ANLP ASSIGNMENT: BERT QUANTIZATION (PTQ & QAT) ON DAIR-AI/EMOTION DATASET

JIYA SINHA (22161)

All the training codes are stored in the github. The inference file is uploaded along with this report to classroom. The trained models are stored in the drive (since the model size exceed 25MB and couldn't be pushed on github).

Github	`https://github.com/sinhajiya/hf_emotion_classifier.git`
Drive (Model Saved here)	`https://drive.google.com/drive/folders/1176dlKl5ZOiFk4ahkHex4pc5lEROz4YZ?usp=sharing`

1. METRICS

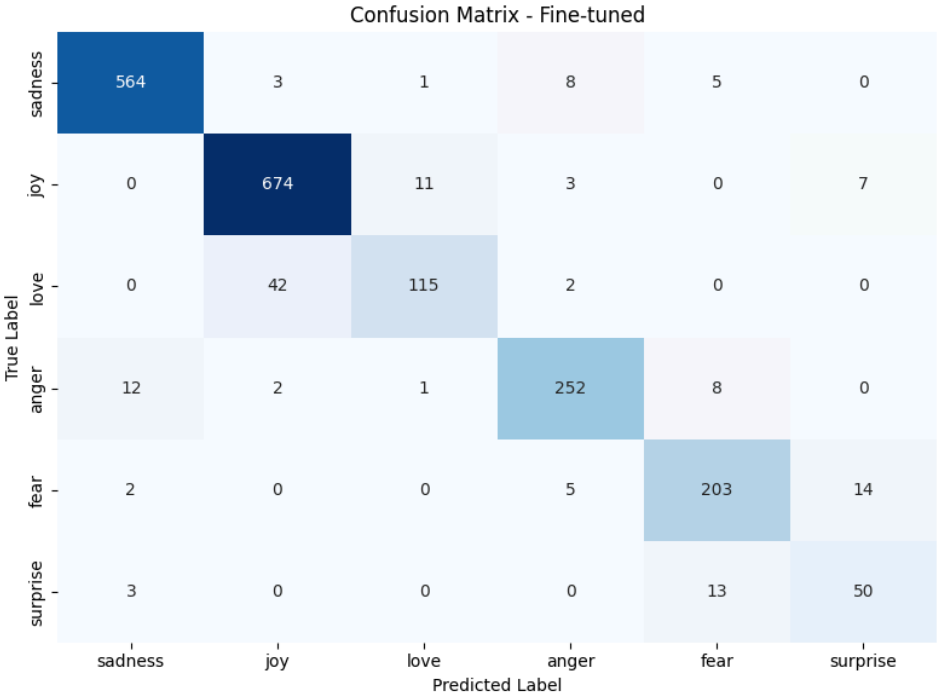
	Accuracy	Macro F1 score	Latency (ms/ sample)	Model Size (in MB)
Fine Tuning	0.93	0.88	33.57	471.67
PTQ	0.92	0.87	257.57	91.09
QAT	0.92	0.88	1684.89	418.95
QLoRA	0.73	0.54	15.49	419.93

2. SOME OBSERVATIONS

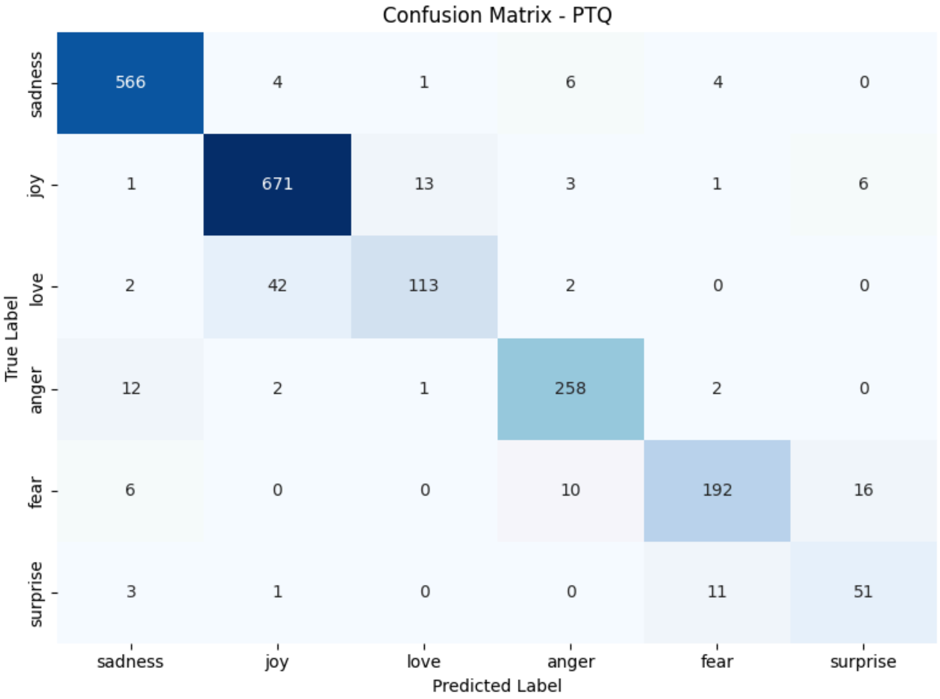
- PTQ shrinks the model size from 471 MB to 91 MB with almost no performance loss (Macro F1: 0.88 to 0.87).
- PTQ has a huge latency spike (33 ms to 257 ms/sample) due to dynamic quant ops on CPU.
- QLoRA is the most efficient at inference (15 ms/sample) but collapses in performance (Macro F1: 0.54), showing low-rank adaptation is unsuitable for this case.

3. Confusion Matrix

3.1. Fine tuned model.

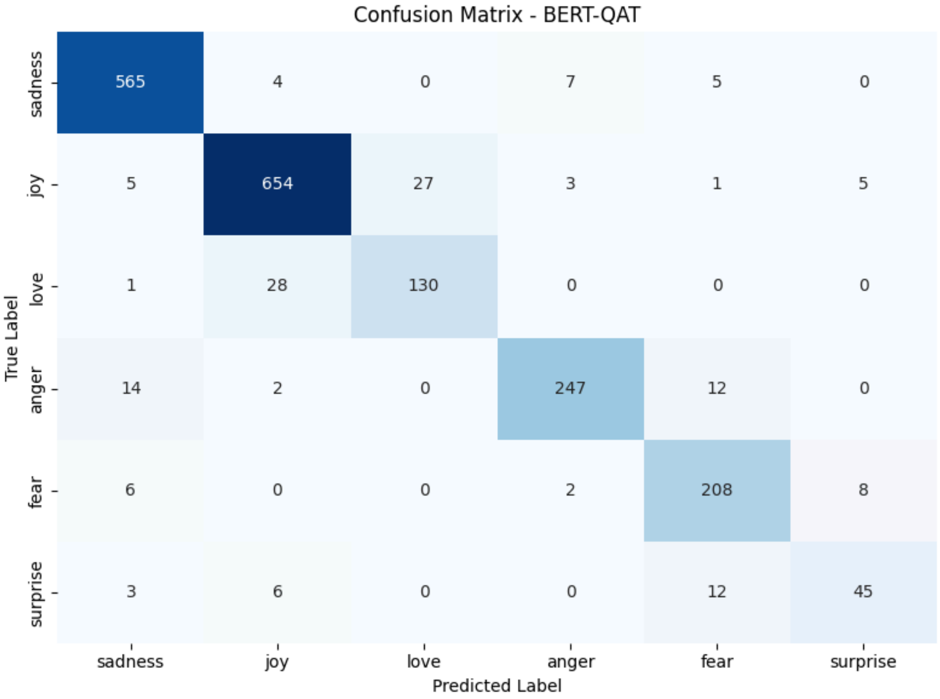


3.2. PTQ Model.



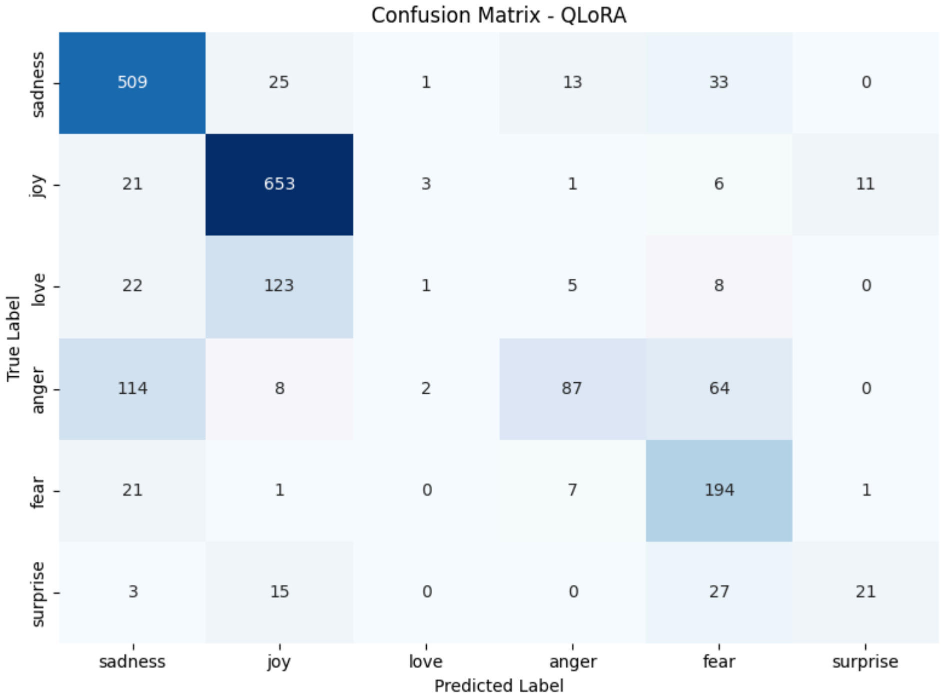
	sadness	joy	love	anger	fear	surprise
sadness	566	4	1	6	4	0
joy	1	671	13	3	1	6
love	2	42	113	2	0	0
anger	12	2	1	258	2	0
fear	6	0	0	10	192	16
surprise	3	1	0	0	11	51

3.3. QAT Model.



	sadness	joy	love	anger	fear	surprise
sadness	565	4	0	7	5	0
joy	5	654	27	3	1	5
love	1	28	130	0	0	0
anger	14	2	0	247	12	0
fear	6	0	0	2	208	8
surprise	3	6	0	0	12	45

3.4. QLoRA.



3.5. Observation:.

All models classify “sadness” and “joy” well, but “love” and “anger” degrade significantly under PTQ and especially QLoRA.