Efficient Financial Language Understanding via Distillation with Synthetic Data

Wen-Fong (Xavier) Huang, Edwin Simpson
School of Engineering Mathematics and Technology, University of Bristol
william.techlover@gmail.com, edwin.simpson@bristol.ac.uk

Abstract

Large instruction-following models are powerful but costly to deploy, particularly in finance, where labelled data are limited by confidentiality and expert annotation cost. We present an efficient framework for financial sentiment analysis through **distillation with synthetic data**, transferring knowledge from a large instruction-tuned teacher to compact student models. The framework is designed for low-resource conditions, where a small set of real examples are collected and labelled by hand. The framework then clusters the examples and uses the clusters to select seeds for generating synthetic examples via structured few-shot prompting. Experiments show that clustering-based seed selection yields more representative synthetic data than random sampling, enabling compact models to achieve strong performance with minimal supervision. Notably, on a more complex and noisy text domain, the compact model trained on the complete synthetic–seed corpus even **outperforms the teacher model**, while remaining competitive on formal text. The framework provides a practical route toward resource-efficient domain adaptation in financial NLP with minimal human labelling effort.

Keywords: synthetic data, instruction distillation, financial NLP, low-resource learning, data selection

1. Introduction

Financial sentiment analysis supports investment decision-making and market monitoring, yet annotated data are often scarce due to confidentiality and expert labelling costs (Lopez-Lira and Tang, 2023; Yang et al., 2023; Kirtac and Germano, 2024). Large language models (LLMs) such as GPT-4o (OpenAI, 2024) offer a potential solution: they exhibit strong instruction-following and reasoning abilities across classification, summarisation, and question answering (Brown et al., 2020), allowing them to be applied to new tasks with minimal labelled data. However, their computational demands, latency, and proprietary nature hinder deployment in cost- and risk-sensitive settings (Miao et al., 2025; Zhen et al., 2025).

A growing line of research transfers instruction-following behaviour from large models to smaller ones through distillation and synthetic supervision. Annotation-efficient pipelines such as *Self-Instruct* (Wang et al., 2023), *Alpaca* (Taori et al., 2023), and *Orca* (Mukherjee et al., 2023) demonstrate that synthetic instruction–response pairs can replace large portions of human annotation, while alignment-oriented approaches like Zephyr (Tunstall et al., 2023) and LIMA (Zhou et al., 2023) show that compact, high-quality prompts can yield strong generalisation. In parallel, lightweight encoder architectures—DistilBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2020), and ModernBERT (Warner et al., 2024)—provide efficient foundations for downstream adaptation. These trends motivate our central question: *Can instruction-following knowledge be distilled from a large teacher into a lightweight, domain-specific student using only minimal labelled data?*

We introduce a reproducible framework for **distillation with synthetic data generation in financial NLP**. Unlike prior studies of financial sentiment analysis that rely on human-annotated corpora to fine-tune pretrained models like FinBERT (Araci, 2019; Yang et al., 2020) and FinGPT (Yang et al., 2023), our approach replaces manual labelling with structured synthetic expansion from minimal seed data and teacher–student distillation. We introduce a novel coreset-style step (Sener and Savarese, 2018) for selecting a small set (12-105) of seed sentences by clustering Sentence-BERT embeddings (Reimers and Gurevych, 2019), then expand these seeds through structured prompting of GPT-4o (zero-, one-, and few-shot) to enhance the diversity of synthetic data. Compact students—DistilBERT and ModernBERT—fine-tuned on this synthetic–seed corpus achieve strong performance on two sentiment benchmarks: *Financial PhraseBank* (Malo et al., 2014) and *Twitter Financial News Sentiment* (on Huggingface, 2022). On the noisier social-media domain, the compact student even **surpasses the GPT-4o teacher** in zero-shot mode. Our contributions are:

- A compact, reproducible framework that transfers instruction-following capability from an LLM to lightweight encoder-based students.
- Strategies for clustering-based seed-selection and multi-prompt data expansion that increase performance with few labelled examples.
- A systematic evaluation for financial sentiment

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10242–10254
11–16 May 2026.

analysis, demonstrating effectiveness of our strategies across two diverse text domains.

- Empirical evidence that, on noisy financial text, a compact student trained on the complete synthetic–seed configuration can outperform its large teacher.

2. Related Work

Distillation and Model Compression LLMs such as GPT-4o (OpenAI, 2024) demonstrate strong instruction-following and reasoning capabilities across domains, yet they remain generalists that can lag behind fine-tuned, task-specific systems on domain benchmarks (Kocon et al., 2023; Liang et al., 2023). Their substantial computational requirements, latency, and closed-weight nature further constrain deployment in scenarios demanding efficiency, transparency, or confidentiality. Knowledge distillation provides a practical way to transfer a teacher model's knowledge to smaller, more efficient students (Sanh et al., 2019; Jiao et al., 2020; Warner et al., 2024). Classical approaches such as DistilBERT (Sanh et al., 2019) and TinyBERT (Jiao et al., 2020) align intermediate representations and soft targets to reduce model size while retaining much of the teacher's performance. ModernBERT (Warner et al., 2024) refines the BERT architecture with rotary embeddings and other efficiency-oriented improvements, making it a strong candidate for student models.

Beyond model compression, in text-level dataset distillation, DiLM (Maekawa et al., 2024) employs clustering within each training loop to select representative and diverse mini-batches for gradient alignment, which helps to stabilise optimisation. In contrast, our framework applies semantic clustering prior to training to identify diverse financial sentences as seeds for data augmentation.

Synthetic Data Generation Building on the distillation methods above, synthetic data generation addresses data scarcity by producing instruction–response pairs through large teacher models. In this paradigm, the teacher not only provides supervision but also generates both the instruction and its corresponding response, creating high-quality synthetic datasets that reduce dependence on human annotation. Frameworks such as Self-Instruct (Wang et al., 2023), Alpaca (Taori et al., 2023), and Orca (Mukherjee et al., 2023) exemplify this approach, demonstrating that synthetic supervision can substantially lower annotation costs. More recent alignment-oriented methods—such as Zephyr (Tunstall et al., 2023) and LIMA (Zhou et al., 2023)—further improve data quality through preference optimisation and *structured prompting*, where prompts follow predefined templates or roles that guide the model to generate consistent, label-aligned, and task-relevant outputs. Recent advances such as DiLM (Maekawa et al., 2024) extend this paradigm by performing text-level dataset distillation, where a language model is trained to generate synthetic text samples directly instead of optimising embedding representations.

However, these studies primarily distil general-purpose instruction-following models rather than developing domain-specific systems. Domain-sensitive applications such as financial sentiment analysis pose additional challenges, including specialised terminology, implicit market cues, and limited expert-labelled data (Rodriguez Inserte et al., 2023; Araci, 2019). To address these challenges, our pipeline expands a small set of domain-representative examples via structured prompting, balancing annotation efficiency with domain fidelity.

Seed Selection and Prompt Design Representativeness and diversity of supervision data are critical for generalisation in low-resource regimes. We employ embedding-based clustering to identify diverse and representative seeds for generating synthetic data, using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to produce sentence-level embeddings that capture semantic similarity more effectively than previous approaches. Applying k-means clustering (Lloyd, 1982) to SBERT embeddings yields semantically diverse centroids, ensuring that generated data span the breadth of financial expressions rather than repeating frequent surface forms. This idea is related to coreset selection, which identifies samples that give good coverage of a larger training set (Sener and Savarese, 2018). Prompt engineering also plays a decisive role in synthetic data quality. Structured templates inspired by Self-Instruct and Alpaca (Wang et al., 2023; Taori et al., 2023) provide predefined prompt formats that control how instructions, examples, and outputs are composed. These templates leverage *in-context learning* (Brown et al., 2020), where a few labelled instances are embedded directly within the prompt to guide model behaviour without parameter updates. This structure enables scalable expansion of small human-labelled sets into coherent, label-consistent corpora. In financial text, where subtle lexical and contextual variations influence sentiment, combining clustering-based seed selection with structured prompting could provide a reliable and reproducible basis for synthetic data generation under constrained supervision.

Research Gap Prior research shows that (1) instruction-following behaviour can be effectively distilled into compact models, (2) synthetic instruc-

tion–response data can substantially reduce manual labelling costs, and (3) domain adaptation requires careful data curation and representativeness. Yet these threads remain largely unconnected in financial NLP. Our work unifies them into a single, framework—integrating clustering-based seed selection, structured synthetic expansion, and compact encoder training—to enable efficient domain adaptation with minimal human supervision.

3. Methodology

We propose a lightweight and reproducible framework for **distillation with synthetic data generation**, transferring instruction-following behaviour from a large teacher model (**GPT-4o**) to compact encoder-based students for **financial sentiment classification**. As shown in Figure 1, the workflow comprises three main stages: (1) embedding-based clustering for representative seed selection, (2) structured prompting for synthetic generation, and (3) student model fine-tuning and evaluation.

3.1. Embedding-Based Seed Selection

To ensure diversity and minimise redundancy, sentences from the financial corpora are encoded using **Sentence-BERT (SBERT)** (Reimers and Gurevych, 2019). The variant *all-MiniLM-L6-v2* (Wang et al., 2021) produces 384-dimensional embeddings optimised for semantic similarity. These embeddings are clustered using k-means (Lloyd, 1982), where the number of clusters k is set to the intended number of seed examples. In practice, this value reflects the available annotation budget—that is, how many instances can feasibly be reviewed or labelled by domain experts. Each cluster therefore corresponds to one candidate seed, and the sentence nearest to the cluster centroid is selected as its representative.

This clustering-based strategy promotes balanced semantic coverage across sentiment categories while avoiding overlap among near-duplicate samples. By constructing the seed set from distinct semantic regions, the framework ensures that synthetic generation originates from diverse and representative financial contexts.

3.2. Structured Prompting for Synthetic Expansion

Synthetic data are generated using the large instruction-following LLM **GPT-4o** (OpenAI, 2024), accessed via the **OpenAI API**. The teacher model expands the clustered seed sentences into diverse, sentiment-consistent variants while preserving the intended polarity and contextual relevance. GPT-4o was chosen for its high reasoning capability and robust text-generation quality across both formal and informal financial language.

Three structured prompting templates are applied to control lexical, syntactic, and contextual diversity, producing a ninefold expansion per seed after deduplication. Each prompt template targets a different aspect of diversity, therefore, by using them in combination, we aim to ensure diverse and domain-consistent generated data. Examples of outputs from each prompt are shown in Table 3.

Template 1 — Few-shot, label-targeted generation. Defines label semantics through in-context examples before requesting a new instance expressing the same sentiment.

```
You are training a sentiment classifi-
cation assistant for financial news.
Instruction: Classify the sentiment
of this sentence.
Examples:
1. {$EX_NEG} → Negative
2. {$EX_NEU} → Neutral
3. {$EX_POS} → Positive
Now generate a new financial news sen-
tence expressing: {$TARGET_LABEL}
```

Template 2 — Single-seed paraphrase expansion. Generates paraphrases of a labelled seed with diverse wording and structure while preserving sentiment.

```
The following sentence has sentiment
{$LABEL}: {$SEED_SENTENCE}
Generate 3 new sentences that:
- Express the same sentiment
- Remain realistic in the finan-
cial/news domain
- Use different wording or phrasing.
```

Template 3 — Multi-seed patterning. Combines multiple seeds of the same class to encourage stylistic and contextual diversity.

```
Below are examples of {$LABEL} finan-
cial news:
1. {$SEED_1}
2. {$SEED_2}
3. {$SEED_3}
Generate 5 more realistic sentences
with the same sentiment.
```

Design rationale. Template 1 establishes label semantics; Template 2 enhances lexical and syntactic diversity; Template 3 promotes intra-class generalisation. Together, these templates aim to produce a compact yet expressive synthetic corpus aligned with financial tone and sentiment.

Figure 1: Overview of the proposed framework: clustering-based seed selection → GPT-4o prompting and synthetic generation → compact student fine-tuning and evaluation.

3.3. Student Model Fine-tuning

Three compact encoder-based models—**DistilBERT** (Sanh et al., 2019), **BERT-Tiny** [1], and **ModernBERT** (Warner et al., 2024)—are fine-tuned on the combined real and synthetic datasets using standard supervised training. Lower transformer layers are optionally frozen to improve stability, and early stopping is applied to mitigate overfitting. The trained student models can then be applied to the task of financial sentiment analysis.

4. Datasets and Experimental Setup

Datasets We evaluate the proposed framework on two complementary English corpora: (i) *Financial PhraseBank* (Malo et al., 2014),(Malo et al., 2014), consisting of formal, expert-authored statements, and (ii) *Twitter Financial News Sentiment* (on Huggingface, 2022), comprising informal, real-time investor discourse. Labels for this second dataset are standardised to {*Negative, Neutral, Positive*} by mapping *Bearish* → *Negative* and *Bullish* → *Positive*. The text data consists only of the Tweets (social media posts) themselves, without metadata such as user handles and timestamps. Examples from the two datasets are shown in the Appendix in Table 5.

Pre-processing Text is pre-processed to preserve sentiment-bearing cues while ensuring compatibility with each model's tokenizer. Automatic cleaning steps comprised decoding escaped Unicode sequences, removing stray quotation marks and leading enumerations, and trimming extraneous whitespace. Aggressive normalisation steps—such as lowercasing, punctuation removal, or contraction expansion—are deliberately avoided, as stylistic features like capitalisation, repeated symbols, emojis, hashtags, and lexical emphasis often encode sentiment polarity and intensity in financial discourse. Stopwords are also retained to preserve syntactic and pragmatic signals (e.g., negation, modality), which are critical for accurate sentiment interpretation.

[1] https://huggingface.co/prajjwal1/bert-tiny

All text is tokenised using the HuggingFace tokenizer corresponding to each model. **DistilBERT** (Sanh et al., 2019) employs an uncased vocabulary, while **BERT-Tiny** and **ModernBERT** (Warner et al., 2024) maintain original casing to preserve distinctions vital to financial language, including stock tickers (`$AAPL`) and acronyms (`GDP`). Sequences are truncated or padded to **512 tokens** for consistency across models, aligning with their maximum supported context length. An **80/10/10** stratified split is applied for training, validation, and testing under fixed random seeds to ensure reproducibility and balanced label distribution. This minimal yet model-consistent normalisation retains the linguistic richness and domain-specific stylistic variation essential for reliable sentiment representation in subsequent embedding and distillation stages.

Seed Sets and Synthetic Expansion Seed selection and prompting follow the procedures described in Section 3. Two configurations are employed: a **105-seed set**, for which we select exactly 105 representative sentences either via k-means clustering (one centroid-nearest sentence per cluster) or via uniform random sampling as a comparison baseline, and a **12-sample set** representing a minimal supervision baseline, also selected using the cluster centroid approach.

Each configuration is expanded using **GPT-4o** (OpenAI, 2024), applying the three structured prompting templates introduced earlier. This process yields approximately a **ninefold** increase in data volume after deduplication and class balancing.

4.1. Training Regimes and Evaluation

We design controlled training regimes to isolate the effects of dataset size, class balancing, layer freezing, and synthetic augmentation. For *Financial Phrasebank,* nine regimes progressively incorporate these factors (Table 1), serving as the main diagnostic setup. For regimes using random selection of seed examples, each model was trained three times with different random seeds to reduce the influence of single-run randomness and ensure fair comparison across models. For *Twitter*

ID	Description
	Financial Phrasebank
(1)	Full training set (1,811/226/227 split)
(2)	105 seeds (* clustered); natural imbalance
(3)	(2) + 4 frozen layers + ES
(4)	(3) + balanced (** random)
(5)	(3) + balanced (* clustered)
(6)	(5) + Synth (** random)
(7)	(5) + Synth (* clustered)
(8)	12 seeds + 4 frozen layers + ES + balanced
(9)	(8) + Synth
	Twitter Financial News Sentiment
(10)	Full training set (9,544/1,193/1,194 split)
(11)	105 seeds + 4 frozen layers + ES + balanced (** random)
(12)	105 seeds + 4 frozen layers + ES + balanced (* clustered)
(13)	Balanced + Synth (** random)
(14)	Balanced + Synth (* clustered)

Table 1: Training regimes for sentiment classification on *Financial PhraseBank* (IDs 1–9) and *Twitter Financial News Sentiment* (IDs 10–14). Each ID corresponds to a specific experimental configuration referenced in Section 5. "ES" = early stopping; "Synth" = ninefold GPT-4o augmentation. ** = random seed selection; * = clustered seeds (ours).

Financial News Sentiment, five regimes evaluate robustness on shorter, noisier financial text with class distribution 65/20/15 (Table 1), mirroring the most informative PhraseBank setups.

All regimes share identical validation and test splits to ensure comparability. Compact student models—**DistilBERT** (Sanh et al., 2019), **BERT-Tiny**, and **ModernBERT** (Warner et al., 2024)—are fine-tuned using standard optimisation settings with early stopping based on validation loss. Lower encoder layers are frozen in selected regimes to enhance stability and reduce computational cost.

We evaluate the student models against the larger teacher model, **GPT-4o**, prompted directly to classify sentiment. We also compare an off-the-shelf sentiment classifier, **FinBERT** (Araci, 2019), which combines domain-specific pretraining and fine-tuning on gold-standard, human-labelled data from Financial Phrasebank. FinBERT uses the standard uncased BERT tokenizer (i.e., tokens are converted to lower case)[2].

Performance is reported as **accuracy** and **macro-F1** on the held-out test set under a fixed random seed.

[2]We used the classifier implementation provided at https://huggingface.co/ProsusAI/finbert

5. Results and Analysis

In this section, we report the performance of the proposed synthetic data distillation framework across datasets and models, assess the contribution of each prompt template, compare clustering-based versus random seed selection, and analyse errors.

5.1. Overall Performance

Table 2 summarises the results of compact student models across datasets and training regimes. For regimes using random seed selection, we report mean results across three runs with different random initialisations. The findings confirm that the proposed framework effectively narrows the performance gap between lightweight encoder models and the large instruction-tuned teacher (**GPT-4o**).

Financial Phrasebank: On *Financial PhraseBank* (IDs 1–9), **ModernBERT** achieves the highest score among student models, reaching **95.15%** accuracy and **94.63%** macro-F1 under the full synthetic–seed configuration (ID 7). This represents a **3.09-point gap** to the off-the-shelf FinBERT classifier (98.24/97.71) and a **2.20-point gap** to the zero-shot teacher (97.35/97.57), while using under 6% of the original human-annotated data. **DistilBERT** performs comparably (93.83/92.76; ID 7), demonstrating that compact encoders can approximate full-scale performance when supported by high-quality synthetic supervision.

On Financial Phrasebank, FinBERT outperforms ModernBERT in the full training regime (ID 1), benefitting from its additional pretraining stage on in-domain financial text, which could be added to our approach in future work.

Twitter Financial News Sentiment: On the more challenging social media dataset (IDs 10–14)—comprising shorter, noisier, and colloquial text—the advantages of synthetic instruction distillation become even more evident. **ModernBERT**, trained on clustered synthetic–seed data (ID 14), achieves **77.14%** accuracy and **71.14%** macro-F1, surpassing the GPT-4o zero-shot teacher (72.78/71.45) on both metrics. This improvement is significant ($p < 0.01$) according to the McNemar's test (McNemar, 1947; Dietterich, 1998) ($\chi^2 = 6.88$ with $p = 0.0087$, see also Appendix 10.2). This indicates that domain-targeted synthetic data can exceed generic large-model reasoning on unstructured financial discourse.

The off-the-shelf FinBERT classifier achieves 71.86% accuracy and 67.39% macro-F1 on the Twitter dataset, which is also notably lower than the student models (IDs 13 and 14). FinBERT was trained on Financial Phrasebank but not on the

Twitter dataset, so it is to be expected that its performance is weaker on the social media dataset.

Seed Selection: Across both datasets, clustering-based seed selection consistently outperforms random sampling—particularly in low-resource regimes such as IDs (4–5) and (13–14)—yielding +3–7 F1 gains on average. Repeated multi-seed runs exhibit minimal variance (<1 F1 deviation), confirming the robustness and reproducibility of these improvements. Overall, structured seed selection coupled with synthetic expansion enables compact encoder models to attain strong accuracy, generalisation, and stability under minimal supervision.

5.2. Examples of Synthetic Data

Table 3 presents three representative real–synthetic pairs to demonstrate the complementary mechanisms of the proposed prompting templates (P1–P3), each targeting a particular sentiment class. In practice, each template is applied to generate data from all of the sentiment classes; we show one class per template here for illustration purposes only. The examples show how each template leverages sentiment cues from seed data to produce synthetic text that is lexically diverse yet sentiment-aligned.

P1 (Neutral) adopts a contrastive design, combining one Bearish, one Neutral, and one Bullish seed in a single prompt. The model observes lexical polarity cues such as *"sues"* (negative), *"moves"* (neutral), and *"investments"* (positive), balancing them to generate a neutral, fact-focused headline. This design encourages the model to learn relational sentiment boundaries.

P2 (Bullish) relies on a single seed and asks for alternative rephrasings of the same sentiment. Here, the model retains domain-specific context (*oil production*, *market movement*) while varying syntax and word choice, demonstrating stable polarity control and lexical flexibility.

P3 (Bullish) provides multiple same-label seeds, prompting generalisation across related contexts. By abstracting common optimistic features such as *price surges*, *upgrades*, and *investor confidence*, the model produces coherent yet varied bullish statements that remain faithful to the domain.

Together, these examples highlight how structured prompting—through cross-sentiment contrast (P1), within-sentiment rephrasing (P2), and multi-seed generalization (P3)—enhances realism and diversity in the synthetic financial corpus.

5.3. Ablation Study on Prompt Templates

To evaluate the contribution of each prompting strategy, we performed an ablation study on *Twitter Financial News Sentiment*. All models were trained under identical conditions with 105 real seeds and 420 synthetic samples, using fixed validation and test splits. Each variant excludes one template from the synthetic pool to isolate its effect.

As shown in Table 4, removing any single template reduces accuracy and macro-F1, confirming that all three templates contribute complementary diversity to the synthetic corpus. The largest drop occurs when Template 3 is removed, indicating that its multi-seed patterning is especially important for broadening contextual coverage and improving generalisation on Twitter text. Nevertheless, the relatively balanced performance across the three ablations demonstrates that each template plays a distinct, non-redundant role, and that combining prompts can improve on using a single template.

5.4. Effect of Seed Selection Strategy

To better understand why clustered sampling improves performance, Figure 2 visualises the spatial distribution of seed coverage for the *Twitter Financial News Sentiment* dataset using a shared t-SNE projection of the SBERT embeddings. Clustered seeds (top row) show broad and evenly dispersed coverage across the embedding space, effectively representing both central and peripheral semantic regions. This balanced selection ensures that subsequent synthetic generation captures diverse lexical, contextual, and stylistic variations characteristic of social-media discourse, contributing to the generalisation improvements observed in Table 2.

In contrast, randomly selected seeds (bottom row) tend to concentrate in high-density regions, overlooking many semantically distinct areas. As a result, synthetic samples generated from random seeds exhibit lower diversity and weaker domain coverage. Overall, both visual and quantitative evidence demonstrate that **clustering-based seed selection enhances semantic diversity and leads to more effective synthetic expansion**, particularly in noisy, low-resource domains such as financial Twitter data.

5.5. Error Analysis Across Datasets

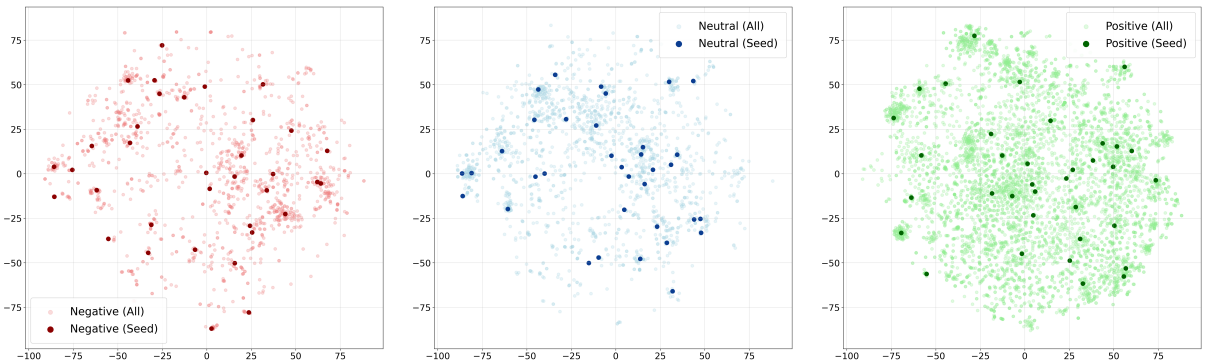
To better understand model behaviour, we conducted a qualitative error analysis across both datasets using the **best-performing model, ModernBERT**. This analysis focuses on ModernBERT because it achieved the highest overall performance among all evaluated student models, making it the most representative case for examining residual classification errors and linguistic patterns.

For **Financial PhraseBank**, ModernBERT achieves consistently strong performance (95.15% accuracy; macro-F1 = 94.63), with most errors arising from the *positive* class. Specifically, 3.5% of

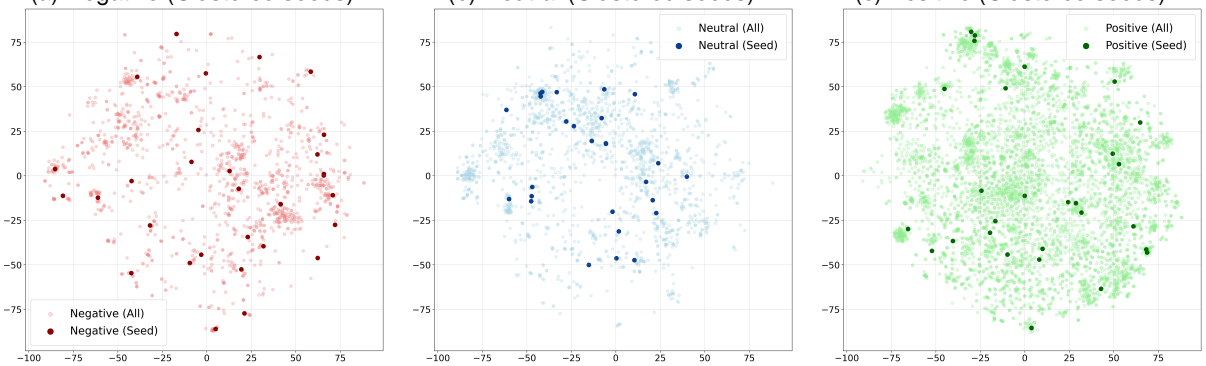
Dataset	ID	Regime	Model	Accuracy (%)	Macro-F1 (%)
PhraseBank	(1)	Full	DistilBERT	96.04	94.95
			ModernBERT	**96.48**	**95.78**
			BERT-Tiny	85.90	85.38
	(2)	105	DistilBERT	80.18	64.48
			ModernBERT	**90.75**	**89.15**
			BERT-Tiny	68.72	53.37
	(3)	105 + Frz+ES	DistilBERT	**89.43**	83.40
			ModernBERT	88.99	**85.79**
			BERT-Tiny	80.62	79.30
	(4)	105 + Frz+ES + Bal (Rand)†	DistilBERT	**91.63**	**88.81**
			ModernBERT	83.99	80.37
			BERT-Tiny	73.32	61.49
	(5)	105 + Frz+ES + Bal (Clust)	DistilBERT	92.07	89.73
			ModernBERT	**92.51**	**91.25**
			BERT-Tiny	80.17	80.27
	(6)	105 + Frz+ES + Bal + Synth (Rand)†	DistilBERT	92.95	90.63
			ModernBERT	**94.57**	**93.29**
			BERT-Tiny	80.03	75.98
	(7)	105 + Frz+ES + Bal + Synth (Clust)	DistilBERT	93.83	92.76
			ModernBERT	**95.15**	**94.63**
			BERT-Tiny	88.99	87.69
	(8)	12 + Frz+ES + Bal	DistilBERT	**79.30**	**67.30**
			ModernBERT	67.40	55.54
			BERT-Tiny	58.59	47.78
	(9)	12 + Frz+ES + Bal + Synth	DistilBERT	**88.99**	85.07
			ModernBERT	88.55	**86.02**
			BERT-Tiny	80.18	74.67
		Publicly available, pretrained classifier	FinBERT	**98.24**	**97.71**
		Teacher (zero-shot)	GPT-4o	97.35	97.57
Twitter	(10)	Full	DistilBERT	86.52	82.49
			ModernBERT	**86.60**	**82.54**
			BERT-Tiny	82.07	76.01
	(11)	105 + Frz+ES + Bal (Rand)†	DistilBERT	**64.13**	**55.30**
			ModernBERT	59.90	50.06
			BERT-Tiny	57.73	47.19
	(12)	105 + Frz+ES + Bal (Clust)	DistilBERT	68.50	61.64
			ModernBERT	**74.20**	**66.14**
			BERT-Tiny	56.87	49.06
	(13)	105 + Frz+ES + Bal + Synth (Rand)†	DistilBERT	70.88	65.40
			ModernBERT	**74.34**	**68.06**
			BERT-Tiny	59.59	56.04
	(14)	105 + Frz+ES + Bal + Synth (Clust)	DistilBERT	74.79	68.22
			ModernBERT	**77.14**	**71.14**
			BERT-Tiny	64.99	56.46
		Publicly available, pretrained classifier	FinBERT	71.86	67.39
		Teacher (zero-shot)	GPT-4o	**72.78**	**71.45**

Table 2: Results across datasets and regimes. **Full** = full training set; **105** = 105 seeds (natural imbalance); **Frz+ES** = frozen lower layers (4; Tiny=2) + early stopping; **Bal** = class-balanced seeds (35/class); **Synth** = +9×GPT-4o augmentation; **Rand**/**Clust** = random vs. clustered seeds. † = average over 3 random runs.

positive samples were predicted as negative and 7.0% as neutral, whereas the *negative* class was perfectly separated and the *neutral* class exhibited minimal confusion. This trend reflects the subtle and sometimes ambiguous tone of corporate financial statements: factual phrasing can obscure



(a) Negative (Clustered seeds) (b) Neutral (Clustered seeds) (c) Positive (Clustered seeds)



(d) Negative (Random seeds) (e) Neutral (Random seeds) (f) Positive (Random seeds)

Figure 2: **t-SNE comparison of seed selection strategies.** Columns correspond to sentiment classes (Negative, Neutral, Positive). Top row: seeds chosen by *K-means clustering*; Bottom row: seeds obtained by *random sampling*. All panels share a fixed t-SNE projection of PhraseBank embeddings, allowing direct spatial comparison between selection strategies.

positive sentiment, while comparative expressions containing explicit terms such as *"negative"* may mislead the classifier. *Representative misclassifications (highest-confidence examples):*

- **Positive** → **Neutral** (4/57 = 7.0% of positive examples): "The company increased its production capacity following the completion of its new plant in Germany." *(confidence 91.8%)*
- **Positive** → **Negative** (2/57 = 3.5%): "Operating profit improved to EUR 2.3 mn compared to a *negative* EUR 5.1 mn a year earlier." *(confidence 89.4%)*

On **Twitter Financial News**, ModernBERT reaches 77.14% accuracy and a macro-F1 of 71.14%. Misclassifications often stem from terse, context-poor posts dominated by tickers or URLs, where limited linguistic cues hinder sentiment inference. Among the misclassified tweets (N=273), `$TICKER` symbols appear in 22.0% (60/273) and URLs in 42.5% (116/273), while emojis are absent. Synthetic augmentation improves robustness—particularly for **DistilBERT** and **BERT-Tiny**—yet residual errors frequently involve neutral macroeconomic or market-commentary headlines flagged as bearish, or brief bullish cues interpreted as neutral. Representative cases include:

- **Neutral** → **Bearish**: "Did Changing Sentiment Drive Mountain Province Diamonds's (TSE:MPVD) Share Price Down A Painful 82%?" *(confidence 94.9%)*
- **Bullish** → **Neutral**: "Stock Market Update: Netflix remains among today's winners" *(confidence 95.1%)*

6. Conclusion and Future Work

This study presented a lightweight framework for **distillation from synthetic data**, enabling compact encoder models to inherit instruction-following behaviour from large teachers using a minimal number of domain-specific seeds. Through clustering-based seed selection and structured prompting, the framework generates semantically diverse, label-consistent data that enhances model performance while reducing annotation effort. With only 12–105 human-labelled sentences, distilled students achieved strong performance, and ModernBERT even *surpassed* the zero-shot GPT-4o teacher on noisy financial text. This result aligns

Template 1 (P1 – For Neutral)

Real Seed:

1. *Bearish: UPDATE 1–California sues e-cigarette maker Juul for selling nicotine products to youth.*
2. *Neutral: Stocks making the biggest moves midday: TD Ameritrade, Tiffany, Uber, Hasbro & more.*
3. *Bullish: $AMZN — Amazon: One Of The Best Long-Term Investments In The Tech Sector.*

Synth. e.g.: *Wall Street opens flat as investors await key economic data releases later this week.*

Template 2 (P2 – For Bullish)

Real Seed: *Oil boosted by renewed hopes for global production cut* `https://t.co/4tAO1U31nz`

Synth. e.g.: *Hopes of a coordinated decrease in global oil production drive prices upward.*

Template 3 (P3 – For Bullish)

Real Seed:

1. *$AMZN – Amazon: One Of The Best Long-Term Investments In The Tech Sector.*
2. *STOCKS SURGE INTO THE CLOSE: Dow up 7.59%, Nasdaq up 7.35%, S&P up 6.95%.*
3. *UnitedHealth stock price target raised to $335 from $310 at SunTrust RH.*

Synth. e.g.: *Apple Inc. ($AAPL) hits a new all-time high as analysts increase price target to $225, citing strong iPhone sales and services growth.*

Table 3: Representative real and synthetic examples grouped by prompt template. Each template example is shown for one sentiment type: this is for the purpose of illustrating all sentiment classes; in practice, each template is used to generate examples of all classes. P1 and P3 use three real seeds to guide synthetic generation, while P2 uses a single seed.

Condition	Accuracy (%)	Macro-F1 (%)
Full model	**74.43**	**67.44**
Without P1	72.92	67.06
Without P2	72.81	67.20
Without P3	71.25	65.96

Table 4: Ablation on prompt templates for the *Twitter Financial News Sentiment* dataset using 105 real seeds and 420 synthetic samples.

with previous findings that show how smaller models with task-specific fine-tuning can out-perform large, generalist LLMs (Koćon et al., 2023; Liang et al., 2023). Ablation results verified the importance of semantically guided seeds and combination of structured prompts, while consistent outcomes across two very different datasets demonstrated generalisability.

Future work could explore active learning-based seed selection (Settles, 2009) to improve coverage, learning from human-annotated or model-generated explanations to enhance transparency (Hase and Bansal, 2022; Wiegreffe and Marasović, 2021), and calibration-aware distillation, which transfers not only predictive accuracy but also uncertainty calibration from teacher to student (Müller et al., 2019; Lee et al., 2022) to further improve reliability. This framework could also be extended beyond sentiment analysis to other finance-related tasks, such as event detection, financial risk signals, and decision reasoning.

7. Limitations

While effective, the proposed framework exhibits several limitations. First, performance can vary with prompt phrasing and the representativeness of selected seeds; domain shift may reduce gains if the seed distribution fails to capture emerging financial expressions. Second, reliance on a single teacher model restricts linguistic variety and may propagate its biases, particularly in sentiment or stance expressions tied to market narratives. Third, although clustering promotes semantic diversity, it may not fully capture pragmatic or temporal nuances in evolving financial discourse. Finally, the framework has not yet been evaluated beyond English or in cross-domain adaptation settings.

To ensure safe and transparent application, future studies should analyse class-wise calibration, document prompt–output dependencies, and consider incorporating human oversight or expert validation when deploying sentiment analysis models in sensitive or high-stakes financial contexts.

Data and Code Availability

The code, trained models, and synthetic datasets used in this study are publicly available at: `https://github.com/XavierHuangWF/efficient-financial-distillation`.

8. Bibliographical References

Dogu Araci. 2019. FinBERT: financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Kemal Kirtac and Guido Germano. 2024. Sentiment trading with large language models. *Finance Research Letters*, 62:105227.

Jan Kocon, Igor Cichecki, Olgierd Kaszyca, Mateusz Kochanek, Małgorzata Maciejewska, Adam Stępień, Katarzyna Pereła, and Adam Radziszewski. 2023. ChatGPT: jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin Zhang. 2022. Hard gate knowledge distillation - leverage calibration for robust and reliable language model. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9793–9803, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *arXiv prepreint arXiv:2211.09110*.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Aru Maekawa, Satoshi Kosugi, Kotaro Funakoshi, and Manabu Okumura. 2024. DiLM: distilling dataset into language model for text-level dataset distillation. In *Findings of the Association for Computational Linguistics: NAACL 2024*. ArXiv:2404.00264 [cs.CL], accepted to Findings of NAACL 2024.

Pekka Malo, Ankur Sinha, Pekka Takala, Pasi Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2025. Towards efficient generative large language model serving: A survey from algorithms to systems. *ACM Computing Surveys*, 58(1):1–37.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4694–4703.

OpenAI. 2024. GPT-4o Technical Report. Technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. 2023. Large language model adaptation for financial sentiment analysis. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 1–10, Bali, Indonesia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, pages 1–6.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin–Madison, Computer Sciences Technical Report 1648.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *35th conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Ranran Zhen, Juntao Li, Yixin Ji, Zhenlin Yang, Tong Liu, Qingrong Xia, Xinyu Duan, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2025. Taming the titans: A survey of efficient LLM inference serving. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 522–541, Hanoi, Vietnam. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

9. Language Resource References

Malo, Pekka and Sinha, Ankur and Takala, Pekka and Korhonen, Pasi and Wallenius, Jyrki. 2014. *Financial PhraseBank*. PID https://huggingface.co/datasets/takala/financial_phrasebank. Sentiment annotations for financial text.

Zeroshot on Huggingface. 2022. *Twitter Financial News Sentiment Corpus*. PID https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment. Three-class sentiment labels for finance-related tweets.

10. Appendix

10.1. Dataset Examples

Examples illustrating the text data from Financial Phrasebank and Twitter Financial News Sentiment are shown in Table 5.

10.2. Statistical and Reproducibility Analysis

To assess the reliability and significance of the proposed framework, we combined statistical testing with empirical consistency analysis. Although each experiment was conducted under a fixed random seed, performance across datasets and configurations remained stable, indicating that the observed gains are not artefacts of random variation. Such consistency is particularly important for low-resource adaptation, where limited supervision can otherwise amplify stochastic effects.

Because only **ModernBERT** surpassed **GPT-4o** in raw accuracy on the Twitter dataset, we further conducted a McNemar's test (McNemar, 1947) to determine whether this improvement was statistically significant. McNemar's test evaluates discordant prediction pairs—instances where one model is correct while the other is not—under the null hypothesis that both models exhibit identical error distributions. It is a standard non-parametric method for paired nominal data, widely used for classifier comparison (Dietterich, 1998).

On the 1,194 aligned Twitter test samples, ModernBERT (distilled) achieved 77.14% accuracy compared to 72.78% for GPT-4o. Among 378 discordant predictions, ModernBERT was correct in 215 cases while GPT-4o was correct in 163 cases, yielding a ratio $b/(b+c) = 0.569$ (95% CI [0.517, 0.619]). The continuity-corrected McNemar statistic was $\chi^2 = 6.88$ with $p = 0.0087$, indicating statistical significance at the 0.05 level. This confirms that ModernBERT's advantage on noisy financial Twitter sentiment classification is unlikely to be due to chance, but rather reflects a genuine improvement in generalisation and robustness.

10.3. Training Hyperparameters

This appendix documents the experimental environment and full training configuration to ensure reproducibility.

10.3.1. Optimisation and Training Configuration

Across all experiments, we fine-tuned the student models using **AdamW** with a fixed weight decay of 0.10 and a fixed random seed (24266). Unless stated otherwise, we used a training batch size of 8 and an evaluation batch size of 32. Sequences were truncated or padded to a maximum length of 512 tokens. A linear learning-rate schedule with warm-up was applied. When enabled, early stopping monitored validation loss with `patience = 10`. Table 6 therefore reports only the hyperparameters that varied across model families and datasets (learning rate and the number of frozen encoder layers).

10.4. Layer Freezing and Stability Notes

DistilBERT and ModernBERT, with deeper stacks (6–12 layers) and wider hidden sizes, achieved stable convergence with a learning rate of 1×10^{-4} and could safely freeze four lower layers in reduced-data regimes.

BERT-Tiny, being shallower and narrower, required a higher learning rate (1×10^{-3}) to avoid underfitting and froze only two layers, as additional freezing removed too much trainable capacity.

On the noisier Twitter dataset, all models used stronger regularisation (weight decay $= 0.10$) and identical batch sizing (8/32) to improve stability under short, variable-length inputs.

10.5. Hardware and Environment

All experiments were conducted on a ROG Strix G533QS laptop equipped with an AMD Ryzen 9 5900HX CPU (3.30 GHz, 8 cores), 32 GB RAM, and an NVIDIA GeForce RTX 3080 Laptop GPU (16 GB VRAM). The system ran Windows 10 Home (Version 22H2) on a 954 GB SSD.

Financial Phrasebank
Positive: *Viking Line 's cargo revenue increased by 5.4 % to EUR 21.46 mn , and cargo volume increased by 2.4 % to 70,116 cargo units .*
Neutral: *At the request of Finnish media company Alma Media 's newspapers , research manager Jari Kaivo-oja at the Finland Futures Research Centre at the Turku School of Economics has drawn up a future scenario for Finland 's national economy by using a model developed by the University of Denver .*
Negative: *Pharmaceuticals group Orion Corp reported a fall in its third-quarter earnings that were hit by larger expenditures on R&D and marketing .*
Twitter Financial News Sentiment
Bullish (positive): *$BHE: Lake Street starts at Buy*
Neutral: *CA$10.60 - That's What Analysts Think Vecima Networks Inc. Is Worth After These Results*
Bearish (negative): *Autodesk downgraded to underweight from neutral at JPMorgan*

Table 5: Examples data from Financial Phrasebank and Twitter Financial News Sentiment, illustrating differences in language.

Table 6: Hyperparameter settings across datasets and models. "Frozen" indicates the number of lower encoder layers frozen in reduced-data regimes.

Dataset	Model	LR	Frozen
PhraseBank	DistilBERT	1×10^{-4}	4
	ModernBERT	1×10^{-4}	4
	BERT-Tiny	1×10^{-3}	2
Twitter	DistilBERT	1×10^{-4}	4
	ModernBERT	1×10^{-4}	4
	BERT-Tiny	1×10^{-3}	2