Assignment: Fine-Tuning a Large Language Model

By: Shubh Patel (002822971)

# 1. Introduction & Motivation

This report presents a comprehensive study on fine-tuning large language models (LLMs) for emotion classification tasks using subsets of the GoEmotions dataset. The motivation lies in evaluating how model architecture, size, and training data volume affect performance, particularly in low- and mid-resource settings. With growing interest in efficient and accessible LLM deployments, this study benchmarks four distinct models (BERT, GPT2-XL, Qwen2.5, and LLaMA-3.2) across twelve dataset configurations.

# 2. Dataset Preparation and Task Design

The GoEmotions dataset consists of 58k Reddit comments annotated with 27 emotion classes. For this project, subsets were carefully curated to reflect varying complexity by altering the number of labels (2, 4, 8, 16) and sample sizes (800, 1600, 2400 per configuration). This leads to 12 evaluation settings. Each subset was balanced and split into train and validation sets. Tokenization was handled using each model's respective tokenizer. Text inputs were padded/truncated to a maximum length of 128 tokens. Below are the preprocessed configuration details and task setup.

# 3. Model Selection and Fine-Tuning Setup

Four pre-trained LLMs were chosen for fine-tuning:  
- BERT-base (110M): Lightweight transformer baseline.  
- GPT2-XL (1.5B): Large autoregressive decoder.  
- Qwen2.5 (1.5B): Instruction-tuned LLM optimized for multi-tasking.  
- LLaMA-3.2 (1B): Compact encoder-decoder hybrid.  
  
Fine-tuning was performed using Hugging Face's `transformers`, `datasets`, and `peft` libraries. Training was optimized for resource efficiency using:  
- 4-bit quantization via BitsAndBytes.  
- LoRA adapters on attention layers.  
- AdamW optimizer with cosine learning rate schedule.  
- bf16 mixed-precision training and early stopping.  
Each model was fine-tuned for 5 epochs with a batch size of 4 and gradient accumulation steps of 1.

# 4. Hyperparameter Tuning & Evaluation

Hyperparameter tuning was performed over learning rates [1e-4, 5e-5, 2e-5]. Evaluation metrics were focused on accuracy across all 12 settings. The best hyperparameters for each model-dataset pair were selected based on validation accuracy.   
  
Error analysis revealed that:  
- Label complexity significantly impacts model accuracy, especially for BERT.  
- Data volume helps in reducing overfitting in high-label tasks.  
- GPT2-XL was prone to instability in low-data, high-label settings.  
Selected training logs and evaluation scripts from the notebook are included below:

Code Snippet (from Fine-tuning Notebook):

from transformers import TrainingArguments, Trainer  
training\_args = TrainingArguments(  
 output\_dir='./results',  
 evaluation\_strategy='epoch',  
 per\_device\_train\_batch\_size=4,  
 per\_device\_eval\_batch\_size=4,  
 num\_train\_epochs=5,  
 learning\_rate=1e-4,  
 logging\_dir='./logs',  
 save\_strategy='epoch',  
 fp16=False,  
 bf16=True,  
 load\_best\_model\_at\_end=True,  
 metric\_for\_best\_model='accuracy',  
)

# 5. Results & Visualization

The following heatmap shows accuracy results across all models and datasets. Qwen2.5 and LLaMA-3.2 exhibit consistent performance, while BERT sharply drops under complex tasks. This reinforces the role of parameter count and instruction tuning in generalization.

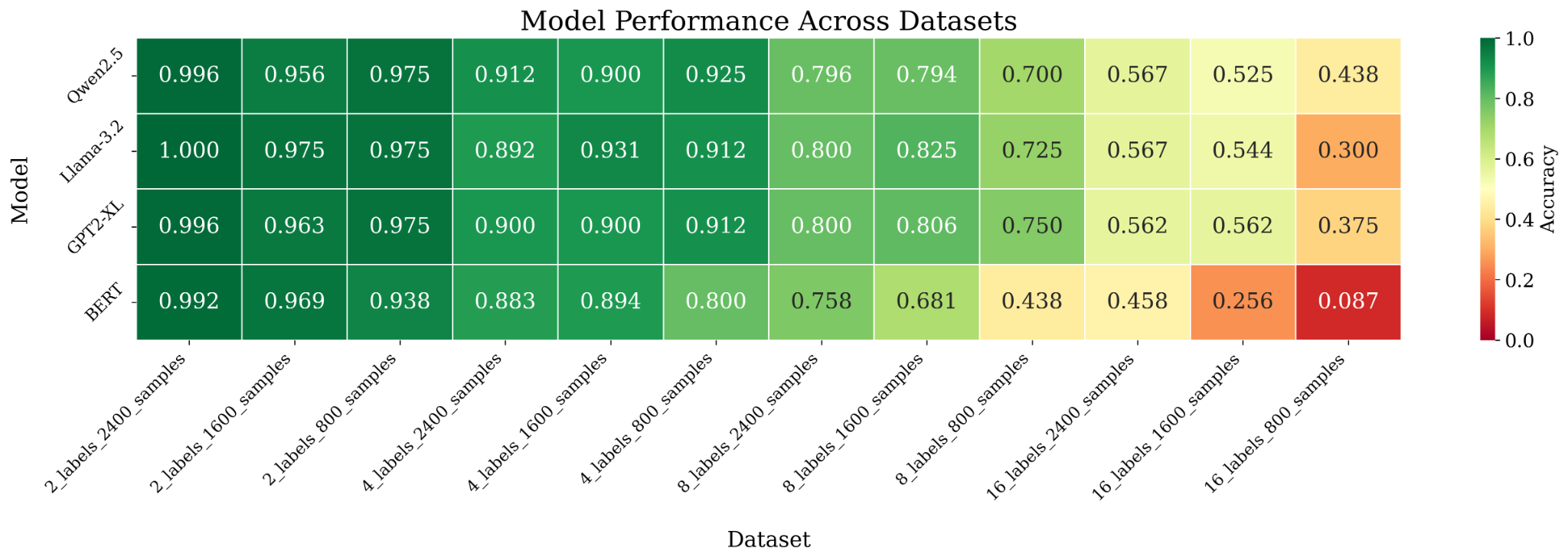


Figure 1: Model Performance Heatmap Across Label and Sample Variants

Table 1: Accuracy Comparison Across All Settings

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unnamed: 0 | eval\_loss | eval\_accuracy | eval\_precision | eval\_recall | eval\_f1 | eval\_runtime | eval\_samples\_per\_second | eval\_steps\_per\_second | epoch | Dataset | Model |
| 0 | 0.1268323361873626 | 0.975 | 0.9761904761904764 | 0.975 | 0.9749843652282676 | 3.9009 | 20.508 | 5.127 | 4.75 | 2\_labels\_800\_samples | meta-llama/Llama-3.2-1B |
| 1 | 0.1507912576198578 | 0.975 | 0.975 | 0.975 | 0.975 | 7.5934 | 21.071 | 5.268 | 2.5 | 2\_labels\_1600\_samples | meta-llama/Llama-3.2-1B |
| 2 | 0.0002776786568574 | 1.0 | 1.0 | 1.0 | 1.0 | 11.3483 | 21.149 | 5.287 | 2.083333333333333 | 2\_labels\_2400\_samples | meta-llama/Llama-3.2-1B |
| 3 | 0.2700769007205963 | 0.9125 | 0.9134920634920636 | 0.9125 | 0.9120988446726572 | 3.7986 | 21.061 | 5.265 | 5.0 | 4\_labels\_800\_samples | meta-llama/Llama-3.2-1B |
| 4 | 0.4807419180870056 | 0.93125 | 0.9332286233602024 | 0.93125 | 0.9309193257722668 | 7.6838 | 20.823 | 5.206 | 3.875 | 4\_labels\_1600\_samples | meta-llama/Llama-3.2-1B |
| 5 | 0.4781498610973358 | 0.8916666666666667 | 0.8994194922978761 | 0.8916666666666667 | 0.8927574522880837 | 11.2972 | 21.244 | 5.311 | 1.75 | 4\_labels\_2400\_samples | meta-llama/Llama-3.2-1B |
| 6 | 0.797638 | 0.725 | 0.729167 | 0.725 | 0.725721 | 3.8527 | 20.765 | 5.191 | 5.0 | 8\_labels\_800\_samples | meta-llama/Llama-3.2-1B |
| 7 | 0.99866 | 0.825 | 0.837546 | 0.825 | 0.825291 | 7.4992 | 21.336 | 5.334 | 3.625 | 8\_labels\_1600\_samples | meta-llama/Llama-3.2-1B |
| 8 | 0.681639 | 0.8 | 0.81347 | 0.8 | 0.801115 | 11.2757 | 21.285 | 5.321 | 2.5 | 8\_labels\_2400\_samples | meta-llama/Llama-3.2-1B |
| 9 | 2.447251 | 0.3 | 0.328968 | 0.3 | 0.290323 | 3.8083 | 21.007 | 5.252 | 5.0 | 16\_labels\_800\_samples | meta-llama/Llama-3.2-1B |
| 10 | 1.77939 | 0.54375 | 0.546076 | 0.54375 | 0.534099 | 7.5813 | 21.105 | 5.276 | 4.5 | 16\_labels\_1600\_samples | meta-llama/Llama-3.2-1B |
| 11 | 1.485324 | 0.566667 | 0.580383 | 0.566667 | 0.549572 | 12.0165 | 19.973 | 4.993 | 2.833333 | 16\_labels\_2400\_samples | meta-llama/Llama-3.2-1B |
| 12 | 0.0639667958021164 | 0.975 | 0.975 | 0.975 | 0.975 | 4.4984 | 17.784 | 4.446 | 5.0 | 2\_labels\_800\_samples | openai-community/gpt2-xl |
| 13 | 0.1864247322082519 | 0.9625 | 0.9636591478696742 | 0.9625 | 0.9624765478424016 | 8.8156 | 18.15 | 4.537 | 2.125 | 2\_labels\_1600\_samples | openai-community/gpt2-xl |
| 14 | 0.0211021658033132 | 0.9958333333333332 | 0.9958677685950412 | 0.9958333333333332 | 0.9958332609941144 | 13.2857 | 18.065 | 4.516 | 1.8333333333333333 | 2\_labels\_2400\_samples | openai-community/gpt2-xl |
| 15 | 0.4365797042846679 | 0.9125 | 0.917929292929293 | 0.9125 | 0.9129072681704262 | 4.4333 | 18.045 | 4.511 | 5.0 | 4\_labels\_800\_samples | openai-community/gpt2-xl |
| 16 | 0.3712025582790375 | 0.9 | 0.9013516970834046 | 0.9 | 0.8994747182066746 | 8.8283 | 18.123 | 4.531 | 2.125 | 4\_labels\_1600\_samples | openai-community/gpt2-xl |
| 17 | 0.4443413913249969 | 0.9 | 0.9038773726273726 | 0.9 | 0.9006520288242975 | 13.4055 | 17.903 | 4.476 | 2.583333333333333 | 4\_labels\_2400\_samples | openai-community/gpt2-xl |
| 18 | 0.996440887451172 | 0.75 | 0.7579004329004329 | 0.75 | 0.7487331296154827 | 4.493 | 17.806 | 4.451 | 5.0 | 8\_labels\_800\_samples | openai-community/gpt2-xl |
| 19 | 0.785417914390564 | 0.80625 | 0.8156397044984001 | 0.80625 | 0.8053535564454452 | 8.98 | 17.817 | 4.454 | 2.75 | 8\_labels\_1600\_samples | openai-community/gpt2-xl |
| 20 | 0.7972381114959717 | 0.8 | 0.8143282047537252 | 0.8 | 0.8026701407583258 | 13.3935 | 17.919 | 4.48 | 3.0 | 8\_labels\_2400\_samples | openai-community/gpt2-xl |
| 21 | 2.0573832988739014 | 0.375 | 0.4104166666666666 | 0.375 | 0.379983660130719 | 4.5211 | 17.695 | 4.424 | 5.0 | 16\_labels\_800\_samples | openai-community/gpt2-xl |
| 22 | 1.5388860702514648 | 0.5625 | 0.5578317515817515 | 0.5625 | 0.5497441085858099 | 8.95 | 17.877 | 4.469 | 5.0 | 16\_labels\_1600\_samples | openai-community/gpt2-xl |
| 23 | 1.3822619915008545 | 0.5625 | 0.604629983306454 | 0.5625 | 0.5497471049342219 | 13.8438 | 17.336 | 4.334 | 2.583333333333333 | 16\_labels\_2400\_samples | openai-community/gpt2-xl |
| 24 | 0.0963472872972488 | 0.975 | 0.975 | 0.975 | 0.975 | 4.4954 | 17.796 | 4.449 | 5.0 | 2\_labels\_800\_samples | Qwen/Qwen2.5-1.5B |
| 25 | 0.1384466588497162 | 0.95625 | 0.9563213002031568 | 0.95625 | 0.9562482909488652 | 8.9811 | 17.815 | 4.454 | 3.625 | 2\_labels\_1600\_samples | Qwen/Qwen2.5-1.5B |
| 26 | 0.0123618934303522 | 0.9958333333333332 | 0.9958677685950412 | 0.9958333333333332 | 0.9958332609941144 | 13.415 | 17.89 | 4.473 | 3.75 | 2\_labels\_2400\_samples | Qwen/Qwen2.5-1.5B |
| 27 | 0.3699818253517151 | 0.925 | 0.9261904761904762 | 0.925 | 0.9252970606629144 | 4.4489 | 17.982 | 4.496 | 5.0 | 4\_labels\_800\_samples | Qwen/Qwen2.5-1.5B |
| 28 | 0.4736730456352234 | 0.9 | 0.9031998123543477 | 0.9 | 0.899465477051684 | 9.0007 | 17.776 | 4.444 | 4.5 | 4\_labels\_1600\_samples | Qwen/Qwen2.5-1.5B |
| 29 | 0.5588087439537048 | 0.9125 | 0.9154871900502196 | 0.9125 | 0.9129650472968184 | 13.4223 | 17.881 | 4.47 | 3.8333333333333335 | 4\_labels\_2400\_samples | Qwen/Qwen2.5-1.5B |
| 30 | 1.3536405563354492 | 0.7 | 0.7332970848595849 | 0.7 | 0.7023573747560825 | 4.4453 | 17.997 | 4.499 | 5.0 | 8\_labels\_800\_samples | Qwen/Qwen2.5-1.5B |
| 31 | 1.065798044204712 | 0.79375 | 0.799970597750203 | 0.79375 | 0.7930854204058424 | 8.9888 | 17.8 | 4.45 | 5.0 | 8\_labels\_1600\_samples | Qwen/Qwen2.5-1.5B |
| 32 | 0.908302366733551 | 0.7958333333333333 | 0.8107484543460153 | 0.7958333333333333 | 0.799674224562392 | 13.464 | 17.825 | 4.456 | 4.416666666666667 | 8\_labels\_2400\_samples | Qwen/Qwen2.5-1.5B |
| 33 | 2.300947904586792 | 0.4375 | 0.4644345238095238 | 0.4375 | 0.4284236596736596 | 4.5175 | 17.709 | 4.427 | 5.0 | 16\_labels\_800\_samples | Qwen/Qwen2.5-1.5B |
| 34 | 1.8477462530136108 | 0.525 | 0.5475837703962704 | 0.525 | 0.5175325655624569 | 8.9408 | 17.896 | 4.474 | 5.0 | 16\_labels\_1600\_samples | Qwen/Qwen2.5-1.5B |
| 35 | 1.6217119693756104 | 0.5666666666666667 | 0.557830209669707 | 0.5666666666666667 | 0.5530889212462922 | 14.431 | 16.631 | 4.158 | 4.666666666666667 | 16\_labels\_2400\_samples | Qwen/Qwen2.5-1.5B |
| 36 | 0.2939144670963287 | 0.9375 | 0.9399748585795098 | 0.9375 | 0.9374119856047568 | 0.7931 | 100.868 | 25.217 | 4.5 | 2\_labels\_800\_samples | google-bert/bert-base-uncased |
| 37 | 0.1736839115619659 | 0.96875 | 0.9688232536333804 | 0.96875 | 0.9687487792491892 | 1.5325 | 104.404 | 26.101 | 2.25 | 2\_labels\_1600\_samples | google-bert/bert-base-uncased |
| 38 | 0.0469948835670948 | 0.9916666666666668 | 0.9918032786885246 | 0.9916666666666668 | 0.9916660879227724 | 2.281 | 105.216 | 26.304 | 2.333333333333333 | 2\_labels\_2400\_samples | google-bert/bert-base-uncased |
| 39 | 0.5674155950546265 | 0.8 | 0.7995670995670995 | 0.8 | 0.7951219512195122 | 0.7802 | 102.54 | 25.635 | 5.0 | 4\_labels\_800\_samples | google-bert/bert-base-uncased |
| 40 | 0.4876578450202942 | 0.89375 | 0.8936360355271005 | 0.89375 | 0.893221082968681 | 1.4626 | 109.394 | 27.348 | 4.125 | 4\_labels\_1600\_samples | google-bert/bert-base-uncased |
| 41 | 0.5397012233734131 | 0.8833333333333333 | 0.8848969948706512 | 0.8833333333333333 | 0.8837278128786668 | 2.1981 | 109.187 | 27.297 | 3.8333333333333335 | 4\_labels\_2400\_samples | google-bert/bert-base-uncased |
| 42 | 1.547277808189392 | 0.4375 | 0.4611441798941799 | 0.4375 | 0.3975441015882192 | 0.7443 | 107.479 | 26.87 | 5.0 | 8\_labels\_800\_samples | google-bert/bert-base-uncased |
| 43 | 1.017117261886597 | 0.68125 | 0.6789498481705831 | 0.68125 | 0.670508793552386 | 1.504 | 106.384 | 26.596 | 5.0 | 8\_labels\_1600\_samples | google-bert/bert-base-uncased |
| 44 | 0.7570485472679138 | 0.7583333333333333 | 0.758992521159635 | 0.7583333333333333 | 0.7552682763805101 | 2.1754 | 110.325 | 27.581 | 3.333333333333333 | 8\_labels\_2400\_samples | google-bert/bert-base-uncased |
| 45 | 2.745312452316284 | 0.0875 | 0.0262880067567567 | 0.0875 | 0.037087912087912 | 0.7624 | 104.93 | 26.233 | 3.75 | 16\_labels\_800\_samples | google-bert/bert-base-uncased |
| 46 | 2.225634813308716 | 0.25625 | 0.1969231997979137 | 0.25625 | 0.1936620486855854 | 1.4963 | 106.93 | 26.733 | 4.75 | 16\_labels\_1600\_samples | google-bert/bert-base-uncased |
| 47 | 1.8312296867370603 | 0.4583333333333333 | 0.4233559158312447 | 0.4583333333333333 | 0.4231779008592211 | 2.1849 | 109.845 | 27.461 | 3.6666666666666665 | 16\_labels\_2400\_samples | google-bert/bert-base-uncased |

# 6. Conclusion & Reflection

This study demonstrates the feasibility of fine-tuning large language models under compute constraints while maintaining performance for specific NLP tasks like emotion classification. Models like Qwen2.5 and LLaMA-3.2 offer strong performance in complex scenarios due to their scaling and pretraining strategies. BERT, though efficient, struggles in fine-grained tasks.  
  
Future work could explore instruction tuning, synthetic data augmentation, or knowledge distillation for boosting performance in low-resource regimes. Ethical filtering and demographic fairness in datasets were also considered during preprocessing. Overall, the project illustrates a scalable and structured approach to LLM fine-tuning.