

BAB V PENUTUP

5.1 Kesimpulan

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa lima model berbasis deep learning, yaitu BERT, DistilBERT, RoBERTa, ALBERT, dan XLNet, dalam analisis sentimen data Twitter. Dengan menggunakan metode pengujian eksperimen komparatif kuantitatif yang konsisten, data dikumpulkan melalui scraping, diproses melalui tahapan pembersihan, pelabelan sentimen menggunakan VADER, serta pembagian dataset ke dalam format Hugging Face. Pelatihan model dilakukan dengan parameter yang seragam untuk memastikan hasil perbandingan yang valid.

Berdasarkan hasil evaluasi keseluruhan, ALBERT menempati peringkat pertama dengan akurasi tertinggi (90.59%), F1-score terbaik, serta nilai ROC AUC yang kuat, menunjukkan konsistensi dan efisiensi kinerja. BERT berada di posisi kedua dengan performa yang sangat kompetitif, namun sedikit di bawah ALBERT dalam aspek F1-score dan waktu pelatihan. Selanjutnya, DistilBERT menempati posisi ketiga karena meskipun metriknya mendekati BERT, model ini sedikit lebih rendah dalam akurasi dan F1-score. RoBERTa dan XLNet berada di posisi keempat dan kelima secara berurutan, karena meskipun tetap menunjukkan hasil yang baik, keduanya tertinggal dalam nilai evaluasi utama seperti *precision*, *recall*, dan ROC AUC. Dengan demikian, ALBERT dapat disimpulkan sebagai model yang paling optimal dalam penelitian ini.

5.2 Saran

Terdapat beberapa hal yang dapat diperbaiki dan dikembangkan lebih lanjut. Pertama, penelitian selanjutnya dapat menggunakan dataset yang lebih beragam dengan menggabungkan data dari berbagai platform media sosial selain Twitter. Hal ini diharapkan dapat meningkatkan generalisasi model untuk berbagai domain aplikasi. Kedua, untuk mengatasi ketidakseimbangan kelas sentimen yang sering ditemukan pada dataset seperti ini, disarankan untuk menerapkan teknik augmentasi data, seperti *oversampling*, *undersampling*, atau menggunakan

algoritma *generative* seperti GAN untuk memperbanyak data pada kelas minoritas. Ketiga, pengujian performa model dapat ditingkatkan dengan mengeksplorasi hyperparameter tuning yang lebih mendalam, seperti variasi dalam *batch size*, *learning rate*, atau jumlah *epoch*.

Selama pengerjaan, salah satu tantangan yang dihadapi adalah proses pelabelan data yang bergantung pada model pre-trained VADER. Penelitian mendatang dapat mempertimbangkan metode pelabelan manual atau semi-otomatis untuk meningkatkan akurasi pelabelan. Terakhir, penerapan model dalam skala produksi memerlukan integrasi dengan *pipeline end-to-end*, termasuk preprocessing otomatis, evaluasi berkala, dan interpretasi hasil, yang dapat menjadi fokus pengembangan di masa depan.

Sebagai tambahan untuk mendukung replikasi dan pengembangan penelitian di masa mendatang, penulis menyarankan agar dataset yang digunakan dalam penelitian ini dapat diakses melalui tautan berikut: [<https://drive.google.com/drive/folders/1RpsmcCkTKPGr5gtMTbRa8cl0Yaylvvg> h?usp=sharing]. Dengan demikian, peneliti lain dapat melakukan eksplorasi lebih lanjut terhadap pendekatan, model, dan teknik analisis yang digunakan.