



Voice Recognition

Gregorius Reynaldi Pratama

10 January 2026

Meet The Team



Gregorius Reynaldi Pratama



<https://github.com/GregReynaldi>

Background

Evolusi Interaksi Manusia dan Komputer (HCI) :

- ✓ Transformasi digital telah menggeser cara manusia berinteraksi dengan mesin, dari antarmuka berbasis teks menuju interaksi berbasis suara (Voice-first interface) yang lebih natural, inklusif, dan efisien.

Tantangan dalam Automatic Speech Recognition (ASR):

- ✓ Meskipun teknologi ASR berkembang pesat, tantangan utama tetap terletak pada variansi bahasa, aksen, dan dialek. Dataset PolyAI/minds14 menjadi sangat relevan karena mencakup 14 kategori maksud (intent) dalam berbagai bahasa, yang menuntut model deep learning untuk memiliki kemampuan generalisasi yang tinggi.

Implementasi Arsitektur State-of-the-Art:

- ✓ Proyek ini mengeksplorasi penggunaan algoritma canggih seperti Wav2Vec2 atau Whisper. Algoritma ini menggunakan teknik Self-Supervised Learning yang mampu mengekstraksi fitur representasi audio secara mendalam, melampaui metode ekstraksi fitur tradisional seperti MFCC.

Peran Teknologi NLP dan Model Canggih:

- ✓ Melalui optimasi model pada dataset MINDS-14, proyek ini bertujuan untuk menjembatani celah antara sinyal akustik mentah dengan pemahaman kontekstual, yang aplikatif untuk asisten virtual, layanan pelanggan otomatis, dan sistem kontrol perangkat berbasis suara.



Tujuan Project

Eksperimentasi Algoritma State-of-the-Art:

- ✓ Mengimplementasikan dan membandingkan performa model Deep Learning tingkat lanjut seperti Wav2Vec2 atau Whisper untuk menangani pemrosesan sinyal audio mentah menjadi teks (ASR).

Pemahaman Dataset Multibahasa & Intent:

- ✓ Menganalisis dan mengoptimalkan kemampuan model dalam mengenali berbagai pola bicara, dialek, dan 14 kategori intent yang terdapat dalam dataset PolyAI/minds14.

Evaluasi Metrik Akurasi:

- ✓ Melakukan evaluasi mendalam menggunakan metrik standar industri seperti Word Error Rate (WER) dan BERT Embedding Cosine Similarity untuk menjamin akurasi transkripsi yang tinggi.

Optimasi Interaksi Manusia & Komputer:

- ✓ Membangun sistem yang mampu menjembatani hambatan komunikasi antara manusia dan mesin melalui pengenalan suara yang responsif dan reliabel.



Cakupan Penelitian

Eksplorasi Dataset Spesifik:

- ✓ Penelitian difokuskan pada penggunaan dataset PolyAI/minds14, yang terdiri dari rekaman audio perbankan (e-banking) yang mencakup 14 kategori intent (maksud) manusia dengan berbagai variasi aksen dan dialek.

Implementasi Sistem ASR:

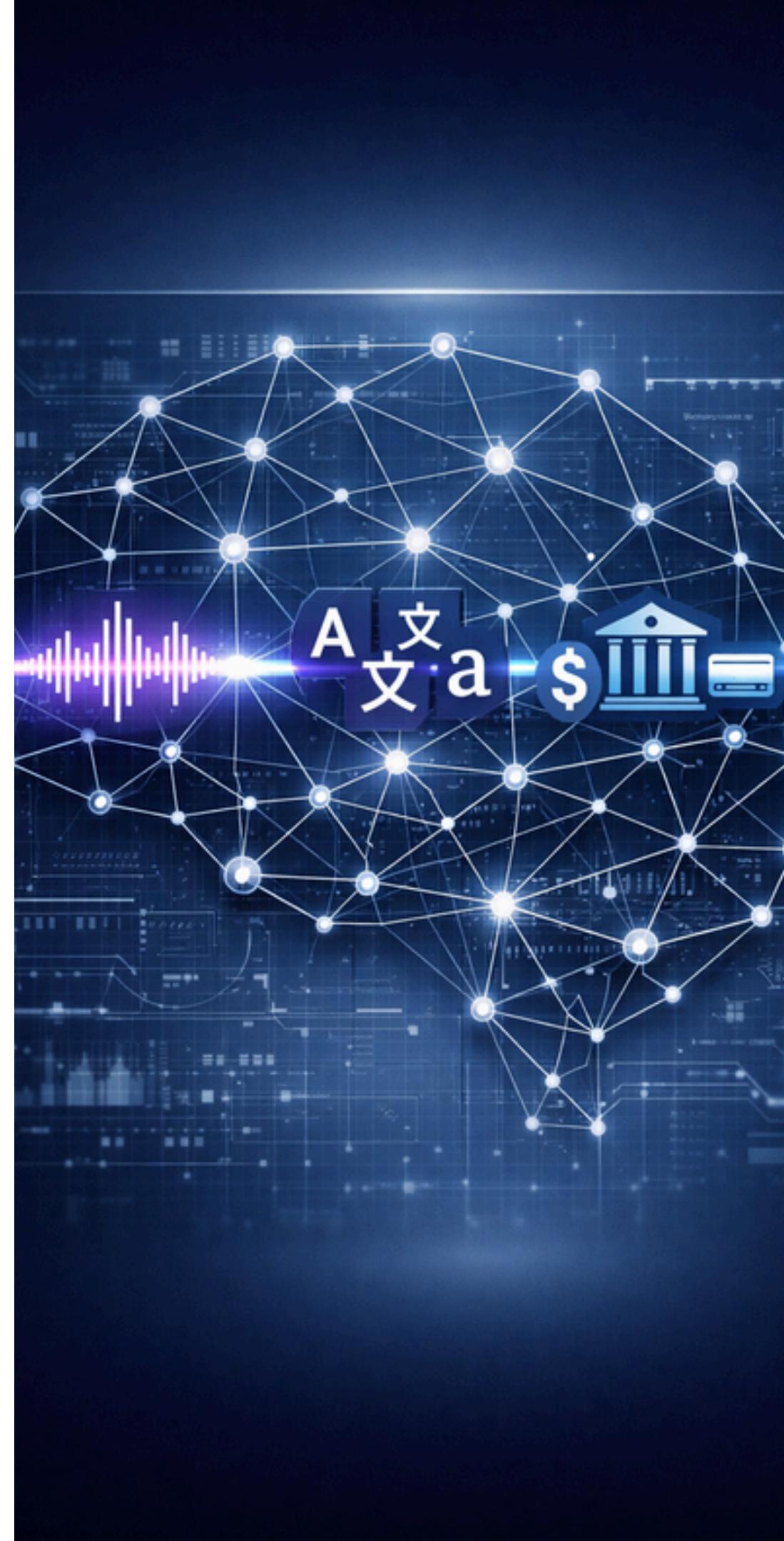
- ✓ Membangun sistem Automatic Speech Recognition (ASR) yang mampu melakukan pemetaan urutan (sequence mapping) dari sinyal audio mentah menjadi output teks secara akurat

Arsitektur Deep Learning:

- ✓ Penggunaan model berbasis Transformer state-of-the-art, secara spesifik bereksperimen dengan arsitektur Wav2Vec2 atau Whisper untuk menangani representasi fitur akustik yang kompleks.

Validasi Performa:

- ✓ Pengujian terbatas pada metrik evaluasi Word Error Rate (WER) dan BERT Cosine Similarity untuk mengecek keberhasilan model dalam mengklasifikasikan intent dari data audio yang diberikan.



Data Collection & Preparation

Sumber Data:

Menggunakan dataset PolyAI/minds14, sebuah dataset benchmark untuk intent classification dan ASR yang berisi rekaman audio asli dari domain layanan perbankan (e-banking)

Struktur Dataset:

- Terdiri dari 14 kategori intent (seperti pay_bill, cash_deposit, joint_account, dll).
- Mencakup variasi bahasa yang beragam, memberikan tantangan nyata pada generalisasi model.
- Format data berupa sinyal audio mentah (raw waveform).

Audio Resampling:

- Melakukan konversi Sampling Rate dari frekuensi asli dataset menjadi 16.000 Hz.
- Penyeragaman ke 16kHz sangat penting karena merupakan standar input yang diterima oleh arsitektur model state-of-the-art (seperti Wav2Vec2 atau Whisper) guna memastikan ekstraksi fitur fitur akustik yang optimal.
- Sinyal audio diproses menggunakan feature extractor untuk mengubah gelombang suara menjadi representasi numerik yang siap diproses oleh algoritma model Whisper dan Wav2Vec.



Exploratory Data Analysis - About Dataset

```
DatasetDict({  
    train: Dataset({  
        features: ['path', 'audio', 'transcription', 'english_transcription', 'intent_class', 'lang_id'],  
        num_rows: 8168  
    })  
})
```

train:

- Ini adalah nama dataset yang berisi data pelatihan. Biasanya, dataset ini digunakan untuk melatih model.

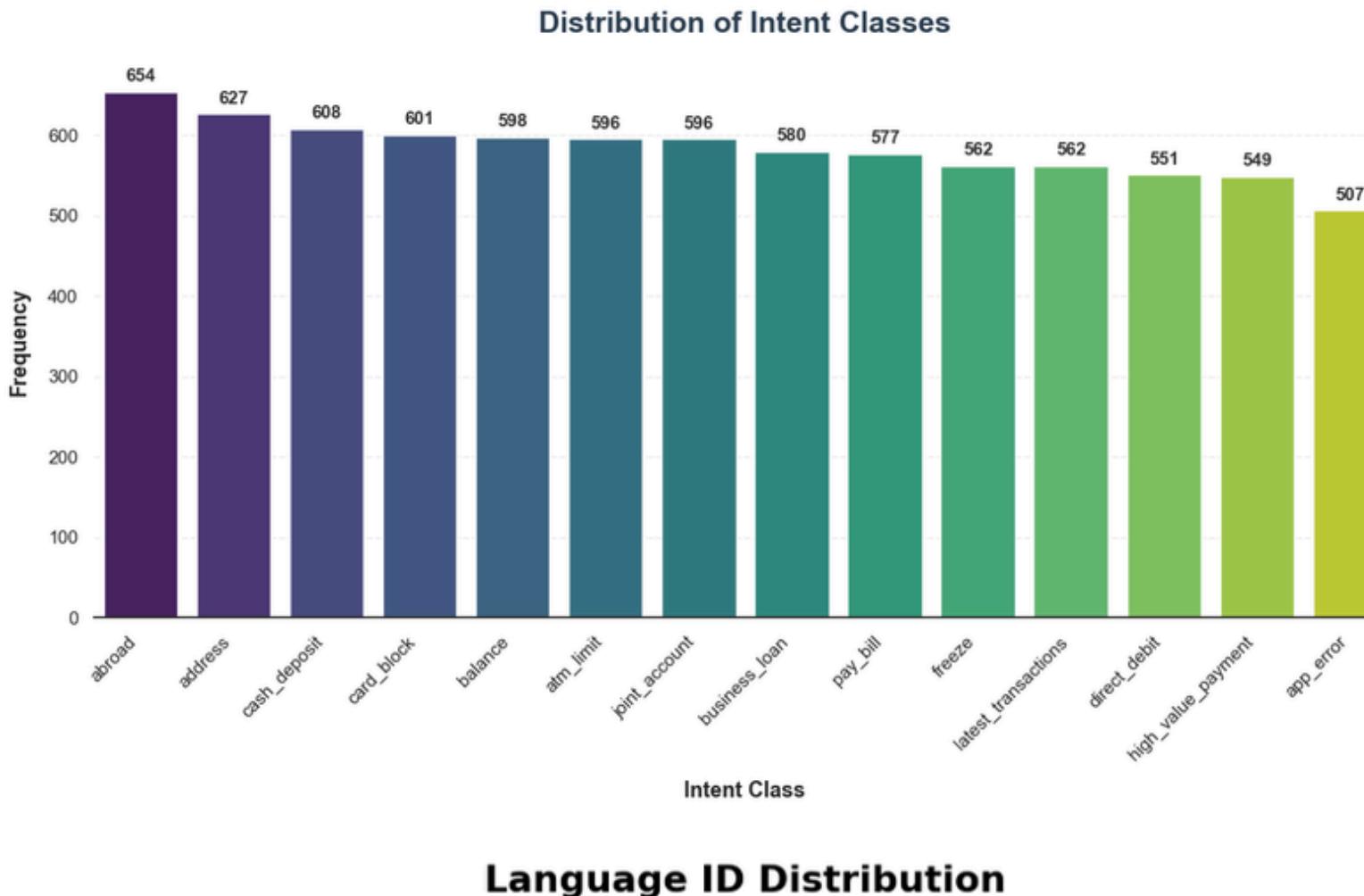
features:

- 'path': Lokasi file audio atau jalur menuju file yang akan diproses.
- 'audio': Terdiri dari 2 key yaitu sampling_rate dan array untuk menunjukkan array wave nya sama seperti jika menggunakan librosa.load(path) yang akan mengembalikan y dan sr
- 'transcription': Teks transkripsi asli dari audio (kemungkinan dalam bahasa asli).
- 'english_transcription': Transkripsi dalam bahasa Inggris dari audio tersebut (mungkin untuk aplikasi multibahasa atau terjemahan).
- 'intent_class': Kelas niat atau kategori dari setiap kalimat, yang dapat digunakan untuk tugas klasifikasi niat dalam pemrosesan bahasa alami (NLP).
- 'lang_id': ID bahasa yang digunakan untuk menandai bahasa apa yang digunakan dalam transkripsi atau audio.

num_rows:

- Ini menunjukkan jumlah baris atau data point dalam dataset. Dalam hal ini, terdapat 8,168 baris data dalam dataset pelatihan.

Exploratory Data Analysis - About Dataset



Dominasi Bahasa :

- Dataset menunjukkan dominasi kuat pada penggunaan Bahasa Inggris (1.809), sementara bahasa lainnya terdistribusi secara kompetitif di bawah 700 entri.

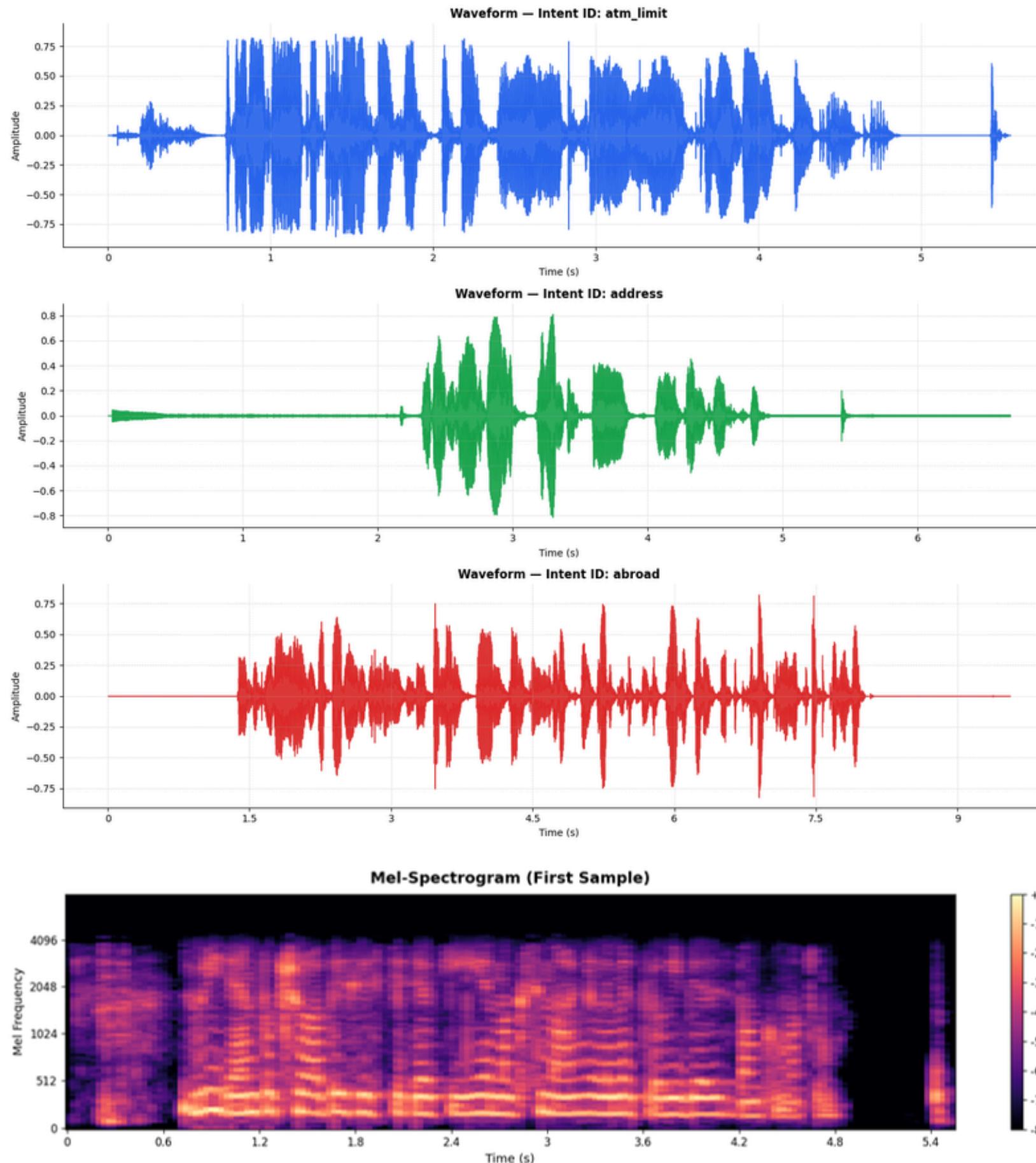
Keseimbangan Kelas :

- Berbeda dengan bahasa, kategori Intent Class memiliki distribusi yang sangat seimbang (balanced), dengan rentang yang rapat antara 507 hingga 654 entri per kelas.

Kesiapan Data:

- Sebaran bahasa yang luas dan kelas intensi yang merata memastikan model memiliki keberagaman data yang cukup untuk mencegah overfitting pada satu topik tertentu.

Exploratory Data Analysis - Wave + Spectrogram



Penggunaan Waveform:

- Waveform menggambarkan perubahan amplitudo audio seiring waktu, menunjukkan pola suara seperti durasi dan intensitas.
- Kelebihan: Berguna untuk menganalisis pola temporal seperti durasi kata dan transisi suara.
- Keterbatasan: Tidak memberikan informasi frekuensi, yang penting untuk pengenalan ucapan atau analisis musik.

Penggunaan Spektrum :

- Spektrum (mel-spectrogram) menunjukkan distribusi energi frekuensi dalam audio seiring waktu.
- Kelebihan: Penting untuk pengenalan ucapan karena menangkap komponen frekuensi yang digunakan dalam kata dan suara.
- Mel-spectrogram: Memanfaatkan frekuensi logaritmik, lebih mirip cara manusia mendengar, dan biasanya tidak digunakan secara langsung sebagai input pada model melainkan hanya digunakan sebagai visualization



ASR

Speech Recognition Part...



Whispers

Komponen	Whisper
Arsitektur	Berbasis Transformer, dilatih untuk multibahasa dan multimodalitas (suara dan teks). Memiliki Encoder dan Decoder
Model Ukuran	Tersedia dalam berbagai ukuran (Tiny, Base, Small, Medium, Large) untuk fleksibilitas akurasi dan kecepatan.
Pre-training	Dilatih pada dataset multibahasa besar, memungkinkan pengenalan suara dari berbagai bahasa tanpa pelabelan bahasa khusus.
Deteksi Bahasa	Mendeteksi bahasa dari input audio secara otomatis tanpa perlu pengaturan manual.
Penggunaan	Ideal untuk aplikasi yang memerlukan pengenalan suara multibahasa atau pengenalan aksen yang beragam.
Kecepatan & Akurasi	Model kecil lebih cepat, sementara model besar memberikan akurasi lebih tinggi, namun lebih memerlukan sumber daya.
Output	Menghasilkan transkrip teks dari input audio dan juga dapat digunakan untuk penerjemahan otomatis.

Model yang Digunakan (openai/whisper-large-v3):

- Jenis Model: Model Automatic Speech Recognition (ASR) berbasis Transformer.
- Ukuran Model: Whisper-large-v3 adalah varian terbesar dengan performa terbaik di antara model Whisper, menawarkan akurasi tinggi untuk pengenalan ucapan.
- Sampling Rate: 16,000 Hz — audio yang masuk harus diproses pada frekuensi sampel ini untuk mendapatkan hasil terbaik.

Arsitektur:

- Encoder-Decoder : Whisper menggunakan arsitektur Transformer dengan encoder-decoder yang digabungkan dalam satu model.
- Multibahasa: Mampu mengenali berbagai bahasa dan aksen tanpa pelatihan ulang untuk bahasa tertentu.

Whispers - ASR Performance (ENGLISH)

	Prediction	Label	Score	Cosine Similarity
0	i m using my banking app and it s not working ...	i m using my banking app and it s not working ...	0.000000	1.000000
1	hi i m just calling up because i have noticed ...	hi i m just calling up because i have noticed ...	0.100000	0.975704
2	i m trying to make a high price payment online...	i m trying to make a high priced payment onlin...	0.047619	0.996596
3	i leave my car abroad can you pay for things w...	i leave my car abroad can you buy things in a ...	0.400000	0.938690
4	i would like to talk about a business loan	i would like to talk about a business loan	0.000000	1.000000
...
95	hi there i m just calling because i m looking ...	hi there i m just calling because i m looking ...	0.313725	0.973692
96	please could you freeze any more transactions ...	please could you freeze any more transactions ...	0.000000	1.000000
97	could i see my account balance please	can i see my account balance please	0.142857	0.963325
98	that you re making a call to a financial organ...	like you re making a call to a financial organ...	0.241379	0.908644
99	we re using an application for my banking acco...	will you think as an application for my bankin...	0.314286	0.920728

Total Test	WER <= 0.1	WER > 0.1	WER<=0.1 + Cos Similar >= 0.9
100	56	44	86

Word Error Rate (WER):

0.1779 Akurasi transkripsi sangat tinggi, menunjukkan model dasar sudah sangat mengenali pola bahasa Inggris.

Cosine Similarity:

0.9491 Menunjukkan tingkat kemiripan semantik yang sangat kuat antara prediksi teks dan data asli. Meskipun memiliki perbedaan penulisan kata namun arti antara prediction dan label yang ada sudah sangat baik.

Transcription Comparison :

Model berhasil mencapai standar kualitas tinggi pada 86% sampel pengujian dengan kombinasi akurasi kata WER<=0.1 dan kemiripan makna Cosine Similarity >= 0.9 yang sangat kuat.

Whispers - ASR Performance (Not ENGLISH)

	Prediction	Label	Score	Cosine Similarity
0	ik wil graag weten wat mijn rekeningsauto is	ik wil graag weten wat mijn rekeningsaldo is	0.125000	0.988598
1	je souhaiterais déposer de l'argent sur mon co...	je souhaiterais déposer de l'argent sur mon co...	0.000000	1.000000
2	goedemiddag ik had even een vraagje ik zou gra...	goedemiddag ik had even een vraagje ik wil gra...	0.080000	0.991223
3	dobrý den chci vložit peníze na můj účet a zaj...	dobrý den chci vložit peníze na můj účet a zaj...	0.000000	1.000000
4	eu queria consultar o saldo do telemóvel	eu queria consultar o saldo do telemóvel	0.000000	1.000000
...
95	hoi ik vroeg me af of ik mijn bankpas ook in h...	hoi ik vroeg me af of ik mijn bankpas ook in b...	0.071429	0.990429
96	według scenariusza potrzebuję wpłacić 15% na s...	według mariusza potrzebuję wpłacić pieniądze n...	0.133333	0.988255
97	sì buongiorno io mi sono trasferita e di conse...	sì buongiorno niente io mi sono trasferita e d...	0.150000	0.993567
98	顯示我的帳戶以為	显示我的账户余额	1.000000	1.000000
99	rád bych se nechal informovat o posledních tra...	rád bych se nešel informovat o poslední transa...	0.272727	0.974685

Total Test WER <= 0.1 WER > 0.1 WER<=0.1 + Cos Similar >= 0.9

100	40	60	96
-----	----	----	----

Word Error Rate (WER):

0.3554 Terdapat kenaikan tingkat kesalahan kata dibandingkan data Inggris, yang dipengaruhi oleh variasi dialek dan aksen

Cosine Similarity:

0.9809 Angka yang sangat tinggi, menunjukkan bahwa meskipun ada kesalahan kata secara literal, makna atau konteks kalimat tetap tertangkap dengan sangat baik oleh model

Transcription Comparison :

Meski variasi bahasa lebih kompleks, model tetap unggul secara semantik dengan 96% sampel pengujian berhasil menjaga kedekatan makna di atas standar 0.9.

Whispers - ASR Performance (Overall)

	Prediction		Label	Score	Cosine Similarity
0	有一笔我不知道的付款	由于我不知道的付款	1.000000	0.968929	
1	hey i wanted to withdraw a significant amount ...	i want to do with at the moment for my atm and...	0.244444	0.963147	
2	hola buenas tardes necesitaba saber si necesit...	hola buenas tardes necesitaba saber si necesit...	0.026316	0.999526	
3	qual é o meu limite na caixa eletrônica posso ...	qual é o alimento na caixa eletrônica não está...	0.428571	0.981229	
4	계좌에 돈을 입금 하려고 하는데 어떻게 해야 될지 모르겠어서요	계좌에 돈을 입금하려고 하는데 어떻게 해야 될지 모르겠어서요	0.250000	0.996656	
-
95	пожалуйста заблокируйте мою карту	пожалуйста заблокируйте мою карту	0.000000	1.000000	
96	¿cómo puedo configurar una punta con punta	cómo puedo configurar la cuenta conjunta	0.833333	0.891716	
97	hola llamo porque quería saber cuál es el lími...	hola llamo porque quería saber cuál es el lími...	0.000000	1.000000	
98	voglio congelare tutti i conti tutte le mie tr...	voglio congelare tutti i conti sulla tutte le ...	0.076923	0.996535	
99	ik wil graag een rekening betalen	ik wil graag een rekening betalen	0.000000	1.000000	

Total Test WER <= 0.1 WER > 0.1 WER<=0.1 + Cos Similar >= 0.9

100	39	61	92
-----	----	----	----

Word Error Rate (WER):

0.4067 Rata-rata kesalahan kata secara keseluruhan sebelum optimasi lebih lanjut

Cosine Similarity:

0.9574 Stabilitas pemahaman konteks tetap terjaga di atas 95% untuk seluruh kategori data

Transcription Comparison :

Secara keseluruhan, sistem menunjukkan performa yang tangguh dengan 92% sampel pengujian berhasil memenuhi kriteria ambang batas akurasi tinggi dan kemiripan semantik yang sangat kuat.

Wav2Vec

Komponen	Wav2Vec2
Arsitektur	Berbasis Transformer, hanya menggunakan encoder untuk memproses input audio menjadi representasi vektor.
Model Ukuran	Tersedia dalam berbagai ukuran (Base, Large), dengan ukuran model lebih kecil untuk efisiensi.
Output	Menghasilkan transkrip teks dari input audio dengan menggunakan CTC loss untuk menghubungkan fitur suara ke teks.
Deteksi Bahasa	Tidak mendeteksi bahasa secara otomatis; pengguna perlu mengatur bahasa input sebelum pemrosesan.
Penggunaan	Ideal untuk aplikasi pengenalan suara dalam bahasa Inggris, khususnya dalam ASR monolingual.
Kecepatan & Akurasi	Model lebih kecil lebih cepat, sementara model besar lebih akurat, namun membutuhkan lebih banyak sumber daya.

Model yang Digunakan (facebook/wav2vec2-base-100h):

- Jenis Model: Model Automatic Speech Recognition (ASR) berbasis Transformer dengan arsitektur encoder-only untuk pemrosesan audio.
- Ukuran Model: Wav2Vec2-base-100h adalah model berbasis Wav2Vec2 dengan pelatihan pada 100 jam data bahasa Inggris. Model ini lebih ringan dan lebih cepat dibandingkan varian large.
- Sampling Rate: 16,000 Hz — audio yang masuk harus diproses pada frekuensi sampel ini untuk menghasilkan pengenalan suara yang optimal.

Arsitektur:

- Encoder-Only: Wav2Vec2 menggunakan encoder berbasis Transformer yang memproses audio menjadi representasi vektor tanpa menggunakan decoder eksplisit.
- CTC (Connectionist Temporal Classification):

Wav2Vec- ASR Performance (ENGLISH) - Before Tune

	Prediction	Label	Score	Cosine Similarity
0	ow i i cant mo	how do i see my account balance	0.857143	0.475091
1	i i m just raing tof he to change my adres in ...	hi i m just ringing to ask you to change my ad...	0.571429	0.708699
2	what is my carent acount balance	what is my current account balance	0.333333	0.566961
3	itei te money	where can i deposit money	0.800000	0.579328
4	wone is torec to ert and how es where can you ...	what is direct debits and how does it work can...	0.533333	0.643115
...
95	halo im wan i to check what my curent acount b...	hello i like to check what my current account ...	0.478261	0.753449
96	how do i to plause the money into my acount	how do i deposit money into my account	0.500000	0.664066
97	ye hi can you tel me how much money have in my...	hey honey can you tell me how much money i hav...	0.315789	0.812176
98	hi there i m having dificulties using your at ...	hi there i m having difficulties using your as...	0.095238	0.920192
99	i wanted to pul at money into my acount	i want to deposit money into my account	0.500000	0.681368

Word Error Rate (WER):

Nilai WER sebesar 61,2% menunjukkan bahwa model Wav2Vec 2.0 dasar (pre-trained) masih mengalami kesulitan signifikan dalam mentranskripsikan audio perbankan secara tepat kata-demi-kata.

Cosine Similarity:

Skor 0.70 mengindikasikan bahwa meskipun banyak kata yang salah tulis, model masih mampu menangkap sekitar 70% konteks atau makna utama dari perintah suara pengguna.

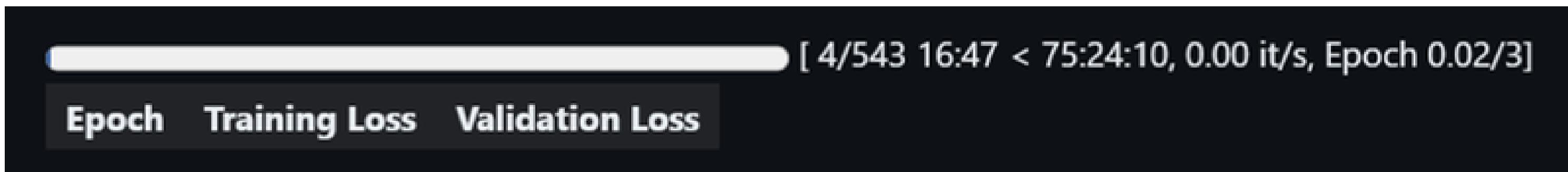
Wav2Vec- ASR Performance (ENGLISH) - After Tune ~ Future Task

Waktu Pelatihan Lama:

Pelatihan model ini membutuhkan waktu yang sangat lama (75+ jam) bahkan dengan dataset kecil, karena model yang digunakan adalah model standar, bukan model yang sudah dioptimasi atau dipercepat untuk tugas tersebut.

Model Tidak Akan Di-Tune Lebih Lanjut:

Mengingat waktu yang dibutuhkan, pelatihan dan tuning lebih lanjut tidak akan dilakukan, terutama untuk dataset yang terbatas. Ini menunjukkan bahwa model saat ini tidak akan mengalami fine-tuning lebih lanjut untuk meningkatkan performa.



Thank You