons yang dihasilkan model CPT memiliki kemiripan makna yang lebih baik terhadap referensi (*ground truth*), sehingga kualitas penjelasan yang dihasilkan menjadi lebih relevan secara semantik.
- Rendahnya performa model Base disebabkan oleh prediksi di luar label yang ditentukan: Berbeda dengan model CPT, model Ministral-3-8B-Base-2512 masih sering menghasilkan label yang tidak termasuk dalam himpunan kelas yang telah didefinisikan. Misalnya, ketika model hanya diperbolehkan memprediksi lima kelas, yaitu Disinformasi, Ujaran Kebencian, Fitnah, Fakta, dan Non-DFK, model Base beberapa kali memberikan keluaran berupa "unknown". Karena label tersebut tidak termasuk dalam target klasifikasi, seluruh prediksi tersebut dihitung sebagai kesalahan (*misclassification*), sehingga nilai Accuracy, Precision, Recall, dan F1-Score menjadi jauh lebih rendah dibandingkan model CPT. Temuan ini menunjukkan bahwa model Base belum mampu mengikuti instruksi klasifikasi secara konsisten, sedangkan proses Continued Pre-Training (CPT)

membantu model menghasilkan prediksi yang lebih sesuai dengan label yang telah ditentukan.

6.3. Evaluasi Sistem

Berdasarkan hasil implementasi dan pengujian, sistem yang dikembangkan sudah mampu menjalankan fungsi utama, yaitu mengklasifikasikan konten teks ke dalam kategori DFK serta memberikan alasan dari hasil klasifikasi tersebut. Model Ministral CPT dipilih sebagai model final karena menunjukkan performa terbaik dibandingkan model pembanding, terutama setelah melalui tahap CPT dan SFT.

Kelebihan utama sistem ini adalah kemampuannya menghasilkan output yang tidak hanya berupa label, tetapi juga reasoning. Hal ini membuat hasil klasifikasi lebih mudah dipahami dan dapat digunakan sebagai bahan pertimbangan oleh pengguna. Selain itu, model sudah di-deploy dalam bentuk API sehingga lebih mudah diintegrasikan dengan aplikasi MVP atau sistem lain.

Namun, sistem masih memiliki beberapa keterbatasan. Model tetap dapat mengalami kesalahan pada kasus yang ambigu, teks yang minim konteks, atau bahasa media sosial yang terlalu tidak baku. Selain itu, reasoning yang dihasilkan model tetap perlu divalidasi, terutama untuk kasus yang sensitif atau berpotensi berdampak hukum.

Secara keseluruhan, sistem sudah memenuhi tujuan utama proyek sebagai layanan klasifikasi DFK berbasis LLM. Meski begitu, sistem masih perlu dikembangkan lebih lanjut melalui pengujian pada data dunia nyata, peningkatan kualitas dataset, serta penerapan validasi manusia untuk kasus-kasus penting.

BAB VII
KESIMPULAN DAN SARAN

7.1. Kesimpulan

Rangkaian proses pengembangan, implementasi, dan pengujian model pada proyek ini menunjukkan bahwa pendekatan *Continued Pre-Training* (CPT) efektif dalam mengadaptasikan pengetahuan model dasarke domain Disinformasi, Fitnah, dan Ujaran Kebencian (DFK). Dengan memanfaatkan 374.930 data bersih atau sekitar 227,4 juta token, model berhasil menurunkan nilai *perplexity* pada domain target dari 7.7082 menjadi 5.8837, yang menunjukkan peningkatan kemampuan adaptasi sebesar 23,67% tanpa mengorbankan pemahaman bahasa umum yang telah dimiliki model. Pada tahap *Supervised Fine-Tuning* (SFT), arsitektur Ministral-3-8B-Base-2512 yang telah melalui proses CPT terpilih sebagai model terbaik karena mampu mencapai nilai akurasi dan F1-score sebesar 0,99 pada data uji, serta menunjukkan performa yang lebih unggul dibandingkan model dasar yang masih sering menghasilkan prediksi di luar label yang ditentukan. Selain itu, peningkatan nilai BERTScore-F1 hingga 0,82 menunjukkan bahwa model tidak hanya mampu melakukan klasifikasi secara akurat, tetapi juga menghasilkan penalaran (*reasoning*) yang memiliki kesesuaian makna yang tinggi dengan jawaban acuan (*ground truth*). Keberhasilan implementasi model ke dalam layanan API berbasis FastAPI menggunakan platform Modal juga menunjukkan bahwa sistem telah siap diintegrasikan sebagai komponen backend cerdas pada purwarupa *Minimum Viable Product* (MVP) untuk mendukung proses moderasi dan

pengawasan ruang digital oleh Kementerian Komunikasi dan Digital.

7.2. Saran

Meskipun sistem yang dikembangkan telah berhasil memenuhi tujuan utama proyek, masih terdapat beberapa aspek yang perlu disempurnakan untuk mendukung pengembangan pada tahap selanjutnya. Model yang digunakan masih memiliki keterbatasan dalam menangani kasus yang ambigu serta variasi bahasa media sosial yang sangat tidak baku. Oleh karena itu, pengumpulan data latih perlu dilakukan secara berkelanjutan dengan memperluas cakupan bahasa nonformal, singkatan, dan istilah slang yang terus berkembang, mengingat realisasi komponen Berita DFK Nonformal pada dataset CPT proyek ini baru mencapai 72% dari target yang ditetapkan. Selain itu, ruang lingkup analisis yang saat ini terbatas pada konten teks berbahasa Indonesia perlu diperluas ke pendekatan multimodal agar sistem mampu mendeteksi media sintetis dalam berbagai bentuk, seperti gambar, audio, maupun video *deepfake*. Di sisi implementasi, pihak Kementerian Komunikasi dan Digital juga disarankan untuk tetap menerapkan mekanisme *Human-in-the-Loop* atau validasi oleh manusia sebelum hasil penalaran model digunakan sebagai dasar penindakan terhadap konten yang bersifat sensitif atau memiliki implikasi hukum. Terakhir, optimalisasi infrastruktur inferensi melalui eksplorasi teknik kuantisasi yang lebih lanjut serta pemanfaatan pustaka *model serving* berperforma tinggi diharapkan dapat meningkatkan efisiensi komputasi sekaligus menekan biaya operasional layanan API *serverless* dalam jangka panjang.

DAFTAR PUSTAKA

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

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

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI.

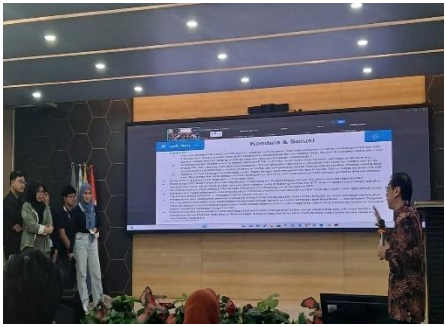
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. International Conference on Learning Representations (ICLR).

LAMPIRAN

Dokumentasi Kegiatan



Realisasi Penggunaan Dana

A	SEWA ALAT/LAYANAN/TEMPAT					
No	**Komponen**	**Jumlah**	**Satuan**	**Volume**	**Biaya Satuan (Rp)**	**Biaya Total (Rp)**
1	Google Colab Pro+	3	akun	1	Rp725.815,67	Rp2.177.447,00
2	Modal Labs	10	USD	1	Rp18.753,07	Rp187.530,65
	SUB TOTAL A (Rp)					Rp2.364.977,65
B	**BELANJA BAHAN HABIS**					
No	**Komponen**	**Jumlah**	**Satuan**	**Volume**	**Biaya Satuan (Rp)**	**Biaya Total (Rp)**
1	Open Router	34,01	USD	1	Rp16.685,20	Rp567.463,65
	SUB TOTAL B (Rp)					Rp567.463,65

BIODATA PENULIS

Nama : Daffa Rinali
NRP : 5025231209
Tempat, Tanggal Lahir : Jakarta, 13 Oktober 2005
Jenis Kelamin : Laki-Laki
Telepon : +6282236053510
Email : 5025231209@student.its.ac.id
AKADEMIS
Kuliah : Departemen Teknik Informatika - FTEIC, ITS
Angkatan : 2023
Semester : 6 (Enam)

Nama : Joe Hanzen Goenawan
NRP : 5054231003
Tempat, Tanggal Lahir : Surabaya, 3 April 2005
Jenis Kelamin : Laki-Laki
Telepon : +62 822-3651-3728
Email : 5054231003@student.its.ac.id
AKADEMIS
Kuliah : Departemen Teknik Informatika - FTEIC, ITS
Angkatan : 2023
Semester : 6 (Enam)