

5. PENGUJIAN

Pada bab ini, akan dijelaskan implementasi dan pengujian sistem analisis sinopsis film menggunakan model pembelajaran mesin. Proses pengembangan dilakukan pada PC dengan spesifikasi prosesor Intel Core i9-11900, RAM 32GB, dan kartu grafis NVIDIA GeForce RTX 3070 Ti, yang memungkinkan pelatihan model berjalan efisien. Untuk deployment, aplikasi di-hosting pada VPS dari NiagaHoster dengan spesifikasi CPU 1 core, memori 4 GB, bandwidth 4 TB, dan ruang penyimpanan 50 GB, yang memadai untuk menjalankan aplikasi secara online.

5.1 Perangkat Lunak yang Digunakan

Pengembangan website untuk menganalisis sinopsis film dan memprediksi rating serta genre dilakukan dengan menggunakan library Python. Pengujian data training menggunakan Python dengan bantuan library Flask, jQuery, dan Bootstrap, serta semua proses ini dilakukan menggunakan Visual Studio Code.

5.2 Tujuan Pengujian

Dalam pengujian ini, fokus utama adalah untuk mengevaluasi akurasi prediksi yang dihasilkan oleh model AI menggunakan metode IndoBERT dan Indonesian RoBERTa dalam memprediksi genre dan rating film berdasarkan sinopsis. Pengujian akan menggunakan data uji yang telah disiapkan, di mana hasil prediksi akan dibandingkan dengan label sebenarnya untuk menilai tingkat akurasi dari masing-masing model.

Selain itu, penting untuk memperhatikan kinerja sistem saat diimplementasikan pada lingkungan produksi yang menggunakan VPS dengan spesifikasi yang lebih rendah. Evaluasi kinerja ini meliputi waktu respons sistem terhadap permintaan pengguna dan stabilitas sistem saat digunakan secara bersamaan oleh beberapa pengguna. Hal ini dilakukan untuk memastikan bahwa sistem dapat beroperasi secara efisien dan dapat diandalkan di lingkungan produksi yang sesungguhnya.

5.3 Pengujian Model

Dalam bab ini, dilakukan pengujian terhadap beberapa model yang telah diimplementasikan dalam proyek, termasuk IndoBERT dan Indonesian RoBERTa. Model-model

tersebut dipilih karena telah terbukti efektif dalam tugas yang serupa. Pengujian dilakukan dengan berbagai parameter yang telah ditetapkan sebelumnya, seperti `max_length=250`, `output_dir="/kfold_bert"`, `num_train_epochs=20`, `per_device_train_batch_size=16`, `per_device_eval_batch_size=16`, `warmup_steps=500`, dan `evaluation_strategy="epoch"`. Total data yang digunakan dalam pengujian ini mencapai 2273 data, yang diharapkan mampu memberikan representasi yang cukup dari berbagai variasi sinopsis cerita yang ada. Metrik evaluasi yang digunakan dalam pengujian ini adalah MAE (Mean Absolute Error) dan Jaccard Similarity.

Dalam pengujian ini, metode 5-cross fold validation digunakan untuk membagi data menjadi subset data latih dan data uji. Pengujian parameter adalah tahap eksperimen untuk menentukan nilai-nilai optimal dari berbagai parameter yang mempengaruhi kinerja model deep learning, dengan tujuan meningkatkan kinerja model melalui kombinasi parameter terbaik. K-fold cross-validation bertujuan untuk memberikan evaluasi yang konsisten dan objektif terhadap performa model dengan membagi dataset menjadi k subset, melatih model pada k-1 subset, dan menguji pada subset yang tersisa. Ini membantu mendeteksi overfitting, memaksimalkan penggunaan data, dan memberikan gambaran yang akurat tentang kemampuan model dalam generalisasi ke data baru.

Tujuan dari pengujian parameter tuning adalah untuk meningkatkan performa model dengan menemukan nilai-nilai parameter yang optimal. Ini dilakukan dengan mencoba berbagai kombinasi parameter untuk menemukan pengaturan terbaik yang menghasilkan kinerja model yang optimal pada data yang tidak terlihat. Proses tuning parameter sangat penting untuk memastikan bahwa model yang dihasilkan dapat memberikan hasil yang maksimal dalam konteks penggunaannya yang sesungguhnya.

5.3.1 Pengujian IndoBERT untuk prediksi *Vote Average*

Pada penelitian ini, model regresi diterapkan dengan menggunakan model IndoBERT untuk memprediksi nilai rata-rata penilaian film berdasarkan deskripsi film (overview). Dataset yang digunakan terdiri dari kolom 'overview' dan 'vote_average', di mana data dengan nilai rata-rata penilaian nol dan data null telah dihapus untuk menjaga kualitas data. Proses tokenisasi dilakukan dengan menggunakan tokenizer dari model "indolem/indobert-base-uncased" dengan panjang maksimum urutan (`max_seq_length`) 250

token. Model dilatih dan dievaluasi pada setiap lipatan menggunakan kelas CustomTrainer, di mana Mean Squared Error (MSE) digunakan sebagai fungsi loss utama. Mean Absolute Error (MAE) dihitung untuk setiap lipatan sebagai metrik evaluasi utama. Model yang memberikan MAE terendah selama proses validasi silang dianggap sebagai model terbaik, yang kemudian disimpan untuk langkah-langkah selanjutnya.

Tabel 5.1 Hasil prediksi IndoBERT untuk vote average

Epoch	Fold 1 MAE	Fold 1 Loss	Fold 2 MAE	Fold 2 Loss	Fold 3 MAE	Fold 3 Loss	Fold 4 MAE	Fold 4 Loss	Fold 5 MAE	Fold 5 Loss
1	0.4771	0.88362	0.46931	0.87816	0.46026	0.86989	0.46997	0.87891	0.48034	0.88755
2	0.44623	0.85073	0.44047	0.84558	0.43186	0.83762	0.44133	0.84646	0.45113	0.8548
3	0.43155	0.83527	0.42685	0.83027	0.41869	0.82244	0.42765	0.83107	0.43702	0.83988
4	0.41977	0.82172	0.41507	0.8169	0.40728	0.80923	0.41573	0.81782	0.42519	0.8263
5	0.41202	0.81306	0.40737	0.80843	0.39971	0.80097	0.40801	0.80967	0.41735	0.8183
6	0.40616	0.80633	0.40154	0.80193	0.3938	0.79447	0.40223	0.80315	0.41121	0.81184
7	0.40178	0.80104	0.39722	0.79671	0.38921	0.78936	0.39781	0.79785	0.40676	0.80692
8	0.39814	0.79651	0.39364	0.7922	0.38538	0.785	0.39408	0.7936	0.40289	0.80244
9	0.39515	0.79264	0.39069	0.78839	0.38216	0.78124	0.39099	0.78988	0.39968	0.79861
10	0.39272	0.78937	0.38829	0.78516	0.37951	0.77796	0.38843	0.78666	0.39697	0.79528
11	0.39075	0.78662	0.38635	0.7824	0.37737	0.77509	0.38631	0.78383	0.39464	0.79242
12	0.38916	0.78432	0.3848	0.78001	0.37568	0.77257	0.38458	0.78138	0.39261	0.79
13	0.38789	0.78243	0.38357	0.77793	0.37439	0.77042	0.3832	0.77929	0.39085	0.78788
14	0.38688	0.78088	0.38259	0.7761	0.37344	0.76856	0.38211	0.77747	0.3893	0.78603
15	0.38608	0.77964	0.38181	0.77451	0.37279	0.76696	0.38128	0.77589	0.38793	0.78442
16	0.38545	0.77866	0.38119	0.77313	0.3724	0.76558	0.38068	0.77452	0.3867	0.78302
17	0.38496	0.7779	0.3807	0.77192	0.37214	0.76438	0.38028	0.77331	0.38558	0.78179
18	0.38458	0.77734	0.38033	0.77087	0.37197	0.76334	0.38001	0.77227	0.38455	0.78071
19	0.3843	0.77695	0.38005	0.76995	0.37187	0.76243	0.3798	0.77134	0.3836	0.77976
20	0.38409	0.7767	0.37985	0.76915	0.37183	0.76164	0.37964	0.77054	0.38271	0.77891

Berdasarkan hasil evaluasi pada setiap epoch dan fold, model terbaik dipilih dari Fold 3 pada Epoch 20. Pada fold ini, model mencapai Mean Absolute Error (MAE) terendah sebesar 0.37183, menunjukkan bahwa model pada fold ini memberikan prediksi paling akurat untuk

nilai `vote_average` berdasarkan deskripsi film. Performanya juga konsisten dengan penurunan nilai loss yang signifikan hingga epoch terakhir, mengindikasikan bahwa model telah terlatih dengan baik dan mampu meminimalkan kesalahan prediksi.

Sehingga, model dari Fold 3 pada Epoch 20 adalah yang paling unggul karena memiliki MAE terendah, yang mencerminkan kemampuan terbaik dalam memprediksi data pada validasi silang. Pemilihan model ini didukung oleh stabilitas dan konsistensi penurunan error, membuatnya menjadi kandidat yang optimal untuk digunakan dalam aplikasi saya.

5.3.2 Pengujian Indonesian RoBERTa untuk prediksi *Vote Average*

Pada penelitian ini, model regresi diterapkan dengan menggunakan model Indonesian RoBERTa untuk memprediksi nilai rata-rata penilaian film berdasarkan deskripsi film (overview). Dataset yang digunakan terdiri dari kolom 'overview' dan 'vote_average', di mana data dengan nilai rata-rata penilaian nol dan data null telah dihapus untuk menjaga kualitas data. Proses tokenisasi dilakukan dengan menggunakan tokenizer dari model "cahya/roberta-base-indonesian-522M" dengan panjang maksimum urutan (max_seq_length) 250 token. Model dilatih dan dievaluasi pada setiap lipatan menggunakan kelas CustomTrainer, di mana Mean Squared Error (MSE) digunakan sebagai fungsi loss utama. Mean Absolute Error (MAE) dihitung untuk setiap lipatan sebagai metrik evaluasi utama. Model yang memberikan MAE terendah selama proses validasi silang dianggap sebagai model terbaik, yang kemudian disimpan untuk langkah-langkah selanjutnya.

Tabel 5.2 Hasil prediksi Indonesian RoBERTa untuk vote average

Epoch	Fold 1 MAE	Fold 1 Loss	Fold 2 MAE	Fold 2 Loss	Fold 3 MAE	Fold 3 Loss	Fold 4 MAE	Fold 4 Loss	Fold 5 MAE	Fold 5 Loss
1	0.78278	1.08754	0.16094	0.52736	0.20569	0.43035	0.17492	0.41792	0.17528	0.43659
2	0.78981	1.07895	0.15369	0.49369	0.20889	0.42155	0.17434	0.40039	0.18022	0.42932
3	0.81875	1.06969	0.15327	0.48515	0.20618	0.41966	0.17075	0.38796	0.17914	0.42819
4	0.79468	1.07822	0.15312	0.47553	0.20754	0.41588	0.1763	0.38436	0.18089	0.42619
5	0.89153	1.06923	0.15466	0.46473	0.20725	0.41889	0.16994	0.37589	0.17964	0.42192
6	0.84697	1.07885	0.15497	0.45777	0.20454	0.41923	0.1723	0.37712	0.17858	0.42431
7	0.88065	1.07116	0.15442	0.45118	0.20458	0.41877	0.17406	0.37421	0.17954	0.42087
8	0.84974	1.07763	0.15508	0.44532	0.20403	0.41958	0.17621	0.37155	0.17812	0.42188
9	0.82439	1.07321	0.15526	0.43955	0.20389	0.42045	0.17295	0.36894	0.17961	0.41964

10	0.82719	1.07568	0.15545	0.43417	0.20382	0.42108	0.17435	0.36612	0.17748	0.41845
11	0.85384	1.07348	0.15564	0.42911	0.20433	0.42256	0.17528	0.36398	0.17795	0.41777
12	0.86891	1.07189	0.15603	0.42436	0.20356	0.42314	0.17456	0.36144	0.17725	0.41666
13	0.84874	1.07512	0.15622	0.42017	0.20325	0.42404	0.17576	0.35911	0.17816	0.41553
14	0.87825	1.07089	0.15635	0.41621	0.2031	0.42479	0.17344	0.35689	0.17798	0.41486
15	0.86521	1.07213	0.15654	0.41252	0.20278	0.42542	0.17512	0.35488	0.17682	0.4141
16	0.86233	1.07347	0.15669	0.40905	0.20266	0.42597	0.17501	0.35285	0.17654	0.41341
17	0.86419	1.07387	0.15685	0.40578	0.20234	0.42656	0.17427	0.35112	0.17621	0.41274
18	0.85439	1.07459	0.15707	0.40271	0.20212	0.42715	0.17592	0.34955	0.17647	0.41199
19	0.86038	1.07276	0.15728	0.39982	0.20199	0.42775	0.17395	0.34808	0.17662	0.41123
20	0.85766	1.07319	0.15746	0.39707	0.20178	0.42834	0.17492	0.34661	0.17598	0.41061

Berdasarkan hasil evaluasi k-fold cross-validation pada model Roberta, terlihat bahwa performa model bervariasi antar fold dan antar epoch. Model terbaik diambil berdasarkan performa pada epoch terakhir (epoch ke-20) dari setiap fold. Dari hasil ini, model pada Fold 2 menunjukkan performa terbaik dengan MAE terendah sebesar 0.1574 dan Loss sebesar 0.39707. Hal ini menunjukkan bahwa model pada fold ini memiliki kemampuan prediksi yang lebih akurat dengan kesalahan rata-rata yang lebih kecil dibandingkan dengan fold lainnya.

Dalam memilih model terbaik, fokus hanya pada performa epoch terakhir penting karena menggambarkan kemampuan model setelah pelatihan penuh. Pengambilan keputusan berdasarkan fold terakhir memungkinkan evaluasi pada titik yang seimbang, menghindari risiko overfitting yang sering terjadi pada epoch sebelumnya. Oleh karena itu, model dari Fold 2 pada epoch ke-20 dipilih sebagai model terbaik untuk tugas regresi ini.

5.3.3 Pengujian IndoBERT untuk prediksi Genre Film

Model yang digunakan dalam proyek ini adalah "indolem/indobert-base-uncased", sebuah model BERT yang telah di-pretrain untuk Bahasa Indonesia. Model ini dirancang untuk memahami konteks bahasa Indonesia secara mendalam dan dapat diaplikasikan pada berbagai tugas pemrosesan bahasa alami (NLP), termasuk klasifikasi teks. Pemodelan berbasis BERT ini memanfaatkan arsitektur Transformer, yang unggul dalam menangkap dependensi panjang dalam teks, menjadikannya sangat efektif untuk tugas klasifikasi teks.

Proyek ini menerapkan *multilabel classification* untuk menentukan genre film berdasarkan sinopsisnya. Berbeda dengan klasifikasi tunggal, *multilabel classification* memungkinkan setiap sinopsis memiliki lebih dari satu genre yang sesuai. Untuk mencapai hal ini, proyek menggunakan *one-hot encoding* pada label genre. *One-hot encoding* mengubah daftar genre setiap sinopsis menjadi vektor biner yang menunjukkan kehadiran atau ketidakhadiran setiap genre di dalam daftar genre unik. Misalnya, jika ada 10 genre unik, vektor biner akan memiliki panjang 10, dengan nilai 1 menunjukkan genre tersebut ada pada sinopsis, dan nilai 0 sebaliknya.

Dalam konteks output model, proyek ini memanfaatkan *sigmoid activation* pada lapisan terakhir. Tidak seperti *softmax* yang digunakan dalam klasifikasi kategori tunggal, *sigmoid activation* memungkinkan model menghasilkan probabilitas terpisah untuk setiap label (genre). Setiap probabilitas menunjukkan seberapa besar keyakinan model bahwa genre tersebut sesuai untuk sinopsis. Probabilitas ini kemudian diubah menjadi prediksi biner menggunakan ambang batas (*threshold*) tertentu, umumnya 0.5. Dengan demikian, setiap sinopsis dapat diklasifikasikan ke dalam satu atau lebih genre berdasarkan probabilitas yang dihasilkan, yang merupakan pendekatan efektif untuk mengatasi tugas *multilabel*.

Tabel 5.3 Hasil prediksi IndoBERT untuk genre

Epoch	Fold 1 Jaccard	Fold 1 Loss	Fold 2 Jaccard	Fold 2 Loss	Fold 3 Jaccard	Fold 3 Loss	Fold 4 Jaccard	Fold 4 Loss	Fold 5 Jaccard	Fold 5 Loss
1	0.0315	0.3298	0.8051	0.0503	0.975	0.0102	0.9933	0.0034	0.9956	0.0025
2	0.1671	0.2771	0.8628	0.0273	0.8984	0.0211	0.9794	0.0136	0.9867	0.003
3	0.356	0.2537	0.926	0.0289	0.9172	0.0102	0.9507	0.0086	0.9694	0.0092
4	0.4937	0.2346	0.9665	0.0188	0.9639	0.0099	0.9404	0.0104	0.95	0.0124
5	0.6029	0.2318	0.9816	0.0168	0.9822	0.008	0.9563	0.0122	0.9409	0.0085
6	0.6798	0.2361	0.9906	0.015	0.9944	0.0054	0.9731	0.0058	0.9588	0.0068
7	0.7411	0.2272	0.9935	0.0141	0.9956	0.005	0.99	0.0052	0.9798	0.0054
8	0.7978	0.2325	0.9941	0.0141	0.9931	0.0046	0.9952	0.0033	0.9892	0.0039
9	0.8312	0.2309	0.9955	0.0123	0.9973	0.0045	0.9953	0.0031	0.9938	0.0032
10	0.8675	0.2411	0.9964	0.0127	0.9964	0.0046	0.9984	0.0032	0.9977	0.0027
11	0.8881	0.2545	0.997	0.0115	0.9989	0.0039	0.9995	0.0028	0.9989	0.0025
12	0.9242	0.2495	0.9982	0.0107	0.9991	0.0035	0.9988	0.0028	0.9995	0.0025
13	0.9523	0.2652	0.9988	0.0105	0.9983	0.0035	0.999	0.003	0.9993	0.0024

14	0.9598	0.2569	0.9978	0.0101	0.9993	0.0034	0.999	0.0025	0.9991	0.0023
15	0.9711	0.2676	0.9973	0.0107	0.9984	0.003	0.9983	0.0026	0.9993	0.0022
16	0.9764	0.2771	0.998	0.0099	0.9983	0.0029	0.999	0.0024	0.9991	0.0021
17	0.9781	0.2851	0.9991	0.0097	0.9993	0.003	0.9992	0.0022	0.999	0.0022
18	0.9851	0.2936	0.9988	0.009	0.9986	0.0028	0.9995	0.0021	0.9995	0.002
19	0.9891	0.2905	0.9982	0.0096	0.9993	0.0027	0.9995	0.0022	0.999	0.0021
20	0.9893	0.3039	0.9986	0.0094	0.9989	0.0027	0.9993	0.0022	0.9989	0.002

Dalam proses pelatihan model ini, terdapat lima fold dalam K-Fold cross-validation, dan setiap fold menjalani 20 epoch. Hasil terbaik diperoleh dari fold yang memiliki rata-rata Jaccard similarity tertinggi pada validasi. Berdasarkan tabel hasil, fold 4 pada epoch ke-20 memberikan hasil terbaik dengan Jaccard similarity sebesar 0.9993.

Fold 4 dipilih karena menunjukkan peningkatan yang konsisten dan stabil pada nilai Jaccard similarity dari epoch awal hingga akhir, mencapai puncaknya pada epoch ke-20. Hal ini menunjukkan bahwa model telah belajar dengan baik dan mampu menggeneralisasi dengan baik pada data validasi, menghasilkan prediksi yang sangat akurat terhadap genre yang sesuai dengan teks overview yang diberikan. Stabilitas dan konsistensi ini memberikan keyakinan bahwa model yang dihasilkan dari fold dan epoch ini adalah yang paling optimal untuk digunakan pada data baru.

5.3.4 Pengujian Indonesian RoBERT untuk prediksi Genre Film

Model yang digunakan dalam proyek ini adalah "cahya/roberta-base-indonesian-522M," sebuah model RoBERTa yang telah di-pretrain untuk Bahasa Indonesia. Model ini dirancang untuk memahami konteks bahasa Indonesia secara mendalam dan dapat diaplikasikan pada berbagai tugas pemrosesan bahasa alami (NLP), termasuk klasifikasi teks. Pemodelan berbasis RoBERTa ini memanfaatkan arsitektur Transformer, yang unggul dalam menangkap dependensi panjang dalam teks, menjadikannya sangat efektif untuk tugas klasifikasi teks.

Dibandingkan dengan model indoBERT yang berbasis BERT, model RoBERTa memiliki beberapa perbedaan signifikan. RoBERTa (Robustly optimized BERT approach) melakukan pelatihan dengan jumlah data yang lebih besar dan menggunakan teknik optimasi yang lebih

agresif, seperti peningkatan ukuran batch dan panjang sequence, serta penggunaan tanpa urutan "Next Sentence Prediction" (NSP). Hal ini membuat RoBERTa seringkali lebih unggul dalam menangkap nuansa dan konteks yang lebih kompleks dibandingkan model BERT standar. Sementara indoBERT juga sangat efektif untuk tugas NLP dalam Bahasa Indonesia, RoBERTa dapat memberikan performa yang lebih baik dalam beberapa kasus berkat optimasi tambahan tersebut.

Tabel 5.4 Hasil prediksi Indonesian RoBERTa untuk genre

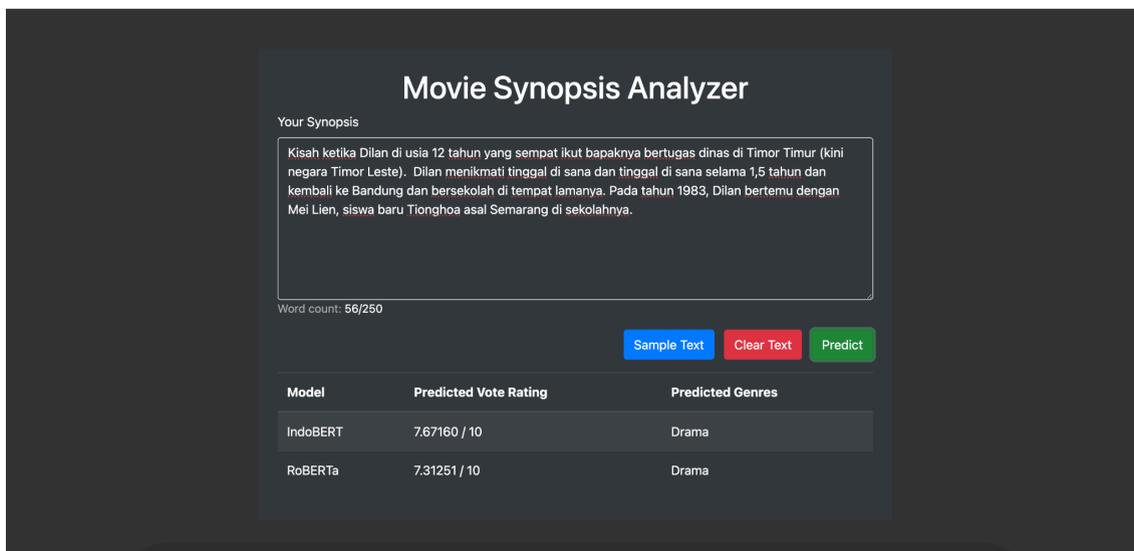
Epoch	Fold 1 Jaccard	Fold 1 Loss	Fold 2 Jaccard	Fold 2 Loss	Fold 3 Jaccard	Fold 3 Loss	Fold 4 Jaccard	Fold 4 Loss	Fold 5 Jaccard	Fold 5 Loss
1	0.2199	0.2844	0.6138	0.1678	0.8989	0.0504	0.9876	0.0077	0.9858	0.0091
2	0.3631	0.2553	0.8705	0.0606	0.8853	0.053	0.9248	0.0414	0.9336	0.0339
3	0.4029	0.2498	0.9295	0.0386	0.9628	0.0267	0.9656	0.0215	0.9088	0.0458
4	0.4434	0.2432	0.9504	0.0292	0.9839	0.0165	0.9742	0.0153	0.9414	0.0305
5	0.4359	0.2514	0.9649	0.0237	0.9912	0.0107	0.9896	0.0102	0.9806	0.0135
6	0.4735	0.246	0.9558	0.0241	0.9957	0.0079	0.9962	0.006	0.9888	0.0079
7	0.4277	0.276	0.9674	0.0192	0.9943	0.0068	0.9969	0.0045	0.9942	0.0058
8	0.4469	0.2821	0.9715	0.019	0.9991	0.0061	0.9979	0.004	0.9955	0.0042
9	0.4638	0.2725	0.976	0.0173	0.9927	0.006	0.9979	0.0036	0.9972	0.0033
10	0.4504	0.2882	0.96	0.0202	0.9948	0.0057	0.9979	0.0033	0.996	0.0034
11	0.4602	0.2878	0.9713	0.0172	0.9955	0.0052	0.9979	0.0035	0.9986	0.0034
12	0.4589	0.3016	0.9682	0.0166	0.9954	0.0054	0.9979	0.0033	0.9967	0.0032
13	0.4638	0.2934	0.9745	0.0163	0.9955	0.0049	0.9979	0.0033	0.9967	0.0032
14	0.4762	0.3054	0.9682	0.0164	0.9938	0.005	0.9979	0.0033	0.998	0.0031
15	0.4693	0.3072	0.9737	0.0154	0.9957	0.0048	0.9979	0.0032	0.998	0.003
16	0.4576	0.3144	0.9739	0.0153	0.9957	0.0047	0.9979	0.0032	0.9972	0.003
17	0.4808	0.3226	0.9714	0.0152	0.9974	0.0043	0.9979	0.0031	0.9986	0.0028
18	0.4736	0.3261	0.9681	0.0162	0.9964	0.0044	0.9979	0.003	0.9972	0.0027
19	0.4634	0.3299	0.9738	0.015	0.9948	0.0043	0.9979	0.003	0.998	0.0029
20	0.4732	0.3373	0.9721	0.0146	0.9967	0.0042	0.9979	0.0029	0.9961	0.0035

Dalam implementasi k-fold cross-validation, fold yang dianggap sebagai yang terbaik adalah fold yang memiliki performa validasi tertinggi. Dalam kasus ini, fold terbaik adalah fold ke-4. Pilihan ini didasarkan pada Jaccard similarity tertinggi yang diperoleh selama proses validasi, yaitu sebesar 0.9979, dengan loss validation sekitar 0.0029. Tingginya nilai Jaccard

similarity menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi label genre untuk data validasi. Selain itu, loss validation yang rendah menandakan bahwa model mampu menghasilkan prediksi dengan tingkat ketepatan yang tinggi. Dengan demikian, fold ke-4 dipilih karena memiliki kombinasi nilai Jaccard similarity yang tinggi dan loss yang rendah, menunjukkan kualitas prediksi yang baik pada data yang belum pernah dilihat sebelumnya.

5.4 Pengujian Aplikasi

Pengujian ini perlu dilakukan untuk memeriksa kembali apakah model sudah belajar dengan cukup baik terhadap data-data yang sudah diberikan sebelumnya. Pengujian ini juga dilakukan di aplikasi final (aplikasi yang digunakan oleh user) sehingga menggambarkan akurasi prediksinya terhadap data yang sudah diajarkan.



Gambar 5.1 Testing website

Dibawah ini merupakan tabel perbandingan antara hasil prediksi dengan nilai sebenarnya untuk “Vote Average” dan “Genre” Film.

Row	Judul Film	Sinopsis	Real Vote Average	IndoBERT	Indonesian Roberta
168698	It Takes a Man and a Woman	Laida dan Miggy dulunya sepasang kekasih, tetapi kini terpaksa bekerja sama dalam kapasitas mereka sebagai pekerja; dan keduanya sama-sama belum bisa	6.6	6.60762	6.74354

		melupakan masa lalu.			
168695	Madre	Hidup Tansen, 27 tahun, berubah dalam sehari hanya dengan sebuah kunci lemari es yang di dalamnya terdapat adonan biang roti berumur 70 tahun bernama Madre. Kakeknya, Tan Sin Gie, memberikan warisan Madre kepada Tansen. Selama ini Tansen tak pernah tahu kalau seperempat darah miliknya berasal dari kakeknya yang beretnis Tionghoa dan seperempat darah India dari neneknya.	8.0	8.33117	8.06093
167346	Identity Thief	Saat dituduh bersalah atas tindak kriminal yang tidak ia lakukan, Sandy Patterson di Colorado harus melacak pelaku sebenarnya yang namanya sama dengan dirinya di Florida.	5.863	5.85428	6.01162
108403	Law Abiding Citizen	Trauma akibat pembunuhan terhadap istri dan putrinya, seorang pria yang berduka menuruti kemarahannya dan merencanakan pembalasan dendam.	7.386	7.64404	7.41059
108401	Zombieland	Ingin bertahan dalam dunia yang dikuasai zombie, seorang mahasiswa konyol bergabung dengan preman kota dan sepasang kakak beradik pecundang.	7.259	7.65182	7.2838

Tabel 5.5 Hasil prediksi vote average

Sesuai dengan tabel 5.5 diatas, hasil prediksi terhadap “Vote Average” sudah cukup presisi, dengan perbedaan paling jauh di angka 0.4 dari rating aslinya di film “Zombieland” pada model IndoBERT.

Tabel 5.6 Hasil prediksi Genre

Row	Judul Film	Sinopsis	Real Genre	IndoBERT	Indonesian Roberta
168698	It Takes a Man and a Woman	Laida dan Miggy dulunya sepasang kekasih, tetapi kini terpaksa bekerja sama dalam kapasitas mereka sebagai pekerja; dan keduanya sama-sama belum bisa melupakan masa lalu.	Comedy, Drama, Romance	Drama, Comedy, Romance	Drama, Comedy, Romance

168695	Madre	Hidup Tansen, 27 tahun, berubah dalam sehari hanya dengan sebuah kunci lemari es yang di dalamnya terdapat adonan biang roti berumur 70 tahun bernama Madre. Kakeknya, Tan Sin Gie, memberikan warisan Madre kepada Tansen. Selama ini Tansen tak pernah tahu kalau seperempat darah miliknya berasal dari kakeknya yang beretnis Tionghoa dan seperempat darah India dari neneknya.	Drama	Drama	Drama
167346	Identity Thief	Saat dituduh bersalah atas tindak kriminal yang tidak ia lakukan, Sandy Patterson di Colorado harus melacak pelaku sebenarnya yang namanya sama dengan dirinya di Florida.	Comedy, Crime	Comedy, Crime	Comedy, Crime
108403	Law Abiding Citizen	Trauma akibat pembunuhan terhadap istri dan putrinya, seorang pria yang berduka menuruti kemarahannya dan merencanakan pembalasan dendam.	Drama, Crime, Thriller	Drama, Crime, Thriller	Drama, Crime, Thriller
108401	Zombieland	Ingin bertahan dalam dunia yang dikuasai zombie, seorang mahasiswa konyol bergabung dengan preman kota dan sepasang kakak beradik pecundang.	Comedy, Horror	Comedy, Horror	Comedy, Horror

Sesuai dengan tabel diatas, hasil prediksi terhadap “Genre” sudah cukup presisi, digambarkan dengan hasil prediksi yang sangat mirip dengan aslinya.

5.5 Pengujian oleh User

Pengujian dilakukan dengan user testing melalui survey kepada tester yang merupakan produser/sutradara/scriptwriter film indie yang pernah membuat film pendek dan layar lebar. Tester yang menjawab survey sebanyak 11 orang. Tester yang mengisi survey merupakan praktisi berpengalaman yang pernah terjun dan terlibat dalam pembuatan film indie maupun layar lebar terutama sinopsis/cerita dari film tersebut.

Survey dilakukan dengan meminta setiap user yang merupakan praktisi langsung untuk menginputkan minimal 3 sinopsis dari film-film yang pernah mereka buat/kerjakan. Dengan menginputkan film yang pernah user kerjakan, mereka lebih paham mengenai film tersebut,

mulai dari *pre production* (pembuatan naskah dan casting) hingga *post production* (editing dan publikasi).

Dari sinopsis yang mereka inputkan, user mendapatkan hasil berupa tabel perbandingan dari kedua model, sehingga user dapat menilai hasil prediksi dari model mana yang paling mendekati dengan film sebenarnya.

Berikut merupakan pertanyaan yang digunakan untuk survey.

Section 1

1. Apakah anda merupakan sutradara/scriptwriter/produser untuk film/series?

Section 2 (Profil Diri)

1. Nama Lengkap
2. Akun Instagram
3. Apa pengalaman anda dalam dunia perfilman? (checkbox)
 - a. Produser
 - b. Director / Sutradara
 - c. Assistant Director / Asisten Sutradara
 - d. Scriptwriter
 - e. Director of Photography (D.O.P) / Penata Kamera
 - f. Cinematographer (Sinematografer)
 - g. Editor
 - h. Set Designer (Perancang Set)
 - i. Music Composer (Komposer Musik)

4. Jumlah Film yang pernah anda kerjakan [**input angka**]

5. Film yang pernah anda buat/kerjakan

Format: Nama Film/Series, Tahun Produksi, Link Tonton

Contoh:

Stargazing, 2021, CGV Cinemas &

<https://youtu.be/xUHfChsTdbM?si=f0Cd2KP8OX714yke>

Menjaga Rasa, 2021, CGV Cinemas &

<https://youtu.be/pZSiwYQEejc?si=ADkJwgCa9r83cjTC>

Section 3 (Coba Aplikasi)

1. [Panduan]

Cara Penggunaan

1. Siapkan 2-3 sinopsis film/series yang pernah anda **BUAT** atau anda **KERJAKAN** (max 250 kata)
2. Buka Situs di <https://gamastar.id/predict> atau <https://srv531535.hstgr.cloud>
(jika muncul connection not secure, klik "show details", lalu tekan "visit this website" atau "proceed anyway")
3. Masukkan sinopsis film anda kedalam textbox
4. Tekan tombol predict
5. Periksa hasil yang muncul di tabel bagian bawah

Gambar 5.2 Alur penggunaan website

Section 4 (Pengalaman Penggunaan)

1. Seberapa mudah Anda menemukan dan mengakses fitur utama dari aplikasi ini?
1 (Sangat sulit) - 5 (Sangat mudah)
2. Seberapa jelas instruksi penggunaan aplikasi ini?
1 (Sangat tidak jelas) - 5 (Sangat jelas)
3. Seberapa cepat aplikasi memberikan hasil prediksi setelah sinopsis diinput?
1 (Sangat lambat) - 5 (Sangat cepat)

Section 5 (Akurasi Prediksi)

1. Menurut Anda, seberapa akurat prediksi genre film yang diberikan oleh model IndoBERT?
1 (Sangat tidak akurat) - 5 (Sangat akurat)
2. Menurut Anda, seberapa akurat prediksi genre film yang diberikan oleh model Indonesian Roberta?
1 (Sangat tidak akurat) - 5 (Sangat akurat)
3. Seberapa akurat prediksi vote average yang diberikan oleh model IndoBERT?
1 (Sangat tidak akurat) - 5 (Sangat akurat)
4. Seberapa akurat prediksi vote average yang diberikan oleh model Indonesian Roberta?

1 (Sangat tidak akurat) - 5 (Sangat akurat)

Section 5 (Perbandingan Model)

1. Mana dari kedua model (IndoBERT atau Indonesian Roberta) yang menurut Anda memberikan hasil prediksi **genre** yang lebih baik?
 - a. IndoBERT
 - b. Indonesian Roberta
 - c. Keduanya sama baik
 - d. Tidak tahu/tidak yakin
2. Mana dari kedua model (IndoBERT atau Indonesian Roberta) yang menurut Anda memberikan hasil prediksi **vote average** yang lebih baik?
 - a. IndoBERT
 - b. Indonesian Roberta
 - c. Keduanya sama baik
 - d. Tidak tahu/tidak yakin

Section 6 (Masukkan dan Saran)

1. Apa kelebihan dari aplikasi ini menurut Anda?
2. Apa kekurangan dari aplikasi ini menurut Anda?
3. Apakah ada fitur tambahan yang Anda ingin lihat di aplikasi ini?
4. Seberapa besar kemungkinan Anda akan merekomendasikan aplikasi ini kepada rekan sesama profesional di industri film?
1 (Sangat tidak mungkin) - 5 (Sangat mungkin)
5. Apakah ada saran atau kritik lain yang ingin Anda sampaikan untuk perbaikan aplikasi ini?

Berikut adalah jawaban dari user dengan profil sebagai berikut:

- User 1: Yosua Hadinata
 - Music Composer
 - Bekerja sebagai Music Composer untuk iklan TV, Film Layar Lebar, dan WebSeries

- User 2: Moreno Nicholas
 - Produser, Director / Sutradara, Scriptwriter, Director of Photography (D.O.P) / Penata Kamera, Cinematographer (Sinematografer), Editor
 - Bekerja sebagai freelance videographer dan DOP Film layar lebar serta webseries
- User 3: Wilson Takhta
 - Director / Sutradara, Editor
 - Lulusan DKV dan bekerja sebagai freelance videographer, editor, dan DOP
 - Bekerja di Production House yang membuat film-film series di Surabaya
- User 4: Austin Gunawan
 - Produser, Director / Sutradara, Assistant Director / Asisten Sutradara
 - Bekerja sebagai produser dan director di film-film layar lebar
- User 5: Tina Dewi Fortuna
 - Produser, Scriptwriter
 - Mahasiswa jurusan perfilman di Universitas Ternama di Jakarta, menjadi produser dan scriptwriter untuk film-film layar lebar
- User 6: Joshua Yordana
 - Produser, Director / Sutradara, Director of Photography (D.O.P) / Penata Kamera, Cinematographer (Sinematografer), Editor
 - Owner dari Production House yang membuat film-film series di Surabaya

Tabel 5.7 Penilaian pengguna terhadap kemudahan, kejelasan, dan kecepatan aplikasi

User	Kemudahan Akses Fitur Utama	Keterangan Instruksi Pengguna	Kecepatan Hasil Prediksi
User 1	5	5	4
User 2	5	5	4
User 3	5	4	4
User 4	5	5	5
User 5	3	4	3
User 6	5	5	4
Average	4.666	4.666	4

Dari tabel tersebut, dapat dilihat bahwa pengguna memberikan penilaian yang bervariasi terhadap kemudahan akses fitur utama, keterangan instruksi pengguna, dan kecepatan hasil prediksi. Dapat dilihat bahwa kemudahan akses dan keterangan instruksi sudah cukup baik di rata-rata 4.666 dari 5, namun untuk kecepatan hasil prediksi masih di 4.0 yang mungkin saja terjadi karena VPS yang saya gunakan hanya memiliki 1 core. Seluruh user tidak ada yang memberikan skala dibawah 3 dari 5 (netral) sehingga dapat disimpulkan fitur dari aplikasi sudah bekerja dengan cukup baik.

Tabel 5.8 Penilaian pengguna terhadap akurasi masing-masing model

User	Genre Film (IndoBERT)	Genre Film (Indonesian Roberta)	Vote Average (IndoBERT)	Vote Average (Indonesian Roberta)
User 1	5	5	5	5
User 2	5	4	5	4
User 3	4	5	4	3
User 4	4	5	5	5
User 5	3	3	3	3
User 6	5	4	5	5
Average	4.333	4.333	4.5	4.166

Dari tabel 5.8 diatas dengan skala 1-5, dapat dilihat bahwa pengguna memberikan skor yang hampir sama untuk prediksi genre, namun untuk prediksi vote average, IndoBERT sedikit lebih unggul.

Tabel 5.9 Kelebihan, kekurangan, dan saran dari pengguna

User	Kelebihan dari aplikasi	Kekurangan dari aplikasi	Saran fitur tambahan
User 1	dapat memberikan informasi rating dari sebuah film, sehingga dapat memberikan rekomendasi film tersebut layak ditonton atau tidak	User interfacenya kurang menarik	tidak

User 2	dapat digunakan untuk menganalisa sinopsis dengan cepat, sehingga dapat dimanfaatkan pada masa pre production sebuah film	-	rekomendasi atau masukan atas sinopsi yang telah kita cantumkan menggunakan ai
User 3	Membantu memprediksi genre film dengan cepat	kadang susah diakses webnya	Penyimpanan Sinopsis Pribadi (history)
User 4	ringkas, cepat, mudah digunakan	-	sejauh ini belum ada
User 5	hal baru dan berbeda	-	-
User 6	Bisa membantu untuk menentukan apa genre yang paling sesuai dan seberapa diminati film tersebut	-	Saran/rekomendasi perubahan sinopsis

Tabel 5.9 diatas menunjukkan kelebihan dan kekurangan dari aplikasi, serta saran dari pengguna. Dapat dilihat bahwa ada beberapa saran yang mungkin bisa dikembangkan, misalnya pengembangan UI/UX, Fitur rekomendasi perubahan sinopsis, dan fitur history.

Tabel 5.10 dibawah ini menunjukkan tingkat kemungkinan pengguna merekomendasikan aplikasi kepada rekan sesama profesional di industri film dalam skala 1-5.

Tabel 5.10 Tingkat rekomendasi aplikasi pada rekan industri

User	Tingkat Rekomendasi
User 1	3
User 2	4
User 3	4
User 4	4
User 5	3
User 6	5
Average	3.833

Dari tabel tersebut, dapat dilihat bahwa sebagian besar pengguna cenderung memiliki kemungkinan yang cukup tinggi untuk merekomendasikan aplikasi kepada rekan sesama profesional di industri film. Dapat dilihat juga tidak ada yang memberikan nilai dibawah 3 dari 5 (netral) sehingga dapat disimpulkan aplikasi cukup layak untuk direkomendasikan.