LLM Project

Beyond Sentiment:
Turning Negative Reviews into Actionable Insights

Korinne Angela White

Project Goals

This project creates a language model capable of efficiently analyzing negative feedback and turning it into actionable tasks, enabling businesses to quickly identify and address areas for service improvement.
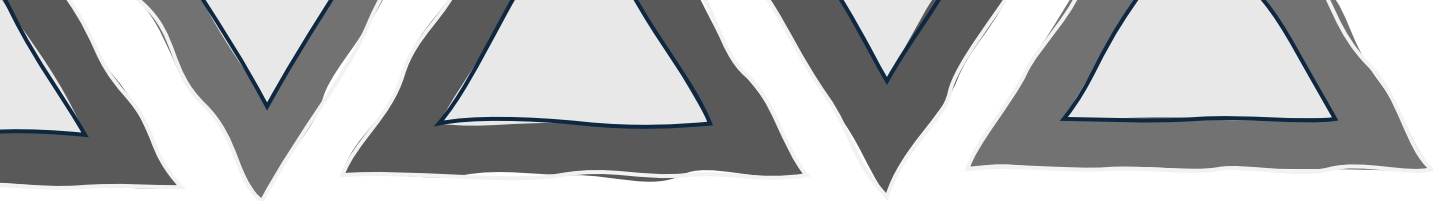
Resources
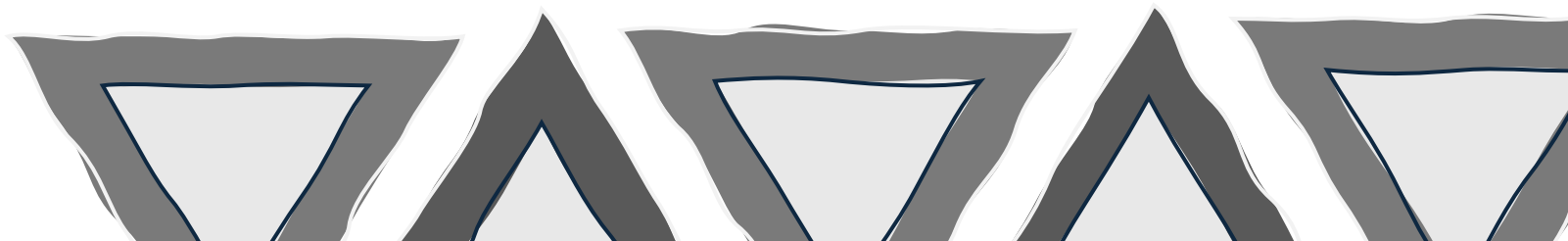
Dataset:

YELP's Full Review Dataset

Pre-trained Models:

- SamLowe/roberta-base-go_emotions

- Mistral-7B-Instruct-v0.1

- Microsoft/phi-2

Process Overview

1. Use SamLowe's RoBERTa-base GoEmotions multi-class classification model on the reviews

2. Use Mistral 7B to generate constructive feedback (Gold Standard)

3. Benchmark base phi-2 to track progress

4. Quantize and fine-tune phi-2 to generate constructive feedback
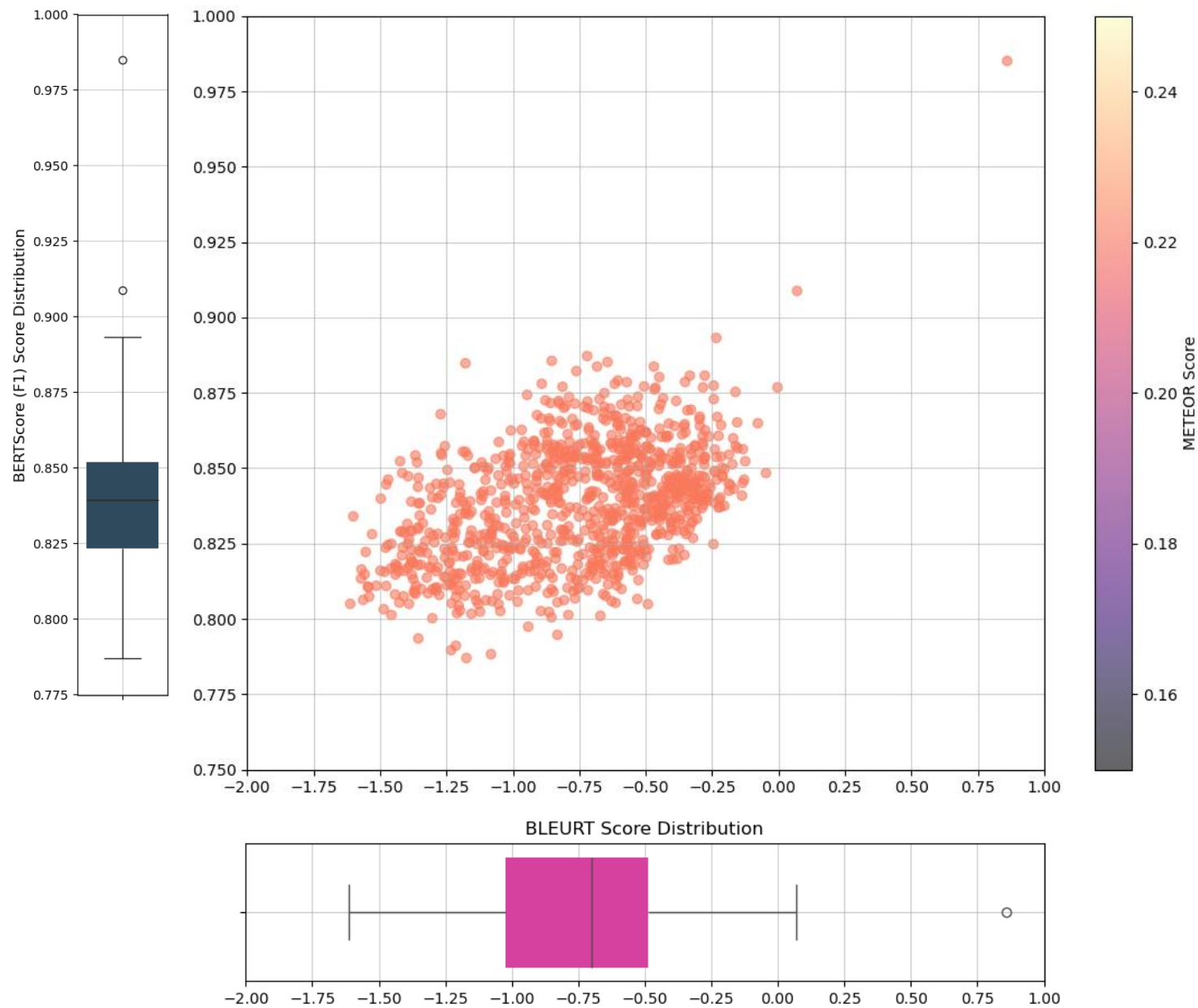
5. Evaluation of the fine-tuned model

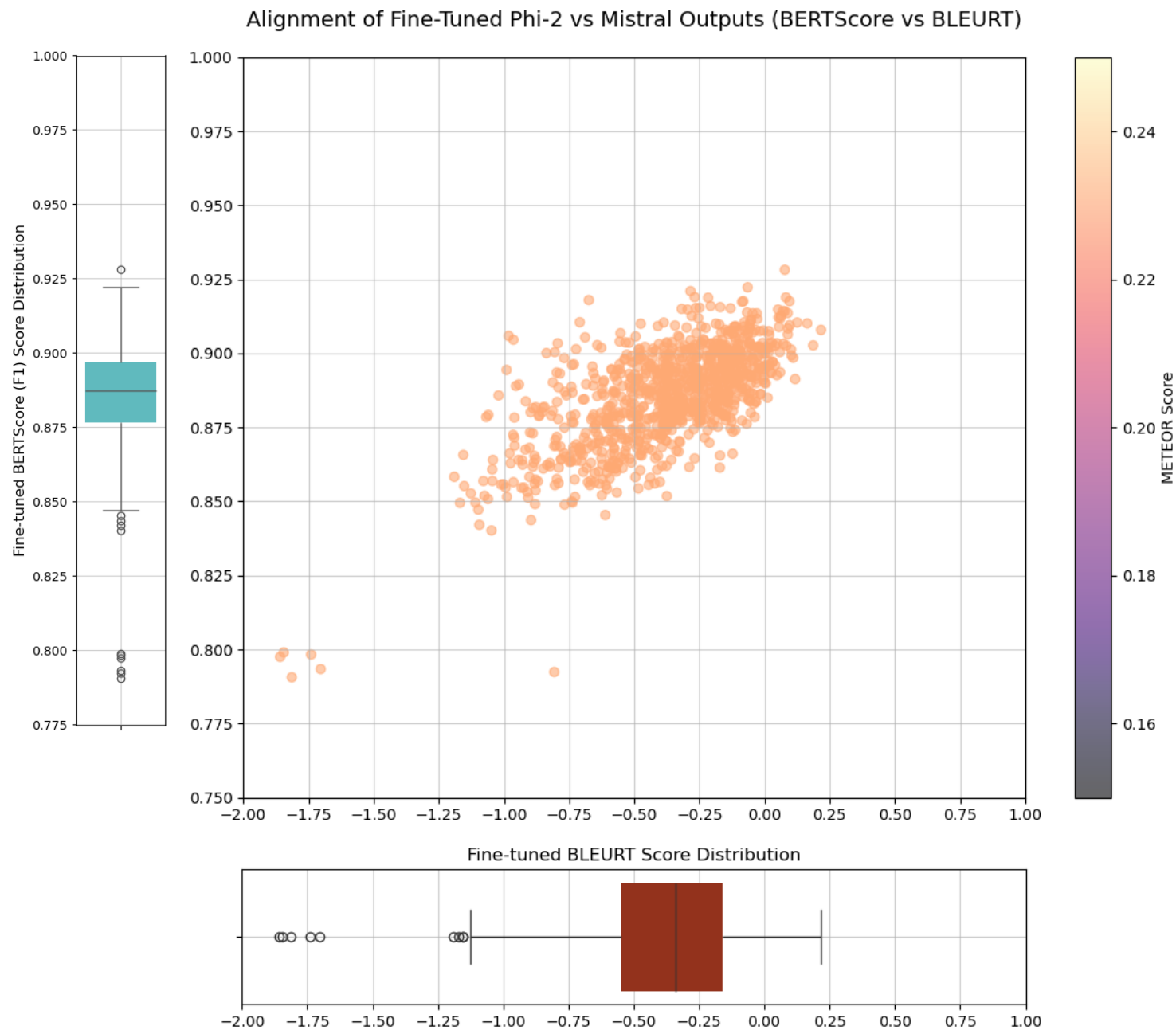# RESULTS

Metrics Overview

1. **BLEURT** -- Bilingual Evaluation Understudy with Representations from Transformers

2. **BERTScore** -- Bidirectional Encoder Representations from Transformers

3. **METEOR** Score -- Metric for Evaluation of Translation with Explicit Ordering

Alignment of Phi-2 vs Mistral Outputs (BERTScore vs BLEURT)

Initial Performance

Fine-Tuned Model
Performance

Alignment of Fine-Tuned Phi-2 vs Mistral Outputs (BERTScore vs BLEURT)

Fine-tuned BERTScore (F1) Score Distribution

METEOR Score

Fine-tuned BLEURT Score Distribution

HYPERPARAMETERS

Hyperparameters

During Training:

- `learning_rate=2e-5`
- `metric_for_best_model="loss" + greater_is_better=False`
- `num_train_epochs=3`
- `weight_decay=0.01`
- `optim="adamw_torch_fused"`

During Generation/Inference:

- `temperature=0.5`
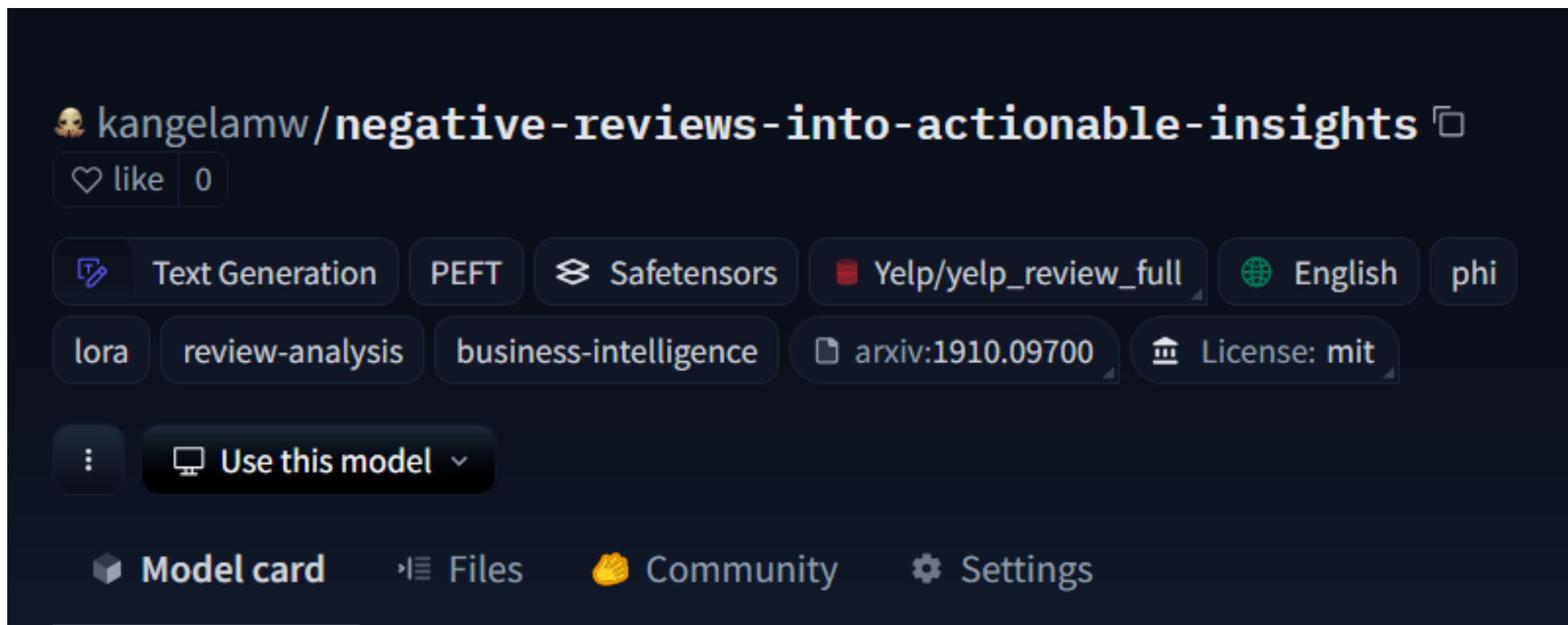- `top_p=0.9`
- `no_repeat_ngram_size=3`

DEPLOYMENT

It's on Hugging Face!

Technical Approach/Requirements

Quantization Method

- QLoRA 4-bit quantization

Hardware

- 12GB VRAM recommended

Library

- PEFT (Parameter-Efficient Fine-tuning)

References

Readings:

- Enable your device for Development

- GoEmotions (Github/Hugging Face)

- Leveraging LLM-as-a-Judge for Automated and Scalable Evaluation

- Reliable Confidence Intervals for Information Retrieval Evaluation Using Generative A.I.

- Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

- BLEURT: a Transfer Learning-Based Metric for Natural Language Generation

- BERTScore

- METEOR Score

- Hugging Face Inference Providers

- Hugging Face Docs (Evaluate, PEFT/LoRA, BitsAndBytes, Trainers)

Videos:

- LoRA & QLoRA Fine-tuning Explained In-Depth

- Simple Training with the Transformers Trainer

- The Trainer API

- Fine-Tuning Large Language Models (LLMs) | w/ Example Code

- What is LoRA? Low-Rank Adaptation for finetuning LLMs EXPLAINED

- What is a Context Window?