

Label Generation and Emotion Classification with Encoders and Decoders

Experiment Tracking Links:

🔗 **W&B Project Overview:** <https://wandb.ai/mourlayetraore120-the-university-of-texas-at-dallas/nlp-emotion-classification/table?nw=nwusermourlayetraore120>

Introduction

This repository explores advanced fine-tuning techniques for **multi-label emotion detection** on tweets, comparing encoder-based and decoder-based language models.

We experimented with both traditional classification heads and text-based label generation to tackle the task of predicting emotions from tweets labeled with 11 possible emotions.

One of the experiments focused on **label generation**, enabling the model to generate the appropriate set of emotions directly as text — effectively reframing multi-label classification as a sequence generation task.

For this, we fine-tuned the compact **Qwen3-0.6B-Base** decoder-only model using **QLoRA** (Quantized Low-Rank Adaptation), a technique that allowed us to efficiently train a large model with 4-bit quantization and adapter layers, reducing memory while maintaining competitive performance.

Workflow

The main workflow of the experiments included:

- Cleaning and preprocessing the tweet dataset
- Converting binary emotion labels into text-based labels (for the label generation approach)
- Formatting input prompts to guide decoder models toward accurate label outputs

- Fine-tuning the models (both encoder and decoder) using **QLoRA** and 4-bit quantization where applicable
- Evaluating model performance with validation metrics and Kaggle public leaderboard scores
- Logging experiments through **Weights & Biases (W&B)** for reproducibility and tracking

These experiments provided hands-on experience with **prompt engineering**, **parameter-efficient fine-tuning (PEFT)**, and understanding how decoder-style language models can be adapted for structured output generation.

Results & Analysis

Several models were fine-tuned and evaluated to assess their effectiveness in predicting emotion labels from tweets.

Encoder-based models — particularly RoBERTa and DistilBERT — clearly outperformed decoder-only models in both internal metrics and Kaggle public scores.

- **Encoder Models**
 - **RoBERTa** delivered the best results, achieving a Kaggle public score of **0.52961**, with strong macro and micro F1 validation scores.
 - **DistilBERT** followed closely, achieving around **0.514**, demonstrating that even smaller, more efficient encoder models perform well for this task.
- **Decoder Models**
 - Despite leveraging **QLoRA** and 4-bit quantization, decoder-only models underperformed compared to encoders.
 - **Qwen2.5-7B** achieved lower scores of **0.23999** and **0.30040**.
 - **Qwen3-0.6B**, using label generation, showed improvement, reaching a Kaggle public score of **0.48605**, but still lagged behind the encoder models.

While the **Qwen3** experiment showed that decoder models can improve with prompt design and tuning, encoders remain better suited for structured multi-label classification tasks.

Conclusion

These experiments highlight the strengths and limitations of different architectures for multi-label emotion classification:

RoBERTa remains the top choice, delivering the highest public score of **0.52961** and excellent validation metrics.

DistilBERT offers a good trade-off between performance and computational efficiency.

⚠️ *Decoder models like **Qwen2.5-7B** and **Qwen3-0.6B** (with label generation) are improving but remain less precise than encoders for this type of task.*

💡 *Advanced fine-tuning methods like **QLoRA** help reduce resource demands but do not fully close the performance gap.*







These findings reinforce that while decoder-only models are promising for generative tasks, encoder-based models remain superior for structured classification challenges like multi-label emotion detection.

Experiment Tracking Links (again for convenience):

🔗 **W&B Project Overview:** <https://wandb.ai/mourlayetraore120-the-university-of-texas-at-dallas/nlp-emotion-classification/table?nw=nwusermourlayetraore120>

Kaggle Public Scores Summary:

Below is a screenshot (attached in the repository) summarizing all submission scores across the experiments for easy comparison.

Submission and Description		Public Score ⓘ	Select
	kaggle_submission.csv Complete · 17s ago · Decoder · Only Qwen3	0.48605	<input type="checkbox"/>
	hw6_qwen_median_submission.csv Complete · 9d ago	0.30040	<input type="checkbox"/>
	hw6_qwen_submission.csv ←	0.23999	<input type="checkbox"/>
	exp3_distilroberta.csv Complete · 17d ago · Exp3_distilroberta model.	0.51404	<input type="checkbox"/>
	exp2_distilbert.csv Complete · 17d ago · Submission of exp2. Distilbert model.	0.51409	<input type="checkbox"/>
	EXP1_roberta.csv ← Complete · 17d ago · HW5_submission. Roberta model comes with the best performance metrics.	0.52961	<input type="checkbox"/>