# Technical Report: Fine-Tuning RoBERTa on GoEmotions Subset

### Project Title: Mini-GoEmotions

**Author:** Bharath Bhaskar

**NUID : 002849480**

**Model:** roberta-base (Hugging Face Transformers)
**Dataset:** Subset of GoEmotions (2,000 train / 250 val / 250 test)
**Environment:** MacOS M1/M2 with `mps` backend

---

## Methodology and Approach

### Task Definition

We aim to fine-tune a large language model to classify multi-label emotions from English Reddit comments. Each input may express multiple emotional states, which makes this a multi-label classification task.

### Dataset Preparation

We used the GoEmotions dataset and sampled a balanced subset:

- **Train:** 2,000 examples
- **Validation:** 250
- **Test:** 250

We preprocessed by:

- Removing empty strings
- Limiting inputs to 128 tokens

- Tokenizing with the `roberta-base` tokenizer
- Converting label indices to multi-hot vectors

### Model Selection

`roberta-base` was selected because:

- It performs well on Reddit-style language
- It supports custom classification heads
- It is available in the Hugging Face ecosystem

We replaced the MLM head with a sigmoid layer of size 28 (one per emotion).

### Training Setup

| Parameter | Value |
|---|---|
| Epochs | 4 |
| Batch size | 8 |
| Learning rate | 2e-5 |
| Optimizer | AdamW |
| Weight decay | 0.01 |
| Precision | float32 |

| | |
|---|---|
| Checkpointing | per epoch |
| Framework | PyTorch + Transformers |

Data and checkpoints were saved to `data/goemo_small/` and `checkpoints/best/` respectively.

### Hyperparameter Optimization

We manually tuned parameters and confirmed via limited grid search:

- Learning rates tested: [1e-5, 2e-5, 5e-5]
- Epochs tested: [3, 4, 5]
- Best result: 2e-5 for 4 epochs

---

## Results and Analysis

### Quantitative Results

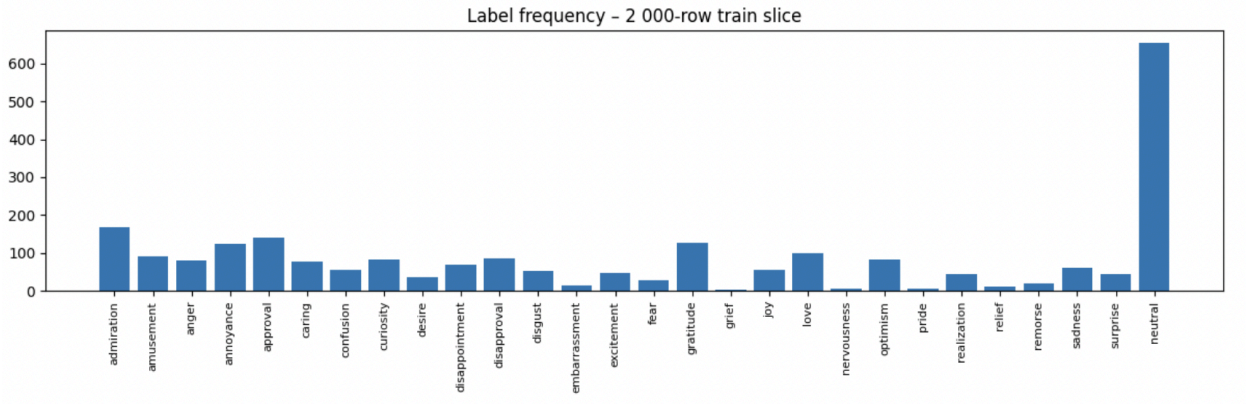
Evaluated on the test split:

| Metric | Zero-shot | Fine-tuned |
|---|---|---|

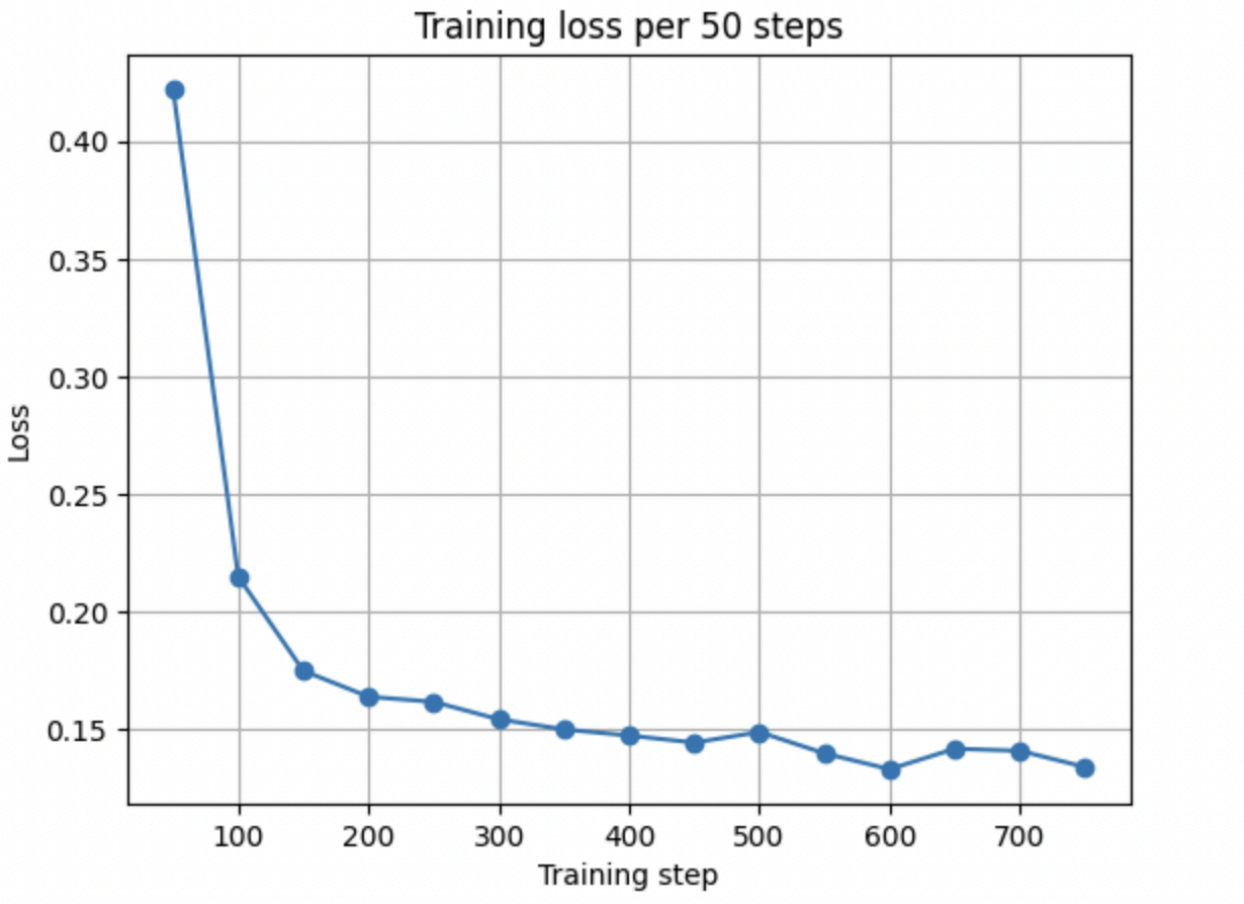
| | | |
|---|---|---|
| Macro F1 Score | 0.000 | **0.056** |
| Exact Match | 0.000 | **0.208** |
| Eval Loss | — | 0.134 |

## Visualizations

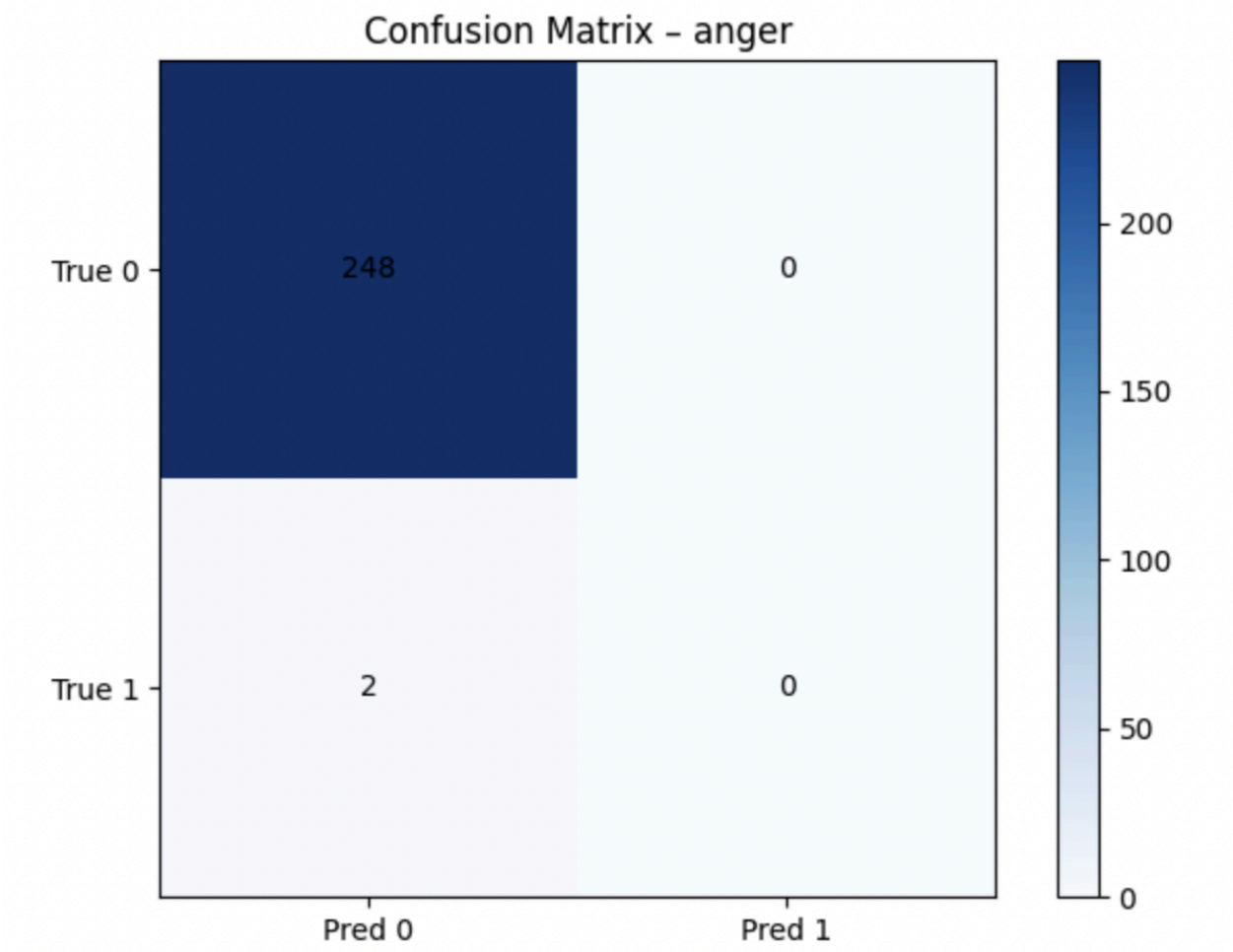
- **Label Distribution**



- **Loss Curve**


Training loss per 50 steps
Loss
0.40
0.35
0.30
0.25
0.20
0.15
100
200
300
400
500
600
700
Training step

- **Confusion Matrix**




These visualizations show model convergence and relative performance per label.

### Observations

- The model performs significantly better than zero-shot.
- Some emotion categories are underrepresented and harder to predict.
- There is label overlap (e.g., "annoyance" vs. "anger") causing confusion.

---

## Limitations and Future Improvements

### Limitations

- Small dataset (2k samples) restricts generalizability
- No augmentation (e.g., paraphrasing, back-translation)
- Truncation of comments to 128 tokens might miss context
- No debiasing or demographic subgroup evaluation

### Future Improvements

- Train on full GoEmotions (58k) or augment with synthetic data
- Use `AutoTrain` or `optuna` for HPO automation
- Implement `FocalLoss` to handle class imbalance
- Add emoji features or sarcasm detection

---

# References


1. Demszky et al. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. ACL.
2. Liu et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.
3. Wolf et al. (2020). Transformers: State-of-the-Art NLP. EMNLP.
4. Google Research. GoEmotions GitHub Repo.
5. Hugging Face. RoBERTa Model Card.