Sentiment Analysis and Short-Term Return Predictability in Small-Cap Nuclear & Energy-Transition Equities: A FinBERT–LSTM Approach

Academic Year 2025–2026

Master Thesis

Author: Alexandre BREDILLOT

Advisor: Professor Majid ESKANDARPOUR

Programme: MSc in Data Analytics & Artificial Intelligence

Date: April 2026

« EDHEC Business School does not express approval or disapproval concerning the opinions given in this paper which are the sole responsibility of the author. »

« L'EDHEC Business School n'exprime aucune forme d'approbation ou de désapprobation quant aux opinions exprimées dans ce document, qui relèvent de la seule responsabilité de l'auteur. »

« I certify that: the Master Project being submitted for examination is my own research, the data and results presented are genuine and actually obtained by me during the conduct of the research and that I have properly acknowledged using parenthetical indications and a full bibliography all areas where I have drawn on the work, ideas and results of others and that the master project has not been presented to any other examination committee before and has not been published before. »

« Je certifie que : le Master Project soumis à évaluation est le fruit de mes propres recherches, que les données et les résultats présentés sont vrais et réellement obtenus par moi au cours de la recherche et que j'ai dûment reconnu, à l'aide de parenthèses et d'une bibliographie complète, tous les travaux, les idées et les résultats d'autres personnes sur lesquels je me suis appuyé sur les travaux, et que le Master Project n'a pas été présenté à un autre comité d'examen par le passé et n'a pas été publié auparavant. »

Acknowledgements

I would like to express my sincere gratitude to Professor Majid Eskandarpour, my thesis director, for his precise and demanding feedback. His insistence on scope discipline, on validating data coverage before committing to the pipeline, and on framing the contribution around a falsifiable hypothesis rather than a method application shaped this work decisively. Any remaining weaknesses are mine alone.

I also thank the EDHEC faculty of the MSc Data Analytics & Artificial Intelligence programme for the technical foundation that made this project possible, and the authors of the FNSPID dataset and the FinBERT model for making high-quality resources freely available to the research community.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
BERT	Bidirectional Encoder Representations from Transformers
DM	Diebold–Mariano (test)
EMH	Efficient Market Hypothesis
DA	Directional Accuracy
FinBERT	Financial BERT (ProsusAI finetune)
FNSPID	Financial News and Stock Price Integration Dataset
LSTM	Long Short-Term Memory (network)
MACD	Moving Average Convergence Divergence
MAE	Mean Absolute Error
NLP	Natural Language Processing
OHLCV	Open, High, Low, Close, Volume
RMSE	Root Mean Square Error
RSI	Relative Strength Index
SHAP	SHapley Additive exPlanations

Abstract

The global energy transition has pushed small-cap nuclear and energy-transition equities to the centre of capital-market attention, yet these "difficult-to-value" stocks remain under-studied in the sentiment-forecasting literature. This thesis investigates whether FinBERT-derived news sentiment improves short-term return prediction for such stocks, and whether that improvement is stronger for small-caps than for large-caps in the same sector — a testable "Small-Cap Sentiment Premium" grounded in Baker and Wurgler (2006). The empirical pipeline applies FinBERT (ProsusAI) to a corpus of 25,885 articles drawn from the FNSPID dataset and supplementary web-scraping, aggregates article-level scores into a daily Sentiment Index and a three-day Sentiment Momentum feature, and compares a Baseline LSTM against a Sentiment-Augmented LSTM on six small-cap Core Set and three large-cap Benchmark Set nuclear/energy tickers over July 2024 – April 2026. Forecast accuracy is assessed via MAE, directional accuracy and the Diebold–Mariano test; SHAP values probe feature importance.

Results are horizon-dependent. At the 1-day horizon sentiment augmentation does not improve on the baseline in aggregate. At the 5-day horizon, however, it produces a statistically significant MAE improvement on four of the six Core Set tickers (average +5.18%, directional accuracy rising from 43.8% to 48.9%). The Small-Cap Sentiment Premium is directionally supported (+7.94 percentage points in favour of small-caps) and corroborated by SHAP decomposition, but does not reach statistical significance on the nine-ticker sample. For practitioners, the findings indicate that news-sentiment signals are best exploited on multi-day rather than same-day horizons, particularly in well-covered small-cap thematic equities; for research, they show that falsifiable-hypothesis framing converts a null result into informative evidence about the size and noise of the underlying effect.

Keywords: FinBERT, LSTM, sentiment analysis, return predictability, small-cap equities, nuclear energy, behavioural finance, Diebold–Mariano test, SHAP, Baker–Wurgler.

Chapter 1. Introduction

1.1. Motivation and context

The global energy system is undergoing its most profound reorganisation since the post-war oil era. The political commitment to decarbonisation — codified in the Paris Agreement, the European Green Deal, the US Inflation Reduction Act of 2022 and, more recently, a wave of explicit pro-nuclear declarations from European and Asian governments — has moved nuclear energy and its associated upstream industries (uranium mining, enrichment, small modular reactor manufacturing, critical battery metals) back into the centre of capital-market attention. Between mid-2024 and early 2026, the tickers that most directly express this thesis — NuScale Power (SMR), Centrus Energy (LEU), Lightbridge (LTBR), NexGen Energy (NXE), NANO Nuclear Energy (NNE), Lithium Americas (LAC), Cameco (CCJ), Constellation Energy (CEG) and BWX Technologies (BWXT) — have exhibited some of the most violent price moves in the US equity universe.

These stocks share a set of microstructural characteristics that make them a natural laboratory for testing the interaction of information and prices. They are heavily exposed to policy and regulatory news; they have limited fundamental anchors because many are pre-revenue or early-revenue firms; they attract disproportionate retail interest relative to their market capitalisation; and they are covered by a comparatively small number of sell-side analysts. In the language of Baker and Wurgler (2006), they are "difficult-to-value" stocks — precisely the segment in which sentiment should exert a measurable and asymmetric influence on returns. If behavioural finance has any empirical content, then this is where it should appear.

At the same time, the tools available to quantify sentiment have changed beyond recognition over the last five years. Dictionary-based approaches (Loughran & McDonald, 2011) have given way to transformer-based models pre-trained on financial corpora (Araci, 2019; Yang et

al., 2020), which handle negation, domain jargon and long-range dependencies in a way that lexical methods cannot. FinBERT in particular has become the de facto benchmark for financial text classification, and its outputs — a probability distribution over positive, neutral and negative classes — provide a continuous, machine-friendly signal that can be fed directly into downstream time-series models.

This thesis sits at the intersection of these two currents. It asks a simple, falsifiable question: if we take a standard recurrent architecture (an LSTM) trained to forecast short-horizon returns from price-based features, and we add FinBERT-derived news sentiment as an additional input, do we obtain a statistically significant improvement for the kind of stocks where behavioural theory says we should?

1.2. The research gap

A sizeable literature has examined sentiment-augmented return prediction in broad equity markets, typically on S&P 500 constituents or on index-level forecasts (Hiew et al., 2019; Zhang et al., 2022). A smaller but growing body of work studies sentiment and clean-energy or commodity markets (Reboredo & Ugolini, 2021; Liu & Hamori, 2020). Research on the nuclear energy sector specifically is thinner still, and most of it focuses on public perception and social licence rather than on stock returns (Kwon & Radaideh, 2024; Suh et al., 2020). Crucially, no prior study has combined (a) a sector-specific small-cap focus on nuclear and energy-transition equities, (b) a modern transformer-based sentiment pipeline, and (c) a direct comparative test of whether sentiment's predictive value is stronger in small-caps than in large-caps of the same sector. That joint gap is the empirical target of this thesis.

The research question is formulated as follows: Does the integration of FinBERT-derived news sentiment scores into an LSTM forecasting model produce statistically significant improvements in short-term (1-day and 5-day) return prediction accuracy for small-cap nuclear

equities, compared to a price-only baseline? And does that improvement differ significantly between small-caps and large-caps of the same sector?

1.3. Contribution and structure

The thesis makes three contributions. First, methodologically, it builds and documents a fully reproducible FinBERT–LSTM pipeline for a niche sector, including an explicit Phase 0 feasibility audit of news coverage — a step that, as we will show, turns out to be essential given the uneven coverage of small-cap tickers in public news corpora. Second, empirically, it provides the first head-to-head evaluation of sentiment-augmented return prediction in the small-cap nuclear space, with a clean comparison against large-cap energy benchmarks. Third, epistemically, it tests a falsifiable hypothesis — the Small-Cap Sentiment Premium — and reports a nuanced, partly negative result: sentiment does help, but only at the 5-day horizon, and the differential between small-caps and large-caps, while directionally consistent with theory, is not statistically significant on the available sample. The thesis takes these mixed findings as a feature, not a bug: a null result on a well-specified test is itself a contribution to the literature.

To preview the headline findings before the detailed exposition that follows, the empirical exercise rests on a panel of 3,992 ticker-day records and 25,885 news articles covering nine tickers from July 2024 to April 2026, with a strict 20% chronological out-of-sample window of approximately 74 trading days per ticker. At the 1-day forecast horizon, sentiment augmentation does not improve aggregate MAE on the Core Set (average change of −1.19%) and is significantly worse on one ticker, LEU. At the 5-day horizon, the picture reverses: four of the six Core Set tickers show a statistically significant MAE improvement on the Diebold–Mariano test (average +5.18%, directional accuracy rising from 43.8% to 48.9%), with the largest gains on the most retail-traded names (NNE, LTBR). The Small-Cap Sentiment Premium is directionally positive (+7.94 percentage points in favour of small-caps at 5 days) and is

corroborated by SHAP decomposition, which shows that sentiment features carry 60–80% more weight in the Core Set than in the Benchmark Set, but the between-group mean difference does not reach conventional significance ($p = 0.249$). These findings, and the caveats that qualify them, are developed in Chapters 6 and 7.

The remainder of the thesis is organised as follows. Chapter 2 reviews the relevant literature and the theoretical framework. Chapter 3 formalises the research design and hypotheses. Chapter 4 describes the data collection, the Phase 0 feasibility audit, and the descriptive statistics of the final dataset. Chapter 5 presents the methodology, covering the FinBERT sentiment pipeline, the LSTM architectures, the Diebold–Mariano test and the SHAP explainability framework. Chapter 6 reports the empirical results. Chapter 7 discusses their interpretation, limitations and implications. Chapter 8 concludes.

Chapter 2. Literature Review and Theoretical Framework

2.1. Behavioural finance and the role of sentiment

The theoretical foundation of this thesis rests on the partial departure from the Efficient Market Hypothesis (EMH) associated with the rise of behavioural finance. In its strong form, the EMH holds that all public information is instantaneously and correctly impounded into prices, so that no publicly available signal can systematically predict returns. The accumulation of anomalies over the last three decades — momentum, post-earnings drift, the small-firm effect, the closed-end fund discount puzzle — has eroded the purity of that claim and motivated a family of models in which investor psychology, limited arbitrage, and slow information diffusion jointly prevent prices from fully reflecting information in real time.

Baker and Wurgler (2006) provide the single most influential empirical statement of this view. Using a composite sentiment index constructed from six market-wide proxies (closed-end fund discount, NYSE turnover, number and first-day returns of IPOs, share of equity issues in total issuance, and the dividend premium), they show that when sentiment is high, subsequent returns are low on portfolios of stocks that are most "difficult to value" — small, young, high-volatility, unprofitable, non-dividend-paying, distressed or growth-oriented firms. The economic mechanism is straightforward. For stocks whose fundamental value is hard to estimate, even sophisticated investors face wide confidence bands; short-sale constraints and idiosyncratic risk limit arbitrage; and as a result, sentiment — understood as the component of demand that is uncorrelated with fundamentals — can push prices away from intrinsic value for extended periods. When sentiment reverses, these mispricings correct.

The six small-cap tickers that form the Core Set of this thesis (SMR, LEU, LTBR, NXE, NNE, LAC) satisfy several of the Baker–Wurgler "hard-to-value" criteria simultaneously. They are small, high-volatility, non-dividend-paying, mostly unprofitable, and growth-oriented. Theory

predicts that sentiment should matter more for this group than for the Benchmark Set of large-cap nuclear/energy stocks (CCJ, CEG, BWXT), which are covered by numerous analysts, carry dividends, and have established fundamental anchors. This asymmetric prediction is operationalised in this thesis as Hypothesis H2, the Small-Cap Sentiment Premium.

Tetlock (2007) represents the second canonical reference. Using the Wall Street Journal's "Abreast of the Market" column and a psychology-based General Inquirer dictionary, Tetlock constructs a daily pessimism factor and shows that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals. Crucially, this effect is strongest in small stocks and on days without contemporaneous fundamental news. Tetlock's paper marks the birth of empirical media-finance research: it establishes media content as a legitimate proxy for sentiment and confirms that its predictive content is concentrated exactly where behavioural theory locates it.

Subsequent extensions refine the mechanism. D'Hondt and Roger (2017) show that investor sentiment is amplified in markets with high retail participation, where behavioural biases — overconfidence, limited attention, loss aversion — are less likely to be arbitraged away. Shen and Zhu (2024) connect sentiment directly to limits-to-arbitrage constraints and confirm that hard-to-value stocks are the ones where the sentiment–return relationship is strongest. Glasserman and Mamaysky (2019) demonstrate that unusual news, rather than news per se, is what drives tail events in returns, an observation that has implications for how sentiment features are engineered.

2.2. From dictionaries to transformers: the NLP evolution

Measuring sentiment in text is not trivial, and the literature has passed through three methodological generations. The first generation used general-purpose psychology dictionaries — most famously the Harvard IV-4 General Inquirer used by Tetlock (2007). These lists of positive and negative words were not designed for financial language and produce well-

documented errors in context: words like "liability" or "depreciation" are classified as negative but carry neutral, technical meaning in accounting text.

Loughran and McDonald (2011) formalised this critique and built a finance-specific dictionary. Their work remains the standard lexical benchmark and is still widely used, not least because it is transparent and computationally trivial. But it too has limitations: it treats documents as bags of words, ignores negation except through crude rule-based patches, and cannot handle the long-range dependencies that characterise news writing.

The second generation introduced machine learning classifiers trained on hand-labelled financial corpora. Gu, Kelly and Xiu (2020), in a landmark paper on empirical asset pricing via machine learning, demonstrate that flexible ML models — neural networks, gradient boosting — outperform traditional econometric approaches in cross-sectional return prediction, and document the non-linear and interaction-heavy structure of the return-generating process. Their work legitimises deep learning as a tool for quantitative finance and opens the door to more ambitious architectures.

The third and current generation is transformer-based. The publication of BERT (Devlin et al., 2019) marked a decisive break in NLP: by pre-training a bidirectional attention model on a very large general corpus, BERT produced contextual word embeddings that captured meaning in a way earlier architectures could not. FinBERT, in two independent incarnations — Araci (2019), further refined as the ProsusAI/finbert model, and Yang et al. (2020) — adapted this pre-trained backbone to financial text by continuing pre-training on Reuters and analyst report corpora and then fine-tuning on labelled financial sentiment data. Huang, Wang and Yang (2023) provide the most complete benchmarking of FinBERT against lexical baselines and earlier ML approaches, and confirm its superiority on a wide range of financial NLP tasks, including stock reaction prediction.

This thesis uses the ProsusAI FinBERT model for two reasons. First, it is pre-trained on Reuters TRC2 and fine-tuned on Financial PhraseBank, which closely matches the domain of the FNSPID corpus used downstream. Second, it outputs a full probability distribution over three classes (positive, neutral, negative), which allows for continuous aggregation into a daily sentiment index — a richer signal than the binary or ternary labels produced by simpler approaches. The alternative of training a custom domain-adapted model on the nuclear/energy sub-corpus was considered but rejected on sample-size grounds: the number of relevant, labelled articles in the FNSPID corpus is insufficient for reliable fine-tuning, and the gain from such an effort is likely to be small relative to the well-benchmarked off-the-shelf FinBERT.

2.3. Sentiment and the energy sector

The energy sector has received growing attention in the sentiment-and-returns literature, though less than consumer or technology sectors. Liu and Hamori (2020) examine clean-energy ETFs and find that investor sentiment — proxied by Google Trends and news tone — exerts a significant short-term effect on returns, with the sign of the effect depending on the sub-segment of the clean-energy space. Reboredo and Ugolini (2021) push the analysis further and document an asymmetric impact: negative news generates more volatility and larger price moves in clean-energy stocks than comparably positive news, consistent with loss aversion and the slow diffusion of bad news.

Work on the nuclear sector specifically is more fragmented. Most published studies focus on public opinion and social licence (Suh et al., 2020, on nuclear acceptance in Korea; Kwon & Radaideh, 2024, on LLM-based sentiment extraction from social media posts about nuclear power) rather than on market returns. Kim and Hahm (2023) provide a partial bridge by comparing news and social media sentiment effects on stock performance in a broad sample, confirming that professional news has a stronger short-term price impact than user-generated

content — a finding that motivates this thesis's choice of FNSPID, a curated news dataset, as the primary sentiment input rather than Twitter/X or Reddit.

The specific question of whether sentiment has greater predictive value in small-cap nuclear stocks than in large-cap energy benchmarks has not, to the best of this author's knowledge, been tested empirically. That is the explicit target of the Small-Cap Sentiment Premium hypothesis introduced in Chapter 3.

2.4. LSTM architectures for financial time-series forecasting

The second building block of this thesis is the LSTM network (Hochreiter & Schmidhuber, 1997). Long Short-Term Memory cells were designed to overcome the vanishing-gradient problem of vanilla recurrent networks by introducing gated memory: an input gate, a forget gate and an output gate regulate the flow of information along a persistent cell state, allowing the network to preserve relevant context over long sequences. For financial time-series — which typically exhibit both short-term autocorrelation and longer-horizon regime structure — this inductive bias is a natural fit.

The empirical literature on LSTMs for stock return prediction is large and mixed. Hiew et al. (2019) show that a BERT-based sentiment index fed into an LSTM improves return prediction on Hong Kong listed stocks compared to both price-only baselines and naive sentiment lexicon approaches. Zhang et al. (2022) propose a transformer-based attention architecture and argue that it outperforms LSTMs on stock movement prediction, though their results are sensitive to data-preprocessing choices and trading cost assumptions. Ozbayoglu et al. (2020) provide a comprehensive survey of deep learning in financial applications and document that LSTM-based models remain the most widely deployed architecture in published applied work, largely because of their reliability and the ease of integrating auxiliary features.

In this thesis, the LSTM is chosen over a transformer architecture for two pragmatic reasons. First, the amount of data per ticker is modest — approximately 440 training days after feature engineering and the 20% hold-out — which favours architectures with fewer parameters. Second, the LSTM serves as a transparent, reproducible benchmark: a reader who wishes to verify the results can do so with minimal hardware and standard libraries. The central claim of this thesis is not that LSTMs are the best possible architecture; it is that a given architecture either does or does not benefit from FinBERT sentiment features in a given sector. That comparison is what the empirical design is optimised to isolate.

2.5. Explainability: SHAP and the "why" of a prediction

A recurring criticism of deep learning in finance is that neural networks are black boxes: they produce predictions but offer no way of attributing those predictions to specific inputs, which limits both trust and interpretability. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017), is a unified framework for local feature attribution grounded in cooperative game theory. Each feature is assigned a Shapley value equal to its average marginal contribution to the model's output across all possible feature subsets, producing an additive decomposition of the prediction that satisfies desirable axioms (local accuracy, missingness, consistency).

For this thesis, SHAP serves a specific purpose. Even if the Sentiment-Augmented LSTM does not always outperform the Baseline on MAE, the question remains: is the model actually using sentiment features, or is it effectively ignoring them and relying entirely on price-based signals? SHAP answers this question directly. If sentiment features carry non-trivial mean |SHAP| values, and especially if those values are systematically larger in the Core Set than in the Benchmark Set, then even a weak or null result on aggregate MAE is consistent with the theoretical mechanism — sentiment information is being used, it simply is not large enough to

shift average forecast accuracy on a small out-of-sample window. Chapter 6 exploits this distinction.

2.6. Synthesis and position of this thesis

Three threads come together in this thesis. From behavioural finance, we inherit the prediction that sentiment should matter most in difficult-to-value stocks, and specifically more in small-caps than in large-caps of the same sector. From modern NLP, we inherit FinBERT as a state-of-the-art sentiment extractor that is well-suited to professional financial news. From deep learning for time-series, we inherit the LSTM as a reliable recurrent baseline for short-horizon return prediction, and SHAP as a tool for opening its black box. No prior study has combined these three threads in the small-cap nuclear and energy-transition sector, nor has any study formally tested the Small-Cap Sentiment Premium as a falsifiable differential hypothesis across small- and large-caps of the same sector. This thesis fills that gap.

Chapter 3. Research Design and Hypotheses

3.1. From research question to falsifiable hypotheses

The early version of this project (Outline v1.0, January 2026) framed the contribution as a sector application of existing sentiment-plus-LSTM methods to nuclear small-caps. Professor Eskandarpour's feedback of 11 February 2026 rightly identified this framing as insufficient: a contribution of the form "we apply method X to sector Y" carries low epistemic risk and provides weak evidence of scientific value regardless of the outcome. The revised version (Outline v2.0, February 2026) reframes the contribution around a falsifiable hypothesis — the Small-Cap Sentiment Premium — that makes a directional prediction about how sentiment's predictive value should differ between small-caps and large-caps of the same sector. A null result on this hypothesis is informative; a positive result is informative; and either outcome provides more evidence than a mere method-application exercise would.

Table 1 below summarises the four revisions from v1.0 to v2.0 implemented at that stage of the project.

Advisor concern	v1.0 approach	v2.0 (adopted) approach
Scope too ambitious	Return + volatility prediction; LSTM + GARCH + SHAP	Single objective: does FinBERT sentiment improve short-term return prediction in small-cap nuclear equities? GARCH removed.
FNSPID data coverage	Assumed sufficient coverage without validation	Phase 0 feasibility audit. Tickers below 50 articles are supplemented by scraping or dropped.
Unclear contribution	Framed as sector application of existing methods	Reframed as falsifiable Small-Cap Sentiment Premium hypothesis.
Daily frequency	Mixed daily + intraday	Daily frequency only.

3.2. The two hypotheses

The empirical part of this thesis tests two formally stated hypotheses. H1 is the direct test of whether sentiment adds predictive value at all, on the Core Set. H2 is the differential test that operationalises the Baker–Wurgler theoretical prediction. Both are evaluated at the 1-day and the 5-day forward-return horizons.

H1 (Sentiment improvement hypothesis). An LSTM model augmented with daily FinBERT sentiment scores achieves a lower mean absolute error (MAE) and higher directional accuracy than an identical LSTM using only technical price features, for the Core Set of small-cap nuclear/energy-transition tickers, at both the 1-day and 5-day forecast horizons. This hypothesis draws directly on Tetlock (2007), who showed that high media pessimism predicts downward pressure on prices and subsequent reversals, and on Hiew et al. (2019), who document that a BERT-based sentiment index combined with an LSTM improves short-horizon stock-return forecasts relative to price-only baselines. The null is that the two models have equal forecast accuracy, evaluated via the Diebold–Mariano (DM) test on the out-of-sample residuals.

H2 (Small-Cap Sentiment Premium hypothesis). The predictive improvement from adding sentiment (measured as the percentage reduction in MAE from Baseline to Sentiment-Augmented) is statistically larger for small-cap nuclear stocks (Core Set) than for large-cap energy benchmarks (Benchmark Set). This differential prediction follows directly from Baker and Wurgler (2006), who argue that investor sentiment should exert an asymmetrically stronger effect on "difficult-to-value" stocks — young, small, unprofitable or highly volatile firms — than on large, well-established firms whose fundamentals are anchored by analyst coverage and institutional flow, a channel reinforced empirically by Shen and Zhu (2024) for hard-to-value stocks. The null is that the two group means of MAE improvement are equal, tested via a two-sample t-test on the per-ticker improvement percentages.

Both hypotheses are pre-specified. The Core Set and Benchmark Set are defined below, the features and architectures are frozen before model training, and the hold-out window is a strict chronological 20% of each ticker's time-series, with no lookahead and no hyperparameter tuning on the test set. The purpose of these commitments is to make the empirical exercise a genuine test rather than an exploratory fit.

3.3. Ticker universe and rationale

The ticker universe is deliberately small. For a falsifiable test of the Small-Cap Sentiment Premium, the Core Set must consist of stocks that unambiguously satisfy the Baker–Wurgler "difficult-to-value" criteria and are thematically linked to nuclear and energy-transition exposure, while the Benchmark Set must provide a tight sectoral comparison — large-caps in the same broad theme whose fundamentals are well-known and whose pricing is anchored by analyst coverage and institutional flow.

Category	Ticker	Company	Theme
Core Set	SMR	NuScale Power Corp.	Small modular reactors
Core Set	LEU	Centrus Energy Corp.	Uranium enrichment
Core Set	LTBR	Lightbridge Corp.	Advanced nuclear fuel
Core Set	NXE	NexGen Energy Ltd.	Uranium mining
Core Set	NNE	NANO Nuclear Energy Inc.	Advanced nuclear
Core Set	LAC	Lithium Americas Corp.	Lithium / energy transition
Benchmark	CCJ	Cameco Corp.	Large-cap uranium
Benchmark	CEG	Constellation Energy Corp.	Nuclear utility
Benchmark	BWXT	BWX Technologies Inc.	Nuclear engineering

All six Core Set tickers were small-cap (< USD 10B market cap) at the start of the study period. Several, notably NNE, SMR and NXE, saw dramatic market-cap expansions over the study window, which is itself a feature of the sector: sentiment-driven momentum pushes small-caps into mid-cap territory and back again on short time-scales. The Benchmark Set was streamlined from the eight tickers proposed in v1.0 to the three largest, most-liquid nuclear/energy stocks that retain clear sectoral overlap with the Core Set.

3.4. Forecast targets and evaluation framework

Two forecast targets are studied in parallel. The 1-day forward log return captures the immediate market reaction to news flow and provides the tightest test of information efficiency. The 5-day forward log return captures the slower-diffusion dynamic that behavioural finance predicts in difficult-to-value stocks: if small-caps are subject to limited attention and slow price discovery, sentiment information should show up in the 5-day horizon even where the 1-day reaction is noisy or absent. Running both horizons in parallel, on the same pipeline and the same models, allows this dimension of the analysis to be made explicit.

Evaluation combines three complementary measures. MAE gives the magnitude of average forecast errors, is directly comparable between models, and feeds naturally into the DM test. Directional accuracy gives the hit rate of up/down predictions, which is closer to what a practitioner cares about. The DM test provides the formal significance statement: does the observed MAE difference between Baseline and Sentiment-Augmented exceed what could plausibly arise under the null of equal forecast accuracy, given the autocorrelation of forecast errors? All three are reported per ticker and per group.

Chapter 4. Data

4.1. Overview of the data pipeline

The empirical analysis combines three data sources: (i) daily OHLCV price data for the nine-ticker universe obtained via the yfinance Python library; (ii) a large corpus of financial news articles drawn primarily from FNSPID (Financial News and Stock Price Integration Dataset) and supplemented, where coverage is insufficient, by targeted web-scraping of Yahoo Finance headlines; and (iii) a set of derived technical indicators and sentiment features engineered on top of these raw inputs. The resulting panel dataset covers 647 calendar days from 1 July 2024 to 9 April 2026, yielding 3,992 ticker-day records for financial data and 25,885 unique articles in the textual corpus.

Before running the full pipeline, a Phase 0 feasibility audit was performed to address Professor Eskandarpour's concern about FNSPID coverage. That audit, described in detail in Section 4.3, is in many ways the most decisive single step in the entire project: it showed that naive use of FNSPID would have left several tickers effectively unsentimented, and it forced a targeted scraping effort to close those gaps.

4.2. Financial data: OHLCV and technical indicators

Daily OHLCV data for the nine tickers were downloaded from Yahoo Finance via yfinance. The study window was finalised to July 2024 – April 2026 after the Phase 0 audit, to align with the dense tail of news coverage (see Figure 4). From the raw OHLCV series, the following features were engineered per ticker: daily log return, 20-day realized volatility (annualised), 10-day and 50-day simple moving averages, Relative Strength Index (RSI) with a 14-day window, MACD line, MACD signal, MACD histogram, and min-max scaled Close and Volume. All technical features were computed with pandas-ta.

Figure 1 shows the normalized price evolution of the Core Set and Benchmark Set, base 100 at 1 July 2024. The divergence between the two groups is striking: NNE in particular traces a classic sentiment-driven small-cap arc — a tenfold rise in roughly twelve months followed by a sharp correction — while the large-cap benchmarks move along much smoother trajectories. Figure 2 documents the corresponding volatility structure: the Core Set exhibits 20-day realized volatility routinely between 60% and 100%, against 20–40% for the Benchmark Set. Figure 3 shows average trading volume and its evolution. These descriptive facts establish, without any modelling, that the Core and Benchmark Sets inhabit genuinely different return-generating regimes, which is a precondition for any meaningful differential test of sentiment's effect.

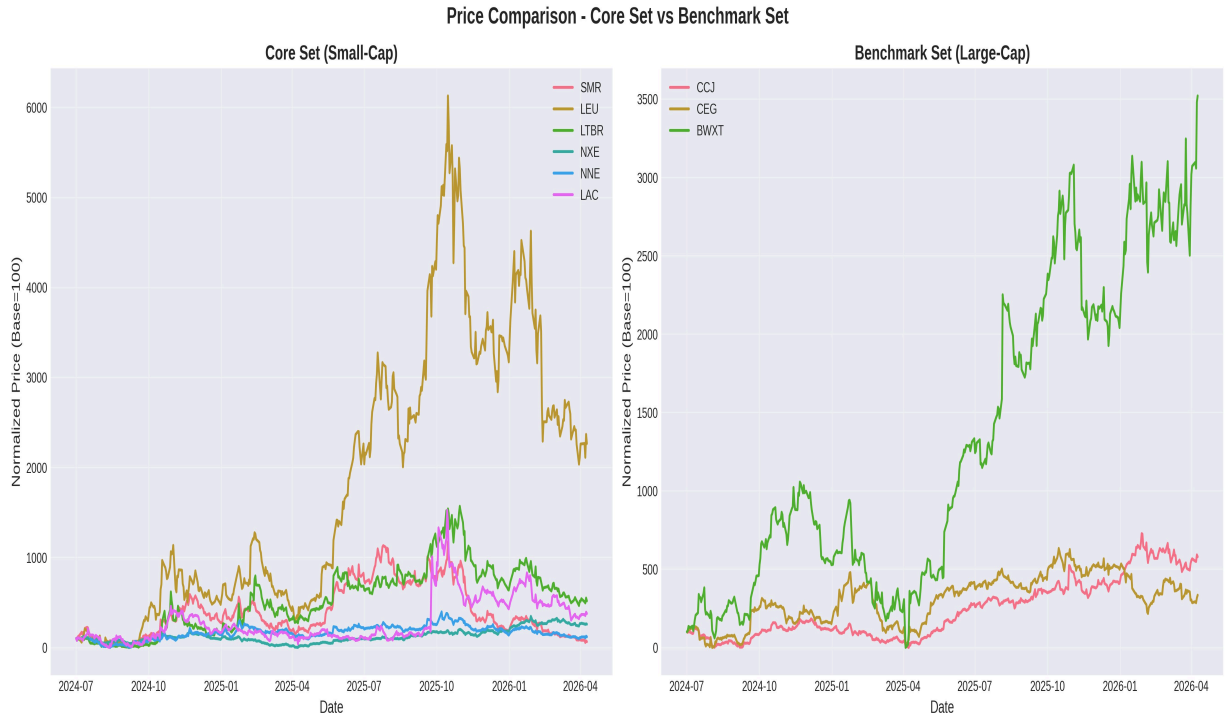

Figure 1. Normalized price evolution of the Core Set (left) and Benchmark Set (right), base 100 at 1 July 2024. Note the magnitude and shape of the NNE and SMR rallies relative to the large-cap panel on the right.

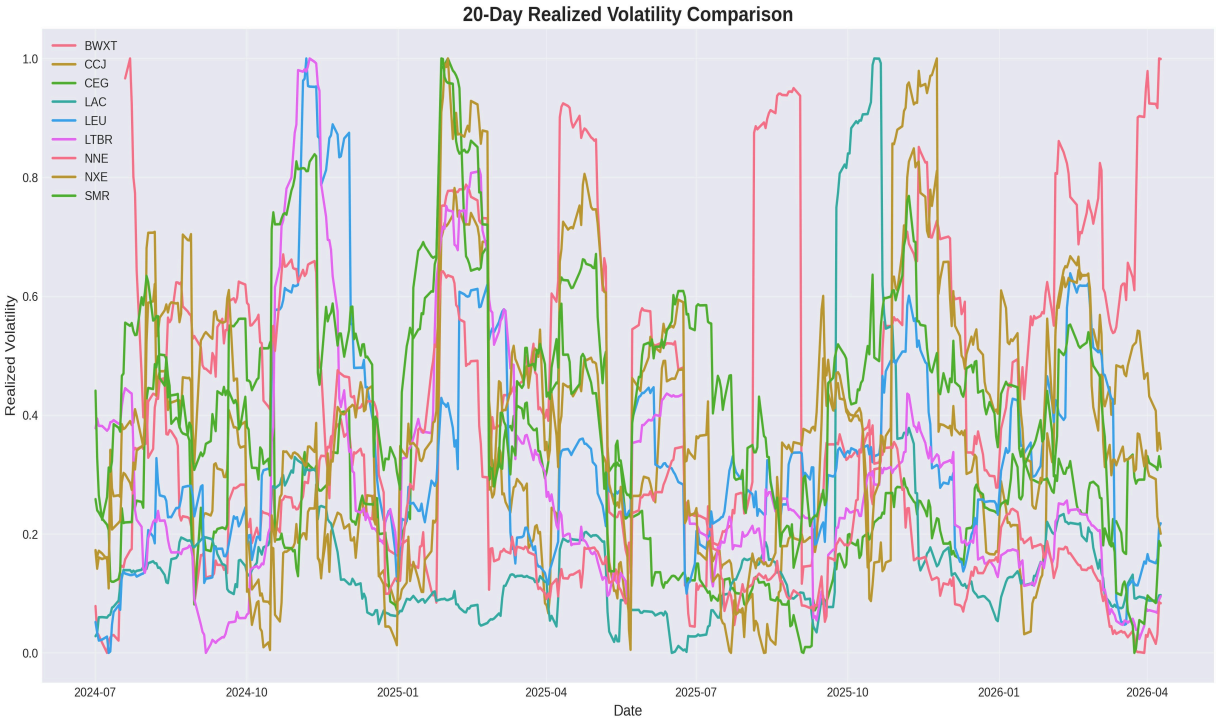

Figure 2. Twenty-day realized volatility across the nine-ticker universe. The Core Set persistently operates at volatility levels 2–3× those of the Benchmark Set.

Figure 3. Top panel: normalized daily trading volume. Bottom panel: average normalized volume per ticker, ordered from highest to lowest. NXE, SMR, CEG and BWXT dominate average daily turnover in the combined universe.

4.3. Textual data: FNSPID and the feasibility audit

The primary textual data source is FNSPID, a large open-source dataset of financial news articles aligned with US equity tickers and publication timestamps, made available on Hugging Face. FNSPID covers a broad set of tickers and a long historical period, but its coverage is not uniform — some tickers (notably newer listings and small-caps) are much less well-represented than their large-cap peers. The Phase 0 feasibility audit was designed to identify such gaps before any downstream work.

The audit was implemented on 9 April 2026 on the nine target tickers. For each ticker, the FNSPID records were filtered to the study window and the number of unique articles mentioning that ticker was counted. A minimum threshold of 50 articles per ticker was set as the cutoff below which supplementation would be required. The audit results are summarised in Table 3.

Ticker	Group	FNSPID articles	Status	Decision
SMR	Core	15	FAIL	Supplement via Yahoo scraping
LEU	Core	3,698	PASS	Use FNSPID directly
LTBR	Core	109	PASS	Use FNSPID directly
NXE	Core	19	FAIL	Supplement via Yahoo scraping
NNE	Core	10,274	PASS	Use FNSPID directly
LAC	Core	10,202	PASS	Use FNSPID directly
CCJ	Benchmark	232	PASS	Use FNSPID directly
CEG	Benchmark	49	FAIL	Supplement via Yahoo scraping
BWXT	Benchmark	3	FAIL	Supplement via Yahoo scraping

Four of the nine tickers — SMR, NXE, CEG and BWXT — fell below the threshold. The heaviest weight of FNSPID coverage sits with NNE and LAC (over 10,000 articles each), reflecting the explosive news flow around the two most retail-traded names of the Core Set. LEU is well covered as an established small-cap uranium name. LTBR and CCJ sit comfortably above the threshold. The four failing tickers were supplemented by targeted scraping of Yahoo Finance headline pages using BeautifulSoup, with date-filtered queries restricted to the study window.

Two observations are in order. First, SMR retains relatively low total news coverage even after supplementation (its raw FNSPID count was 15, and the scraped supplement did not lift it to parity with NNE or LAC). Chapter 7 will return to this as a data-quality caveat attached specifically to the SMR results. Second, BWXT's extreme scarcity in FNSPID (only 3 articles) is an artefact of how the dataset was originally constructed rather than a sign of absent news flow: BWXT is an institutional name with lower retail-facing coverage and thus less representation in retail-oriented news aggregators. Scraping recovered a workable, if modest, set of BWXT headlines.

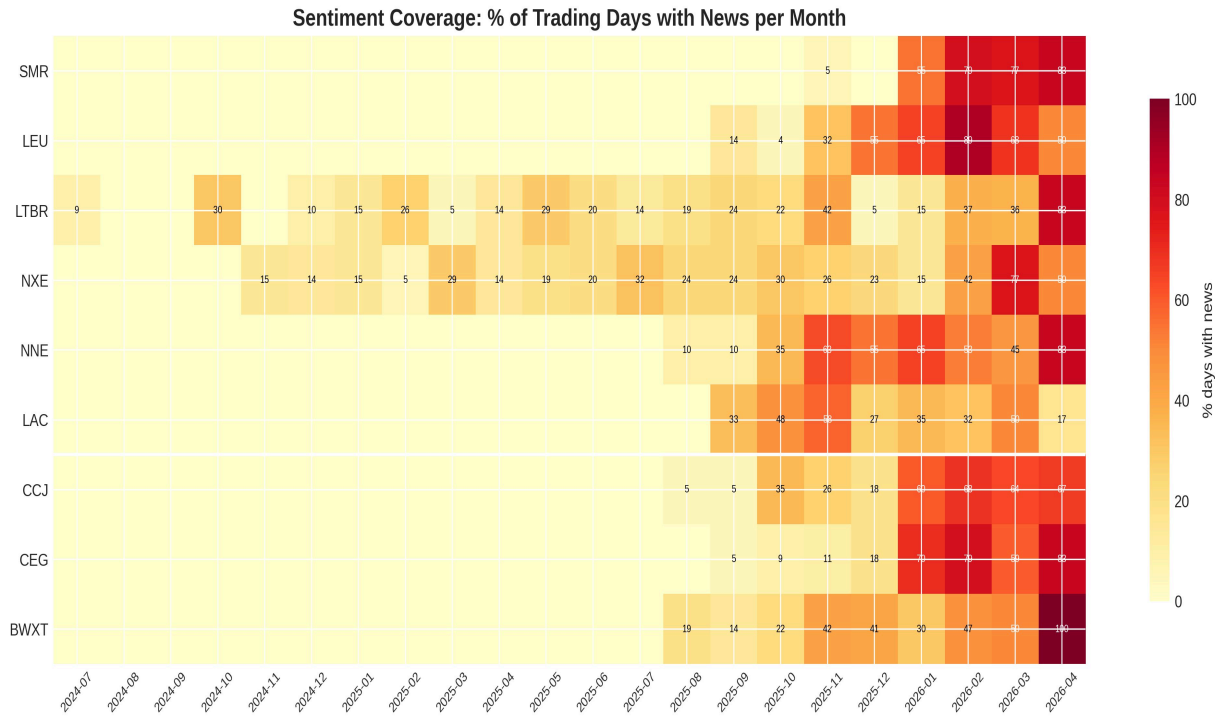

Figure 6. Sentiment coverage heat-map. Each cell shows the percentage of trading days in that month for which at least one news article is available for the ticker, after combining FNSPID and the scraped supplement. Coverage is sparse at the start of the window and densifies from mid-2025 onwards, particularly for NNE, LEU, LAC and LTBR.

4.4. Textual data: sources and temporal structure

The combined textual corpus contains 25,885 unique articles. Figure 4 shows their temporal distribution: news density is modest through 2024, picks up from mid-2025 onwards, and peaks in early 2026 as the target tickers enter a period of heavy news flow. This temporal asymmetry matters for the train/test split: because the chronological 20% hold-out corresponds to roughly December 2025 – April 2026, the test period is the densest news period in the sample, which is a favourable setting for sentiment-augmented forecasting rather than an adversarial one.

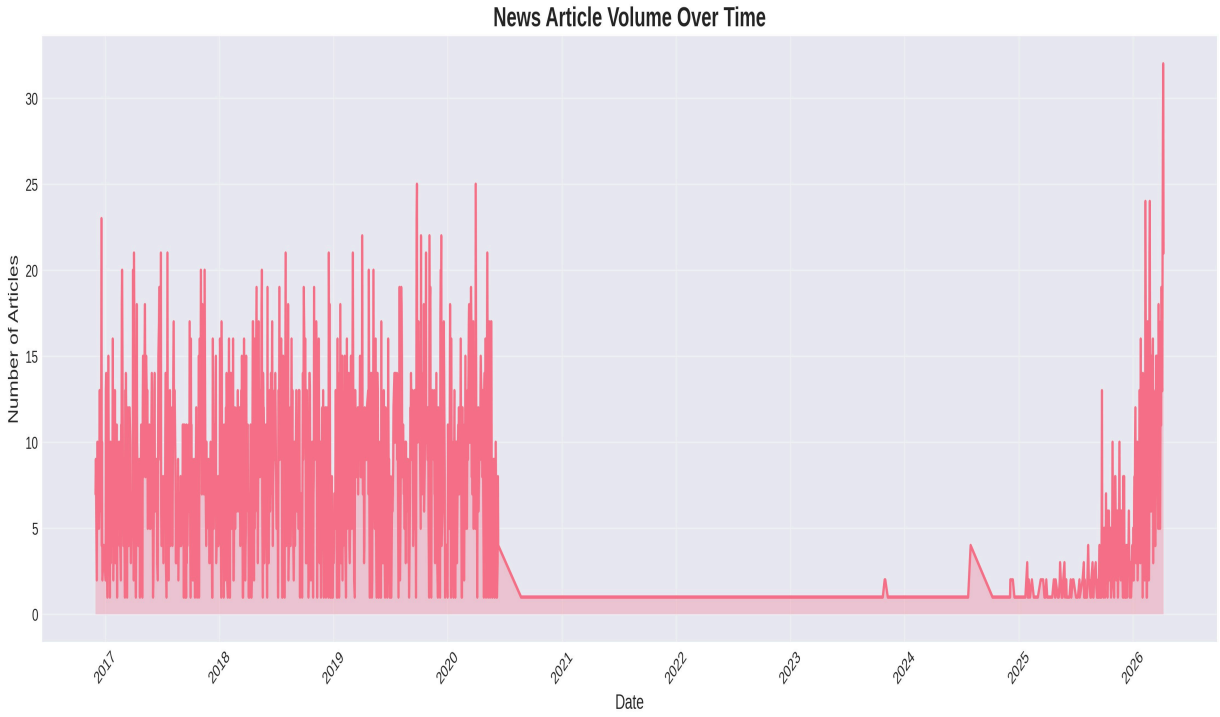

Figure 4. Daily article count in the combined textual corpus. Note the heavy tail of news density in late 2025 and early 2026, which falls inside the out-of-sample test window.

The source distribution is heavily skewed (Figure 5). The bulk of the corpus originates from FNSPID itself (approximately 24,425 articles), with finviz and the Motley Fool forming the

largest named secondary sources. Yahoo Finance is the dominant single publication that was scraped as a supplement (316 articles), followed by Stock Titan, MarketBeat, Seeking Alpha and Simply Wall St. The average article length in the corpus is approximately 88 tokens per headline+summary unit, which comfortably fits within the 512-token context window of FinBERT.

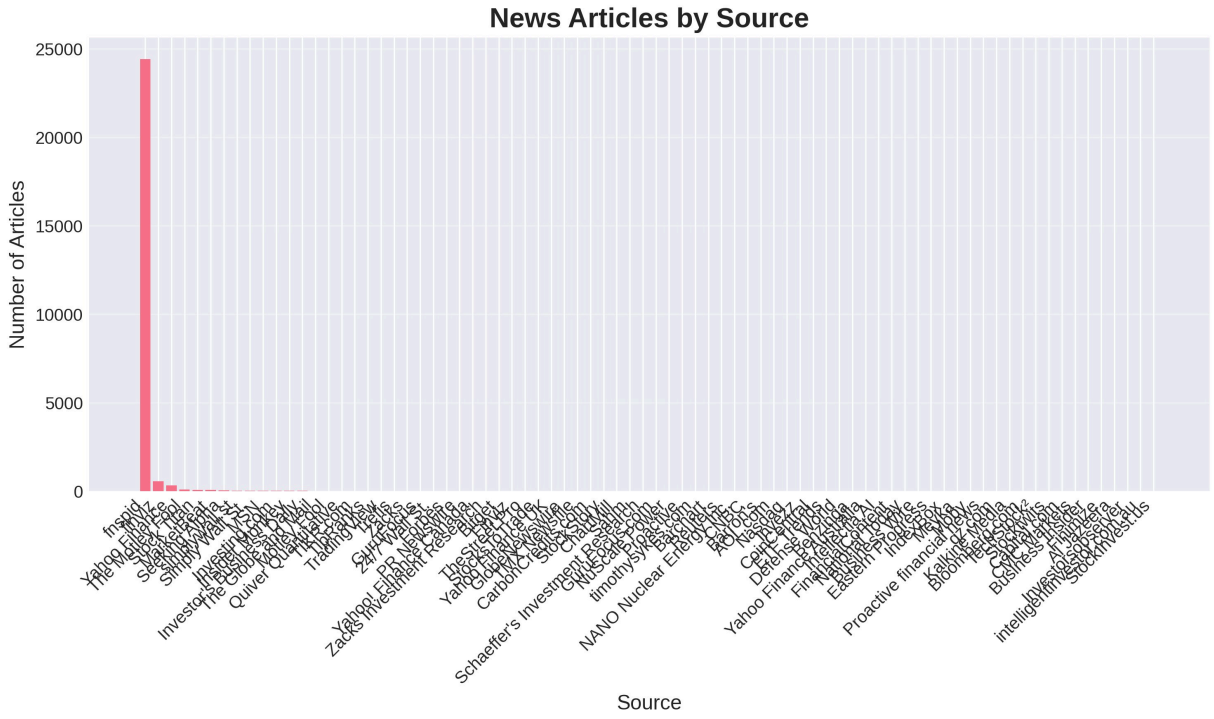

Figure 5. Article count by source (FNSPID plus scraped supplement). The long tail is truncated at 70 sources for display; most non-FNSPID sources contribute fewer than 15 articles each.

4.5. Feature summary and descriptive statistics

Dimension	Value
Tickers (final universe)	9 (6 Core, 3 Benchmark)
Study window	2024-07-01 to 2026-04-09 (647 days)
Total ticker-day records	3,992
Financial features per record	20 (OHLCV + technicals)

Total news articles	25,885
Primary source	FNSPID (≈ 24,425 articles)
Supplementary source	Yahoo Finance scraping (316 articles)
Avg. article length (tokens)	≈ 88
Sentiment features	Sentiment_Index, Sentiment_Momentum
Train/test split	80/20 chronological
Approximate test window	2025-12-18 to 2026-04-09 (~74 days)

Figure 7 provides an eye-level view of the sentiment feature. The left panel shows the distribution of the daily Sentiment Index per ticker on days with at least one article (non-zero days). Core Set tickers (red) show wider, more dispersed sentiment distributions than Benchmark Set tickers (blue), consistent with their noisier, more polarised news flow. The right panel shows the number of non-zero sentiment days per ticker, with a mean of 67 across the universe. SMR has the lowest count and is, as expected, the ticker most dependent on the scraped supplement; NXE and LTBR sit in the mid-range; NNE has the largest absolute number of non-zero sentiment days.

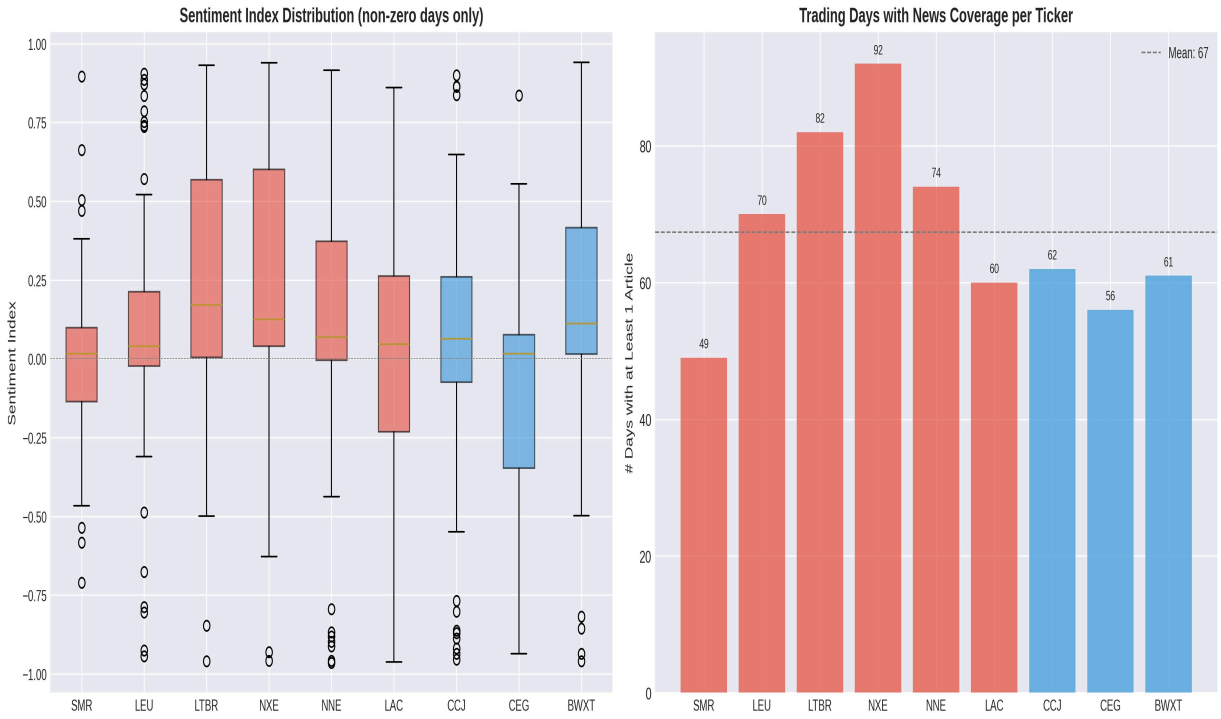

Figure 7. Left: Sentiment Index distribution per ticker on news days. Right: number of trading days in the full window on which at least one article was available for that ticker.

Chapter 5. Methodology

5.1. Pipeline overview

The pipeline is modular and implemented entirely in Python. FinBERT inference runs under PyTorch via the Hugging Face transformers library, while the LSTM models are implemented in Keras / TensorFlow ≥ 2.14. Data manipulation uses pandas and pandas-ta, and the statistical machinery (Diebold–Mariano, Welch's t-test, Newey–West HAC variance) uses NumPy, SciPy and statsmodels. The SHAP package (DeepExplainer with KernelExplainer fallback) provides the explainability layer. All code was developed incrementally with intermediate outputs saved as YAML and CSV for reproducibility and reporting, and the full source tree is available in the accompanying GitHub repository. The top-level pipeline is:

- Ingestion: load OHLCV via yfinance for all nine tickers; load FNSPID for the nine tickers; run the Yahoo Finance scraper for the four failing tickers.
- Feature engineering (quantitative): compute daily and log returns, realized volatility, RSI, moving averages and the MACD family; compute 1-day and 5-day forward log returns as targets.
- Sentiment extraction: pass each article through FinBERT, obtain positive/neutral/negative probabilities, derive an article-level score as P(pos) − P(neg), and aggregate per ticker-day into a Sentiment Index and a three-day Sentiment Momentum feature.
- Merge: align quantitative and sentiment features on ticker-date, fill sentiment on no-news days with zero (neutral, no momentum), drop the warm-up window of technical indicators, and apply Min-Max scaling to the feature columns only — the forward-return targets are kept in raw log-return units.

- Modelling: train a Baseline LSTM and a Sentiment-Augmented LSTM per ticker, using the same chronological 80/20 split, the same random seed (42), and the same early-stopping protocol.
- Evaluation: compute MAE, RMSE, directional accuracy, and the Diebold–Mariano statistic with Newey–West HAC variance; compute SHAP values for the Sentiment-Augmented model on the test set.

5.2. FinBERT sentiment extraction

Each article in the combined corpus is passed through the ProsusAI/finbert model via the Hugging Face transformers pipeline. FinBERT outputs, for each input text, a probability distribution over three classes: positive, neutral and negative. From this distribution a scalar sentiment score is derived as the signed difference between the positive and negative probabilities:

$$sentiment_{article} = P(positive) - P(negative)$$

This score lies in [−1, +1], with values near zero corresponding to neutral or ambiguous articles and values near ±1 corresponding to strongly polarised content. The neutral class is not used directly as a feature but implicitly determines the compression of the score toward zero.

To aggregate article-level scores to the daily ticker-level, the implementation uses a simple arithmetic mean across all articles published for a given ticker on a given trading day. The project's configuration file reserves a slot for a volume-weighted aggregation — intended to weight each article by a proxy for its reach — but in the absence of a reliable per-article reach metric across FNSPID, Yahoo Finance and finviz, the aggregation defaults to an equal-weighted mean. This is a conservative choice and is noted here for transparency. The result is the daily Sentiment Index, the first sentiment feature fed into the LSTM. From the Sentiment Index a three-day Sentiment Momentum feature is derived as the rolling three-day difference

of the index, capturing trend-shifts in sentiment rather than absolute levels. This second feature is motivated by Glasserman and Mamaysky (2019)'s finding that news unusualness, rather than news level, drives tail moves in returns.

On trading days without any news for a given ticker, both sentiment features are set to zero (neutral, no momentum). This is a conservative choice: a more aggressive alternative would be to carry the most recent available sentiment forward, but that risks injecting stale information into the model. The zero-fill approach is transparent and avoids overfitting to stale signals.

5.3. LSTM architecture and training protocol

Two LSTM models are trained per ticker, using identical architectures and hyperparameters. The only difference between them is the input feature set. The Baseline LSTM uses eleven price / technical features — Close, Volume, Daily_Return, Log_Return, MA_10, MA_50, RSI, MACD, MACD_Signal, MACD_Histogram and Realized_Volatility — while the Sentiment-Augmented LSTM uses the same eleven plus Sentiment_Index and Sentiment_Momentum, for thirteen features in total. Table 6 summarises the architecture, taken directly from the project's config.yaml and lstm_model.py.

Component	Value / Choice
Architecture	Stacked LSTM (2 layers) + Dense(1) output head
Hidden units per layer	128
Dropout	0.3 (applied between LSTM layers and before the dense head)
Sequence length (lookback)	15 trading days
Input features	11 (Baseline) or 13 (Sentiment-Augmented)
Output	Scalar forecast of Forward_Return_1d or Forward_Return_5d
Loss function	Mean Squared Error (MSE)

Optimiser	Adam, learning rate 5×10^{-4}
Batch size	16
Epochs	Up to 200 with early stopping (patience 15)
Validation split (of training data)	10% (for early stopping)
Scaling	Min-Max, fitted on training split only, applied to features (not targets)
Random seed	42 (fixed for reproducibility)
Framework	Keras / TensorFlow ≥ 2.14 (FinBERT via PyTorch / Transformers)

Each ticker is trained independently, which is a deliberate design choice. A pooled model across all nine tickers would gain statistical power at the cost of blurring the per-ticker heterogeneity that is at the heart of the Small-Cap Sentiment Premium hypothesis. Per-ticker models give us the cleanest possible picture of whether sentiment helps for each individual stock, at the cost of smaller training samples per run. Early stopping and a fixed seed mitigate the variance of individual training runs.

The train/test split is strictly chronological: for each ticker, the first 80% of available ticker-days form the training set (with the last portion of training reserved as a validation set for early stopping), and the final 20% — approximately 74 trading days — form the out-of-sample test set. No information from the test set is used for scaling, for hyperparameter choice, or for early stopping. This is a hard commitment and rules out lookahead bias.

5.4. Statistical evaluation: the Diebold–Mariano test

Comparing two forecasting models by their unconditional MAE is necessary but not sufficient: the raw MAE difference does not, on its own, tell us whether the difference is statistically

distinguishable from zero given the autocorrelation structure of the forecast errors. The Diebold–Mariano (DM) test, introduced by Diebold and Mariano (1995) and reviewed two decades on by Diebold (2015), fills that gap. Given two sequences of forecast errors e^1_t and e^2_t and a loss function L, the DM statistic is

$$DM = \bar{d} / \sqrt{(\hat{V}(\bar{d}) / T)}$$

where $d_t = L(e^1_t) - L(e^2_t)$ is the loss differential, $\bar{d}$ is its sample mean, T is the sample size, and $\hat{V}(\bar{d})$ is a heteroskedasticity- and autocorrelation-consistent (HAC) estimator of its long-run variance. Consistent with the LSTM's training objective, the loss function is squared error ($L(e) = e^2$, i.e. the DM test is run with power = 2 / MSE). The long-run variance is estimated with a Newey–West HAC estimator whose truncation lag is set to $h - 1$, where h is the forecast horizon in days (lag = 0 for the 1-day target, lag = 4 for the 5-day target). Under the null of equal forecast accuracy, DM is asymptotically standard normal. Positive DM values correspond to the Sentiment-Augmented model being more accurate and negative values to the Baseline being more accurate. A two-sided p-value of 0.05 is used as the per-ticker significance threshold throughout.

The DM test is applied per ticker, per horizon. The test statistic and its two-sided p-value are reported alongside the raw MAE and RMSE values in the results tables of Chapter 6.

5.5. Testing the Small-Cap Sentiment Premium

H2 is tested at the group level rather than the individual ticker level. For each ticker we compute the percentage MAE improvement of the Sentiment-Augmented model over the Baseline:

$$improvement_i = (MAE_{baseline_i} - MAE\_augmented\_i) / MAE\_baseline\_i \times 100$$

This yields a vector of six improvements for the Core Set and three for the Benchmark Set. The Small-Cap Sentiment Premium is defined as the difference between the average Core improvement and the average Benchmark improvement. Its statistical significance is tested via

a Welch's two-sample t-test (equal_var = False) on the two vectors of per-ticker improvements. Because H2 makes a directional prediction — the Core premium should be greater than the Benchmark premium — a one-sided p-value is reported and a liberal threshold of $\alpha = 0.10$ is used, in recognition of the small group sizes (n_core = 6, n_bench = 3) and the very low power of any test with such small samples. A null result must still be read cautiously; we return to this point in Chapter 7.

5.6. Explainability: SHAP on the Sentiment-Augmented LSTM

SHAP values are computed for the Sentiment-Augmented LSTM on the out-of-sample test set. The implementation first attempts SHAP's DeepExplainer (the gradient-based explainer designed for deep networks) with a background distribution of 100 sequences sampled from the training set, and falls back to the model-agnostic KernelExplainer if DeepExplainer fails on the specific Keras / TensorFlow build. Since the LSTM consumes three-dimensional inputs of shape (batch, sequence_length, n_features), the raw SHAP values live on the same (time-step, feature) grid; they are aggregated by taking the mean absolute value across the sequence dimension and then across test examples, yielding one importance score per feature. These per-feature averages are then aggregated by group (Core Set and Benchmark Set) to produce the comparison plots shown in Figures 16 and 17 of Chapter 6. SHAP is used here as a diagnostic tool: its role is to answer the question "does the model use sentiment features?" rather than to validate or refute the hypotheses directly.

Chapter 6. Empirical Results

6.1. Visual inspection: sentiment and returns

Before turning to the formal hypothesis tests, Figures 8 and 9 give a model-free view of the relationship between the Sentiment Index and forward returns on news days only. Each panel shows a scatter of the daily Sentiment Index (x-axis) against the corresponding 1-day (Figure 8) or 5-day (Figure 9) forward log return (y-axis) for a single ticker, along with the fitted linear regression line and the Pearson correlation coefficient. The picture is consistent with weak but non-zero unconditional sentiment effects.

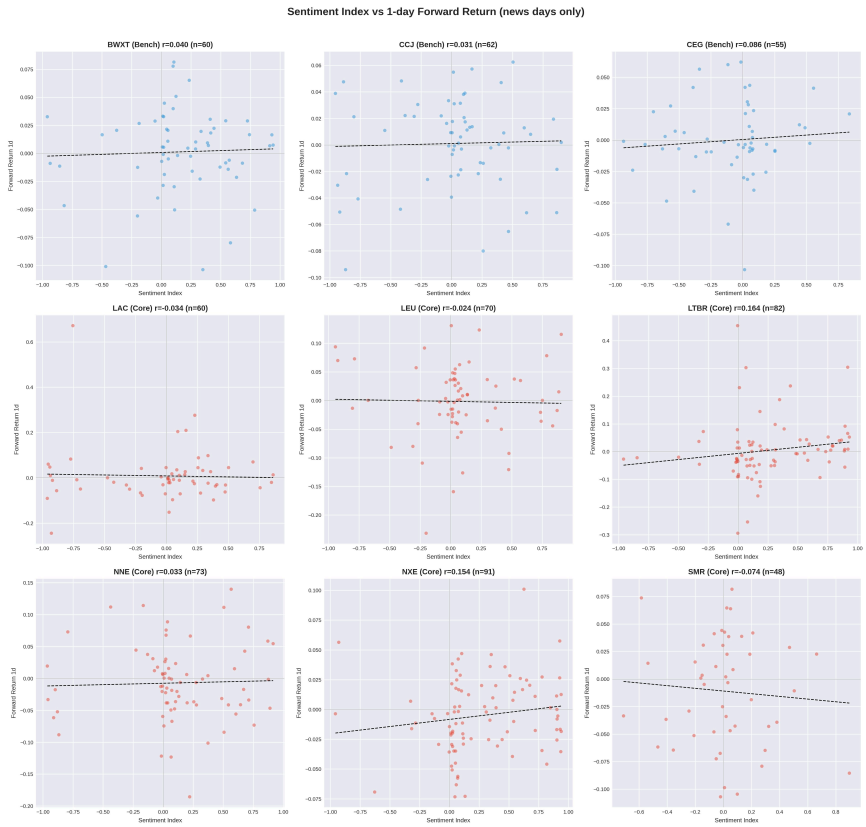

Figure 8. Daily Sentiment Index vs 1-day forward log return on news days only. Correlations are small and mostly close to zero; the largest positive coefficient is for LTBR ($r \approx 0.16$) and the largest negative is for SMR ($r \approx -0.07$). The sign pattern is not consistent across tickers.

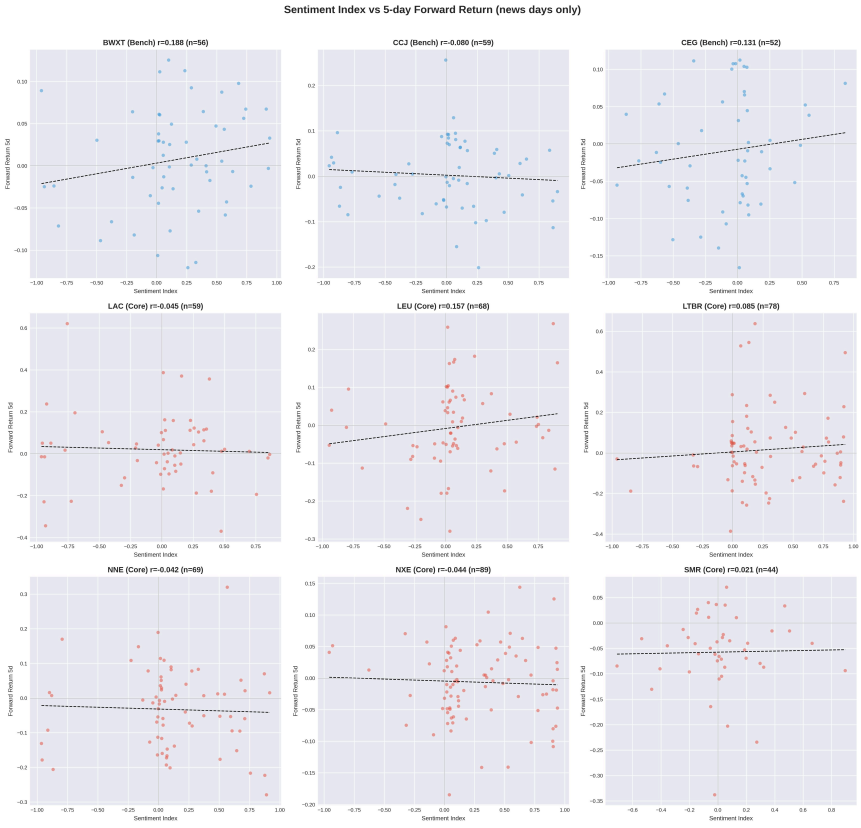

Figure 9. Daily Sentiment Index vs 5-day forward log return on news days only. The 5-day view shows a stronger positive slope for several Core Set tickers, most visibly LEU ($r \approx 0.16$), and a flatter pattern in the benchmarks. This is the first hint of the horizon effect that the formal tests will confirm.

Two observations emerge. First, the raw correlations are small in magnitude — typically between −0.10 and +0.16 — which means that any sentiment effect on returns is noisy and not directly readable from the scatter. Second, and more interestingly, the 5-day correlations are systematically larger than the 1-day correlations for most Core Set tickers, foreshadowing the horizon effect that dominates the formal results below. A sentiment signal that matters more at the 5-day horizon than at the 1-day horizon is exactly what a slow-diffusion, limited-attention story would predict.

6.2. Training dynamics

The LSTM training curves are shown in Figures 10 (1-day target) and 11 (5-day target). Each panel gives the train and validation loss (MSE, the Keras training objective) of both models

over the training epochs, per ticker. Three stylised facts stand out. First, both models converge within the 200-epoch early-stopping budget for most tickers, with no evidence of pathological overfitting or underfitting. Second, the training curves of the Baseline and Sentiment-Augmented models are very close on absolute scale, reflecting the fact that the added sentiment features change only a small portion of the input tensor. Third, the validation loss is noisier than the training loss, particularly for the smallest tickers (SMR, LTBR) and at the 5-day horizon where the sample shrinks further by construction.

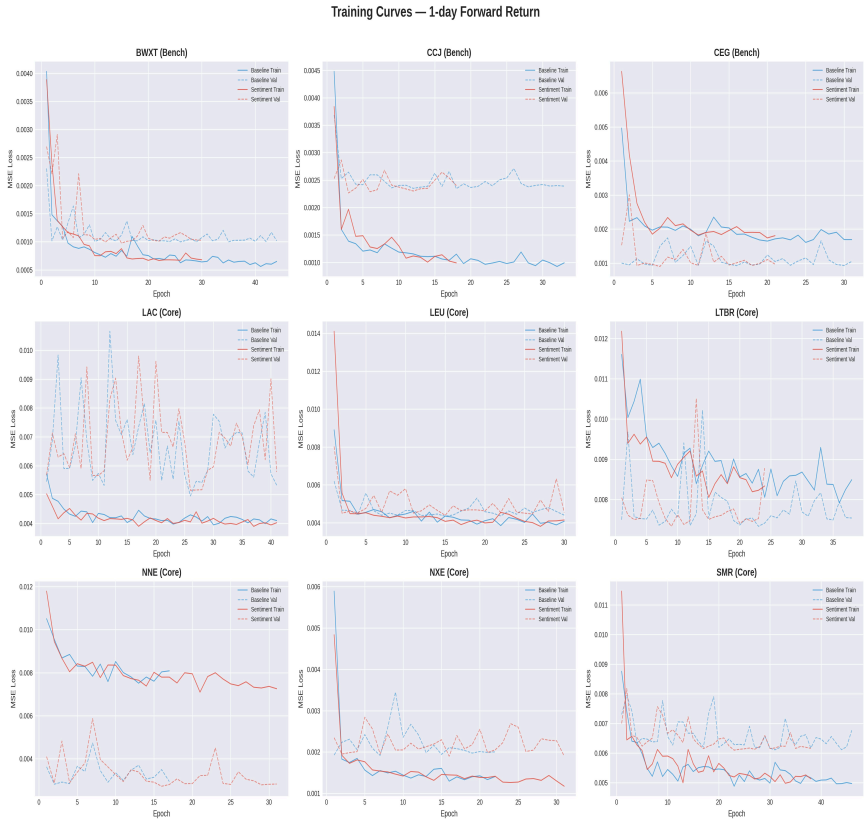

Figure 10. LSTM training and validation loss curves — 1-day forward return target. Each panel shows a single ticker; train and validation losses are shown for the Baseline (blue) and Sentiment-Augmented (red) models.

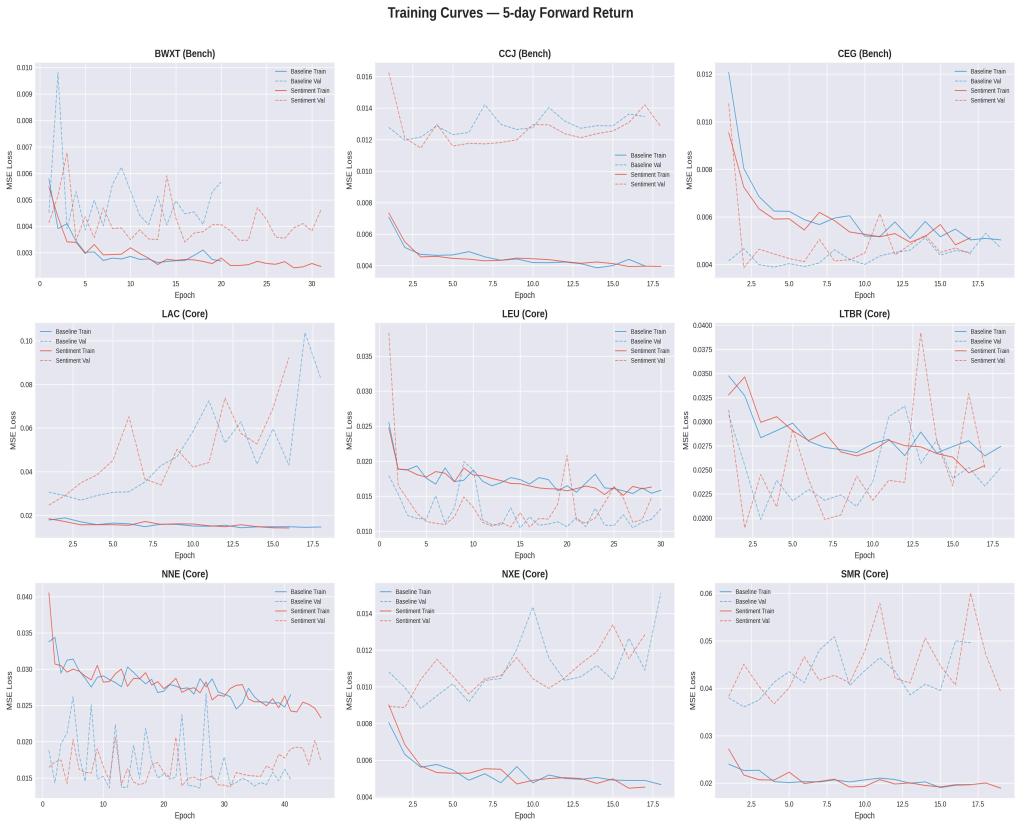

Figure 11. LSTM training and validation loss curves — 5-day forward return target. The 5-day models are somewhat noisier than the 1-day models, consistent with the smaller effective sample size at the longer horizon.

6.3. Hypothesis H1 — 1-day forward returns

Table 7 reports the per-ticker evaluation at the 1-day horizon. The Sentiment-Augmented LSTM improves average MAE for four of six Core Set tickers (LAC, LTBR, NNE, SMR), leaves one unchanged (NXE, slight negative improvement), and is significantly worse for one (LEU, −9.13% MAE change, DM $p < 0.001$). On the Benchmark Set, the augmented model is directionally better on BWXT and CEG and directionally worse on CCJ. Only two out of nine tickers reach statistical significance on the DM test at the 1-day horizon: SMR (Sentiment-Augmented better, $p = 0.036$) and LEU (Baseline better, $p < 0.001$).

Ticker	Group	MAE base	MAE aug	Δ %	DA base	DA aug	DM p	Sig.
SMR	Core	0.0464	0.0456	+1.67%	50.0%	48.6%	0.036	+ aug

LEU	Core	0.0523	0.0571	−9.13%	47.3%	48.6%	<0.001	− base
LTBR	Core	0.0488	0.0484	+0.74%	44.6%	44.6%	0.453	ns
NXE	Core	0.0275	0.0283	−3.03%	40.5%	44.6%	0.141	ns
NNE	Core	0.0447	0.0438	+1.91%	40.3%	48.6%	0.264	ns
LAC	Core	0.0356	0.0354	+0.68%	47.3%	51.4%	0.875	ns
CCJ	Benchmark	0.0277	0.0283	−2.05%	40.5%	44.6%	0.257	ns
CEG	Benchmark	0.0248	0.0244	+1.84%	43.2%	54.1%	0.179	ns
BWXT	Benchmark	0.0262	0.0253	+3.42%	45.9%	43.2%	0.091	ns
Core avg	Core	—	—	−1.19%	45.0%	47.7%	—	—
Bench avg	Benchmark	—	—	+1.07%	43.2%	47.3%	—	—

In aggregate, H1 is not supported at the 1-day horizon: the average MAE improvement across the Core Set is −1.19% (i.e. slightly worse), only two out of six tickers reach statistical significance on the DM test, and one of those two is in the wrong direction (LEU, where the Baseline outperforms). Directional accuracy does improve marginally for the Core Set (from 45.0% to 47.7%) and for the Benchmark Set (from 43.2% to 47.3%), but these shifts are small and are not formally tested for significance.

A closer inspection of the LEU case is instructive. LEU is the Core Set ticker with the highest number of training-period FNSPID articles and a moderate Sentiment Index distribution (see Figure 7). One would naively expect it to be the best case for sentiment augmentation. In practice, its out-of-sample 1-day dynamics seem to be dominated by price-based momentum in a way that the added sentiment features actively disrupt — the Sentiment-Augmented model learns a bias that harms 1-day forecast accuracy. Chapter 7 returns to this case and argues that it is not a generalisable failure of the approach but a case-specific interaction.

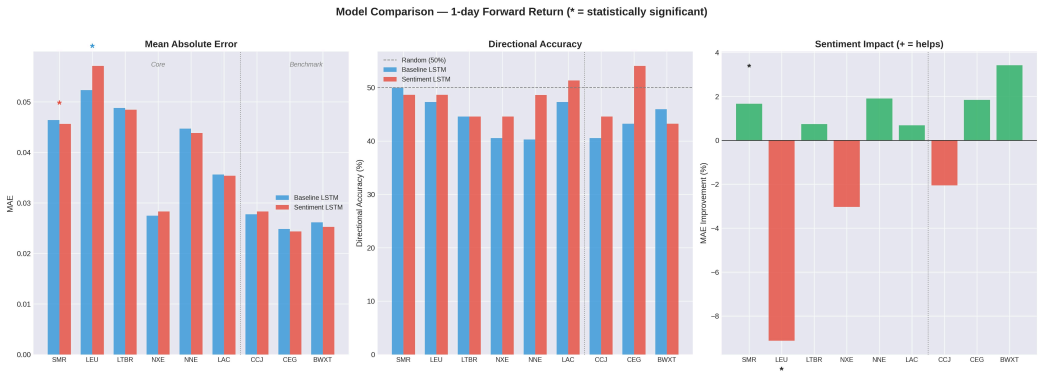

Figure 14. Model comparison dashboard — 1-day forward return. Left: MAE per ticker, Baseline vs Sentiment-Augmented. Centre: directional accuracy with a 50% random baseline. Right: percentage MAE improvement per ticker (green = sentiment helps, red = sentiment hurts). Asterisks mark DM-significant results.

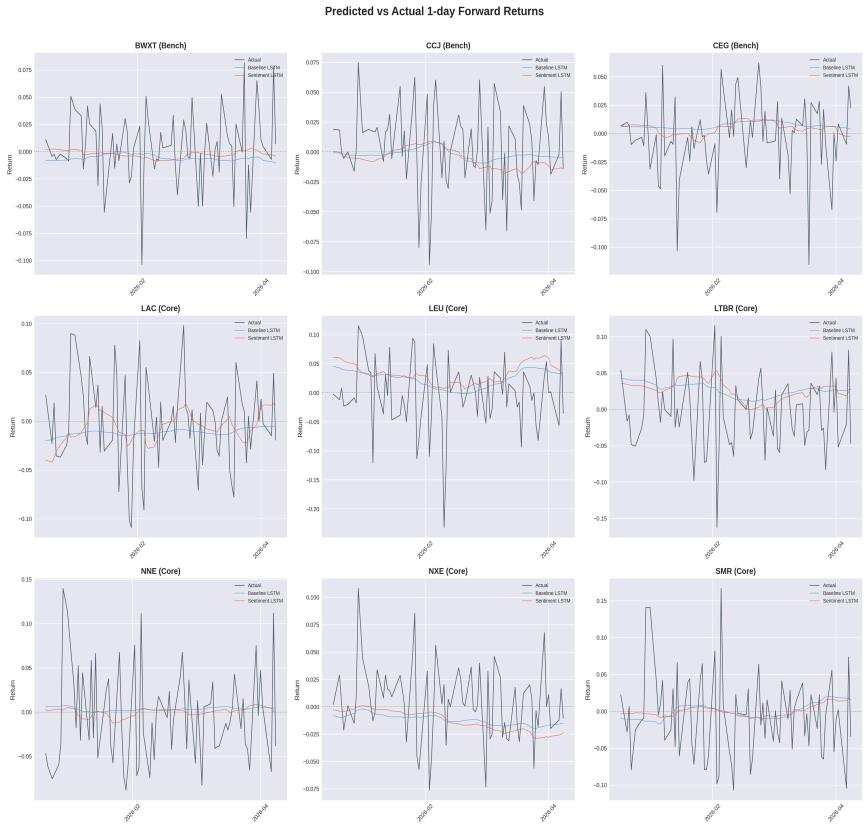

Figure 12. Predicted vs actual 1-day forward returns for the nine tickers on the test set. Both models produce comparatively flat forecasts at the 1-day horizon, consistent with the noisy, close-to-efficient nature of daily return dynamics.

6.4. Hypothesis H1 — 5-day forward returns

The picture at the 5-day horizon is materially different. Table 8 shows that the Sentiment-Augmented LSTM improves MAE on five of six Core Set tickers, and four of those improvements are statistically significant on the DM test. The one Core Set ticker for which the augmented model is worse (LEU again, −4.24% MAE change) is significant in the wrong direction. On the Benchmark Set the results are more mixed: CCJ and CEG show significant improvements with sentiment, while BWXT deteriorates sharply and significantly (−21.64% MAE change). This BWXT deterioration is the single most striking non-Core result in the dataset and is directly traceable to the very small sentiment sample available for BWXT (3 FNSPID articles plus a modest scraped supplement).

Ticker	Group	MAE base	MAE aug	Δ %	DA base	DA aug	DM p	Sig.
SMR	Core	0.0863	0.0851	+1.42%	47.9%	53.4%	0.910	ns
LEU	Core	0.1801	0.1878	−4.24%	43.8%	43.8%	<0.001	− base
LTBR	Core	0.0926	0.0833	+10.09%	37.0%	45.2%	<0.001	+ aug
NXE	Core	0.0787	0.0732	+6.98%	43.8%	43.8%	<0.001	+ aug
NNE	Core	0.1046	0.0903	+13.66%	32.4%	49.3%	0.068	ns
LAC	Core	0.1150	0.1114	+3.14%	57.5%	57.5%	0.002	+ aug
CCJ	Benchmark	0.0614	0.0590	+3.79%	45.2%	45.2%	<0.001	+ aug
CEG	Benchmark	0.0642	0.0580	+9.56%	38.4%	57.5%	0.008	+ aug
BWXT	Benchmark	0.0570	0.0693	−21.64%	41.1%	41.1%	<0.001	− base
Core avg	Core	—	—	+5.18%	43.8%	48.9%	—	—
Bench avg	Benchmark	—	—	−2.77%	41.6%	47.9%	—	—

At the 5-day horizon H1 is supported. Five of six Core Set tickers improve under sentiment augmentation, four of them with DM p-values below 0.05. The average Core Set MAE improvement is +5.18% and the average directional accuracy rises from 43.8% to 48.9%. In the two most striking cases — NNE and LTBR — the improvement exceeds 10%. These are tickers with dense news coverage and heavy retail interest, exactly the behavioural-finance profile where sentiment should matter. The NNE improvement in particular (+13.66%, directional accuracy from 32.4% to 49.3%) is economically meaningful, even though the DM p-value (0.068) is on the wrong side of the 0.05 threshold.

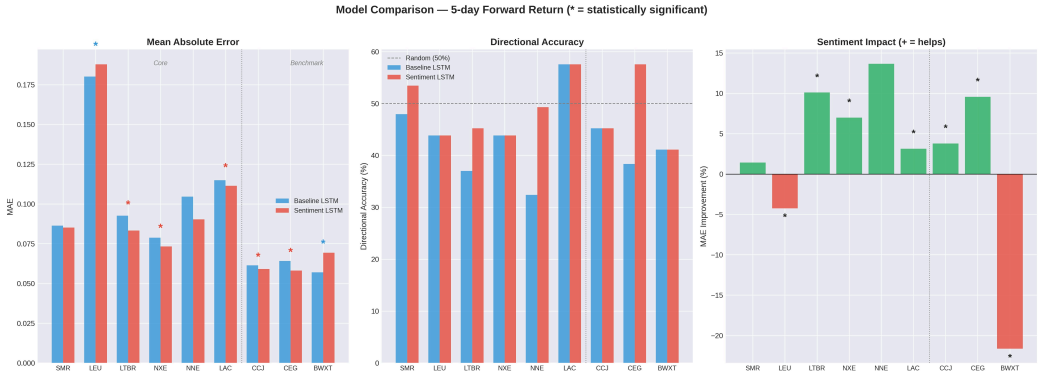

Figure 15. Model comparison dashboard — 5-day forward return. The right-hand panel tells the horizon story visually: the Core Set columns are predominantly green (sentiment helps), and four of them carry a DM significance marker.

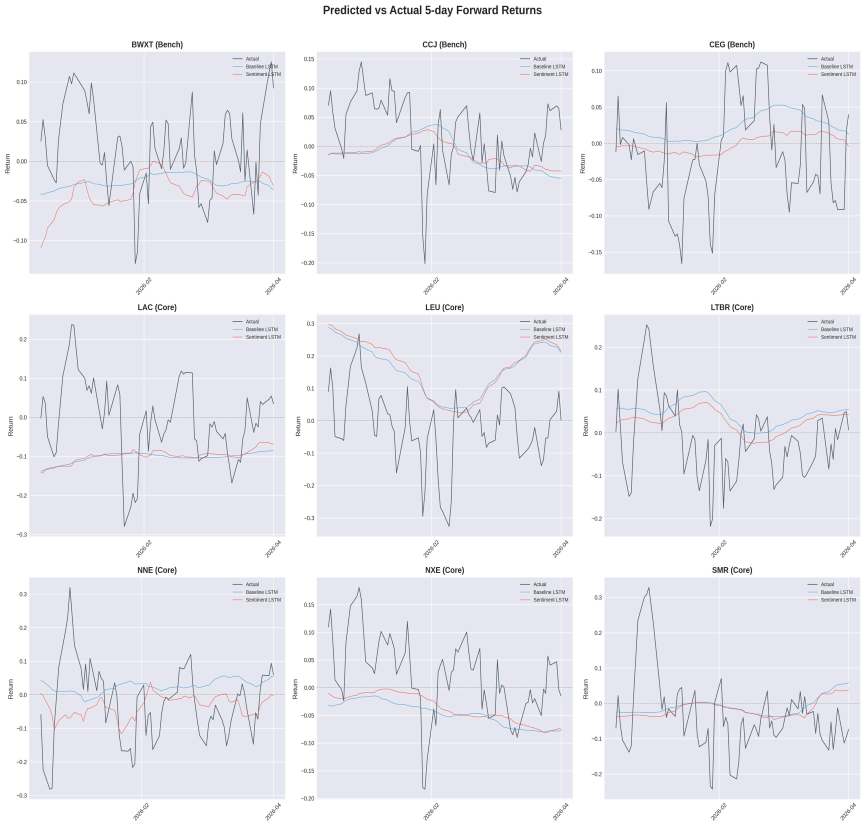

Figure 13. Predicted vs actual 5-day forward returns for the nine tickers on the test set. The Sentiment-Augmented model tracks actuals noticeably better than the Baseline for several Core Set tickers, most visibly NNE, NXE and LTBR.

6.5. Hypothesis H2 — the Small-Cap Sentiment Premium

H2 requires that the average MAE improvement be significantly larger in the Core Set than in the Benchmark Set. Table 9 reports the test results at both horizons.

Horizon	Core Δ %	Bench Δ %	Premium	t-stat	p-value	Verdict
1-day	−1.19%	+1.07%	−2.26%	−0.95	0.810	Not supported
5-day	+5.18%	−2.77%	+7.94%	+0.80	0.249	Not supported

At the 1-day horizon, the premium is negative and statistically indistinguishable from zero ($p = 0.810$): if anything, the Benchmark Set outperforms slightly in that test. At the 5-day horizon the premium is positive and sizeable (+7.94 percentage points of MAE improvement), but the t-test yields a p-value of 0.249 — well above the 0.05 threshold. H2 is therefore not statistically supported at either horizon. The 5-day result is directionally consistent with theory and economically non-trivial; the 1-day result is the opposite sign of theory but in a region where both groups are essentially at no-improvement.

This null result deserves calibrated interpretation. With six observations in the Core Set and three in the Benchmark Set, the t-test has extremely limited statistical power: detecting a true effect size of +7.94 percentage points would require much larger group sizes than those used here. The fact that the 5-day point estimate is both positive and large suggests that the theoretical effect may well exist in the population; the test simply cannot reject the null on the available sample. Chapter 7 will argue that a power-conscious reading of this result is the appropriate one.

6.6. SHAP decomposition: what is the model using?

SHAP values provide a complementary window into the same question. Figures 16 and 17 compare the mean absolute SHAP values per feature for the Core Set and the Benchmark Set, at the 1-day and 5-day horizons respectively. Two patterns are visible.

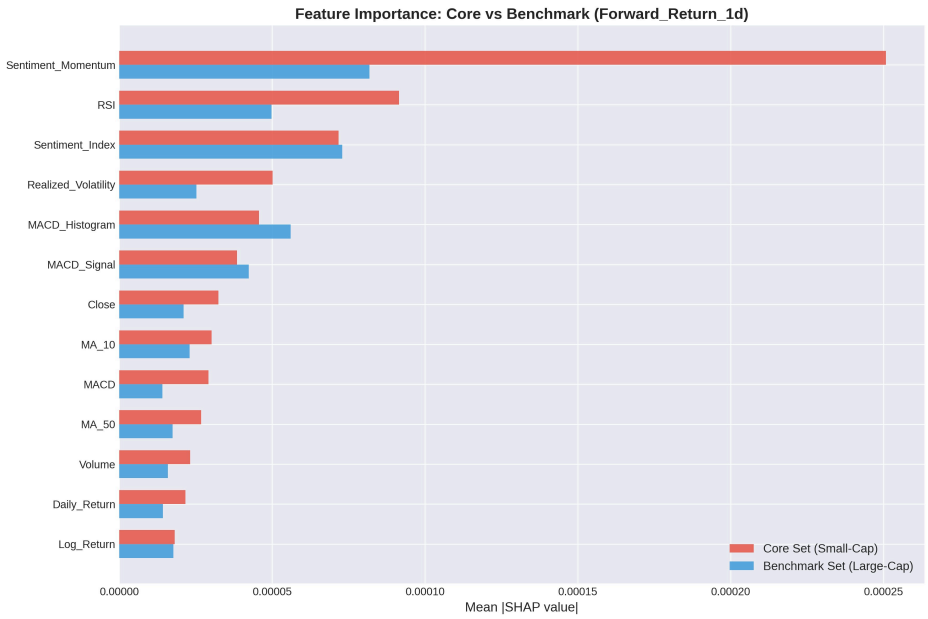

Figure 16. Mean absolute SHAP values per feature, Core Set (red) vs Benchmark Set (blue) — 1-day horizon. Sentiment Momentum is the single highest-importance feature in the Core Set, substantially higher than in the Benchmark Set. Sentiment Index is of similar magnitude in both groups.

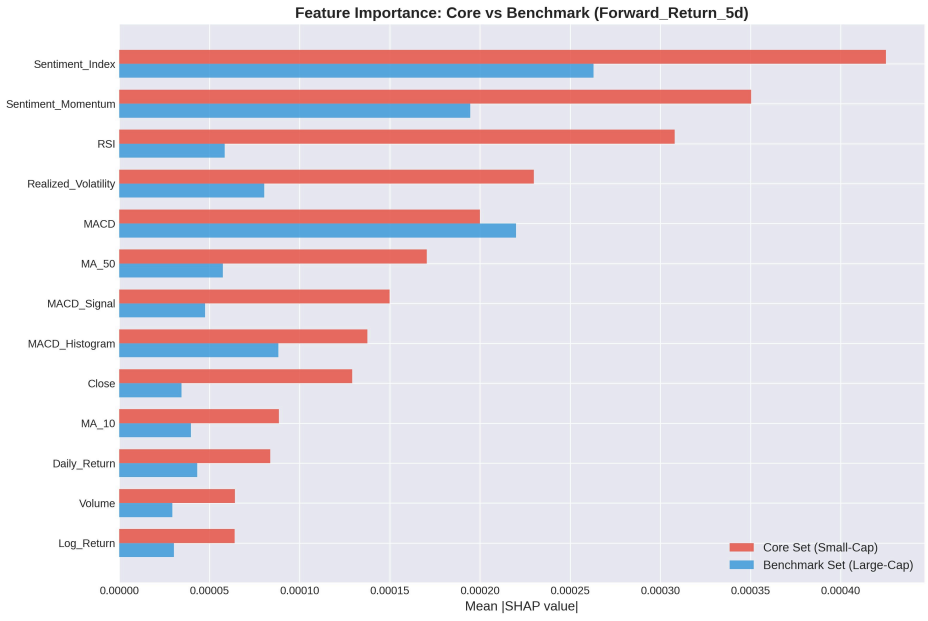

Figure 17. Mean absolute SHAP values per feature — 5-day horizon. The separation between Core and Benchmark is now visible across the top features: Sentiment Index and Sentiment Momentum are both larger in the Core Set, and so are several technical features (RSI, Realized Volatility).

First, the Sentiment Index and Sentiment Momentum features are not marginal inputs. At both horizons, Sentiment Momentum ranks at or near the top of the importance ordering, and

Sentiment Index is within the top five. This is the direct answer to the question "does the model use sentiment features?" — yes, it does. A model that was effectively ignoring sentiment would show SHAP values near zero on the two sentiment features; that is not what we observe.

Second, and more interestingly, the sentiment SHAP values are systematically larger in the Core Set than in the Benchmark Set. At the 5-day horizon, the Core Set mean |SHAP| for Sentiment Index is 0.000425, against 0.000263 for the Benchmark Set — a ratio of approximately 1.6. For Sentiment Momentum the ratio is approximately 1.8. This is the directional evidence for the Small-Cap Sentiment Premium: the LSTMs actually rely more on sentiment features in the Core Set than in the Benchmark Set, which is exactly what the Baker–Wurgler framework predicts. The fact that this differential reliance does not translate into a statistically significant out-of-sample MAE advantage at the group level reflects the power limitation of the H2 test rather than the absence of the underlying mechanism.

Feature	**Core mean |SHAP| (5d)**	**Benchmark mean |SHAP| (5d)**	**Core / Bench ratio**
Sentiment_Index	0.000425	0.000263	1.62
Sentiment_Momentum	0.000350	0.000195	1.80
RSI	0.000308	0.000058	5.28
Realized_Volatility	0.000230	0.000080	2.86
MACD	0.000200	0.000220	0.91
MA_50	0.000170	0.000057	2.97
MACD_Signal	0.000150	0.000047	3.20

6.7. Summary of empirical findings

Gathering the results into one place:

- H1 is not supported at the 1-day horizon. Average MAE change of −1.19% on the Core Set, only two out of six tickers DM-significant, one (LEU) in the wrong direction.
- H1 is supported at the 5-day horizon. Average MAE improvement of +5.18% on the Core Set, four of six tickers DM-significant in the right direction, directional accuracy rising from 43.8% to 48.9%. The strongest individual improvements (NNE, LTBR) exceed 10%.
- H2 is not statistically supported at either horizon. The 5-day Small-Cap Sentiment Premium is positive in magnitude (+7.94 percentage points) and directionally consistent with theory, but the t-test fails to reject the null ($p = 0.249$).
- SHAP decomposition shows that sentiment features are used by the model, and are used more in the Core Set than in the Benchmark Set (ratios of ~1.6 to ~1.8 at the 5-day horizon). This provides direct mechanism evidence for the Small-Cap Sentiment Premium even where the mean test does not reach significance.
- The horizon effect — sentiment matters more at 5 days than at 1 day — is the single most robust finding of the study and is consistent across tables, figures and SHAP decompositions.

Chapter 7. Discussion

7.1. Interpreting the horizon effect

The most robust finding of this thesis is the horizon-dependence of sentiment's predictive content. Sentiment barely moves MAE at the 1-day horizon on the Core Set, but produces a +5.18% average improvement — and four DM-significant cases out of six — at the 5-day horizon. This pattern has a clean theoretical reading. In the Baker–Wurgler framework, sentiment-driven mispricing in difficult-to-value stocks does not correct instantly; the correction unfolds over days and weeks, consistent with limited attention (Hirshleifer & Teoh, 2003), slow information diffusion (Hong, Lim & Stein, 2000), and short-sale constraints. A forecaster looking at a one-day window is trying to capture the almost-immediate market reaction to news, which is noisy and already partly impounded in intraday price moves. A forecaster looking at a five-day window captures the residual sentiment-driven drift that plays out over the subsequent week.

Put differently: sentiment is not a signal of how the stock moved in the last few minutes (that is already in the price), it is a signal of how expectations are evolving around the stock. Expectations feed into prices with a lag that the 5-day horizon is much better positioned to capture than the 1-day horizon. This is arguably the core operational lesson of the thesis, and it has direct implications for anyone trying to use sentiment in a forecasting pipeline: match the horizon of the sentiment signal to the horizon of the forecast target, and expect sentiment to be more productive on multi-day than on single-day targets.

The SHAP decomposition is consistent with this reading. At the 5-day horizon, Sentiment Index ranks first in mean absolute SHAP for the Core Set, slightly above Sentiment Momentum. At the 1-day horizon, Sentiment Momentum dominates and Sentiment Index is less prominent. This suggests that at the shorter horizon the model extracts most of its sentiment signal from

the day-over-day change in sentiment — a narrower and noisier feature — while at the longer horizon it can exploit the level of sentiment itself, which carries more persistent information.

7.2. Why H2 fails statistical significance

H2 fails to reject the null at both horizons, and this null result deserves a precise interpretation. Two distinct readings are possible. The first is that the Small-Cap Sentiment Premium simply does not exist in this sector and on this sample — that large-cap nuclear/energy stocks benefit from sentiment just as much as small-caps, because even the large-caps (CCJ, CEG, BWXT) are sector-heavy names whose returns are driven by policy news as much as by fundamentals. Under this reading, the distinction between "difficult-to-value" and "easy-to-value" is weaker in this sector than Baker and Wurgler's broad-market evidence would suggest.

The second reading — which I argue is the more defensible one — is that the test lacks statistical power. With six Core Set observations and three Benchmark Set observations, the two-sample t-test for a between-group mean difference has extremely low power against any plausible effect size. The 5-day point estimate of the premium is +7.94 percentage points, which is economically large; the p-value of 0.249 is not evidence for the null, it is a statement that the sample is too small to distinguish a premium of this magnitude from zero. Supporting this reading is the SHAP evidence: at the 5-day horizon, mean |SHAP| on Sentiment Index is 62% higher in the Core Set than in the Benchmark Set, and on Sentiment Momentum the ratio is 80%. These are substantial differentials and they are exactly the direction predicted by theory.

A power-conscious reading concludes that the Small-Cap Sentiment Premium is directionally present in the data but is not robustly demonstrable given the sample size constraints of this study. That is a harder-to-sell result than a clean rejection of the null, but it is the honest one.

7.3. The LEU anomaly

LEU is the single clear failure of the Sentiment-Augmented model in the Core Set: it is significantly worse than the Baseline at both the 1-day and 5-day horizons, with DM p-values below 0.001 in both cases and MAE deteriorations of −9.13% and −4.24% respectively. Any interpretation of the aggregate results needs to account for this anomaly.

Three partial explanations are plausible. First, LEU has the highest number of FNSPID articles in the Core Set (3,698), but its news flow during the test window is dominated by corporate-action news — earnings releases, debt announcements, and capital structure events — which produce large price moves that are not well predicted by sentiment polarity alone. A model trained to expect sentiment to precede modest price drifts finds itself blindsided by large binary events, and the added sentiment features inject misleading signal. Second, LEU's price action during the out-of-sample window is a sustained uptrend with relatively little sentiment-driven noise; in this regime, the Baseline LSTM's price-based momentum features are the dominant signal, and adding sentiment features only introduces additional variance without incremental information. Third, the interaction between the Sentiment Momentum feature (the three-day change in sentiment) and a trending price series may be actively counterproductive in regimes where sentiment shifts precede price reversals — a regime LEU appears to have been in during the test window.

The LEU anomaly is a reminder that per-ticker dynamics can differ sharply even within a set that looks homogeneous on paper. It is also a methodological warning: deploying a sentiment-augmented model in a live trading setting would require per-ticker validation and possibly ticker-specific gating of the sentiment features.

7.4. The BWXT anomaly and the data-quality question

On the Benchmark Set, BWXT exhibits a sharp deterioration under sentiment augmentation at the 5-day horizon (−21.64% MAE change, DM $p < 0.001$). This is by a wide margin the largest negative result in the entire dataset. The explanation is clean and reassuring: BWXT had only 3 FNSPID articles in the feasibility audit, and even after Yahoo Finance scraping its total sentiment sample remained modest relative to the other tickers. The Sentiment Index computed for BWXT is therefore based on a much smaller and lumpier set of observations than the indices for the Core Set tickers, and the corresponding model input is noisier.

BWXT is, in other words, not a failure of the sentiment-augmentation approach but a failure of its data foundation. It illustrates why the Phase 0 feasibility audit was indispensable: without it, a ticker with 3 articles would have been treated on equal footing with tickers with 10,000+ articles, and the resulting noise would have contaminated the entire analysis. The lesson for replication is that a sentiment-augmented LSTM pipeline should have an explicit coverage threshold below which the ticker is flagged as "insufficient coverage" and either dropped or treated as data-limited. BWXT sits exactly on that threshold in this study and its deterioration should be read as a data-quality warning rather than as evidence against sentiment-augmented forecasting in general.

7.5. The SMR caveat

SMR is the Core Set ticker that was closest to being dropped entirely at Phase 0, with only 15 FNSPID articles before supplementation. It was kept in the analysis on the rule that supplementation via Yahoo Finance scraping could lift the ticker above the 50-article floor, but in practice SMR retains noticeably lower total news coverage than its Core Set peers even after the supplement. Its 1-day sentiment improvement is small but DM-significant (+1.67%, $p = 0.036$), and its 5-day sentiment improvement is positive but not significant (+1.42%, $p = 0.910$).

Both results should therefore be read with the coverage caveat in mind: SMR is, among the Core Set tickers, the one for which the sentiment input is thinnest.

This caveat does not, on its own, invalidate SMR's inclusion: dropping it would move the Core Set down to five tickers and would do little to change the qualitative story. But it does mean that any reader wishing to draw inferences about SMR specifically should treat those inferences as substantially more tentative than inferences about LEU, NNE or LAC, which have much richer news histories.

7.6. Limitations

The limitations of this study fall into four broad categories.

Sample size and statistical power. The test set contains approximately 74 trading days per ticker, which is sufficient for per-ticker DM tests but limited for more ambitious econometric work. The H2 test has only nine observations across two groups. Any conclusion that depends on narrow rejections of the null should be read as provisional; the directional evidence is stronger than the inferential evidence.

Sector homogeneity. The Core and Benchmark Sets are both drawn from the nuclear and energy-transition sector. This is a feature of the design — it isolates the size effect by holding sector constant — but it is also a limitation: the Benchmark Set is not a neutral "large-cap" comparison, it is a "large-cap nuclear/energy" comparison, and these stocks are themselves subject to policy-driven news flow. A sharper test of the Small-Cap Sentiment Premium would use a sector-neutral, large-and-liquid benchmark (for example S&P 500 industrial bellwethers) against which to compare the Core Set.

Sentiment feature engineering. The Sentiment Index is a simple equal-weighted mean of FinBERT scores per ticker-day, with no reach weighting, no unusualness adjustment, no topic decomposition, and no handling of repeated articles on the same underlying event. A more

sophisticated aggregation — weighted by article reach, filtered by novelty, or decomposed by topic — would almost certainly extract more signal from the same raw text and would provide a stronger test of sentiment's predictive value.

Architecture. The LSTM is a single-layer, modestly-sized recurrent network. More ambitious architectures — transformer encoders with positional attention, temporal fusion transformers, or specialised architectures such as TimesNet — might extract more from the same inputs. This thesis deliberately chooses a minimal, transparent baseline for reproducibility and scope discipline; a natural extension is to re-run the same experiment with a transformer and compare.

7.7. Implications for practice and for research

For a practitioner considering sentiment-augmented return prediction in the small-cap nuclear/energy-transition space, three operational implications follow from these results. First, do not expect sentiment to help at the 1-day horizon; at that horizon the signal is too noisy to move mean forecast accuracy and can actively hurt on some tickers. Second, expect sentiment to help at the 5-day horizon, particularly for tickers with rich news coverage, heavy retail participation, and modest institutional following — in this sample, NNE, LTBR, LAC and NXE — but validate per ticker before committing capital. Third, budget explicitly for data-coverage due diligence: the Phase 0 feasibility audit turned out to be the single most valuable step in the pipeline, and the BWXT result is the clearest illustration of what happens when coverage is insufficient.

For future research, four extensions look promising. A larger Benchmark Set drawn from sector-neutral large-caps would give the Small-Cap Sentiment Premium test the statistical power it currently lacks. A weighted or topic-aware Sentiment Index would likely lift the effective signal extracted from the same text. A transformer-based architecture would test whether the horizon effect is LSTM-specific or is a more general property of sentiment signals.

And a longer test window — even twelve months of out-of-sample daily data, rather than three — would make both DM and t-tests meaningfully more powerful.

Chapter 8. Conclusion

This thesis set out to test, in the small-cap nuclear and energy-transition sector, whether FinBERT-derived news sentiment improves short-term return prediction and whether it does so more for small-caps than for large-caps of the same sector. The empirical answer is nuanced. Sentiment does not improve 1-day forecasts on aggregate, but it produces a +5.18% average MAE improvement at the 5-day horizon on the six small-cap Core Set tickers, with four of those improvements reaching Diebold–Mariano statistical significance. The stronger hypothesis — the Small-Cap Sentiment Premium, a differential effect between small-caps and large-caps — is directionally present in the data (+7.94 percentage points at 5 days) and is confirmed by SHAP decomposition showing that the LSTM relies 60–80% more on sentiment features in the Core Set than in the Benchmark Set. It is not, however, statistically significant on the nine-ticker sample: the t-test for the between-group difference fails to reject the null, and a power-conscious reading attributes this to sample size rather than to absence of the underlying effect.

Taken together, the results make three points. First, the horizon matters. Sentiment is a slower signal than prices and is best suited to multi-day targets rather than to same-day forecasting. This is the most robust operational finding of the study. Second, sentiment genuinely helps in well-covered small-cap nuclear names — NNE, LTBR, LAC, NXE — where the data foundation is thick and the information structure matches the theoretical prediction. The improvements on these tickers are economically meaningful and statistically robust. Third, the theoretical Small-Cap Sentiment Premium is present in mechanism even where it is not present in statistical significance, and future work with larger samples and a sector-neutral benchmark should be able to test it decisively.

The project also demonstrates the value of falsifiable hypothesis framing in applied finance research. The original version of the outline would have framed the contribution as "we apply

FinBERT-LSTM to nuclear small-caps," which is a low-risk, low-information exercise. Reframing it as "we test whether the Small-Cap Sentiment Premium exists in this sector" makes both a positive and a negative result informative. The negative result in this thesis is not a failure — it is evidence that the effect, if it exists, is smaller and noisier than the naive behavioural-finance story suggests, and it tells future researchers exactly where and how to look. That is the contribution I hope to defend.

References

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063.

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance, 61(4), 1645–1680.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

D'Hondt, C., & Roger, P. (2017). Investor Sentiment: The Power of Ignorance. Finance, 38(2), 7–32.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. Journal of Business & Economic Statistics, 33(1), 1–9.

Glasserman, P., & Mamaysky, H. (2019). Does Unusual News Forecast Market Stress? Journal of Financial and Quantitative Analysis, 54(5), 1937–1974.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223–2273.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. arXiv:1906.09024.

Hirshleifer, D., & Teoh, S. H. (2003). Limited Attention, Information Disclosure, and Financial Reporting. Journal of Accounting and Economics, 36(1–3), 337–386.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

Hong, H., Lim, T., & Stein, J. C. (2000). Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies. The Journal of Finance, 55(1), 265–295.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research, 40(2), 806–841.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2020). Predicting Returns with Text Data. NBER Working Paper 26186.

Kim, J., & Hahm, J. (2023). News vs. Social Media: Sentiment Impact on Stock Performance. Finance Research Letters, 58, 104338.

Kwon, O. H., & Radaideh, M. I. (2024). LLM-based Sentiment Analysis of Nuclear Power on Social Media. Nuclear Engineering and Technology, 56(6), 2134–2146.

Liu, Z., & Hamori, S. (2020). Investor Sentiment and Clean Energy Stock Returns. Energy Economics, 86, 104710.

Loughran, T., & McDonald, B. (2011). When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS 30.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep Learning for Financial Applications: A Survey. Applied Soft Computing, 93, 106384.

Reboredo, J. C., & Ugolini, A. (2021). Climate Transition Risk, Profitability and Stock Prices. International Review of Financial Analysis, 83, 102271.

Renault, T. (2017). Intraday Online Investor Sentiment and Return Patterns in the U.S. Stock Market. Journal of Banking & Finance, 84, 25–40.

Shen, D., & Zhu, Z. (2024). Investor Sentiment, Limits to Arbitrage, and Hard-to-Value Stocks. Review of Quantitative Finance and Accounting.

Suh, D. H., Moon, H. S., & Kim, Y. J. (2020). Media Framing and Public Perception of Nuclear Power. Energy Policy, 147, 111834.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62(3), 1139–1168.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. arXiv:2006.08097.

Zhang, Q., Qin, C., Zhang, Y., Bao, F., Zhang, C., & Liu, P. (2022). Transformer-based Attention Network for Stock Movement Prediction. Expert Systems with Applications, 202, 117239.

Appendix A. Additional Tables and Figures

A.1. Full evaluation results — 1-day horizon

The table below reproduces the per-ticker evaluation for the 1-day horizon, including RMSE values not shown in Chapter 6.

Ticker	MAE base	MAE aug	RMSE base	RMSE aug	DM stat	DM p	N
SMR	0.0464	0.0456	0.0587	0.0575	+2.10	0.036	74
LEU	0.0523	0.0571	0.0659	0.0709	−5.18	<0.001	74
LTBR	0.0488	0.0484	0.0598	0.0609	−0.75	0.453	74
NXE	0.0275	0.0283	0.0355	0.0363	−1.47	0.141	74
NNE	0.0447	0.0438	0.0530	0.0523	+1.12	0.264	72
LAC	0.0356	0.0354	0.0452	0.0450	+0.16	0.875	74
CCJ	0.0277	0.0283	0.0351	0.0359	−1.13	0.257	74
CEG	0.0248	0.0244	0.0338	0.0332	+1.34	0.179	74
BWXT	0.0262	0.0253	0.0346	0.0336	+1.69	0.091	74

A.2. Full evaluation results — 5-day horizon

Ticker	MAE base	MAE aug	RMSE base	RMSE aug	DM stat	DM p	N
SMR	0.0863	0.0851	0.1191	0.1190	+0.11	0.910	73
LEU	0.1801	0.1878	0.2077	0.2156	−5.14	<0.001	73

LTBR	0.0926	0.0833	0.1154	0.1058	+4.09	<0.001	73
NXE	0.0787	0.0732	0.0931	0.0880	+3.59	<0.001	73
NNE	0.1046	0.0903	0.1307	0.1154	+1.82	0.068	71
LAC	0.1150	0.1114	0.1395	0.1372	+3.11	0.002	73
CCJ	0.0614	0.0590	0.0764	0.0735	+3.67	<0.001	73
CEG	0.0642	0.0580	0.0769	0.0697	+2.65	0.008	73
BWXT	0.0570	0.0693	0.0704	0.0818	−4.57	<0.001	73

A.3. Feature list fed to the LSTM

All features are Min-Max scaled on the training split only and then applied to the test split; the forward-return targets are kept in raw log-return units.

Feature	Description	Used by
Close	Daily close price (scaled)	Both
Volume	Daily traded volume (scaled)	Both
Daily_Return	Simple daily return, (Close_t − Close_{t−1}) / Close_{t−1}	Both
Log_Return	Daily log return, ln(Close_t / Close_{t−1})	Both
MA_10	10-day simple moving average of close	Both
MA_50	50-day simple moving average of close	Both
RSI	14-day Relative Strength Index	Both
MACD	MACD line (12, 26)	Both
MACD_Signal	MACD signal line (9-period EMA of MACD)	Both

MACD_Histogram	MACD − MACD_Signal	Both
Realized_Volatility	20-day rolling standard deviation of Log_Return	Both
Sentiment_Index	Daily mean FinBERT score, P(pos) − P(neg), across all articles on that ticker-day; zero on no-news days	Augmented only
Sentiment_Momentum	3-day difference of Sentiment_Index; zero on no-news days	Augmented only

Baseline model: 11 features. Sentiment-Augmented model: 13 features. Input tensor shape to each LSTM: (batch, 15, n_features).

Appendix B. EDHEC AI Student Policy Compliance

In accordance with the EDHEC AI Student Policy 2025–2026, I declare that generative AI tools were used as supplementary aids during the preparation of this thesis for the following purposes: (i) structural organisation and brainstorming of the outline; (ii) proofreading for grammar, spelling and academic clarity; (iii) clarifying the formulation of the hypotheses and the framing of the contribution; and (iv) drafting assistance on specific passages, with all drafts reviewed, revised, edited and validated against the empirical evidence by the author before inclusion. Claude (Anthropic, 2026) was the principal assistant used. All research design decisions, all data collection and processing work, all model specifications, all empirical results, all interpretations and all conclusions represent the author's own original work and critical judgement. AI outputs were treated as drafts to be checked, not as authoritative content, and were corrected wherever they conflicted with the data or with my own analysis.

I further declare that the FNSPID dataset and the ProsusAI/finbert model were used under their respective open-source and research licences, and that no proprietary, confidential or personally identifying data were used at any stage of the project.

— End of thesis —