Financial Sentiment Analysis& Knowledge Distillation

Benchmarking and Improving Financial Text-Classification

ArthurMrv

December 2025

1 Executive Summary

This project aims to **benchmark and improve financial sentiment classification** on financial text (news-like summaries and filings). Accurate financial sentiment analysis is critical for algorithmic trading and market monitoring, yet labeled datasets are often scarce or noisy.

This work serves as a completion to a previous study that visualized the link between news analysis and stock prices. While the initial study relied on a standard FinBERT model, this project advances performance by combining a **fine-tuned DeBERTa-v3** model with a **knowledge distillation** roadmap for real-time use. The fine-tuned model reaches a weighted **F1 score of 0.92** (vs. $\sim$0.5 previously), using evaluation practices adapted to a highly imbalanced label distribution (stratified splits and F1-based reporting).

All artifacts are **accessible on Hugging Face**: the transformed/labeled dataset, the trained weights, and deployment options (inference endpoint and direct weight download).

2 Research Methodology

Our approach is built on four key pillars designed to address the limitations of existing off-the-shelf models.

2.1 Label Generation (Silver Labels)

The original news dataset lacked consistent, high-quality sentiment labels. To resolve this, we employed a **Large Language Model (LLM)**—specifically **DeepSeek-V3.2**—to act as an expert annotator on the **summary** column of `kritsadaK/EDGAR-CORPUS-Financial-Summarization` (10.6k rows; columns: `input`, `summary`, `model`).

- **Process:** We generated a stable identifier (`hashed_input`) and requested an LLM sentiment score in $\{-2, -1, 0, 1, 2\}$ with **reasoning-first** output for explainability and improved annotation consistency.
- **Outcome:** A transformed and consistently labeled dataset, published on Hugging Face as `ArthurMrv/EDGAR-CORPUS-Financial-Summarization-Labeled`.

2.1.1 Why score from -2 to +2?

We intentionally used a 5-level scale (-2 to $+2$) instead of a strict 3-class scheme:

- **Reasoning-before-scoring:** Prior benchmarking experience suggests that letting the model explain its reasoning before assigning a score can improve both **explainability** and **accuracy**.
- **Avoid over-granularity:** Too fine a numeric scale can increase confusion/hallucination (e.g., distinguishing "97/100" from "98/100" is not meaningful for either the model or the annotator).
- **Better distribution diagnostics:** The 5-level scale enables bias analysis (pessimistic/optimistic drift) and supports prompt tuning for extreme cases ($-2/+2$).

After manual inspection, the LLM displayed a **mostly neutral** point of view overall; therefore, we collapsed to a 3-class setting for modeling: **bearish** $\leftarrow \{-2, -1\}$, **neutral** $\leftarrow \{0\}$, **bullish** $\leftarrow \{1, 2\}$.

2.2 Class Imbalance & Human Verification

The silver labels are **highly imbalanced**. Figure 1 and Table 1 summarize the class counts prior to collapsing:

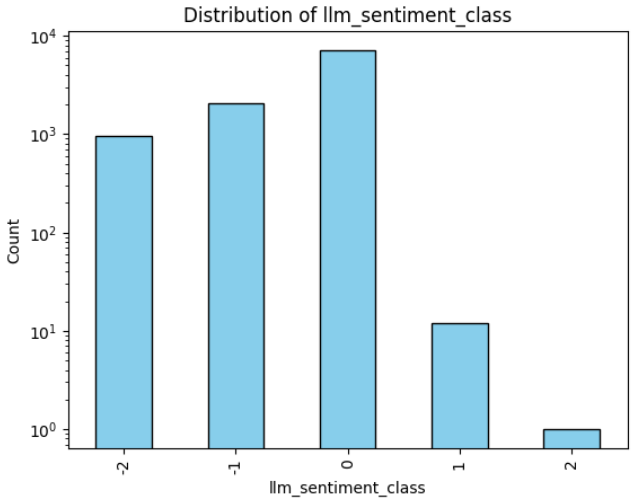

Figure 1: Distribution of `llm_sentiment_class` (log scale).

`llm_sentiment_class`	count
0	7121
-1	2061
-2	952
1	12
2	1

Table 1: Raw LLM label counts (before collapsing to 3 classes).

Because knowledge distillation and supervised fine-tuning are sensitive to noisy or skewed labels, we performed **human verification on 50+ randomly sampled rows** to sanity-check label quality and confirm the imbalance. This motivated the use of **stratified sampling/splitting** and **F1-based metrics** throughout evaluation.

2.3 Model Benchmarking

We evaluated several pretrained transformer models to establish a strong baseline before fine-tuning. We focused on the **DeBERTa-v3** architecture due to its superior performance on NLU tasks.

- **Baselines Evaluated:**
  - `nickmuchi/deberta-v3-base-finetuned-finance-text-classification`
  - `mrm8488/deberta-v3-ft-financial-news-sentiment-analysis`

2.4 Fine-Tuning

We selected the best-performing baseline (`mrm8488`) and further fine-tuned it on our curated, LLM-labeled dataset.

- **Technique:** Weighted cross-entropy loss to address class imbalance, with **stratified** train/validation/test splits.
- **Metrics:** We report **F1** (macro/weighted) rather than accuracy, which can be misleading under heavy imbalance.
- **Result:** The fine-tuned model (`ArthurMrv/deberta-v3-ft-financial-news-sentiment-analysis-fi`) achieved a weighted **F1 score of 0.92**, significantly outperforming the baseline macro F1 of 0.60.

2.5 Knowledge Distillation

To ensure the model is viable for real-time applications, we are implementing a Knowledge Distillation pipeline to compress the heavy "Teacher" model (DeBERTa-v3) into a lighter "Student" model (e.g., DistilBERT) without significant performance loss.

3 Implementation Details

3.1 Dataset

- **Original Source:** Hugging Face `kritsadaK/EDGAR-CORPUS-Financial-Summarization`
- **Transformed Dataset (Hugging Face):** `ArthurMrv/EDGAR-CORPUS-Financial-Summarization-Labe`
- **Links:** original dataset, labeled dataset, fine-tuned model
- **Raw LLM labels:** `llm_sentiment_class` $\in \{-2, -1, 0, 1, 2\}$
- **Collapsed labels for modeling:** bearish $\leftarrow \{-2, -1\}$, neutral $\leftarrow \{0\}$, bullish $\leftarrow \{1, 2\}$

3.2 Training Pipeline

The research utilized a series of Jupyter Notebooks:

1. `init_hg.ipynb`: Initializes the dataset structure.
2. `update_refined.ipynb`: Runs the DeepSeek-V3.2 inference loop to label data.
3. `fine_tune_model.ipynb`: Loads data, runs training using Hugging Face `Trainer`, evaluates, and saves the model.

3.3 Evaluation Results

The improvement is quantitatively visible in confusion matrices, showing a drastic reduction in misclassified instances after fine-tuning.

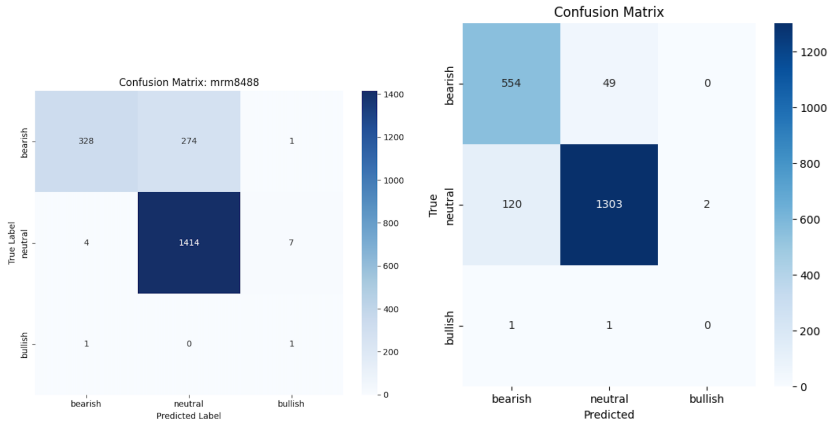

Figure 2: Confusion Matrix Comparison: Before (Left) vs. After Fine-Tuning (Right)

4 Application & Deployment

To demonstrate practical utility, we developed a production-ready application.

- **Tech Stack:** Python, Streamlit (Frontend/UI), FastAPI (service layer), Hugging Face (model hosting / inference).
- **Features:**
  - **Custom Text Analysis:** Users can input raw text to get sentiment predictions.
  - **Ticker Mode:** Users can choose a ticker for automatic news retrieval (e.g., AAPL, TSLA) via `yfinance`, then run sentiment analysis end-to-end.
  - **API Connectivity:** A FastAPI layer exposes the model to other applications and services.
- **Deployment:** Containerized via Docker for easy distribution; the model is also accessible via Hugging Face (inference endpoint) or by downloading weights directly.

5 Future Work

- **Distillation Completion:** Finalize the student model training for ultra-low latency.
- **Data Expansion:** Incorporate informal sources (Twitter/X, Reddit) to improve robustness.
- **Live Feeds:** Integrate professional news feeds (Bloomberg/Reuters).

6 References

- He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. arXiv:2111.09543.
- Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv:1908.10063.
- Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network*. arXiv:1503.02531.
- Dataset: `kritsadaK/EDGAR-CORPUS-Financial-Summarization`. Hugging Face.
- Dataset (labeled): `ArthurMrv/EDGAR-CORPUS-Financial-Summarization-Labeled`. Hugging Face.
- Baseline model: `nickmuchi/deberta-v3-base-finetuned-finance-text-classification`. Hugging Face.
- Baseline model: `mrm8488/deberta-v3-ft-financial-news-sentiment-analysis`. Hugging Face.
- Fine-tuned model: `ArthurMrv/deberta-v3-ft-financial-news-sentiment-analysis-finetuned`. Hugging Face.
- DeepSeek-V3.2 (teacher LLM used for silver labels). Hugging Face organization page.