Emotion Detection with Multi-Label Classification: A Comparative Study Using Transformer Models

Project Overview:

This project focuses on multi-label emotion detection from text data using state-of-the-art transformer models. It employs a comparative approach, leveraging various pre-trained transformer architectures including DistilBERT, Gemma-1.1-2b-it, and Mistral-7B. The models are fine-tuned with advanced techniques like LoRA (Low-Rank Adaptation), PEFT (Parameter-Efficient Fine-Tuning), and quantization to enhance model efficiency, reduce resource consumption, and ensure scalable performance for real-world deployment.

Key Highlights:

- **Multi-label Classification:** Classifies text data into multiple emotion categories such as anger, joy, sadness, and more.
- **Comparative Analysis of Models**: A systematic comparison of DistilBERT, Gemma-1.1-2b-it, and Mistral-7B, showcasing the strengths and trade-offs in terms of performance, computational efficiency, and scalability.
- **Advanced Fine-Tuning Techniques**: Utilizes LoRA and PEFT for efficient training, minimizing the number of trainable parameters while maintaining or improving model accuracy.
- **Quantization Techniques**: Implements 4-bit quantization using BitsAndBytes to optimize large models like Mistral-7B, allowing for more efficient inference with reduced memory and computational cost.
- **Experiment Tracking and Visualization**: Integrated with Weights & Biases (wandb) for real-time logging, experiment tracking, and visualizing training and validation metrics such as accuracy, precision, recall, and F1-score.

Model Comparison:

The project uses a series of models and techniques, providing a comprehensive comparison:

Aspect	DistilBERT	Gemma-1.1-2b-it with LoRa	Mistral-7B with LoRa	Mistral-7B with Quantization & LoRa
Model Type	DistilBERT (DistilBERT-base-uncased)	Gemma-1.1-2b-it	Mistral-7B	Mistral-7B Instruct with Quantization

Fine-Tuning	Standard Fine-Tuning	LoRA-based Fine-Tuning	LoRA and PEFT Fine-Tuning	LoRA with Quantization (4-bit) and PEFT
Quantization	Not utilized	Not utilized	4-bit Quantization	Advanced Quantization with BitsAndBytes
Metrics	Accuracy, F1 (Micro, Macro)	Accuracy, F1 (Micro, Macro)	Accuracy, F1 (Micro, Macro)	Accuracy, F1 (Micro, Macro), Efficient Inference
Optimizers	AdamW	AdamW	AdamW with Learning Rate Scheduler	AdamW with Learning Rate Scheduler
Scalability	Suitable for smaller-scale models	Suitable for moderately large models	Scalable for large models like Mistral-7B	Optimized for large-scale models with Quantization
Inferences	Standard inference	Standard inference	Standard inference with enhanced efficiency	Fast Inference with Quantized Models

Technologies & Skills:

- **Machine Learning & Deep Learning**: Transformer-based models, BERT, DistilBERT, Mistral-7B, HuggingFace Transformers
- **Natural Language Processing (NLP):** Text Classification, Emotion Detection, Multi-Label Classification
- **Model Optimization**: LoRA, PEFT, BitsAndBytes, Quantization
- **Experiment Tracking & Visualization**: Weights & Biases (wandb), Model Metrics, Loss Logging
- **Programming Languages**: Python, PyTorch, HuggingFace Libraries, scikit-learn
- **Libraries & Frameworks**: HuggingFace Transformers, Datasets, Accelerate, evaluate, matplotlib, seaborn, pandas
- **Data Science & Analytics**: Multilabel Confusion Matrix, Precision, Recall, F1-Score, Data Preprocessing, Data Visualization

Project Structure:

1. **Data Preprocessing**: Loads and preprocesses emotion-labeled datasets, converts text into tokenized datasets compatible with transformer models.
2. **Model Fine-Tuning**: Fine-tunes models (DistilBERT, Gemma-1.1-2b-it, and Mistral-7B) for emotion detection, applying LoRA and PEFT for optimized training.
3. **Metrics Calculation & Visualization**: Defines and calculates key performance metrics (accuracy, F1-score, confusion matrix), visualizing results through plots and heatmaps.
4. **Inference & Model Deployment**: Pushes the fine-tuned models to the HuggingFace Hub for easy access and inference on new datasets.
5. **Comparison and Analysis**: Compares models' performance in terms of efficiency, accuracy, and scalability, providing insights into best practices for emotion detection.

Conclusion:

This project demonstrates the effectiveness of fine-tuning large transformer models for multi-label emotion classification while optimizing computational efficiency through quantization and parameter-efficient fine-tuning (PEFT) methods. By leveraging advanced techniques like LoRA and BitsAndBytes, it shows how to effectively scale large models without compromising performance.