GoEmotions Baseline Report

2025-09-26 16:30 UTC

Run Summary

Model: distilbert-base-uncased
Dataset: go_emotions (raw)
Training: epochs=3, batch_size=32, max_length=128, threshold=0.5

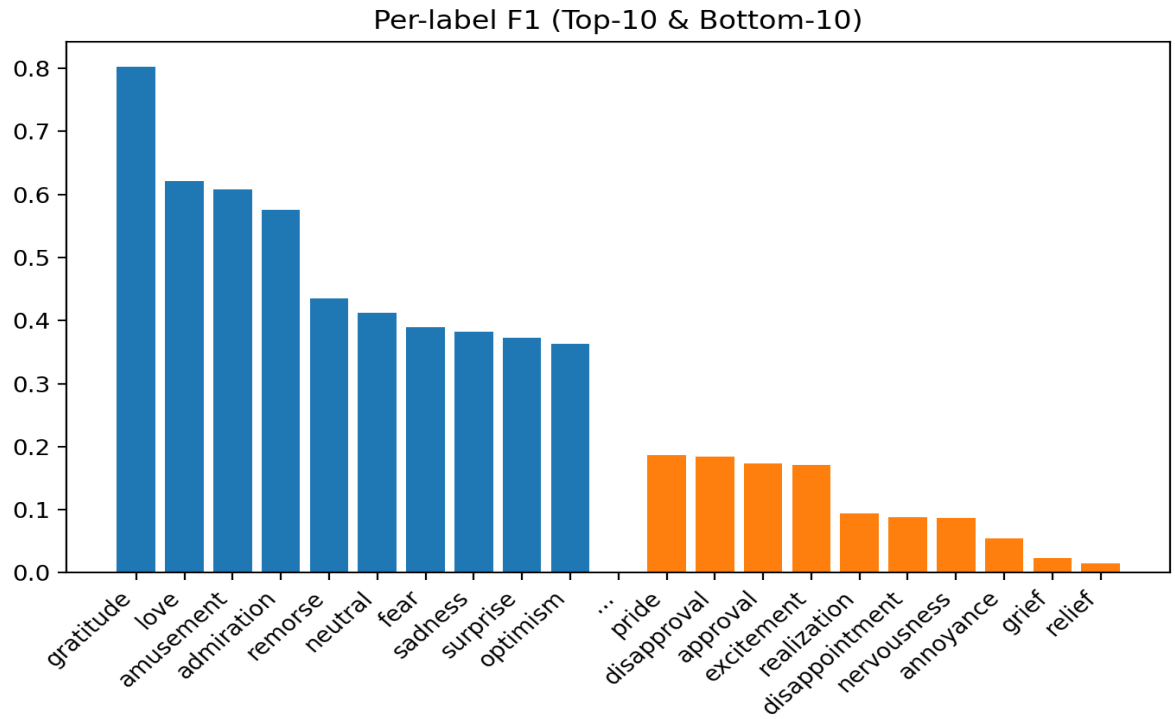
Evaluation Metrics

	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

Efficiency Snapshot

Trainable parameters: 66,975,004
Avg latency (ms) per batch of 32: 10.23

Per-label F1 — Top & Bottom 10



Debug

Debug: val keys=['eval_loss', 'eval_f1_micro', 'eval_f1_macro', 'eval_f1_weighted', 'eval_runtime', 'eval_samples_per_second', 'eval_steps_per_second', 'epoch'], test keys=['eval_loss', 'eval_f1_micro', 'eval_f1_macro', 'eval_f1_weighted', 'eval_runtime', 'eval_samples_per_second', 'eval_steps_per_second', 'epoch']

Notes

Notes:
• Accepts either eval_* or plain metric keys; if missing, backfills from classification_report.json.
• Scores use a fixed threshold (default 0.5). Threshold sweeps can improve macro-F1.
• Latency is a simple forward pass; serving latency varies by hardware and batch size.