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

JIYA SINHA (22161)

Link to github: `https://github.com/sinhajiya/hf_emotion_classifier.git`

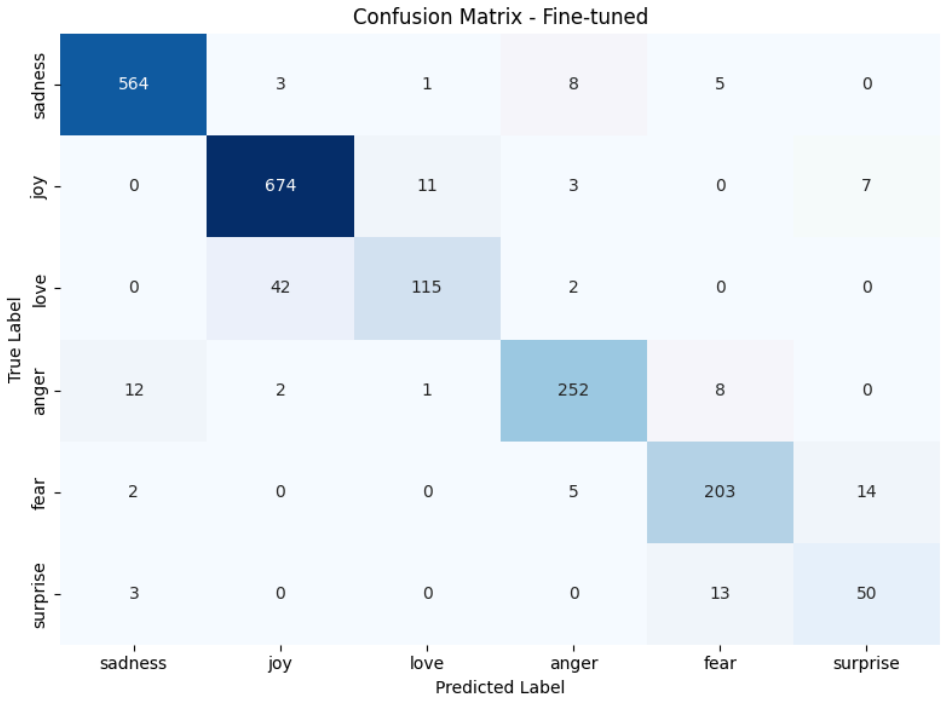
`Link to model metrics saved in google drive`

1. METRICS

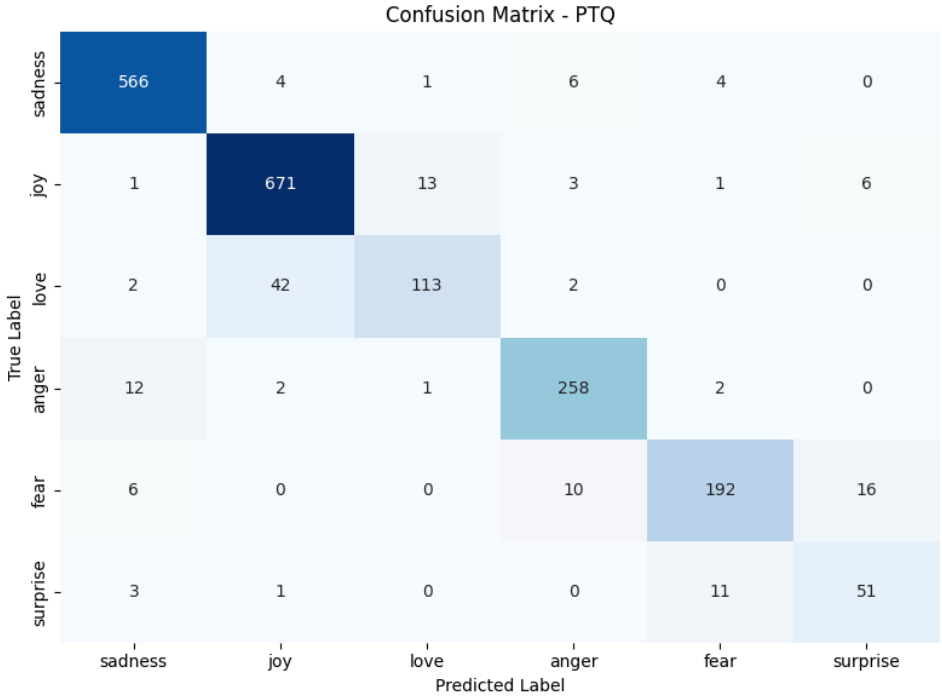
	Accuracy	Macro F1 score	Latency (ms/ sample)	Model Size (in MB)
Fine Tuning	0.93	0.88	33.57	471.67
PTQ	0.92	0.87	257.57	91.09
QAT	0.92	0.88	1684.89	418.95
QLoRA	0.73	0.54	15.49	419.93

2. Confusion Matrix

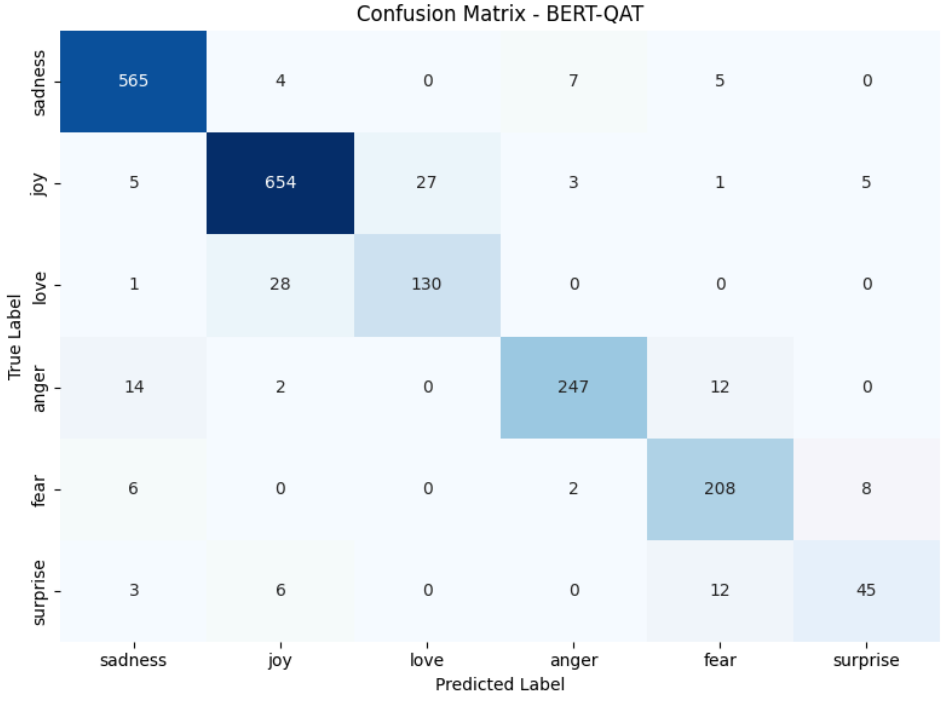
2.1. **Fine tuned model.**


Confusion Matrix - Fine-tuned
True Label
Predicted Label
sadness
joy
love
anger
fear
surprise
564 3 1 8 5 0
0 674 11 3 0 7
0 42 115 2 0 0
12 2 1 252 8 0
2 0 0 5 203 14
3 0 0 0 13 50

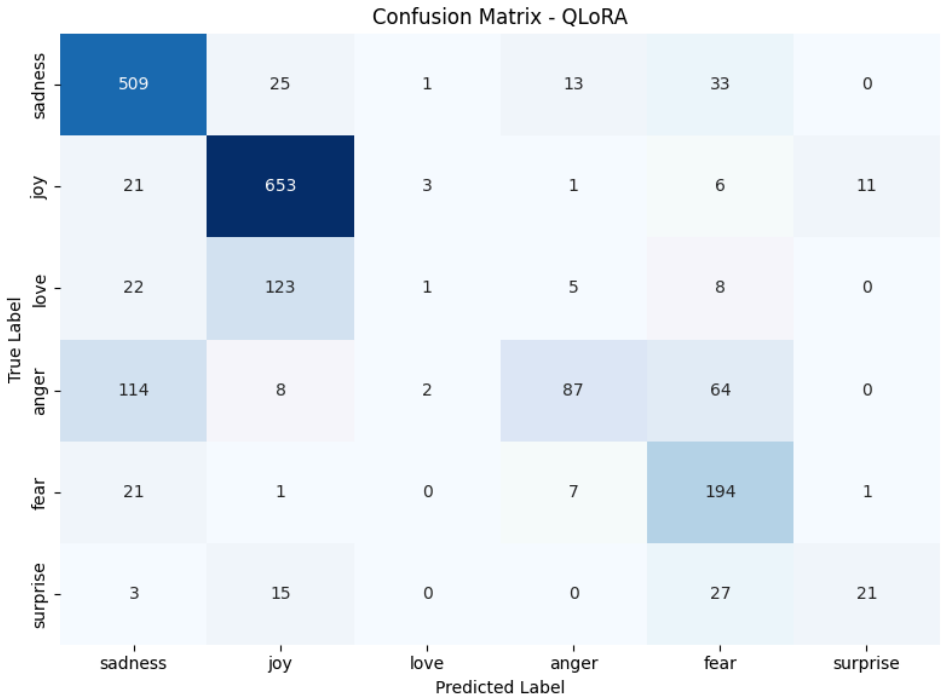
2.2. **PTQ Model.**



2.3. **QAT Model.**



2.4. **QLoRA.**


Confusion Matrix - QLoRA
True Label
Predicted Label
sadness
joy
love
anger
fear
surprise
509 25 1 13 33 0
21 653 3 1 6 11
22 123 1 5 8 0
114 8 2 87 64 0
21 1 0 7 194 1
3 15 0 0 27 21