Pendahuluan

Dalam tugas ini, dilakukan eksperimen untuk membandingkan performa tiga arsitektur model jaringan saraf yaitu RNN (SimpleRNN), LSTM, dan GRU dalam melakukan klasifikasi sentimen menggunakan dataset IMDb yang terdiri dari ulasan film dengan label biner (positif atau negatif). Dataset ini dimuat menggunakan fungsi imdb.load_data dari Keras, dan hanya menggunakan 40.000 kata teratas. Setiap ulasan dipadatkan menjadi panjang maksimal 400 kata menggunakan pad_sequences.

Arsitektur Model

Setiap model dibangun dengan struktur dasar yang sama, yaitu lapisan embedding, dua lapisan sel RNN sesuai jenis yang diuji (SimpleRNN, LSTM, atau GRU), diikuti dengan dropout, dense layer berukuran 64 dengan aktivasi ReLU, dan output layer sigmoid. Optimizer yang digunakan adalah Adam dengan fungsi loss binary_crossentropy.

Proses Pelatihan

Model dilatih selama 3 epoch dengan batch size 128 dan validation split sebesar 20% dari data latih. Hasil evaluasi model kemudian diukur menggunakan akurasi, classification report (precision, recall, f1-score), dan Area Under Curve (AUC).

Hasil Evaluasi

1. RNN (SimpleRNN)

   Model RNN menunjukkan hasil performa yang kurang memuaskan. Pada akhir epoch ketiga, akurasi validasi hanya sebesar 50.54%, dengan nilai AUC 0.5360. Berdasarkan classification report, diketahui bahwa model ini gagal mendeteksi kelas 1 dengan baik (recall hanya 0.05) meskipun cukup baik dalam mengenali kelas 0. Akurasi keseluruhan pada data uji adalah 50.11%, hampir setara dengan tebakan acak.

2. LSTM

   Model LSTM memberikan performa yang jauh lebih baik. Akurasi validasi pada akhir epoch ketiga mencapai 87.14%, dengan AUC 0.9300. Akurasi pada data uji juga tinggi, yaitu 85.80%. Model mampu mengenali kedua kelas dengan baik, dengan precision dan recall masing-masing sekitar 0.84–0.88. F1-score keseluruhan menunjukkan bahwa model ini seimbang dan akurat dalam klasifikasi sentimen.

3. GRU

   Model GRU juga menunjukkan hasil yang baik, meskipun sedikit lebih rendah dibanding LSTM. Akurasi validasi akhir adalah 83.44% dengan AUC 0.9183. Pada data uji, akurasi yang dicapai adalah 82.31%, dengan f1-score yang cukup tinggi di kedua kelas (sekitar 0.81–0.84). Model ini tetap menunjukkan kemampuan generalisasi yang baik, meski sedikit menurun dibanding LSTM.

Kesimpulan

Berdasarkan hasil eksperimen, dapat disimpulkan bahwa LSTM memiliki performa terbaik dalam klasifikasi sentimen ulasan film IMDb dibandingkan RNN dan GRU. Hal ini terlihat dari nilai akurasi, AUC, serta konsistensi dalam precision dan recall.

Model RNN tidak direkomendasikan untuk tugas ini karena menunjukkan performa mendekati acak. Di sisi lain, GRU bisa menjadi alternatif ringan yang kompetitif jika ingin menyeimbangkan performa dan efisiensi.

Visualisasi hasil pelatihan memperlihatkan bahwa LSTM dan GRU mengalami peningkatan akurasi dan penurunan loss yang signifikan selama pelatihan, sedangkan RNN stagnan dan overfitting sangat awal.