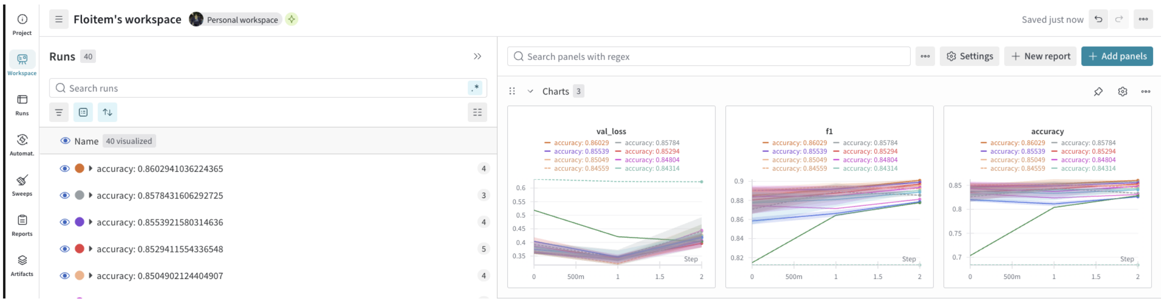
Machine Learning Operations - Project 1: Hyperparameter Tuning Report

1 INTRODUCTION

The goal of this project was to maximize validation accuracy and F1 score of a pretrained DistilBERT (base uncased) model on the MRPC dataset through systematic hyperparameter tuning. MRPC is a binary paraphrase detection task from the GLUE benchmark. The optimization process compared manual and automatic hyperparameter search approaches.

2. Experiment Tracking

All experiments were performed on google collab. I chose Weights & Biases (`wandb`) for experiment tracking. The tool was used to log essential metrics like accuracy, F1 score, and loss for every run. This allowed for straightforward comparison and ranking of the different hyperparameter configurations to determine the best-performing model.



Name 40 visualized	
accuracy: 0.8602941036224365	4
accuracy: 0.8578431606292725	3
accuracy: 0.8553921580314636	4
accuracy: 0.8529411554336548	5
accuracy: 0.8504902124404907	4

Figure 1: W&B dashboard showing a best accuracyruns from the manual tuning.

3. Week 1: Initial Hyperparameter Exploration (20 Runs)

Methodology: An initial exploratory phase was conducted using a greedy search approach across 10 hyperparameters. One parameter was varied at a time to isolate its impact, with the best-performing value being carried forward to the next experiment. The findings are summarized in Table 1.

Table 1: Summary of Week 1 Hyperparameter Exploration

Hyperparameter	**Tested Values**	**Best Value Found**
Learning Rate	1×10^{-5}, 2×10^{-5}, 3×10^{-5}, 5×10^{-5}	5×10^{-5}
Weight Decay	0.0, 0.01, 0.1	0.0
Warmup Steps	0, 17, 25, 33, 50	0
Train Batch Size	16, 32	32
Optimizer	AdamW, Adam, SGD	AdamW
Max Sequence Length	64, 128, 256	256
Parameters kept constant: Adam Epsilon, Adam Betas, Grad Accumulation, Seed (initially)		

Analysis of Findings: The results from Table 1 clearly show that **Learning Rate** and **Optimizer** choice had the most significant impact on performance. Settings for Weight Decay and Warmup Steps, when tested individually, either decreased or had a negligible effect on the final accuracy.

Top 3 Hyperparameters for Week 2: Based on this exploration, the three most promising and interactive hyperparameters were selected for detailed tuning: (1) **Learning Rate**, (2) **Weight Decay**, and (3) **Warmup Steps**.

4. Week 2: Detailed Manual Tuning (20 Runs)

Methodology: The 20 manual runs were structured into a multi-phase search to efficiently explore the interactions between the top 3 hyperparameters.

- **Phase A - Grid Search (12 Runs):** A grid search was performed over learning rates $\{3\times10^{-5}, 5\times10^{-5}\}$, weight decays $\{0.01, 0.1\}$, and warmup steps $\{0, 25, 50\}$. An intersting finding was that every single run with **0 warmup steps** outperformed its counterparts with warmup, confirming this was a key parameter to fix.
- **Phase B - Middle Ground Search (4 Runs):** Based on the success of learning rates 3×10^{-5} and 5×10^{-5}, a middle value of 4×10^{-5} was tested with various weight decays. This did not yield a new best result but helped confirm the optimal learning rate was likely not between these values.
- **Phase C - Validation & Final Tuning (4 Runs):** The best-performing configuration from all previous runs ('lr: 5e-5, wd: 0.0, ws: 0') was re-run with a different random seed ('2024') to ensure its stability. This reproducibility

run confirmed the result and became the overall champion of the manual search. Maximum sequence length was a critical to make it too small(at least 128), but not one for deep tuning.

Best Manual Configuration: After all 20 runs, the best-performing set of hyperparameters was identified.

Table 2: Comparison of the Best Manual Configuration Across Seeds

Metric / Parameter	Best Run (Seed 42)	Validation Run (Seed 2024)
Learning Rate	5×10^{-5}	
Weight Decay	0.0	
Warmup Steps	0	
Max Sequence Length	256	
Validation Accuracy	**0.8603**	**0.8603**
Validation F1 Score	0.9005	**0.9022**
Validation Loss	0.4331	**0.3960**

5. Week 3: Automatic Hyperparameter Tuning

Methodology: The goal of Week 3 was to compare the manual search against an automated approach. I used **W&B Sweeps** with a **Bayesian optimization** strategy, given the same 20-run budget. The search space was defined broadly around the promising values from Week 1.

- **Learning rate:** Log-uniform distribution between 1×10^{-5} and 9×10^{-5}.
- **Weight decay:** Uniform distribution between 0.0 and 0.1.
- **Warmup steps:** Integer distribution between 0 and 50.

Results: The automated sweep explored the parameter space but failed to surpass the best manual result. Table 3 compares the best configuration found by the automatic sweep against the best manual configuration using the same baseline seed (42).

Table 3: Comparison of Best Manual vs. Automatic Run (Seed 42)

Metric / Parameter	Best Manual Run	Best Automatic Run
Learning Rate	5×10^{-5}	5.18×10^{-5}
Weight Decay	0.0	0.059
Warmup Steps	0	27
Validation Accuracy	**0.8603**	0.8554
Validation F1 Score	**0.9005**	0.8991
Validation Loss	**0.4331**	0.4397

6. Reflection

The greedy search is very time intense if you do not program it to automatically take the best value of the last run. Therefore in the second week I choose a hybrid search to find the best combination of the 3 most important hyperparameters. The structured, phased approach to manual tuning proved highly effective. By first identifying influential hyperparameters and then performing a grid search to understand their interactions, I was able to find a robust and high-performing configuration. The use of `wandb` was perfect for organizing results and making data-driven decisions at each stage.

The most important lesson was that my manual search found a better result than the automated tool. The automatic sweep wasted some of its 20 runs on bad settings I had already ruled out. For future projects, I would do a few manual runs first to find a good starting area, and then use an automated tool for the final fine-tuning.