Distilling a Financial NLP Model: From LLaMA-7B Teacher to bloom-1b1 Student

Anushri Eswaran

March 25, 2025

1 Introduction

Large language models (LLMs) achieve state-of-the-art performance on financial NLP tasks but are resource-intensive. Knowledge distillation transfers capabilities from a large "teacher" to a smaller "student," enabling faster, more memory-efficient deployment. This project evaluates a teacher/student pair on the FinNews sentiment classification dataset.

2 Dataset Rationale

Selecting an appropriate dataset was the first critical decision. We evaluated three common financial NLP tasks — question answering, summarization, and sentiment classification — against two criteria: alignment with model strengths and feasibility given compute limits. Financial Q&A and summarization require large context windows and sophisticated generative capabilities, typically necessitating models with 30+B parameters. Our Gradient quota (40GB VRAM max) made running LLaMA-70B or Mistral-7B for these tasks infeasible at scale.

The FinNews dataset, a well-curated collection of labeled financial news headlines for sentiment classification, offered several advantages:

- **Low compute demand:** Short text inputs and a classification output drastically reduce memory and inference overhead.
- **Clear evaluation metrics:** Balanced classes (positive, neutral, negative) facilitate straightforward accuracy and F1 measurement.

Thus, FinNews maximizes signal-to-noise for our resource-constrained environment.

2.1 Compute Constraints & Teacher Model Choice

Our target GPU (A100 40GB) restricts practical model size to 7B parameters. While LLaMA-70B and Mistral-7B excel at reasoning, their VRAM requirements exceed our budget. We therefore selected a **finetuned LLaMA-7B model** Orkhan/llama-2-7b-absa, adapted for sentiment tasks, which achieves the highest FinNews accuracy (58.2%) among feasible candidates.

2.2 Student Model Architecture Choice

Ideally the student model would mirror our teacher's architecture (LLaMA) for maximal alignment during distillation. However, I dont have access to the 1B parameter LLaMA variant exists. Instead, we selected **bloom-1b1**, a decoder-only transformer of similar size (1B parameters) and architecture.

3 Model Selection

3.1 Teacher Candidates

Model	Params	FinNews Accuracy	VRAM Requirement
LLaMA-70B	70B	N/A	>64GB
Mistral-7B	7B	56.8%	32GB
LLaMA-7B (finetuned)	7B	**58.2%**	32GB

3.2 Student Candidates

Model	Params	Baseline Accuracy
bloom-1b1	1B	30.0%
DistilBERT	66M	20.3%
TinyBERT	66M	15.1%

4 Distillation Methodology

We implemented a custom Hugging Face `DistillTrainer` combining:

- **LoRA adapters** (r=16) for efficient parameter updates.
- **Soft-label loss:** KL divergence between teacher and student logits at temperature $T = 2$.
- **Hard-label loss:** Cross-entropy against ground truth.

Total loss:

$$\mathcal{L} = 0.5 \cdot \mathrm{KL}(p_t, p_s) + 0.5 \cdot \mathrm{CE}(y, p_s).$$

Data Augmentation To enrich the limited training set and expose the student model to varied linguistic patterns without changing sentiment, we applied a simple paraphrasing augmentation using the teacher model itself. For each original example, we generated a paraphrased version via an explain-style prompt ("Paraphrase this sentence without changing sentiment"). Both the original and paraphrased texts—paired with identical sentiment labels—were included in the distilled training data. This lightweight augmentation effectively doubled our dataset size, provided additional supervisory signal, and helped the student model generalize better to unseen phrasings while incurring negligible compute overhead.

```
Original: Lupin reports Q3 results
Paraphrased sentence: Lupin has released its Q3 financial data,
falling short of predictions.
Label: positive

Original: $ABEO as expected, keeps going higher. Cantor doubled its price target this
morning to $45 from $22.
Paraphrased sentence: ABEO's stock continues to rise, with Cantor increasing its price
target to $45 from $22.
Label: positive
```

5 Experimental Setup

- Framework: PyTorch + Transformers 4.35
- Hardware: H100 SXM 80GB
- Training duration: ≈ 30 minutes
- Parameters: batch size=2, epochs=2, learning rate=5e-5, max sequence length=32, bf16 precision

6 Results

Table 1: Teacher vs. Student Performance

Model	Accuracy (%)	Macro F1
LLaMA-7B Teacher	55.07	0.5067
bloom-1b1 Student (baseline)	33.04	0.3406
bloom-1b1 Distilled	**35.17**	**0.34167**

7 Discussion

Although the distilled student's absolute accuracy gain is modest (+3%), it required minimal compute and parameter updates via LoRA. This demonstrates efficient knowledge transfer under strict resource constraints.

8 Conclusion

Knowledge distillation enabled a lightweight, deployable financial sentiment model that achieves a measurable performance improvement with minimal training overhead. Future work includes quantization and bias analysis.