Emotion Detection Model Comparison Report

1. Introduction

This project aims to compare the performance of three encoder-only NLP models on a multi-label emotion detection task. The dataset contains tweets labeled with emotions like anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. We experimented with three encoder-only models:

- **Experiment 1:** RoBERTa Base
- **Experiment 2:** DistilBERT
- **Experiment 3:** DistilRoBERTa

2. Model Descriptions and Comparison to BERT Architecture

2.1 RoBERTa Base

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is an optimized version of BERT, removing Next Sentence Prediction and training on larger mini-batches and longer sequences. Its base version has 12 layers, 768 hidden units, and 110 million parameters. This model retains BERT's core architecture but enhances performance with improved pretraining.

2.2 DistilBERT

DistilBERT is a lighter, smaller, and faster version of BERT. Distilled using knowledge distillation, it retains 97% of BERT's language understanding but is nearly 40% smaller, with just 66 million parameters. The architecture remains similar but with reduced layers (6 instead of 12), making it more efficient for tasks with limited resources.

2.3 DistilRoBERTa

DistilRoBERTa is a distilled version of RoBERTa, sharing the same lightweight benefits as DistilBERT with roughly the same number of parameters. It provides a faster and more resource-efficient alternative while preserving most of RoBERTa’s language understanding.

3. Experiment Setup and Methodology

3.1 Data Preparation

- **Tokenization:** We used specific tokenizers for each model (e.g., `RobertaTokenizer` for RoBERTa).
- **Class Weights:** Calculated inverse class frequencies for each emotion to handle class imbalance and applied these weights to the loss function.

- **Train-Validation Split:** The training data was split 80-20 for training and validation.

3.2 Evaluation Metrics

The primary evaluation metrics were:

- **F1 Score (Macro):** Averages the F1 scores across all classes, giving equal weight to each emotion.
- **Accuracy:** Proportion of correct predictions overall.

3.3 Weights & Biases (W&B) Tracking

Each model's training and evaluation metrics were tracked using W&B, with training loss, validation accuracy, and F1 score logged across epochs.

4. Results and Performance Comparison

Model	Validation Accuracy	Validation F1 Score (Macro)	Observations
RoBERTa Base	0.1773	0.5897	Highest F1 due to larger capacity and optimized architecture.
DistilBERT	0.1961	0.5806	Reasonable performance with faster training; slight drop in F1 score.
DistilRoBERTa	0.1508	0.5703	Similar to DistilBERT but lower F1, suggesting RoBERTa's larger pretraining set was beneficial.

4.1 Observations

- **Model Complexity vs. Performance:** RoBERTa Base performed best due to its larger capacity, suggesting it could better capture nuances in the emotion detection task.
- **Distillation Impact:** Both DistilBERT and DistilRoBERTa performed well given their smaller size, with DistilBERT showing slightly better performance due to its familiarity with BERT's extensive pretraining.
- **Class Imbalance Handling:** Adding class weights to the loss function significantly improved performance on underrepresented emotions (e.g., "surprise" and "trust").

5. Challenges and Observations

- **Class Imbalance:** Minority classes like "surprise" and "trust" were challenging, but using class weights helped improve their recall.
- **Training Efficiency:** Distilled models, particularly DistilBERT, trained faster and required less memory, making them practical for limited-resource environments.
- **Multi-label Complexity:** Multi-label classification increased computational demands, as multiple labels needed prediction per sample.

6. Conclusion

The best-performing model for this task was **RoBERTa Base**, which achieved the highest validation F1 score. Although more computationally demanding, it demonstrated superior understanding of nuanced emotions in the dataset. DistilBERT and DistilRoBERTa provided efficient alternatives with competitive results, proving valuable for resource-constrained settings.

6.1 Final Recommendations

- **For Accuracy and Depth:** RoBERTa Base is preferred.
- **For Efficiency:** DistilBERT offers a balanced trade-off between speed and performance.

7. Weights & Biases (W&B) Project Link

You can access the project's W&B logs, training metrics, and model comparisons here:

RoBERTa Base: https://wandb.ai/pxy230011-the-university-of-texas-at-dallas/Exp1/workspace

DistilBERT: https://wandb.ai/pxy230011-the-university-of-texas-at-dallas/Exp2?nw=nwuserpxy230011

distilroberta-base: https://wandb.ai/pxy230011-the-university-of-texas-at-dallas/Exp3?nw=nwuserpxy230011