

Nama : Azriel Dzaki Ravideva

NIM : 24917032

## Laporan Proyek

### 1. Pendahuluan

#### 1.1 Latar Belakang

Dalam industri perfilman, klasifikasi genre sangat krusial untuk sistem rekomendasi dan pengarsipan data. Namun, sebuah film jarang hanya memiliki satu genre (misalnya: sebuah film bisa berupa Action sekaligus Thriller). Hal ini memunculkan tantangan dalam Multi-label Text Classification. Penggunaan Large Language Model (LLM) seperti Qwen3-0.6B menawarkan kemampuan pemahaman semantik yang lebih dalam dibandingkan metode tradisional untuk memahami konteks sinopsis film yang kompleks.

#### 1.2 Pembahasan Masalah

Masalah utama dalam tugas ini adalah bagaimana melakukan adaptasi (fine-tuning) pada model bahasa berukuran besar agar dapat memprediksi beberapa label sekaligus secara akurat. Kendala yang dihadapi meliputi:

- **Ketidakeimbangan Data:** Distribusi genre yang tidak merata (beberapa genre sangat dominan, sementara yang lain langka).
- **Efisiensi Komputasi:** Melakukan full fine-tuning pada LLM memerlukan sumber daya besar, sehingga diperlukan pendekatan Parameter-Efficient Fine-Tuning (PEFT).

### 2. Deskripsi Dataset

Dataset yang digunakan adalah 25k IMDb Movie Dataset.

- **Ukuran Dataset:** Diambil sampel sebanyak 5.000 data untuk efisiensi eksperimen.
- **Fitur:** overview (sinopsis film) sebagai input teks.
- **Target:** genres (multi-label) sebagai output.
- **Jumlah Kelas:** Terdapat total 23 variasi label (setelah proses binarisasi), mencakup genre utama seperti Drama, Action, Comedy, hingga genre spesifik.

### 3. Preprocessing dan Tokenisasi

Langkah-langkah pengolahan data meliputi:

1. **Pembersihan Data:** Menghapus baris dengan nilai null dan string kosong.
2. **Normalisasi Genre:** Mengubah format genre (misal: "Action, Drama") menjadi format pipa ("Action|Drama") untuk memudahkan parsing.
3. **Multi-label Binarization:** Menggunakan MultiLabelBinarizer untuk mengubah label teks menjadi vektor biner (0 dan 1).
4. **Splitting:** Data dibagi menjadi:
  - **Training Set:** 70% (3.500 data)
  - **Validation Set:** 15% (750 data)

- **Testing Set:** 15% (750 data)
5. **Tokenisasi:** Menggunakan tokenizer bawaan Qwen/Qwen3-0.6B dengan max\_length=128 dan padding ke EOS token.

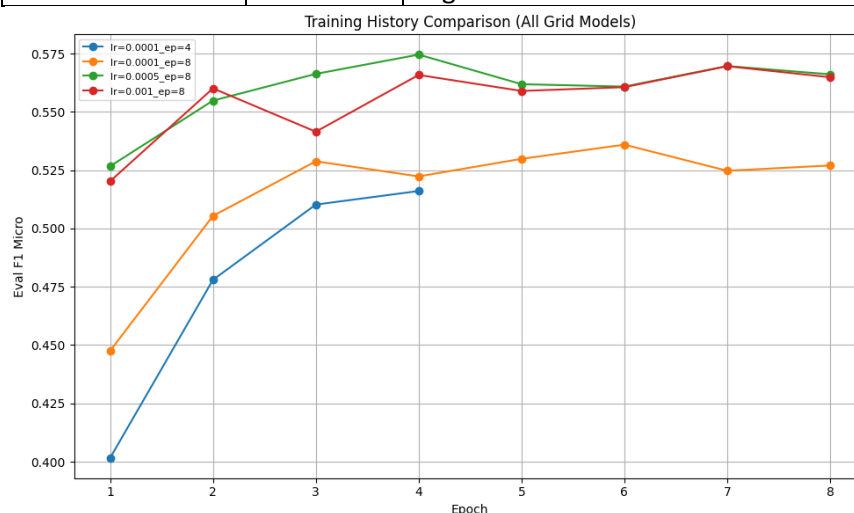
#### 4. Pemilihan Model dan Arsitektur

- **Model Dasar:** Qwen/Qwen3-0.6B. Model ini dipilih karena ukurannya yang ringan (0.6 miliar parameter) namun memiliki performa bahasa yang kuat.
- **Metode Fine-tuning: LoRA (Low-Rank Adaptation).**
  - **Justifikasi:** LoRA memungkinkan untuk melatinkan Model secara Sebagian sehingga menghemat waktu dan computing resource.
  - **Konfigurasi LoRA:** r=8, lora\_alpha=16, menargetkan modul q\_proj dan v\_proj.
- **Custom Head:** Menambahkan lapisan linear (nn.Linear) di atas hidden state terakhir untuk memetakan output LLM ke jumlah label genre (23 kelas).

#### 5. Eksperimen Hyperparameter Tuning

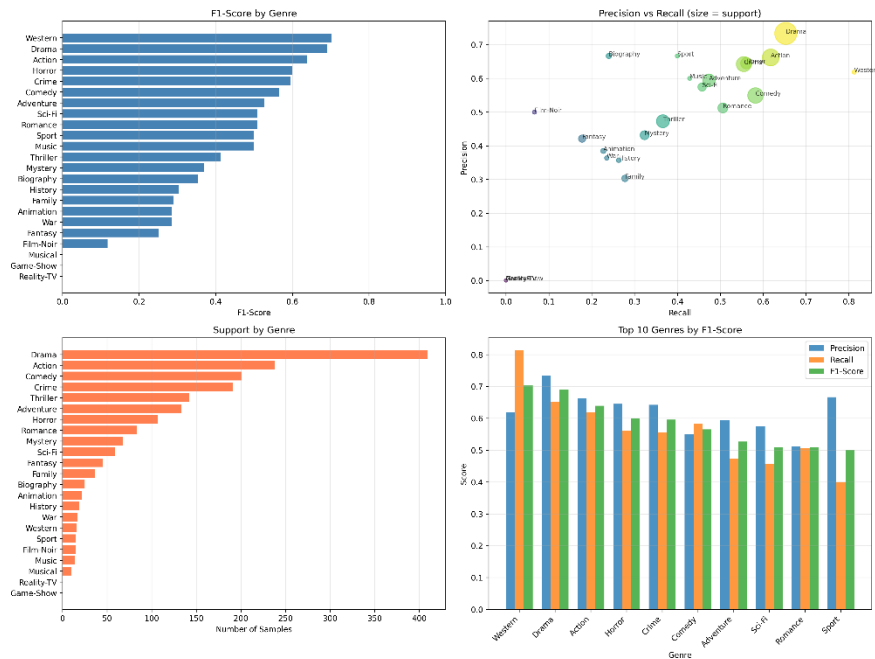
Eksperimen dilakukan dengan nilai berikut:

Hyperparameter	Nilai	Justifikasi
<b>Learning Rate</b>	1E-04, 5E-04, 1E-05	Dilakukan Pencarian menggunakan Grid search untuk melihat learning rate terbaik.
<b>Batch Size</b>	8, 12, 16	Menyesuaikan pemakaian RAM dan CPU. Ditemukan bahwa untuk token max length 128 learning paling cepat berada di batch size 16.
<b>Number of Epochs</b>	4, 8	Epoch 4 dinilai kurang sehingga pelatihan diperpanjang sampai 8 epoch.
<b>Optimizer</b>	AdamW	Memiliki mekanisme weight decay (0.01) yang baik untuk regularisasi model NLP.



Terlihat pada grafik diatas. Bahwa skor F1 Mikro terbaik didapatkan menggunakan lr=5E-04 dan ep=8

#### 6. Hasil dan Visualisasi



Setelah dilakukan evaluasi pada test set, didapatkan metrik sebagai berikut:

### 6.1 Metrik Keseluruhan

- **Micro F1-Score:** 0.56
- **Precision (Micro):** 0.61
- **Recall (Micro):** 0.52
- **Hamming Loss:** 0.0886 (Menunjukkan kesalahan klasifikasi per label cukup rendah secara rata-rata).

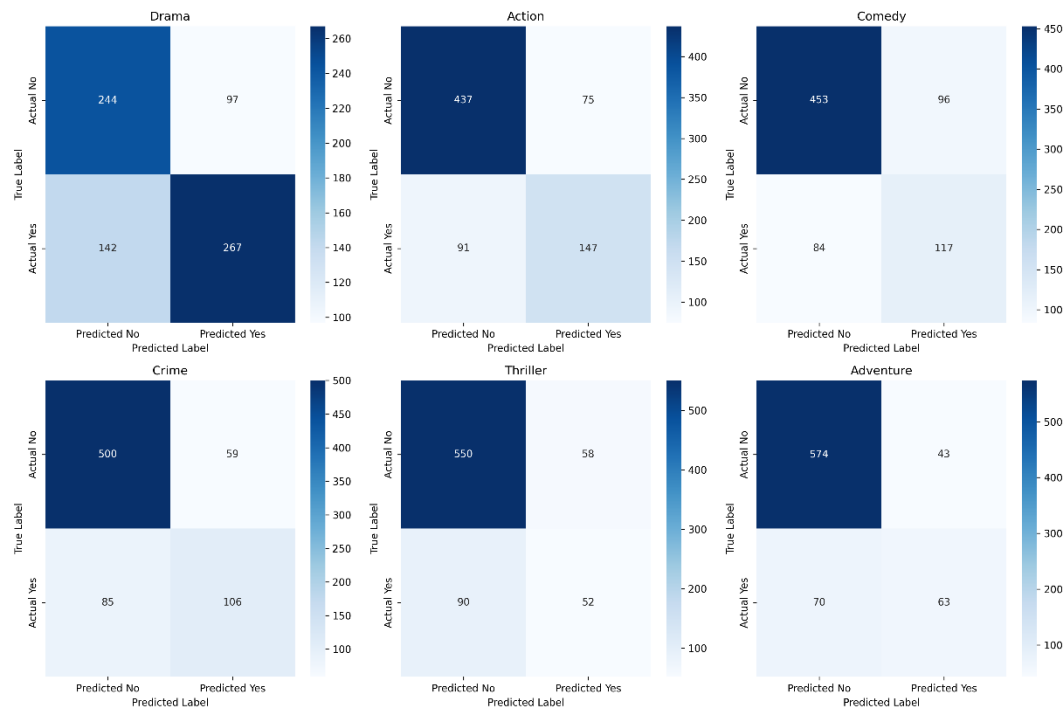
### 6.2 Analisis Per-Class

Berdasarkan hasil evaluasi, model bekerja paling baik pada genre dengan dukungan data (support) yang besar:

- **Western:** F1-Score 0.70
- **Drama:** F1-Score 0.69
- **Action:** F1-Score 0.64

Namun, model kesulitan pada genre langka (seperti Reality-TV, Game-Show, dsb) yang mendapatkan skor 0. Ini disebabkan oleh class imbalance yang ekstrem pada dataset IMDB.

### 6.3 Visualisasi Confusion Matrix



Confusion matrix menunjukkan bahwa model cenderung sangat konservatif dalam memberikan prediksi "Yes" (True Positive rendah) karena ambang batas sigmoid 0.5 dan kurangnya data pelatihan untuk genre spesifik.

## 7. Analisis dan Kesimpulan

### 7.1 Analisis Hasil

1. **Kemampuan Ekstraksi:** Model berhasil menangkap fitur untuk genre umum (Action, Comedy) dengan presisi yang cukup baik ( $>0.5$ ).
2. **Kelemahan:** Model memiliki Recall yang rendah artinya banyak genre yang "luput" diprediksi. Hal ini dikarenakan model membutuhkan lebih banyak data atau durasi training yang lebih lama untuk memahami genre yang jarang muncul.
3. **Efisiensi:** Implementasi LoRA terbukti berhasil menjalankan proses training pada model 0.6B dengan penggunaan memori yang minimal.

### 7.2 Kesimpulan

Fine-tuning Qwen3-0.6B dengan LoRA cukup efektif untuk klasifikasi teks multi-label. Untuk meningkatkan performa di masa depan, disarankan untuk menerapkan teknik Oversampling pada genre minoritas atau menggunakan fungsi rugi (loss function) yang berbobot (Weighted BCE Loss) untuk menangani ketidakseimbangan data.