

ABSTRAK

Joshua Yordana:

Skripsi

Prediksi Genre dan Rating Penonton terhadap Sinopsis Film Menggunakan Bidirectional Encoder Representations from Transformers (BERT)

Proses penulisan dan pengembangan sinopsis film membutuhkan evaluasi yang mendalam agar dapat sesuai dengan minat pasar. Penelitian ini bertujuan untuk membandingkan akurasi dua model kecerdasan buatan, yaitu IndoBERT dan Indonesian RoBERTa, dalam memprediksi genre dan rating penonton terhadap sinopsis film. Model ini diimplementasikan dalam sebuah aplikasi web yang dikembangkan menggunakan bahasa pemrograman Python dan framework Flask.

Dataset yang digunakan dalam penelitian ini terdiri dari 1.824 sinopsis film berbahasa Indonesia yang diambil dari API The Movie Database (TMDb) dan Kaggle. Setiap film memiliki atribut seperti judul, rating, genre, dan sinopsis. Data diolah menggunakan Auto Tokenizer dari *huggingface*.

Hasil pengujian menunjukkan bahwa kedua model mampu memprediksi genre dan rating dengan akurasi yang tinggi. Untuk prediksi rating, model IndoBERT menunjukkan sedikit keunggulan dengan rata-rata perbedaan sebesar 0.4 dari rating asli. Untuk prediksi genre, kedua model menunjukkan hasil yang hampir identik dengan genre sebenarnya.

Aplikasi ini diuji oleh produser, sutradara, dan penulis naskah film indie melalui survei yang menunjukkan bahwa mereka menemukan aplikasi ini bermanfaat dan akurat dalam membantu evaluasi sinopsis film. Dengan nilai rata-rata 4.333 untuk akurasi prediksi genre dan 4.5 untuk prediksi rating menggunakan model IndoBERT, aplikasi ini terbukti dapat memberikan rekomendasi yang relevan dan berguna bagi para profesional di industri film.

Kata kunci:

IndoBERT, Indonesian RoBERTa, prediksi genre, prediksi rating, sinopsis film, aplikasi web

ABSTRACT

Joshua Yordana:

Undergraduate Thesis

Prediction of Genre and Audience Rating for Movie Synopsis Using Bidirectional Encoder Representations from Transformers (BERT)

The process of writing and developing movie synopsis requires thorough evaluation to align with market interests. This study aims to compare the accuracy of two artificial intelligence models, namely IndoBERT and Indonesian RoBERTa, in predicting the genre and audience rating of movie synopsis. These models are implemented in a web application developed using the Python programming language and the Flask framework.

The dataset used in this study consists of 1,824 Indonesian movie synopsis sourced from The Movie Database (TMDb) API and Kaggle. Each movie has attributes such as title, rating, genre, and synopsis. The data is processed using the Auto Tokenizer from Huggingface.

The testing results show that both models are capable of predicting genre and rating with high accuracy. For rating prediction, the IndoBERT model shows a slight advantage with an average difference of 0.4 from the actual rating. For genre prediction, both models demonstrate results that are almost identical to the actual genres.

The application was tested by producers, directors, and indie film script writers through a survey, which showed that they found the application useful and accurate in helping to evaluate movie synopsis. With an average score of 4.333 for genre prediction accuracy and 4.5 for rating prediction using the IndoBERT model, the application proves to be able to provide relevant and useful recommendations for professionals in the film industry.

Keywords:

IndoBERT, Indonesian RoBERTa, genre prediction, rating prediction, movie synopsis, web application

DAFTAR ISI

HALAMAN JUDUL.....	1
LEMBAR PENGESAHAN.....	2
KATA PENGANTAR.....	4
ABSTRAK.....	5
ABSTRACT.....	6
DAFTAR ISI.....	7
DAFTAR GAMBAR.....	10
DAFTAR SEGMENT PROGRAM.....	11
DAFTAR TABEL.....	12
1. PENDAHULUAN.....	1
1.1 Latar Belakang Masalah.....	1
1.2 Perumusan masalah.....	3
1.3 Tujuan Penelitian.....	3
1.4 Ruang Lingkup.....	3
1.5 Metodologi Penelitian.....	4
1.6 Sistematika Penulisan.....	5
2. LANDASAN TEORI.....	6
2.1 Tinjauan Pustaka.....	6
2.1.1 Peringkat TMDb.....	6
2.1.2 Bidirectional Encoder Representations from Transformers (BERT).....	6
2.1.3 IndoBERT.....	8
2.1.4 Indonesian RoBERTa.....	8
2.1.5 Confusion Matrix.....	9
2.1.6 Mean Absolute Error (MAE).....	10
2.2 Tinjauan Studi.....	11
3. ANALISIS DAN DESAIN SISTEM.....	14
3.1 Analisis Permasalahan dan Kebutuhan.....	14
3.2 Analisis Data.....	15
3.3 Desain Sistem.....	18
3.3.1 Model.....	18
3.3.2 Scraping Data.....	21
3.3.3 Data Cleaning dan Preprocessing.....	21
3.3.4 Tokenization.....	23
3.4 Desain Website.....	25
4. IMPLEMENTASI SISTEM.....	26
4.1 Implementasi Perangkat Lunak yang Digunakan.....	26
4.2 Model Vote Average.....	26
4.2.1 Library yang Digunakan.....	26

4.2.2 Custom Trainer Class dan Cek GPU.....	27
4.2.3 Load Data dan Preprocessing.....	27
4.2.4 Load Tokenizer dan Model.....	27
4.2.5 Konversi dan Tokenisasi Dataset.....	28
4.2.6 Definisi Argumen Pelatihan & Inisialisasi Variabel Evaluasi.....	28
4.2.7 Pelaksanaan K-fold Cross-Validation.....	29
4.2.8 Hitung Rata-rata Hasil Evaluasi.....	29
4.3 Model Genre.....	30
4.3.1 Library yang Digunakan.....	30
4.3.2 Definisi Fungsi.....	30
4.3.3 Load Data dan Preprocessing.....	30
4.3.4 Definisi KFold Cross-Validator.....	31
4.3.5 Load Tokenizer dan Model.....	31
4.3.6 Training Loop.....	32
4.3.7 Menyimpan Model Terbaik.....	35
4.4 Website.....	36
4.4.1 Library yang Digunakan.....	36
4.4.2 Inisialisasi Flask dan CORS.....	36
4.4.3 Menambahkan header CORS.....	36
4.4.4 Load Tokenizer dan Model.....	36
4.4.5 Definisi Threshold untuk Prediksi.....	37
4.4.6 Preprocessing Input.....	37
4.4.7 Route utama aplikasi Flask.....	37
4.4.8 Route untuk prediksi rating.....	37
4.4.9 Route untuk prediksi genre.....	38
4.4.10 Menjalankan Aplikasi.....	40
4.4.11 HTML Header dan Metadata.....	40
4.4.12 Form Input Sinopsis.....	41
4.4.13 Tabel Hasil Prediksi.....	42
4.4.14 Program Javascript.....	43
5. PENGUJIAN.....	46
5.1 Perangkat Lunak yang Digunakan.....	46
5.2 Tujuan Pengujian.....	46
5.3 Pengujian Model.....	46
5.3.1 Pengujian IndoBERT untuk prediksi Vote Average.....	47
5.3.2 Pengujian Indonesian RoBERTa untuk prediksi Vote Average.....	49
5.3.3 Pengujian IndoBERT untuk prediksi Genre Film.....	50
5.3.4 Pengujian Indonesian RoBERT untuk prediksi Genre Film.....	52
5.4 Pengujian Aplikasi.....	54
5.5 Pengujian oleh User.....	56
6. KESIMPULAN DAN SARAN.....	63
6.1 Kesimpulan.....	63

6.2 Saran.....	64
DAFTAR REFERENSI.....	65

DAFTAR GAMBAR

Gambar 2.1 Struktur model BERT.....	7
Gambar 2.2 Perbedaan BERT Base dan BERT Large.....	8
Gambar 2.3 Visualisasi Confusion Matrix.....	10
Gambar 3.1 Bentuk dataset.....	16
Gambar 3.2 Distribusi Vote Average.....	17
Gambar 3.3 Analisis Vote Average.....	17
Gambar 3.4 Distribusi Genre.....	18
Gambar 3.5 Flowchart Training Model.....	19
Gambar 3.6 Flowchart Data Preprocessing.....	22
Gambar 3.7 Design UI.....	25
Gambar 5.1 Testing website.....	54
Gambar 5.2 Alur penggunaan website.....	58

DAFTAR SEGMENT PROGRAM

Segmen Program 4.2.1 Import Library.....	26
Segmen Program 4.2.2 Class Custom Trainer dan pengecekan GPU.....	27
Segmen Program 4.2.3 Load Data dan Preprocessing.....	27
Segmen Program 4.2.4 Load Tokenizer dan Model untuk IndoBERT.....	27
Segmen Program 4.2.5 Load Tokenizer dan Model untuk Indonesian RoBERTa.....	28
Segmen Program 4.2.6 Konversi dan Tokenisasi.....	28
Segmen Program 4.2.7 Inisialisasi variabel pelatihan.....	28
Segmen Program 4.2.8 Training Model.....	29
Segmen Program 4.2.9 Hitung MAE.....	30
Segmen Program 4.3.1 Import Library.....	30
Segmen Program 4.3.2 Class Custom Trainer dan pengecekan GPU.....	30
Segmen Program 4.3.3 Load Data dan Preprocessing.....	31
Segmen Program 4.3.4 Load Data dan Preprocessing.....	31
Segmen Program 4.3.5 Load Tokenizer dan Model untuk IndoBERT.....	32
Segmen Program 4.3.6 Load Tokenizer dan Model untuk Indonesian RoBERTa.....	32
Segmen Program 4.3.7 Konversi dan Tokenisasi.....	35
Segmen Program 4.3.8 Inisialisasi variabel pelatihan.....	35
Segmen Program 4.4.1 Import Library.....	36
Segmen Program 4.4.2 Inisialisasi Flask dan set CORS.....	36
Segmen Program 4.4.3 Header CORS.....	36
Segmen Program 4.4.4 Memuat Model.....	37
Segmen Program 4.4.5 Set ambang batas.....	37
Segmen Program 4.4.6 Trim Whitespace.....	37
Segmen Program 4.4.7 Route untuk halaman utama.....	37
Segmen Program 4.4.8 Route untuk prediksi vote_average.....	38
Segmen Program 4.4.9 Route untuk prediksi genre.....	40
Segmen Program 4.4.10 Run aplikasi dengan SSL.....	40
Segmen Program 4.4.11 Header HTML.....	41
Segmen Program 4.4.12 Form HTML.....	41
Segmen Program 4.4.13 Tabel hasil prediksi.....	42
Segmen Program 4.4.14 Code Javascript.....	45

DAFTAR TABEL

Tabel 5.1 Hasil prediksi IndoBERT untuk vote average.....	48
Tabel 5.2 Hasil prediksi Indonesian RoBERTa untuk vote average.....	49
Tabel 5.3 Hasil prediksi IndoBERT untuk genre.....	51
Tabel 5.4 Hasil prediksi Indonesian RoBERTa untuk genre.....	53
Tabel 5.5 Hasil prediksi vote average.....	55
Tabel 5.6 Hasil prediksi Genre.....	55
Tabel 5.7 Penilaian pengguna terhadap kemudahan, kejelasan, dan kecepatan aplikasi.....	60
Tabel 5.8 Penilaian pengguna terhadap akurasi masing-masing model.....	61
Tabel 5.9 Kelebihan, kekurangan, dan saran dari pengguna.....	61
Tabel 5.10 Tingkat rekomendasi aplikasi pada rekan industri.....	62