Report: Emotion Detection in Tweets

# **Overview**

For a detailed view of all experiments and results, please refer to my Weights & Biases project <https://wandb.ai/akshgad01/projects>

In this experiment, we compared the performances of three language models. The experiment involved data preparation, model training, performance evaluation, and documentation using Google Colab. We used the Weights & Biases (W&B) platform for logging training progress and documenting results.

# Methodologies:

**Data Preparation:** We used a dataset with approximately 7,000 rows containing tweets labeled with emotions such as 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'love', 'optimism', 'pessimism', 'sadness', 'surprise', and 'trust'. The dataset was preprocessed to convert emotion labels into binary indicators for each tweet.

**Model Training:** We trained each model using Hugging Face's Transformers library.

Training parameters included the number of epochs, batch sizes, learning rate, weight decay, and

optimizer choice.

# Part 1a: Gemma Model Training

**Gemma Model Training Configuration:**

Model: RoBERTa with checkpoint 'google/gemma-1.1-2b-it'

Tokenizer: RoBERTaTokenizer

Training Parameters:

- Epochs: 4

- Batch Size: 8

- Learning Rate: 5e-6

- Optimizer: AdamW

- Loss Function: BCEWithLogitsLoss with class weights

- Evaluation Strategy: Steps with evaluation every 50 steps

- Model Saving Strategy: Steps with saving every 50 steps

- Logging Strategy: Steps with logging every 20 steps

**Gemma Model Evaluation Results for Part 1a:**

Accuracy: 0.23

F1 Score (Micro): 0.71

F1 Score (Macro): 0.61

Loss: 0.48

**Gemma Model Interpretive Results**

- The Gemma model trained with RoBERTa achieved an accuracy of 0.23, indicating that it correctly predicted emotions in tweets about 23% of the time.

- The F1 score (Micro) of 0.71 suggests a good overall performance in terms of precision and recall across all classes.

- The F1 score (Macro) of 0.61 indicates that the model performs reasonably well in classifying emotions, considering the macro-average across all classes.

- The loss of 0.48 indicates the error or discrepancy between the predicted and actual values, with lower values indicating better model performance in terms of loss minimization.

# Part 1b: Gemma Model Training and Evaluation

**Gemma Model Training Configuration:**

Model: DistilBERT-base-uncased

Tokenizer: AutoTokenizer with default settings

Training Parameters:

-Epochs: 4

-Batch Size: 8

-Learning Rate: 5e-6

-Optimizer: AdamW with gradient accumulation

-Loss Function: BCEWithLogitsLoss with class weights

-Evaluation Strategy: Steps with evaluation every 50 steps

-Model Saving Strategy: Steps with saving every 50 steps

-Logging Strategy: Steps with logging every 20 steps

**Gemma Model Evaluation Results for Part 1b:**

Accuracy: 0.19

F1 Score (Micro): 0.66

F1 Score (Macro): 0.50

Loss: 0.57

**Gemma Model Interpretive Results**

- In contrast to Part 1a, the Gemma model trained with DistilBERT achieved a lower accuracy of 0.18835, indicating a decrease in the model's ability to accurately predict emotions compared to Part 1a.

- The F1 score (Micro) of 0.65667 shows a slightly lower overall performance compared to Part 1a but still maintains a reasonable level of precision and recall.

- The F1 score (Macro) of 0.49967 is lower than that of Part 1a, suggesting that the model's performance in classifying emotions across different classes has decreased.

- The higher loss value of 0.57061 compared to Part 1a indicates increased error or discrepancy in the model's predictions, leading to a less optimal performance.

# Part 2: SFR-Embedding-Mistral Model Training and Evaluation

**SFR-Embedding-Mistral Model Training Configuration:**

Model: SFR-Embedding-Mistral

Tokenizer: AutoTokenizer with padding

Training Parameters:

- Epochs: 2

- Batch Size: 8

- Learning Rate: 1e-4

- Optimizer: AdamW with gradient accumulation

- Loss Function: BCEWithLogitsLoss with class weights

- Evaluation Strategy: Steps with evaluation every 50 steps

- Model Saving Strategy: Steps with saving every 50 steps

- Logging Strategy: Steps with logging every 50 steps

- Quantization: BitsAndBytes with 4-bit quantization

**SFR-Embedding-Mistral Model Evaluation Results:**

Accuracy: 0.23

F1 Score (Micro): 0.72

F1 Score (Macro): 0.63

Loss: 0.47

**SFR-Embedding-Mistral Model Interpretive Results**

- The SFR-Embedding-Mistral model achieved an accuracy of 0.22654, indicating a moderate ability to accurately classify emotions in tweets.

- The F1 score (Micro) of 0.7175 and F1 score (Macro) of 0.62697 demonstrate a reasonable level of precision and recall across all classes, highlighting the model's capability in emotion classification tasks.

- The lower loss value of 0.46706 compared to other parts indicates reduced error or discrepancy in the model's predictions, leading to improved performance.

- The model's runtime during evaluation was 126.5225 seconds, with a samples-per-second rate of 12.211, indicating a relatively efficient processing speed for predictions.

# Challenges

**Training Time**: Managing training time for larger models proved demanding requiring substantial computational resources and prolonged training periods. To ensure efficient execution I had to buy Colab Pro to enabling timely model completion.

**GPU/ Storage**: I kept running out of Cuda memory and eventually had to upgrade to Google Collab Pro Plus which was even more expensive and cost $75. Had to purchase 500 additional computing units as well. Manually deleted large files so my Google drive would not run out of storage after every run, I do not have more money to spend on Google One for more storage.

**Performance Variability**: I noticed variability in model performance across different training runs, necessitating multiple runs for reliable results.

**Test set**: Extracting a CSV file from trainer results for the test set presented a complex challenge, necessitating extensive coding, refinement, and debugging to ensure accurate data handling.

**Mental Stress**: This homework assignment took a lot out of me physically and mentally. Had to dedicate lots of hours per day for sometimes to yield 0 results but I am proud of my accomplishments and glad this HW is done and over.

# Comparison and Observations

**Accuracy**: The SFR-Embedding-Mistral model and the Gemma model with RoBERTa achieved similar accuracies, around 0.23, while the Gemma model with DistilBERT performed slightly lower at 0.19. This indicates that both Gemma models outperformed the DistilBERT model in terms of accuracy.

**F1 Scores**: The Gemma model with RoBERTa achieved the highest F1 scores, with Micro and Macro scores of 0.71 and 0.61, respectively. The SFR-Embedding-Mistral model also performed well with F1 scores of 0.72 (Micro) and 0.63 (Macro). In contrast, the Gemma model with DistilBERT had lower F1 scores, indicating a decrease in performance compared to the other models.

**Loss**: The Gemma model with RoBERTa had the lowest loss of 0.48, followed by the SFR-Embedding-Mistral model with 0.47 loss. The Gemma model with DistilBERT had the highest loss at 0.57, indicating higher error or discrepancy in its predictions.

# Conclusions

The Gemma model trained with RoBERTa showed the best overall performance among the models evaluated, with higher accuracy, F1 scores, and lower loss compared to other configurations.

The SFR-Embedding-Mistral model also demonstrated strong performance, particularly in F1 scores, suggesting its suitability for emotion detection tasks.

The Gemma model with DistilBERT, while still achieving reasonable results, showed a decrease in performance compared to the RoBERTa-based Gemma model, highlighting the impact of model architecture on task performance.