

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

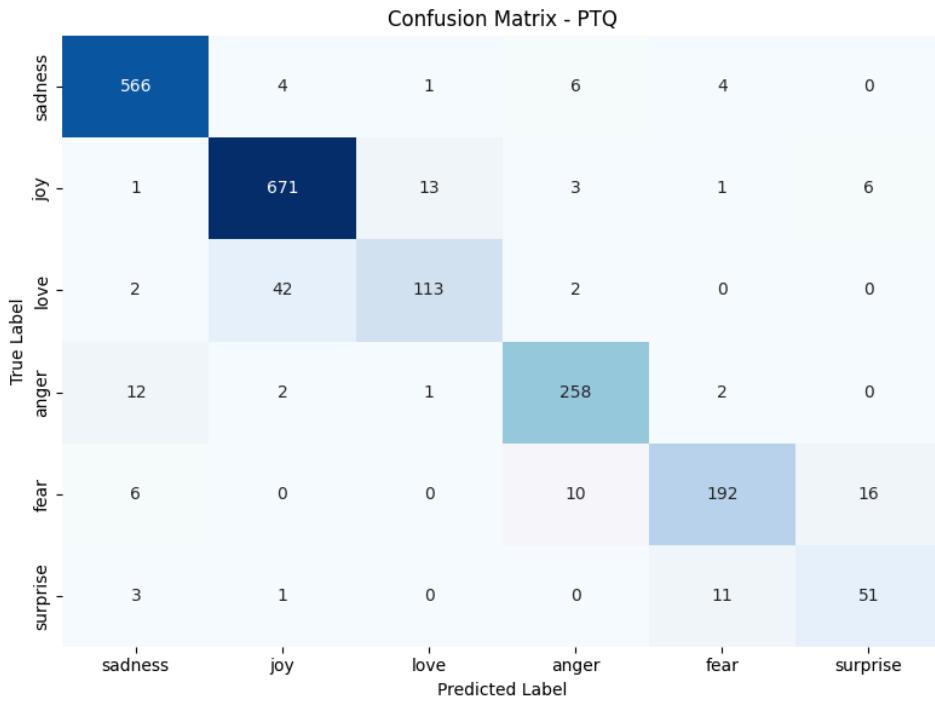
| | Accuracy | Macro F1 score | Latency (ms/sample) | Model Size (in MB) |
|-------------|----------|----------------|---------------------|--------------------|
| Fine Tuning | 0.93 | 0.88 | 33.57 | 471.67 |
| PTQ | 0.92 | 0.87 | 257.57 | 91.09 |
| QAT | 0.92 | 0.88 | 1684.89 | 418.95 |
| QLoRA | 0.73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

2. CONFUSION MATRIX

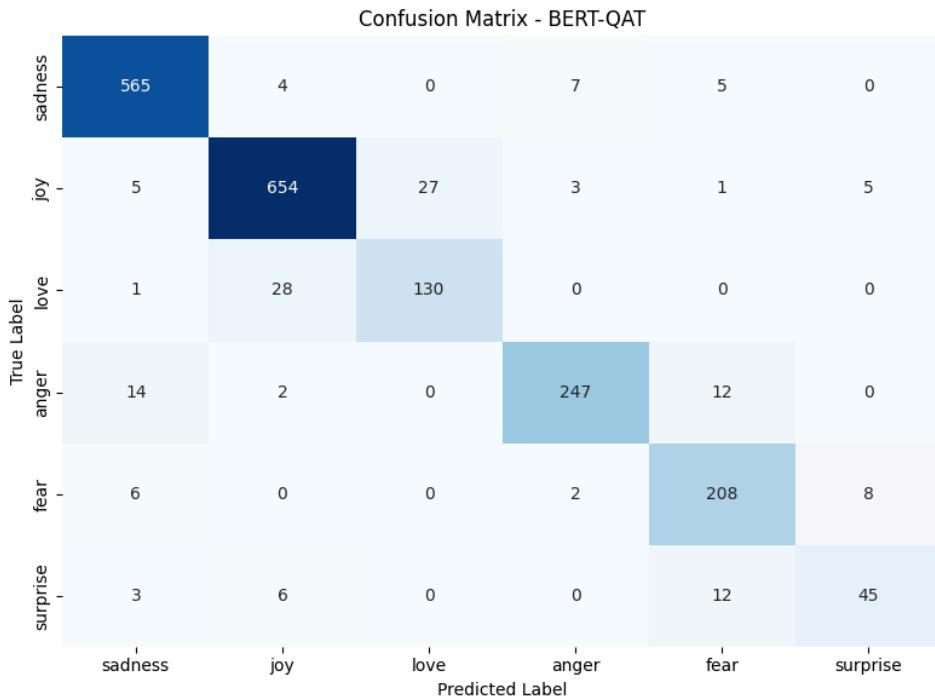
2.1. Fine tuned model.

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

2.2. PTQ Model.



2.3. QAT Model.



2.4. QLoRA.

Confusion Matrix - QLoRA

A confusion matrix heatmap comparing predicted emotions against true emotions. The y-axis is labeled "True Label" and the x-axis is labeled "Predicted Label". The categories are sadness, joy, love, anger, fear, and surprise. The diagonal shows correct predictions. Other cells show the count of misclassifications.

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