Assignment 3: Fine-Tuning Pretrained Transformers for Emotion Classification
A Comparison Between Full and LoRA Fine-Tuning

Jia Cheng Chung

October 15, 2025

Abstract

This project compares two fine-tuning strategies for adapting pretrained Transformer models to emotion classification. Using the GoEmotions dataset of Reddit comments labeled with 28 emotions, I evaluate **full fine-tuning** and **LoRA (Low-Rank Adaptation)** on a BERT-based model. Motivated by my final project on course recommender systems, I explore whether emotion detection can enhance the analysis of student course reviews. The results show that full fine-tuning achieves substantially higher accuracy and F1 scores, while LoRA significantly reduces computational cost but at the expense of performance.

Code Availability. All source code, training notebooks, and result files for this project are publicly available on GitHub: `github.com/JiachengCJC/Emotion-Classification-using-BERT---Full-Fine-Tuning-vs-LoRA`

1 Introduction

Pretrained Transformer models such as BERT enable effective transfer learning across many NLP tasks. Fine-tuning allows these models to adapt quickly to specific applications without large labeled datasets. Among these tasks, **emotion classification** is valuable for understanding human affect and improving systems that respond to user sentiment.

In this project, I use the **GoEmotions** dataset to study how Transformers can recognize emotions in text, with the long-term goal of improving course review recommendations. I compare two methods: (1) **full fine-tuning**, which updates all model parameters, and (2) **LoRA fine-tuning**, a parameter-efficient technique that trains small low-rank adapters. This comparison highlights the trade-off between performance and efficiency, showing how lightweight tuning can support scalable emotion-aware applications

2 Methodology

2.1 Dataset

The experiments use the **GoEmotions** dataset, which contains over 58,000 Reddit comments labeled with 27 distinct emotion categories and a neutral label. Texts were tokenized and truncated to a maximum sequence length of 128. Figure 1 shows the emotion distribution, revealing strong label imbalance where *neutral*, *admiration*, and *approval* dominate the dataset.

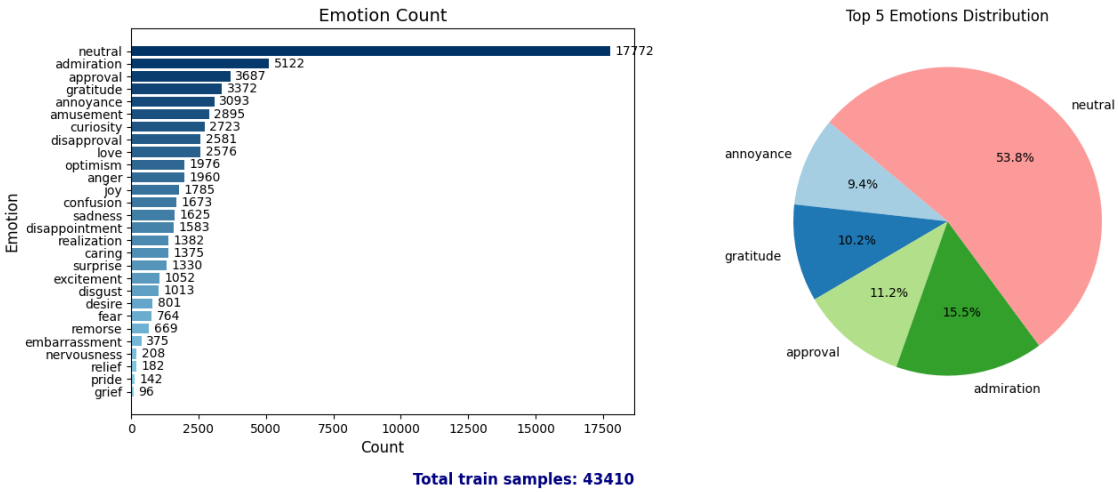


Figure 1: Emotion label distribution in the GoEmotions training set (top 5 emotions highlighted).

Comment: The dataset is highly skewed toward the *neutral* emotion, which accounts for over half of all samples. This imbalance can bias the model toward predicting neutral or positive emotions, making minority categories such as *grief* and *pride* more difficult to learn effectively.

2.2 Model and Fine-Tuning Strategies

A **BERT-base-uncased** model was used for both experiments.

- **Full Fine-Tuning:** All model parameters were updated using cross-entropy loss.
- **LoRA Fine-Tuning:** Only low-rank adapter weights were trained, with base model weights frozen. LoRA rank was set to 8.

2.3 Experimental Setup

Both models were trained for 5 epochs using the AdamW optimizer with a batch size of 16 and a learning rate of 2×10^{-5}. The weight decay was set to 1×10^{-2}, and training used mixed-precision (FP16) to improve efficiency. Evaluation metrics included the macro F1-score and accuracy. The **full fine-tuning** experiment was conducted on Google Colab using a single NVIDIA T4 GPU, while the **LoRA fine-tuning** experiment was performed locally on my laptop equipped with an NVIDIA GeForce RTX 3050Ti GPU. This setup demonstrates LoRA's ability to run efficiently on consumer-grade hardware with limited resources.

3 Results and Analysis

3.1 Overall Performance

Table 1: Comparison between LoRA and Full Fine-Tuning on GoEmotions.

Method	Epoch	Train Loss	Val Loss	F1	Accuracy
LoRA	5	0.1328	0.1303	0.1226	0.0616
Full Fine-Tuning	5	0.0547	0.0897	0.5589	0.4419

3.2 Full Fine-Tuning Evaluation

The best model achieved a **macro F1 of 0.5115**, **micro F1 of 0.6018**, and **subset accuracy of 0.4295**. Performance was strong across common labels but weaker for rare emotions.

Table 2: Top-5 and Bottom-5 Labels by F1 (Full Fine-Tuning).

Label	Precision	Recall	F1	Support
L15: gratitude	0.962	0.872	0.915	352
L1: amusement	0.746	0.902	0.816	264
L18: love	0.733	0.866	0.794	238
L0: admiration	0.696	0.696	0.696	504
L27: neutral	0.611	0.777	0.684	1787
L16: grief	0.000	0.000	0.000	6
L22: realization	0.252	0.207	0.227	145
L21: pride	0.571	0.250	0.348	16
L9: disappointment	0.314	0.404	0.354	151
L3: annoyance	0.319	0.434	0.368	320

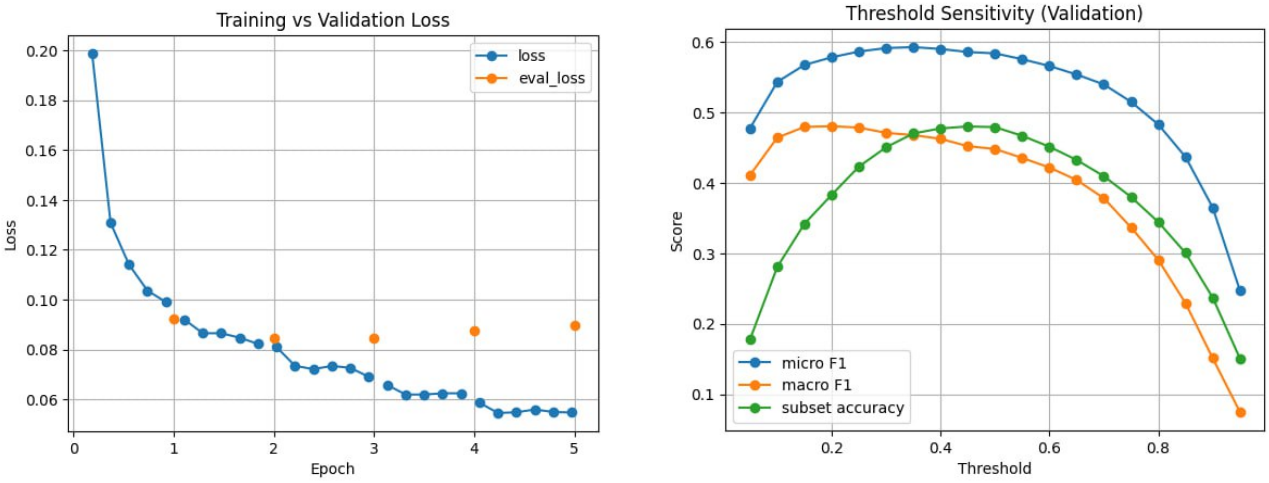
3.2.1 Threshold Optimization

Threshold tuning was performed to identify the decision probability that maximized validation F1. The optimal thresholds varied across labels, ranging roughly between 0.17 and 0.77. Figure 2b illustrates the sensitivity of macro F1, micro F1, and subset accuracy across thresholds. The best overall performance occurred near a threshold of **0.35**, yielding a validation macro F1 of 0.5115.

Table 3: Top 5 Labels by Validation F1 at Best Threshold (Full Fine-Tuning).

Label	Best Threshold	Val F1 at Best Thr.
L15: gratitude	0.77	0.9165
L1: amusement	0.24	0.8066
L18: love	0.30	0.7811
L24: remorse	0.20	0.7703
L0: admiration	0.51	0.7428

Note: This table presents the top five labels ranked by their validation F1 scores at the optimal threshold for each class. Higher-performing labels such as *L15 (gratitude)* and *L1 (amusement)* demonstrate stronger confidence calibration compared to others.



(a) Training vs Validation Loss (Full Fine-Tuning).

(b) Threshold Sensitivity Curve (Full Fine-Tuning).

Figure 2: Full fine-tuning performance curves.

Comment: In Figure 2a, the training and validation losses decrease steadily and converge by epoch 3, indicating efficient learning and minimal overfitting during full fine-tuning. Meanwhile, Figure 2b shows that the model achieves peak macro- and micro-F1 scores around thresholds 0.3–0.4, demonstrating that moderate probability cut-offs yield the best precision–recall balance for multi-label emotion classification.

3.3 LoRA Fine-Tuning Evaluation

The LoRA fine-tuned model achieved **macro F1 of 0.1512**, **micro F1 of 0.2869**, and **subset accuracy of 0.0601**. Although accuracy was lower, LoRA completed training faster and used less GPU memory.

Table 4: Top-5 and Bottom-5 Labels by F1 (LoRA Fine-Tuning).

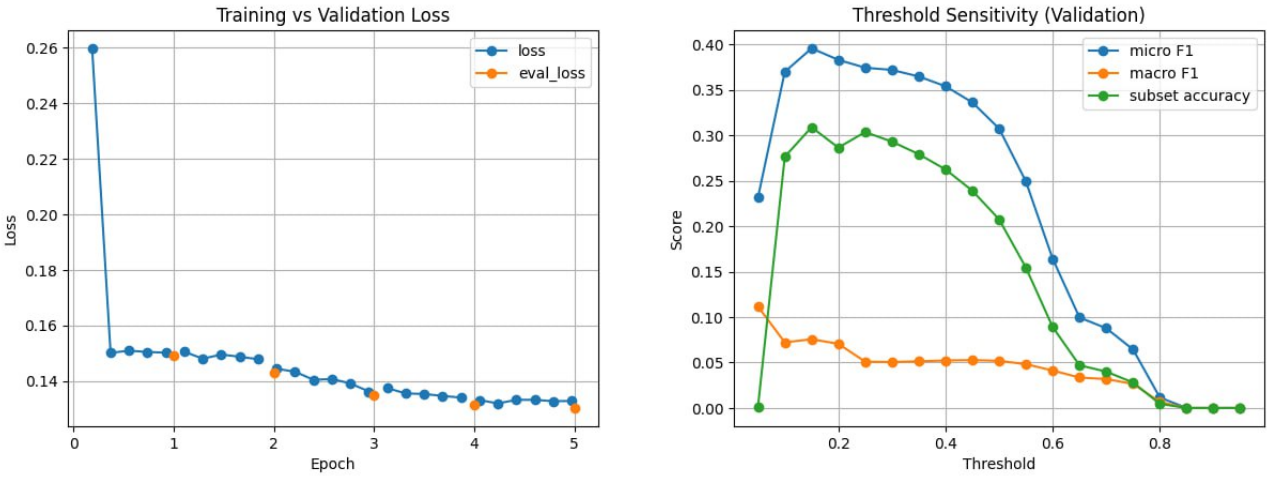
Label	Precision	Recall	F1	Support
L15: gratitude	0.946	0.855	0.899	352
L27: neutral	0.515	0.823	0.633	1787
L18: love	0.512	0.550	0.530	238
L0: admiration	0.225	0.766	0.349	504
L6: confusion	0.197	0.242	0.217	153
L5: caring	0.000	0.000	0.000	135
L14: fear	0.000	0.000	0.000	78
L11: disgust	0.000	0.000	0.000	123
L8: desire	0.000	0.000	0.000	83
L12: embarrassment	0.000	0.000	0.000	37

3.3.1 Threshold Optimization

A similar per-label threshold search was conducted for the LoRA model. The optimal thresholds were generally lower (0.05–0.64), indicating weaker confidence calibration. As shown in Figure 3b, performance peaked at a threshold of approximately **0.25**, achieving a validation macro F1 of 0.1512. This low score reflects limited learning capacity under restricted parameter updates.

Table 5: Example of Best Thresholds (LoRA Fine-Tuning).

Label	Best Threshold	Val F1 at Best Thr.
L15: gratitude	0.64	0.9069
L27: neutral	0.32	0.6410
L18: love	0.17	0.5451
L0: admiration	0.13	0.3467
L7: curiosity	0.07	0.2143



(a) Training vs Validation Loss (LoRA Fine-Tuning).

(b) Threshold Sensitivity Curve (LoRA Fine-Tuning).

Figure 3: LoRA fine-tuning performance curves.

Comment: In Figure 3a, the training and validation losses decrease sharply in the first epoch and then plateau, indicating limited but stable learning due to the small number of trainable parameters in LoRA. Figure 3b shows that performance peaks at a low threshold range (0.2–0.3), where micro-F1 reaches about 0.29. Beyond

this point, all metrics drop rapidly, reflecting weaker confidence calibration and reduced discriminative ability compared to full fine-tuning.

3.4 Hyperparameter Sensitivity

To further examine optimization behavior, both models were retrained using a higher learning rate of 5×10^{-5} and 10 epochs. The new setup slightly improved full fine-tuning performance (macro F1 0.5739), suggesting that the model was already close to saturation at five epochs. However, LoRA achieved a substantial performance increase from macro F1 0.12 to 0.43 and accuracy 0.29. This confirms that parameter-efficient methods such as LoRA require longer training and higher learning rates to adequately update their smaller set of weights. The validation losses for both models remained stable across epochs, indicating no signs of overfitting even under the extended schedule.

Table 6: Effect of Increasing Learning Rate and Training Epochs.

Method	**Epochs**	**Learning Rate**	**F1**	**Accuracy**
Full Fine-Tuning (baseline)	5	2×10^{-5}	0.5589	0.4419
Full Fine-Tuning (new)	10	5×10^{-5}	0.5739	0.4513
LoRA (baseline)	5	2×10^{-5}	0.1226	0.0616
LoRA (new)	10	5×10^{-5}	0.4290	0.2890

Comment: Raising the learning rate and extending training to ten epochs yielded only modest gains for full fine-tuning but produced a large relative improvement for LoRA, which reached an F1 of 0.43. This demonstrates that LoRA benefits from longer and more aggressive training schedules due to its smaller number of trainable parameters.

(a) Full Fine-Tuning performance across 10 epochs

Epoch	**Train Loss**	**Val Loss**	**F1**	**Accuracy**
1	0.0924	0.0878	0.4833	0.3378
2	0.0775	0.0831	0.5361	0.4010
3	0.0611	0.0869	0.5424	0.4178
4	0.0482	0.0978	0.5532	0.4449
5	0.0338	0.1109	0.5515	0.4499
6	0.0237	0.1199	0.5546	0.4502
7	0.0172	0.1279	0.5663	0.4548
8	0.0121	0.1372	0.5681	0.4495
9	0.0090	0.1443	0.5724	0.4502
10	0.0065	0.1472	0.5739	0.4513

(b) LoRA Fine-Tuning performance across 10 epochs

Epoch	**Train Loss**	**Val Loss**	**F1**	**Accuracy**
1	0.1411	0.1341	0.0300	0.0168
2	0.1211	0.1146	0.2284	0.1367
3	0.1084	0.1059	0.3186	0.2007
4	0.1053	0.1011	0.3648	0.2392
5	0.1005	0.0984	0.3929	0.2582
6	0.0977	0.0966	0.4228	0.2862
7	0.0974	0.0958	0.4191	0.2807
8	0.0959	0.0949	0.4318	0.2930
9	0.0970	0.0943	0.4268	0.2875
10	0.0954	0.0942	0.4290	0.2890

Figure 4: Training and validation trends for both methods (LR $= 5 \times 10^{-5}$).

Comment: Both models exhibit smooth convergence. Full fine-tuning shows steady but diminishing gains after epoch 7, while LoRA continues to improve up to around epoch 8–10, confirming that it benefits from longer training schedules due to its limited parameter capacity.

4 Discussion

Performance vs. efficiency. Full fine-tuning clearly outperformed LoRA on GoEmotions (macro F1 ≈ 0.56 vs. 0.12 at the baseline), confirming that updating all weights yields stronger task adaptation for nuanced, multi-label signals. However, LoRA's profile improved substantially when trained longer and with a higher learning rate (macro F1 $\uparrow$ to ≈ 0.43 at 10 epochs, 5×10^{-5}), highlighting a key insight: *parameter-efficient methods need more aggressive schedules* to unlock their capacity.

Calibration and thresholds. Both approaches benefited from per-label threshold tuning. Full fine-tuning peaked around a global threshold of $\sim$0.35, whereas LoRA peaked lower ($\sim$0.25), suggesting weaker confidence calibration for LoRA. High-scoring labels (e.g., *gratitude*, *amusement*, *love*) also had higher optimal thresholds, reflecting more separable logit distributions. In practice, *label-wise thresholds* are essential for multi-label emotion tasks.

Label imbalance dominates outcomes. Frequent labels such as *neutral*, *admiration*, and *approval* drove much of the performance, whereas rare emotions (e.g., *grief*, *pride*) remained difficult. This shows up as near-zero F1 for the rarest classes (especially under LoRA), despite reasonable overall micro-F1. Remedies include class-weighted/focal loss, re-sampling (e.g., tempered oversampling), and targeted data augmentation for minority emotions.

Learning dynamics. Loss curves show smooth convergence and no obvious overfitting; full fine-tuning saturates early (by ∼epoch 3–5), while LoRA keeps improving through ∼epoch 8–10. This supports using *shorter, conservative schedules for full FT* and *longer, higher-LR schedules for LoRA*.

Application relevance (course reviews). For downstream use in course-review recommendation, full fine-tuning currently offers stronger detection of positive/neutral affect (useful for ranking "helpful" or "encouraging" reviews). With schedule tuning, LoRA becomes a viable on-device or low-cost option, especially if paired with threshold calibration and imbalance mitigation.

Key Takeaways

- Full fine-tuning delivers the best accuracy and F1 out-of-the-box; LoRA trades accuracy for speed and memory.
- LoRA *needs* longer training and higher LR to be competitive; schedule matters more than usual.
- Per-label thresholding is not optional—it's a major lever for macro/micro-F1 gains.
- Label imbalance is the main failure mode; without rebalancing, rare emotions underperform.
- For practical deployment, choose full FT when resources allow; choose LoRA for scale/latency, but budget for schedule tuning and calibration.

Limitations

- **Single backbone.** Results are on `bert-base-uncased`; other backbones (e.g., RoBERTa, DeBERTa) or larger models may shift conclusions.
- **Loss/metrics.** Standard BCE with macro/micro-F1; alternative objectives (focal loss, class-balanced loss) and ranking-aware metrics were not explored.
- **Imbalance handling.** No dedicated re-weighting/augmentation pipeline; rare-label performance likely underestimates the attainable ceiling.
- **Threshold search scope.** Thresholds were tuned simply; more principled calibration (temperature scaling, isotonic regression) may further help.
- **Hardware variability.** Runs were split across T4 and RTX 3050 Ti; wall-clock comparisons are indicative, not strictly controlled.

5 Conclusion

This study shows a clear accuracy–efficiency trade-off between full and LoRA fine-tuning for multi-label emotion classification on GoEmotions. Full fine-tuning achieves the strongest performance with modest schedules, while LoRA—though initially weaker—closes a large part of the gap when trained longer at a higher learning rate and paired with label-wise thresholding. For resource-rich, quality-critical applications (e.g., evaluating course reviews to surface nuanced sentiment), full fine-tuning is preferable. For edge or large-scale deployments, LoRA is attractive, provided one invests in schedule tuning, calibration, and imbalance mitigation.

Future work. Combine LoRA with class-balanced or focal loss, perform systematic calibration (temperature scaling per label), and model label dependencies (classifier chains or structured prediction). Explore stronger backbones and prompt/adapter variants (prefix-tuning, IA^3), and integrate active learning to acquire more examples for rare emotions. These steps should raise minority-label F1 while preserving LoRA's efficiency.