

GoEmotions – DistilBERT Baseline: Metrics Summary

Model: distilbert-base-uncased | Dataset: go_emotions (raw) | epochs=3 batch_size=32
max_len=128 threshold=0.5

	F1 micro	F1 macro	F1 weighted	Loss
Validation	0.3757	0.2916	0.3440	0.1101
Test	0.3757	0.2916	0.3440	0.1101

- Micro-F1 \approx 0.376 and Macro-F1 \approx 0.292 on validation/test at threshold 0.5.
- This is below typical BERT-base baselines on GoEmotions (micro-F1 \approx 0.45–0.50 on 27+neutral).
- Likely causes: fixed threshold=0.5, no class weighting, only 3 epochs, and using the wide schema variant.
- Action items: tune threshold, train longer with LR schedule, consider class-weighted loss or focal loss, check label mapping integrity.

Recommended Next Steps

- Threshold sweep (e.g., 0.2–0.7) and pick best Macro-F1.
- Run simplified config (6+neutral) for sanity and faster iteration.
- Replicate run with ALBERT and ModernBERT (same settings).
- Add post-training INT8 quantization and log size/latency deltas.