**Homework 5: Model Evaluation Report**

In this assignment, I have used roberta-base, distilbert-based-uncased and albert-base-v2 encoder models.

**Here’s the description of all the three models:**  
RoBERTa-base is an improved version of BERT that uses the same transformer architecture but is trained on much more data and for longer. It removes the next sentence prediction task and uses dynamic masking instead of static masking. These changes make RoBERTa more robust and accurate than BERT, especially on tasks like text classification and question answering

DistilBERT-base-uncased is a lighter and faster version of BERT, created by distilling knowledge from the original BERT model. It has about half the layers, making it much smaller and more efficient, but it still keeps over 95% of BERT’s accuracy. DistilBERT is ideal when you need good performance but lower resource usage

ALBERT-base-v2 is a compact version of BERT that reduces the number of parameters by sharing them across layers and using smaller embeddings. Despite being much smaller, ALBERT keeps strong performance by using techniques like sentence order prediction instead of next sentence prediction. It’s highly memory-efficient and works well for tasks that need whole-sentence understanding

**Comparison of Models’ Performances**

Here’s a comparison of the models’ performances using the main metric (**F1 score**) and a discussion of other metrics based on your outputs and the literature:

F1 Score (Main Metric):

* RoBERTa-base achieved the highest validation F1 score, with a best F1 of 0.59032.
* DistilBERT-base-uncased followed with a best F1 of 0.56498.
* ALBERT-base-v2 had the lowest best F1 at 0.53475.

**Validation Loss:**  
RoBERTa-base also had the lowest validation loss (0.39209), indicating better generalization. DistilBERT and ALBERT had slightly higher losses (0.42417 and 0.43104, respectively), which aligns with their lower F1 scores.

**DistilBERT** is much faster and more efficient than the other two, making it a good choice if you need quick predictions or have limited resources, though with a slight trade-off in accuracy. ALBERT is the most compact in terms of memory but with the lowest performance.

Summary:

* RoBERTa-base is the most accurate and robust for emotion detection, with the best F1, precision, recall, and lowest validation loss.
* DistilBERT-base-uncased is a strong performer and much faster, offering a good balance between speed and accuracy.
* ALBERT-base-v2 is the most lightweight and memory-efficient, but its accuracy and F1 are lower.

**Significant Observations and Challenges**

One major observation is that model size and architecture have a clear impact on both performance and resource requirements. For example, RoBERTa-base generally achieves the highest accuracy and F1 scores, but it also requires more memory and computational power.

DistilBERT-base-uncased is much faster and more efficient, but you may notice a slight drop in performance compared to RoBERTa. However, it is often the best choice when speed or resource constraints are important. Still, it’s possible to run into issues with batch sizes—if you set them too high, even DistilBERT can cause memory errors.

ALBERT-base-v2 is designed to be very lightweight and memory-efficient, but the trade-off is sometimes lower accuracy. One challenge with ALBERT is that, despite its small size, it can be sensitive to hyperparameters and may require more tuning to reach its best performance. Also, its parameter sharing can make debugging or interpreting layer-wise behavior more difficult.

**Validation Loss Fluctuations and Overfitting:**  
One of the main challenges was that the validation loss did not always decrease smoothly. Sometimes, it fluctuated or even started increasing while the training loss kept going down. This is a classic sign of overfitting, where the model learns the training data too well but fails to generalize to new, unseen data. Larger models like RoBERTa were especially prone to this, and it made it tricky to decide when to stop training.

**Metric Instability:**  
The F1 score, which balances precision and recall, sometimes showed only modest improvements or even plateaued after a few epochs. This made it difficult to judge whether further training would help or just waste resources. In some cases, the F1 score was also sensitive to class imbalance in the dataset, so a high overall score could hide poor performance on minority classes.

RoBERTa-base performed the best among the three models. It achieved the highest validation F1 score, indicating superior accuracy and generalization on the emotion classification task. DistilBERT-base-uncased was the next best, offering a good balance between performance and efficiency, while ALBERT-base-v2, although more lightweight, had the lowest F1 score and did not perform as well as the others. Therefore, for this task, RoBERTa-base is the top-performing model.

**Weights and Biases Project Link:**

https://wandb.ai/ritinwadekar6-onpoint-insights/emotion-classification?nw=nwuserritinwadekar6

Screenshot of the project, incase the link doesn’t work



The blue line represents the RoBERTa-base encoder model

The red line represents the DistilBERT-base-uncased encoder model

The green line represents the ALBERT-base-v2 encoder model