Bintang Nabiil Lukman

1103223126

TK-46-06

Tugas Week 4 – Analisis

Dalam eksperimen ini, telah dilakukan pelatihan dan evaluasi terhadap tiga jenis model jaringan saraf berulang (Recurrent Neural Networks), yaitu **RNN**, **LSTM**, dan **GRU**, menggunakan dua framework berbeda: **TensorFlow** dan **PyTorch**. Masing-masing model diuji berdasarkan akurasi, loss, serta metrik evaluasi seperti precision, recall, F1-score, dan AUC. Hasilnya menunjukkan perbedaan yang cukup signifikan baik dari sisi performa maupun kestabilan pelatihan.

1. **Model RNN**
   Model RNN menunjukkan performa paling rendah dibandingkan dua model lainnya. Pada TensorFlow, akurasi validasi tertinggi hanya mencapai **0.5399**, dengan **recall sangat rendah (0.0846)** dan **F1-score sebesar 0.1554**, meskipun precision-nya tinggi (0.9455), yang mengindikasikan adanya **imbalance atau kesalahan dalam prediksi kelas minoritas**. AUC-nya berada di angka **0.7776**, dengan kurva ROC yang tidak terlalu melengkung, menandakan **daya klasifikasi yang kurang baik**. Sementara pada PyTorch, performa RNN sedikit lebih baik dengan akurasi mencapai **0.7118**, F1-score sebesar **0.6698**, dan AUC **0.7806**. Namun secara keseluruhan, **RNN rentan terhadap permasalahan vanishing gradient**, dan hal ini terlihat dari peningkatan performa yang tidak terlalu signifikan selama proses training.

2. **Model LSTM**
   LSTM secara signifikan mengungguli RNN dalam semua aspek. Di TensorFlow, akurasi validasi mencapai **0.8619** dengan **F1-score sebesar 0.8632** dan **AUC 0.9324**. Model ini menunjukkan kestabilan selama training, meskipun val loss sempat mengalami kenaikan setelah epoch ketiga. Hal ini bisa menandakan **tanda overfitting ringan**, namun performa klasifikasinya tetap tinggi. Pada PyTorch, hasilnya hampir serupa dengan akurasi **0.8643**, F1-score **0.8642**, dan AUC **0.9349**. Grafik ROC menunjukkan kurva yang sangat baik, melengkung tajam menuju titik (0,1), menandakan performa yang sangat solid. LSTM terbukti mampu **mengatasi permasalahan long-term dependencies** yang tidak dapat ditangani oleh RNN.

3. **Model GRU**
   Model GRU menunjukkan performa yang bahkan **sedikit lebih baik dari LSTM**, terutama pada implementasi PyTorch. Dengan akurasi **0.8862**, precision **0.8605**, recall **0.9220**, dan F1-score **0.8902**, serta **AUC tertinggi sebesar 0.9555**, GRU menjadi model terbaik dalam eksperimen ini. Model ini tidak hanya stabil selama proses pelatihan (loss menurun konstan, akurasi meningkat tajam), tetapi juga efisien dari segi kompleksitas dibandingkan LSTM. Sementara pada TensorFlow, GRU juga menunjukkan performa yang sangat baik dengan akurasi **0.8617**, F1-score **0.8606**, dan AUC **0.8617**, meskipun terlihat gejala overfitting pada peningkatan val loss setelah epoch ke-2. Secara umum, **GRU memberikan trade-off terbaik antara performa dan efisiensi model**, menjadikannya pilihan yang sangat layak digunakan dalam task analisis sentimen seperti ini.

Model	**Framework**	**Akurasi**	**Precision**	**Recall**	**F1 Score**	**AUC Score**	**Catatan**
RNN	TensorFlow	0.5399	0.9455	0.0846	0.1554	0.7776	Akurasi & recall sangat rendah; overfitting terjadi
RNN	PyTorch	0.7118	0.7842	0.5846	0.6698	0.7806	Performa lebih stabil dibanding TF, tapi masih di bawah LSTM/GRU
LSTM	TensorFlow	0.8619	0.8549	0.8718	0.8632	0.9324	Performa tinggi, val loss naik → potensi overfitting ringan
LSTM	PyTorch	0.8643	0.8652	0.8631	0.8642	0.9349	Performa sangat konsisten & stabil
GRU	TensorFlow	0.8617	0.8676	0.8537	0.8606	0.8617	Sangat baik di awal, val loss meningkat setelah epoch 2
GRU	PyTorch	**0.8862**	0.8605	**0.9220**	**0.8902**	**0.9555**	Performa terbaik dari semua model

Kesimpulan

Dari hasil evaluasi seluruh model, dapat disimpulkan bahwa **LSTM dan GRU secara konsisten mengungguli RNN** dalam hal akurasi dan kemampuan generalisasi. **GRU sedikit lebih unggul dari LSTM**, terutama pada hasil evaluasi di PyTorch, dan menunjukkan **stabilitas dan efisiensi pelatihan yang lebih baik**. Model RNN meskipun lebih sederhana, memiliki performa yang kurang memadai terutama pada task yang memerlukan pemrosesan informasi dalam jangka panjang. Oleh karena itu, GRU dapat direkomendasikan sebagai model terbaik dalam kasus ini.