DOKUMENTASI DATASET MOZILLA COMMON VOICE

Untuk Proyek Speech Recognition - Deep Learning

INFORMASI UMUM DATASET

Nama Dataset: Mozilla Common Voice Dataset v19.0

Bahasa: English (en)

Tanggal Rilis: September 2024

Sumber: https://commonvoice.mozilla.org/

License: CC0 (Creative Commons Zero)

DESKRIPSI DATASET

Mozilla Common Voice adalah dataset speech recognition open-source yang dikumpulkan melalui crowdsourcing dari sukarelawan di seluruh dunia. Dataset ini berisi rekaman suara manusia yang membaca kalimat dalam berbagai bahasa, disertai dengan transkripsi teks yang akurat.

Karakteristik Dataset:

- **Format Audio:** MP3
- **Sampling Rate:** 48kHz (akan dikonversi ke 16kHz untuk training)
- **Durasi Audio:** Bervariasi dari 1-10 detik per clip
- **Jumlah Speaker:** Ribuan speaker dengan berbagai aksen dan demografis
- **Kualitas:** Telah divalidasi oleh komunitas untuk akurasi transkripsi

---

STRUKTUR DATASET

```
cv-corpus-19.0-delta-2024-09-13/
└── en/
    ├── clips/                    # Folder utama berisi file audio
    │   ├── common_voice_en_41227191.mp3
    │   ├── common_voice_en_41227192.mp3
    │   └── ... (sekitar 18.538+ file)
    ├── validated_sentences.tsv   # File metadata utama
    ├── train.tsv                 # Split training data
    ├── test.tsv                  # Split testing data
    ├── dev.tsv                   # Split development/validation data
    ├── other.tsv                 # Data tambahan
    ├── invalidated.tsv           # Data yang tidak valid
    └── reported.tsv              # Data yang dilaporkan bermasalah
```

PENJELASAN FILE-FILE PENTING

1. Folder `clips/`

- Berisi semua file audio dalam format MP3
- Nama file mengikuti pola: `common_voice_en_[ID].mp3`
- Setiap file berukuran antara 30-70 KB
- Total ukuran folder: ~900 MB - 1.2 GB

2. File `validated_sentences.tsv`

File metadata utama yang berisi informasi setiap audio clip:

Kolom	Deskripsi
client_id	ID unik pembicara
path	Nama file audio (misal: common_voice_en_41227191.mp3)
sentence	Transkripsi teks dari audio
up_votes	Jumlah vote positif dari validator
down_votes	Jumlah vote negatif dari validator
age	Kelompok usia pembicara (teens, twenties, thirties, dll)
gender	Jenis kelamin pembicara (male, female, other)
accents	Aksen pembicara (us, england, australia, dll)
locale	Kode bahasa (en untuk English)
segment	Informasi segmentasi
variant	Varian bahasa

3. File Split Data

- **train.tsv:** Data untuk training model (~80% dari total data)
- **test.tsv:** Data untuk testing/evaluasi final (~10% dari total data)
- **dev.tsv:** Data untuk validation selama training (~10% dari total data)

STATISTIK DATASET

Volume Data (Estimasi):

- **Total Audio Files:** ~18.538 file
- **Total Durasi:** ~25-30 jam audio
- **Total Size:** ~1.5 GB (keseluruhan dataset)
- **Ukuran per file:** Rata-rata 50-60 KB

- **Jumlah Kalimat Unik:** ~15.000 kalimat
- **Jumlah Speaker:** ~8.000+ kontributor **Distribusi Data:**
- **Training Set:** ~14.800 samples (80%)
- **Validation Set:** ~1.850 samples (10%)
- **Test Set:** ~1.850 samples (10%) **Demografi Speaker:**
- **Gender:** Seimbang antara male/female
- **Age Groups:** Dari teens hingga seventies
- **Accents:** Mayoritas US English, dengan Australia, England, dll

---

MENGAPA DATASET TIDAK DISERTAKAN DI GITHUB

Alasan Teknis:

1. **Ukuran File:** Dataset total ~1.5 GB melebihi batas GitHub (100 MB per file)
2. **Jumlah File:** 18.000+ file audio akan membuat repository sangat berat
3. **Bandwidth:** Download/clone repository akan sangat lambat
4. **Storage Costs:** GitHub memiliki batasan storage untuk repository

Alasan Legal:

1. **License Compliance:** Meskipun CC0, tetap perlu redistribusi yang proper
2. **Attribution:** Perlu memberikan kredit yang tepat kepada Mozilla
3. **Version Control:** Dataset besar tidak efektif untuk version control

Best Practice:

- Dataset besar biasanya disimpan di external storage (Google Drive, OneDrive, etc.)
- Repository hanya berisi kode, dokumentasi, dan sample data kecil
- Instruksi download disediakan dalam dokumentasi

CARA MENDAPATKAN DATASET

Opsi 1: Download Resmi

1. Kunjungi: https://commonvoice.mozilla.org/en/datasets
2. Pilih bahasa "English"
3. Pilih "Common Voice Corpus 19.0"
4. Download file ZIP (~1.5 GB)

---

5. Extract ke folder lokal

Opsi 2: Menggunakan Script (Opsional)

```python
# Script untuk download otomatis (jika tersedia)
import requests
import zipfile

def download_common_voice():
    url = "https://mozilla-common-voice-datasets.s3.dualstack.us-west-2.amazonaws.com/cv-corpus-19.0-delta-2024-09
    # Download dan extract otomatis
    pass
```

SETUP DATASET UNTUK PROYEK

Langkah-langkah Setup:

1. Persiapan Google Drive:

- Pastikan memiliki space minimal 2 GB di Google Drive
- Buat folder khusus: `Deep Learning/Speech Recognition/`

2. Upload Dataset:

- Upload folder `cv-corpus-19.0-delta-2024-09-13` ke Google Drive
- Struktur final di Drive:

```
My Drive/
└── cv-corpus-19.0-delta-2024-09-13/
    └── en/
        ├── clips/
        └── validated_sentences.tsv
```

3. Konfigurasi Path di Notebook:

```python
# Sesuaikan path ini dengan lokasi Anda
DATASET_PATH = "/content/drive/MyDrive/cv-corpus-19.0-delta-2024-09-13/en"
CLIPS_PATH = os.path.join(DATASET_PATH, "clips")
VALIDATED_TSV = os.path.join(DATASET_PATH, "validated_sentences.tsv")
```

4. Verifikasi Dataset:

```python
# Cek keberadaan file
print(f"Dataset folder exists: {os.path.exists(DATASET_PATH)}")
print(f"Clips folder exists: {os.path.exists(CLIPS_PATH)}")
print(f"Metadata file exists: {os.path.exists(VALIDATED_TSV)}")

# Hitung jumlah file
clip_count = len([f for f in os.listdir(CLIPS_PATH) if f.endswith('.mp3')])
print(f"Total audio files: {clip_count}")
```

PREPROCESSING YANG DIPERLUKAN

1. Audio Processing:

- **Resampling:** Dari 48kHz ke 16kHz untuk kompatibilitas model
- **Normalization:** Normalisasi amplitudo audio
- **Duration Filtering:** Filter audio 1-10 detik untuk efisiensi
- **Format Conversion:** Jika perlu, konversi MP3 ke WAV

2. Text Processing:

- **Lowercasing:** Konversi ke huruf kecil
- **Punctuation Removal:** Hapus tanda baca jika perlu
- **Character Filtering:** Hanya karakter alfabet dan spasi
- **Tokenization:** Menggunakan Wav2Vec2Tokenizer

3. Data Validation:

- **File Existence Check:** Pastikan file audio ada
- **Audio Quality Check:** Verifikasi audio dapat dibaca
- **Text-Audio Alignment:** Pastikan panjang teks sesuai audio
- **Duplicate Removal:** Hapus data duplikat jika ada

PERTIMBANGAN ETHICAL DAN LEGAL

Ethical Considerations:

- **Privacy:** Dataset tidak berisi informasi personal identifiable
- **Consent:** Semua kontributor telah memberikan consent

- **Diversity:** Dataset mencakup berbagai demografi dan aksen
- **Bias:** Perhatikan potential bias dalam distribusi data

Legal Compliance:

- **License:** CC0 memungkinkan penggunaan bebas tanpa atribusi
- **Commercial Use:** Diizinkan untuk penggunaan komersial
- **Redistribution:** Dapat didistribusi ulang dengan proper attribution
- **Academic Use:** Sangat cocok untuk research dan pendidikan

---

TROUBLESHOOTING DATASET

Masalah Umum dan Solusi:

1. File tidak ditemukan:

```
Error: [Errno 2] No such file or directory: 'common_voice_en_xxxxx.mp3'
```

Solusi: Verifikasi path dan pastikan file tidak corrupt saat upload

2. Audio tidak bisa dibaca:

```
Error: Could not load audio file
```

Solusi: Install codec yang diperlukan atau gunakan librosa sebagai fallback

3. Metadata tidak match:

```
Error: Audio file exists but not in TSV
```

Solusi: Gunakan only files yang ada di both clips folder dan TSV file

4. Memory error saat loading:

```
OutOfMemoryError: Dataset terlalu besar
```

Solusi: Batasi ukuran dataset atau gunakan data sampling

SAMPLE DATA UNTUK TESTING

Jika ingin testing tanpa download dataset lengkap, gunakan sample ini:

```python
# Sample data untuk testing cepat
sample_files = [
    "common_voice_en_41227191.mp3",
    "common_voice_en_41227192.mp3",
    "common_voice_en_41227193.mp3"
]

sample_texts = [
    "the quick brown fox jumps over the lazy dog",
    "hello world this is a test",
    "machine learning is fascinating"
]
```

KESIMPULAN

Mozilla Common Voice Dataset adalah pilihan ideal untuk proyek Speech Recognition karena:

- ✅ **Open Source:** Gratis dan legal untuk digunakan
- ✅ **High Quality:** Data sudah divalidasi komunitas
- ✅ **Diverse:** Mencakup berbagai aksen dan demografi
- ✅ **Well-Structured:** Format dan metadata yang jelas
- ✅ **Active Maintenance:** Reguler update dan perbaikan
- ✅ **Community Support:** Dokumentasi dan forum yang aktif

Dataset ini memungkinkan pencapaian target WER < 30% dengan proper preprocessing dan fine-tuning model Wav2Vec2.

REFERENSI

1. **Mozilla Common Voice Paper:** https://arxiv.org/abs/1912.06670
2. **Dataset Official Page:** https://commonvoice.mozilla.org/
3. **Wav2Vec2 Paper:** https://arxiv.org/abs/2006.11477
4. **Hugging Face Documentation:** https://huggingface.co/transformers/
5. **Speech Recognition Best Practices:** Various academic papers
6. **Keseluruhan Dataset:** https://drive.google.com/drive/folders/1YXHwGj04-BuaxOeNwxLBeHq6fpT73TIQ?usp=sharing