**Logbook Final Project Natural Language Processing**

**"Pengembangan Sistem Terjemahan Otomatis Cerita Rakyat Indonesia ke Bahasa Jawa untuk Pelestarian Budaya Lokal"**

| 5 mei 2025 | mencari referensi ide, dataset dan memilih Translasi |
| --- | --- |
| 12 mei 2025 | membuat dataset dan mencoba beberapa model |
| 19 mei 2025 | menyelesaikan Dataset, menggunakan model cendol, menambah BLEU score |
| 26 mei 2025 | dataset harus sudah siap,menambahkan metriks lain,dan mencoba model TTS  Presentasi 5- translasi bahasa jawa (bu dini)  Tim: Rafif, Dimas, Humam  - sudah training dengan cendol, tambahkan model lain: llama atau lainnya untuk perbandingan  - rencana dibuat Text2Speech |
| 27 mei - 01 Juni 2025 | * mencoba beberapa model TTS * menyiapkan dataset audio untuk fine tuning TTS * mencari perbandingan model Translation   model yang berhasil speechT5, meskipun logatnya bahasa inggris tetapi kata atau kalimat bahasa jawanya sudah terbaca. (sudah di fine tuning)  tidak seperti model seamlessM4T, kita inputnya cerita basa jawa tapi outputnya suara bahasa indonesia (belum fine tuning) |
| 02 juni - 11 juni 2025 | * Translation sudah aman mendapatkan bleu score 30 an, dengan cara nge split dataset lagi karena model memiliki batas tokenisasi * Text-to-Speech kemarin sudah berhasil tetapi model hanya mendukung bahasa inggris, suara tidak cocok   jadi dari kami masih mencari model yang mendukung bahasa indonesia dan bisa di fine tuning |
| 12 juni - 15 juni 2025 | menjalankan model TTS untuk mendapatkan output semua cerita & Mulai untuk Deploy |

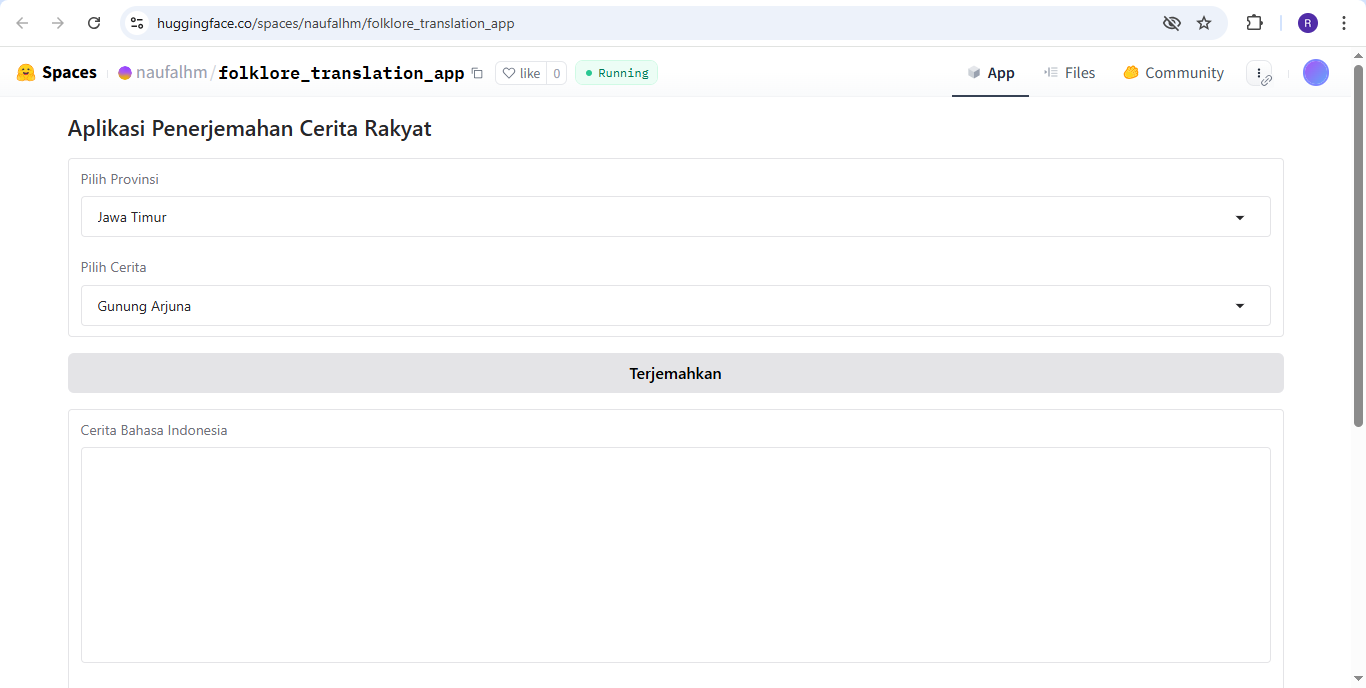
**TRANSLATION**

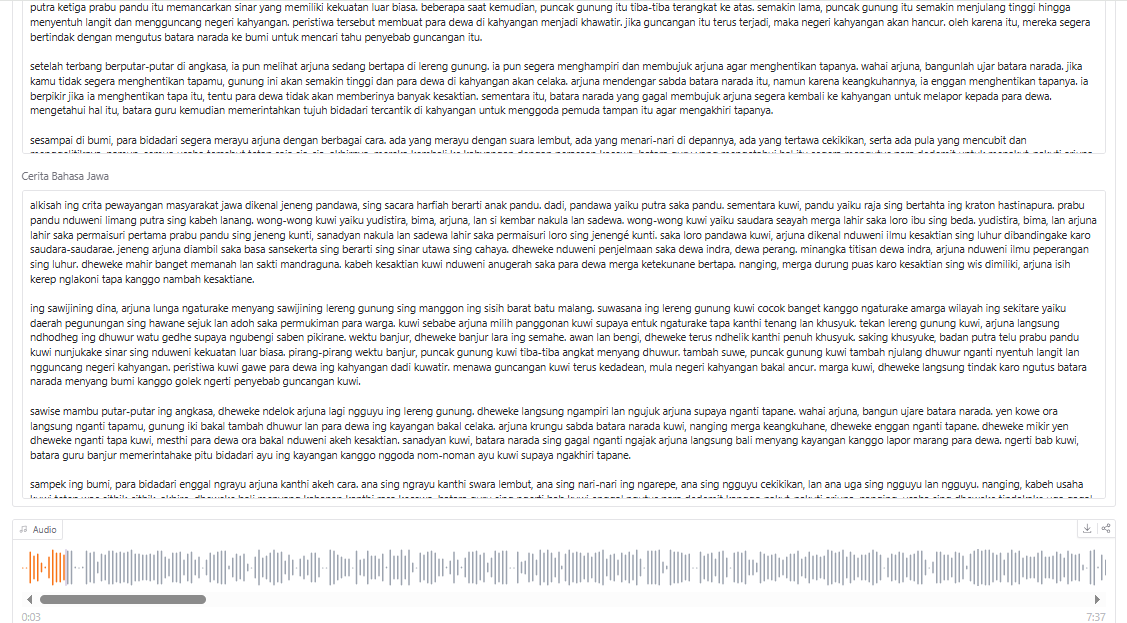
| **Tahap** | **Metode** | **BLEU Score** |
| --- | --- | --- |
| Fine-tuning awal | MT-5 Cendol | 8% |
| Fine-tuning data splitting | MT-5 Cendol, LoRA | 25% |
| Hyperparameter Tuning | MT-5 Cendol, LoRA | 32% |

| **Evaluasi Model** | **BLEU** | **ROUGE-L** | **METEOR** |
| --- | --- | --- | --- |
| Sebelum Split Data | 8.15 | 0.3853 | 0.2406 |
| Setelah Split Data & Tuning | 32.23 | 0.5762 | 0.5175 |

**Text to Speech**

| **MODEL** | **ALASAN** |
| --- | --- |
| Coqui TTS | Menjalankan di terminal |
| seamlessM4T | inputnya cerita basa jawa tapi outputnya suara bahasa indonesia |
| Microsoft SpecchT5 | Model cenderung membaca teks Bahasa Jawa dengan intonasi dan fonetik atau Bahasa Inggris, sehingga terdengar tidak alami dan sulit dipahami. |
| facebook/mms-tts-ind(Digunakan) | Output suara dapat membaca Bahasa Jawa dengan pelafalan yang dapat dipahami, Aksen masih terpengaruh oleh pelafalan Bahasa Indonesia, namun tidak mengganggu makna. |





## 

## **Alur Sistem :**

1. Input teks cerita rakyat Bahasa Indonesia & Bahasa Jawa.
2. Terjemahan ke Bahasa Jawa (menggunakan model Cendol).
3. Ubah hasil terjemahan menjadi suara (menggunakan TTS Bahasa Jawa).
4. Output suara dalam Bahasa Jawa(Deploy ke Hugging Face).

Ketika judul dipilih, sistem otomatis menampilkan teks cerita, menerjemahkannya (pakai Cendol), lalu hasilnya diubah menjadi suara Bahasa Jawa (pakai MMS-TTS).

### **Daftar Pustaka**

BLEU SCORE :

### **SRPH-Konvergen AI – WMT21 Submission**

* + **Referensi**:  
     Agustinus Kristiadi, et al. *Data Processing Matters: SRPH-Konvergen AI's Machine Translation System for WMT'21*. arXiv:2111.10513  
    <https://arxiv.org/abs/2111.10513>
  + **BLEU Score**:  
     Rata-rata BLEU untuk Indonesia–Jawa: **15.86**  hingga **22.30**

1. <https://ar5iv.labs.arxiv.org/html/2404.06138v1>
2. **Statistical Machine Translation Jawa–Indonesia**
   * **Referensi**:  
      Titis Mulat Argo et al. *Phrase-Based Statistical Machine Translation Javanese-Indonesian*, 2023.  
     <https://www.researchgate.net/publication/367868050>
   * **BLEU Score**:  
      Dengan data 300 kalimat hingga 5000 kalimat: skor BLEU berkisar diantara **39.79**, tergantung juga parameter SMT-nya.

**Nilai Rata-rata BLEU**: Skor BLEU yang realistis untuk model terjemahan mesin dari Bahasa Indonesia ke Bahasa Jawa biasanya berkisar antara 15 hingga 30, tergantung pada kualitas model dan dataset yang digunakan.

**Ukuran Dataset Minimal**: Untuk evaluasi yang stabil dan representatif, disarankan menggunakan dataset dengan ukuran minimal 500 kalimat. Namun, dataset yang lebih besar akan memberikan hasil yang lebih akurat.

BLEU memiliki keterbatasan, seperti ketergantungan pada n-gram dan tidak mempertimbangkan kesalahan semantik. disarankan untuk menggunakan BLEU bersama dengan matrik lain, seperti **METEOR** atau **ROUGE**, untuk evaluasi yang lebih komprehensif.

MODEL CENDOL :

### **Xu, N., Mulia, F., Akbari, M., et al. (2024). *Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages*. arXiv preprint arXiv:2404.06138.**

### **Link:** [**https://arxiv.org/abs/2404.06138**](https://arxiv.org/abs/2404.06138)

### **Kutipan:**

### **"Cendol models demonstrate strong capabilities in understanding and generating text across various Indonesian regional languages, providing a foundation for language translation tasks including Indonesian to Javanese."**

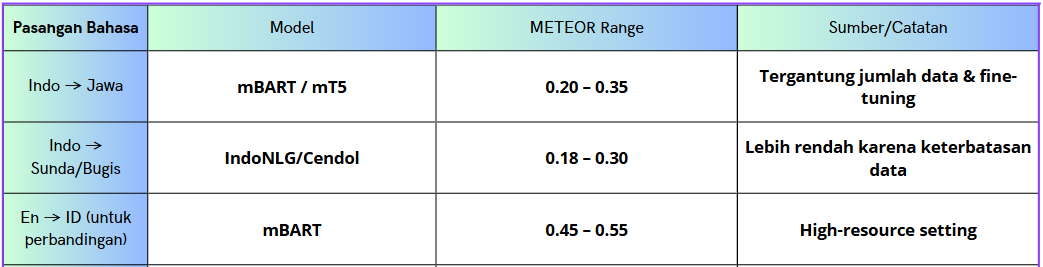
Model Cendol merupakan kumpulan model bahasa besar (LLM) yang dirancang khusus untuk bahasa-bahasa Indonesia, termasuk bahasa daerah seperti Jawa. Model ini menunjukkan kemampuan dalam memahami dan menghasilkan teks dalam berbagai bahasa daerah, meskipun belum secara eksplisit diterapkan untuk penerjemahan Indonesia-Jawa. Namun, potensi adaptasinya untuk tugas tersebut cukup besar.

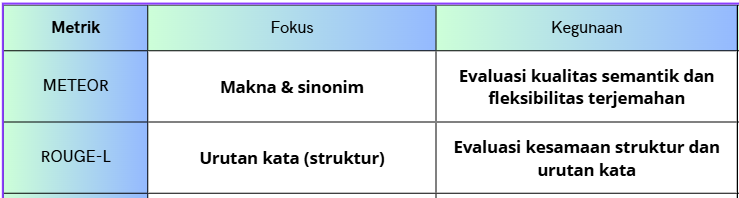
1. **Cahyawijaya, S., Mulia, R., Saputra, A., dkk. (2024). *Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages*. arXiv:2404.06138.** Link:<https://arxiv.org/abs/2404.06138> **Bukti kutipan:**

“Cendol menunjukkan performa terbaik pada berbagai benchmark Bahasa Indonesia, melampaui model open-source dan komersial dalam tugas seperti penerjemahan, tanya-jawab, dan rangkuman.”  
 (hal. 5)

1. **Tim IndoNLP. (2024). *Cendol: Kumpulan Model LLM Bahasa Indonesia Open-source*. GitHub Repository.** Link:<https://github.com/IndoNLP/cendol> **Bukti kutipan:**

“Cendol mengungguli bahkan model komersial tertutup seperti GPT-3.5 dan Claude untuk tugas-tugas dalam Bahasa Indonesia.”  
 (dikutip dari README proyek GitHub)





**Flowchart Translation & TTS :**

1. Input

- Cerita Bahasa Indonesia (untuk Translation Model)

- Cerita Basa Jawa (Translate sendiri, untuk Fine tuning model)

↓

2. Preprocessing Teks

- Normalisasi (huruf kecil, tanda baca)

- Tokenisasi

↓

3. Translasi Indonesia → Jawa

- Pilihan:

• Model Cendol fine-tuned

↓

4. Evaluasi Model

- Bleu

- Rouge

- Meteor

↓

5. Text-to-Speech Bahasa Jawa

- Menggunakan model TTS yang mendukung Bahasa Jawa

• model facebook/mms-tts-ind

↓

6. Output: Audio Bahasa Jawa (WAV/MP3)

↓

7. Deploy Hugging Face