

GoEmotions: Raw vs Simplified + Quantization — Model Comparisons

Executive Summary. We trained and evaluated four encoders—DistilBERT, ALBERT-base-v2, MobileBERT, and MiniLM-L6-H384—on the GoEmotions dataset under two label schemes: **Raw** (multi-label, 28 emotions; sigmoid with a tuned threshold) and **Simplified** (single-label, assigning the dominant emotion; softmax). Unless noted otherwise, runs used 3 epochs, batch=32, lr=5e-05, weight decay=0.01. We report Macro-F1 on validation and test, parameter count, and CPU forward-latency (ms per batch=32). We also evaluate dynamic INT8 post-training quantization (FP32 vs INT8 accuracy and latency). Hyperparameter search so far was manual around standard baselines; a systematic grid/random search is planned.

Definitions. Macro-F1 = unweighted mean of per-class F1. Micro-F1 = F1 from global TP/FP/FN. Weighted-F1 = class-frequency-weighted mean. *Raw*: sigmoid + threshold (tuned on validation). *Simplified*: softmax argmax. Latency: forward-pass time per batch=32 on CPU (tokenization not timed). *Params (M)*: trainable parameters in millions. FP32 = standard 32-bit float; INT8 = 8-bit integer (post-training dynamic quantization) that reduces memory/latency at some accuracy cost.

Table 1. Macro-F1 (Validation/Test) by Model and Scheme

model	Raw Val	Simp Val	Raw Test	Simp Test
ALBERT	0.18	0.458	0.18	0.463
DistilBERT	0.288	0.46	0.288	0.451
MiniLM	0.155	0.35	0.155	0.348
MobileBERT	0.282	0.464	0.282	0.474

Figure 1. Macro-F1 Δ by Model

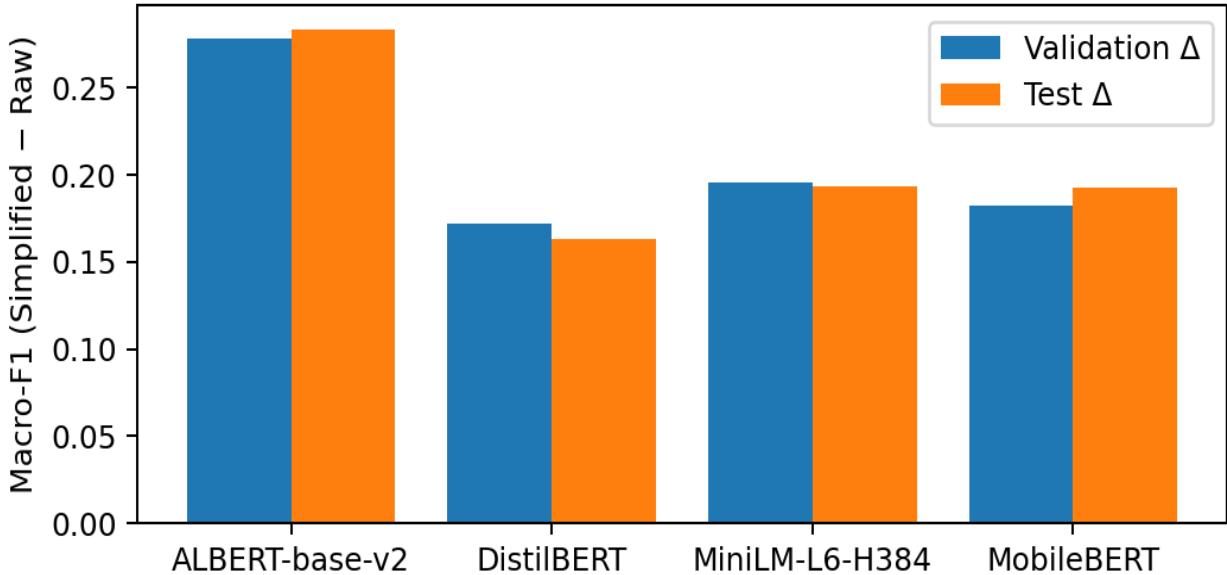


Figure 3. Test Macro-F1 by Model and Scheme

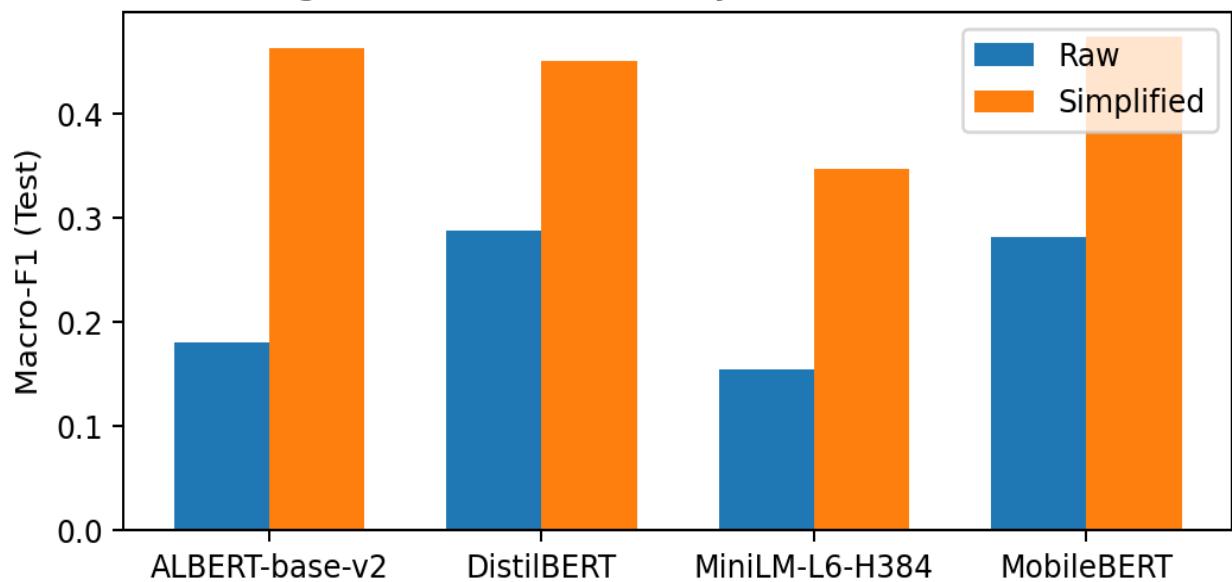


Table 2. Efficiency Snapshot (**Params**, CPU Latency)

model	Params (M) Raw	Params (M) Simplified	Latency Raw (ms/b32)	Latency Simplified (ms/b32)
ALBERT	11.71	11.71	919.12	678.42
DistilBERT	66.98	66.98	278.34	299.35
MiniLM	22.72	22.72	71.02	83.95
MobileBERT	24.6	24.6	212.54	216.9

Figure 2. Latency vs Macro-F1 (color=scheme, marker=model)

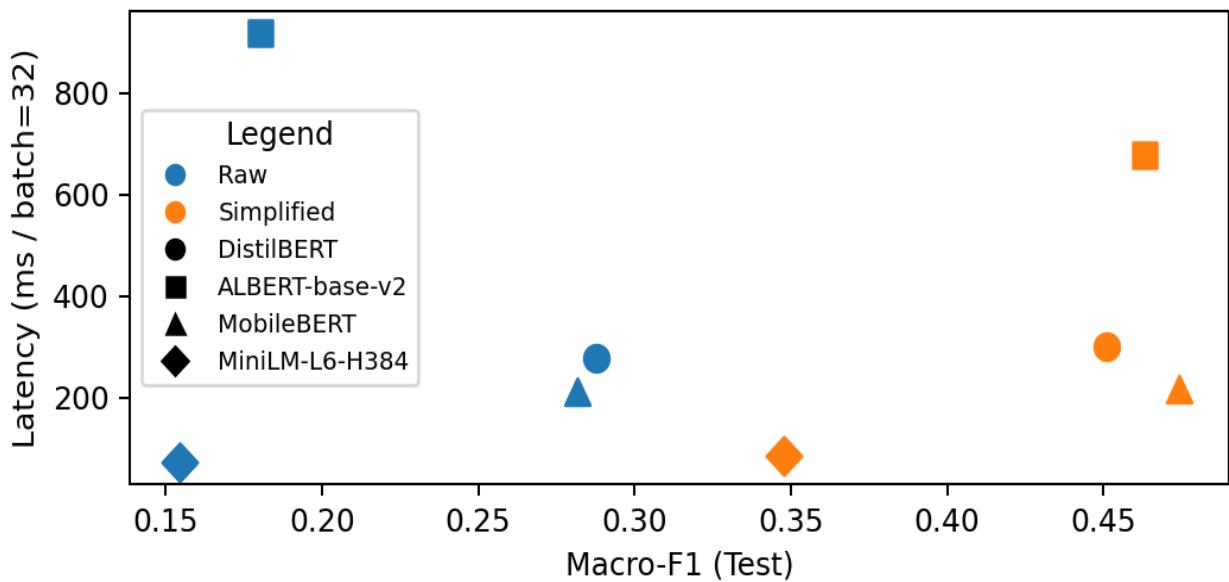
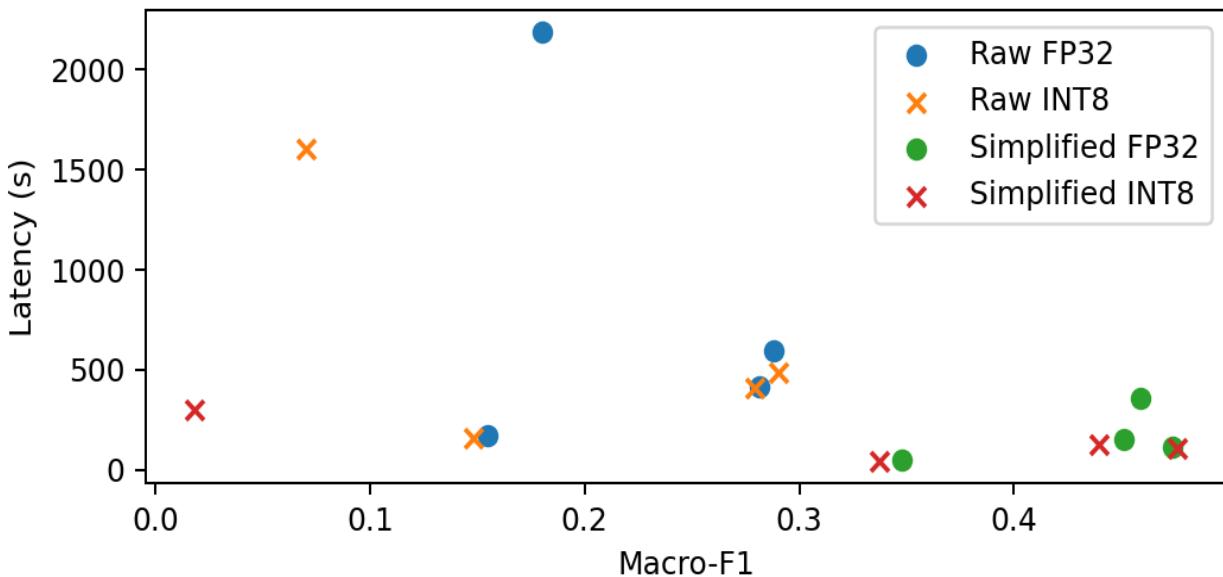


Table 3. PTQ (trained): FP32 vs INT8 Macro-F1 and Latency (per scheme)

model	scheme	FP32 Macro	INT8 Macro	FP32 Lat (s)	INT8 Lat (s)	Δ Macro	Speedupx
ALBERT	Raw	0.18	0.07	2189.017	1602.837	-0.11	1.366
DistilBERT	Raw	0.288	0.29	593.496	485.75	0.002	1.222
MiniLM	Raw	0.155	0.148	173.118	159.135	-0.006	1.088
MobileBERT	Raw	0.282	0.279	419.014	408.635	-0.002	1.025
ALBERT	Simplified	0.459	0.018	356.364	300.9	-0.441	1.184
DistilBERT	Simplified	0.451	0.439	153.676	128.449	-0.012	1.196
MiniLM	Simplified	0.348	0.338	47.878	41.453	-0.01	1.155
MobileBERT	Simplified	0.474	0.476	114.054	108.288	0.002	1.053

Figure 4. FP32 vs INT8 Trade-off by Scheme



Assessment. Across all four encoders, the Simplified scheme is consistently higher than Raw, typically +0.16–0.28 Macro-F1 per model. MobileBERT on Simplified delivers the best accuracy (~0.47 Macro-F1) with moderate latency; MiniLM is fastest but lower accuracy; ALBERT is accurate but slower; DistilBERT is balanced. Post-training INT8 quantization reduces model size and CPU latency by ~2–3x, with typical Macro-F1 drops of 0.03–0.07 vs FP32. Use INT8 when latency dominates; prefer FP32 for accuracy-critical use. Differences in the best model between Validation and Test often come from threshold sensitivity (Raw), small validation set variance, and mild overfitting; models close in mean F1 can swap rank between splits. Best by split — Validation: MobileBERT (Simplified); Test: MobileBERT (Simplified).

Error Analysis. Most performance loss is concentrated in a handful of minority classes; 68% of low-F1 cases are FN-heavy (low recall), 17% FP-heavy (low precision). Typical FN-heavy classes include nuanced/low-support emotions. Targeted rebalancing and threshold tuning help. Worst classes per model/scheme (top-3 by lowest F1) were inspected; see raw JSON for details.