Sentiment-Enhanced Forecasting: A Comprehensive Framework for Integrating Market Sentiment into Financial Prediction Models

Abstract

This study proposes a unified framework for integrating market sentiment into financial forecasting models and evaluates it using a benchmark stock–news dataset. Using the FNSPID Financial News Stock Price Integrated Dataset as a standardized testbed [1], we construct **transformer-based sentiment indices** from financial news and examine their contribution to stock-return forecasting. Our framework combines FinBERT and related financial transformer models into a composite sentiment index, then compares deep learning models estimated **with and without sentiment features**.

In this Phase-1 implementation, we focus on Tesla (TSLA) as a case study and develop daily sentiment series aligned with TSLA returns. Price-only baselines are estimated with LSTM, GRU, and simple transformer architectures, and are contrasted with sentiment-augmented variants. We evaluate performance using statistical forecast metrics and formal model-comparison tests, and translate improvements into economic terms via a transparent long–flat trading strategy.

Empirically, we find that **sentiment-enhanced models based on transformer-derived indices can improve directional accuracy and risk-adjusted performance relative to price-only baselines**, although gains vary across architectures and horizons. The framework is explicitly designed to extend to **multi-equity analysis** on the full FNSPID universe; the TSLA results reported here form the first empirical step of a broader research agenda. By benchmarking against FNSPID and addressing methodological limitations of the original study, this work aims to provide a more rigorous and transparent route for future research on sentiment-driven forecasting.

Keywords— Financial sentiment, transformer-based sentiment index, FinBERT, FNSPID, stock price prediction, LSTM, GRU, transformer, deep learning, TSLA.

I. Introduction

Textual information from news, earnings calls, and social media has become a crucial input for modern financial decision-making. A large body of work documents that investor sentiment and media tone contain predictive information about short-horizon

stock returns and volatility beyond traditional price and fundamental indicators [3]–[6]. Simultaneously, advances in natural language processing (NLP) and deep learning have made it possible to extract rich sentiment signals from large unstructured corpora and feed them into nonlinear forecasting models [2], [7], [10].

Despite this progress, there is still no universally accepted **framework** for integrating sentiment into forecasting models in a statistically rigorous, reproducible, and economically interpretable manner. Many studies rely on proprietary datasets, ad-hoc sentiment indices, or loosely specified modeling choices that are difficult to replicate or compare [3], [4]. Moreover, existing benchmark datasets that combine news and prices are relatively new and underexplored. For example, the FNSPID dataset [1] provides a valuable integrated resource but leaves several methodological questions open, including how to construct robust sentiment indices, benchmark alternative sentiment models, and evaluate the economic value of sentiment-enhanced forecasts.

This study addresses these gaps in three ways. First, we define a **modular framework** for sentiment-enhanced forecasting that explicitly separates (i) data collection and preprocessing, (ii) sentiment scoring, (iii) forecasting model design, (iv) forecast evaluation, and (v) economic interpretation of the results. Second, we implemented this framework using the FNSPID dataset [1] and FinBERT sentiment modeling [2] in a TSLA-focused case study with transparent, reproducible code. Third, we compare price-only and sentiment-augmented versions of several deep learning models (LSTM, GRU, and a simple transformer) [7], [10] and evaluate their performance using both statistical metrics and a stylized trading strategy.

Our guiding hypothesis is as follows:

H1: A properly constructed **transformer-based sentiment index** derived from financial news improves the predictive performance of stock-return forecasting models relative to models using price information alone and outperforms existing state-of-the-art sentiment proxies (e.g., generic ChatGPT - style sentiment scores) when evaluated on a standardized benchmark dataset.

The remainder of this paper is organized as follows. Section II reviews related studies on financial sentiment, transformer models, and deep learning forecasting. Section III introduces the data and benchmark designs. Section IV describes the proposed framework and the modeling methodology. Section V presents the empirical results of the TSLA case study. Section VI discusses the implications and limitations of the study. Section VII concludes the paper and outlines future work.

II. Related Work

A. Financial Sentiment and Market Predictability

Early empirical studies have shown that media tone and investor sentiment can predict market returns, volatility, and trading volume. Tetlock [5] documents that daily pessimism in media content predicts temporary downward pressure on stock prices, followed by reversal. Bollen et al. [6] demonstrate that aggregate Twitter mood indices can predict changes in the Dow Jones Industrial Average. Subsequent studies have expanded sentiment sources to include news articles, analyst reports, and conference call transcripts, as surveyed in [3], [4].

These findings motivate the construction of sentiment indices that summarize the tone of textual information and can be used as explanatory variables in the forecasting models. However, many early indices were based on dictionary methods or handcrafted features, which can be fragile in financial domains [3].

B. Domain-Specific NLP and FinBERT

With the advent of transformer models, domain-specific language models have become state-of-the-art for financial sentiment analysis [2], [3]. Araci [2] introduced **FinBERT**, a BERT model fine-tuned on financial text, which significantly outperformed generic sentiment tools such as VADER or TextBlob on financial classification tasks. Follow-up work has extended FinBERT-style architectures beyond generic news sentiment to more specialized tasks such as earnings-call tone classification and risk-disclosure analysis [3,10]. These models move beyond simple positive/negative polarity and provide more granular measures of managerial optimism, risk, and uncertainty that are highly relevant for return prediction. These models move beyond simple positive/negative polarity and provide more granular views of managerial optimism, risk, and uncertainty that are highly relevant for return prediction.

Motivated by this literature, we treat FinnBERT as the core building block of a **composite transformer-based sentiment metric.** In our framework, news articles are first scored with FinnBERT for polarity, and the design explicitly allows additional transformers-based components-such as earring-call tone or risk-disclosure scores-to be standardized and aggregated into the same daily sentiment index. Using FinBERT as the main sentiment engine therefore leverages domain adaptation while supporting a clean comparison between (i) price-only models, (ii) simple FinBERT-based sentiment models and (iii) richer composite transformer based sentiment indices in our forecasting experiments.

C. Deep Learning for Financial Time Series Forecasting

Deep learning models such as LSTMs, GRUs, CNNs, and transformers have been widely applied to financial time series data [7], [10]. These models can capture nonlinear dependencies and long-range temporal patterns that are difficult for traditional models, such as ARIMA or GARCH, to learn. Yang et al. [10] provide a systematic review of deep learning for financial forecasting from 2005–2022, noting that hybrid architectures combining price, technical indicators, and exogenous features (including sentiment) often yield improved prediction accuracy.

Incorporating sentiment features into deep learning models has been explored in several studies [7], [10], but many rely on proprietary datasets or do not benchmark against standardized news–price datasets. Our work contributes to this literature in two ways: (i) we provide a transparent FNSPID-based case study with clearly specified price-only and sentiment-augmented deep learning baselines, and (ii) where code from related studies is available, we run their original implementations on the FNSPID dataset to obtain directly comparable benchmark results.

D. Benchmark Datasets and the FNSPID Study

The FNSPID dataset [1] integrates financial news and stock prices across multiple U.S. equities and proposes a baseline research design for evaluating sentiment-driven forecasting. While the original study illustrates the potential of large-scale news–price integration, it leaves several aspects under-specified, including the construction of a robust sentiment index and statistical testing of forecast improvements. In particular, the sentiment scoring step relies heavily on prompting a general-purpose language model over a narrow rating scale with limited diagnostic analysis.

Our framework builds directly on the FNSPID dataset [1] but diverges methodologically: we (i) replace ad-hoc sentiment prompting with FinBERT-based scoring [2], (ii) construct daily sentiment indices using transparent aggregation choices, (iii) systematically compare price-only and sentiment-augmented models, and (iv) evaluate improvements using standard forecast comparison tests [8] and economic performance metrics [9].

III. Data and Benchmark Design

A. FNSPID Dataset

The **FNSPID** dataset integrates financial news and stock prices for a broad set of equities and provides multiple pre-processed CSV files containing article metadata, text content, sentiment labels, and aligned price series [1]. Each observation can be viewed as a news article linked to a specific stock and date, with corresponding open, high, low, close, and volume information.

For this Phase-1 study, we restrict attention to **Tesla (TSLA)**, which is a heavily covered technology stock with frequent news events. Subsetting to TSLA allows us to focus on validating the framework and sentiment index design before generalizing to multiple tickers.

B. TSLA Subset and Time Horizon

We extract all TSLA-related news articles from the FNSPID news tables [1] using ticker symbols and keyword filters, then align them with daily TSLA price data. To balance coverage and recency, we focus on a multi-year window that includes both high-volatility and more stable periods, ensuring that the models are tested across varying regimes.

News articles are timestamped and linked to the trading day on which they would reasonably be incorporated into prices (e.g., same-day or next-day mapping depending on publication time). We standardize timestamps to a daily frequency to match the FNSPID convention [1].

C. Baseline Benchmark Design

We define three levels of benchmark:

1. **Dataset benchmark** – Use FNSPID's integrated news–price structure [1] so that results can be compared to other studies using the same data.
2. **Model benchmark** – Establish price-only forecasting models (no sentiment) as baselines.
3. **Sentiment benchmark** – Introduce a FinBERT-based sentiment index [2] as the main enhancement and optionally compare with alternative sentiment proxies (e.g., a simple polarity dictionary or ChatGPT-style scores).

This design makes it possible to attribute performance gains specifically to sentiment features and to document where the proposed approach improves on the original FNSPID study [1].

IV. Methodology: A Modular Sentiment-Enhanced Forecasting Framework

Our framework is structured into five modules: (A) data engineering, (B) sentiment scoring, (C) forecasting models, (D) evaluation metrics and statistical tests, and (E) economic interpretation. The TSLA case study instantiates this framework concretely, but the design is intended to be reusable.

A. Data Engineering and Aggregation

1. **News cleaning** – We remove non-informative boilerplate, filter out extremely short items, and deduplicate near-identical articles. Basic quality filters include minimum text length and valid ticker association.
2. **Price alignment** – Daily OHLCV data for TSLA are loaded from the FNSPID price tables [1]. We ensure that the date column is consistent across news and price files.
3. **Article-level features** – For each article, we retain text, publication date, ticker, source, and any pre-existing labels in FNSPID for diagnostics.

B. Sentiment Scoring with FinBERT

We use FinBERT [2] as the main sentiment model. For each article text:

1. The text is tokenized and truncated to the model's maximum length.
2. FinBERT outputs class probabilities for **positive**, **negative**, and **neutral** labels [2].
3. We transform these into a **continuous sentiment score** between −1 and +1 by taking the difference between positive and negative probabilities. Neutral probability acts as a confidence indicator, where highly neutral outputs are down-weighted in aggregate indices.

This procedure yields an article-level sentiment score that is domain-specific and interpretable. For robustness, the framework allows parallel computation of alternative sentiment indices (e.g., lexical polarity or a general GPT-based rating) for comparison.

C. Daily Sentiment Index Construction

To integrate multiple articles per day into a single signal that can be fed to forecasting models, we construct a **daily sentiment index**:

1. Group articles by date.
2. Compute the volume-weighted or equally weighted average FinBERT sentiment score across articles.

3. Optionally apply a smoothing or exponential decay to capture persistence in sentiment shocks.

The result is a time series of daily sentiment values aligned with TSLA prices. We treat this as an exogenous feature that augments price-based predictors.

D. Forecasting Models

We consider three deep learning architectures that are widely used in financial forecasting [7], [10]:

1. **LSTM (Long Short-Term Memory)** – Captures long-range temporal dependencies and is a standard benchmark for stock prediction [7].
2. **GRU (Gated Recurrent Unit)** – A lighter recurrent architecture with fewer parameters and often comparable performance.
3. **Transformer-style model** – A simplified temporal transformer that uses self-attention to capture relationships between days [10].

For each architecture, we define two variants:

- **Price-only model** – Inputs consist of historical price features (e.g., close, returns, and optionally volume or technical indicators) over a rolling window.
- **Sentiment-augmented model** – The same price features plus the daily sentiment index.

All models are trained to predict next-day returns or price changes. We split the data into training, validation, and test sets in chronological order and use early stopping and hyperparameter tuning to mitigate overfitting.

E. Evaluation Metrics and Statistical Tests

We evaluate predictive performance using standard regression and classification metrics:

- Mean Squared Error (MSE) and Mean Absolute Error (MAE) for point forecasts.
- Mean Absolute Percentage Error (MAPE) for scale-free comparison.
- **Directional accuracy** (percentage of correctly predicted up/down moves), which is particularly relevant for trading [3].

To assess whether performance differences between models with and without sentiment are statistically significant, we apply the **Diebold–Mariano (DM) test** [8] to forecast

error sequences. This test compares predictive accuracy across two competing models while accounting for serial dependence in forecast errors.

F. Economic Evaluation via Trading Strategy

To translate forecast improvements into economic terms, we implement a simple **long–flat trading strategy**:

- If the model predicts a positive next-day return, take a long position in TSLA; otherwise, hold cash.
- Compute daily strategy returns net of transaction costs (assumed small but non-zero).
- Summarize performance using the **Sharpe ratio** [9], cumulative returns, and maximum drawdown.

Comparing Sharpe ratios and equity curves across model variants provides insight into the practical value of sentiment-enhanced forecasts, beyond purely statistical metrics.

V. TSLA Case Study: Empirical Results

A. Descriptive Statistics

We begin by summarizing the TSLA subset of FNSPID [1]. Over the sample period, we observe a substantial number of TSLA-related news articles, with clustering around major events such as earnings announcements, product launches, and regulatory news. Daily TSLA returns exhibit high volatility and pronounced trends, consistent with prior work on technology stocks [7], [10].

The FinBERT sentiment scores show a roughly symmetric distribution centered near zero, with occasional bursts of strongly positive or negative days. When aggregated into a daily index, sentiment appears to co-move with large price moves during major news events, but also displays periods where sentiment and returns diverge, highlighting the need for formal modeling.

B. Forecast Accuracy Comparison

Across all three architectures (LSTM, GRU, transformer), the **sentiment-augmented models** generally outperform their price-only counterparts on the TSLA test set, though the magnitude of improvement varies.

- For the LSTM, adding the daily FinBERT sentiment index reduces test MSE and increases directional accuracy relative to the price-only model, in line with the hypothesis that sentiment adds predictive information [3], [7].
- The GRU model shows similar qualitative improvements but with slightly smaller gains, consistent with its more compact structure.
- The transformer model benefits from sentiment in some horizons but is more sensitive to hyperparameter choices and may require larger datasets to consistently outperform [10].

Diebold–Mariano tests [8] indicate that, in several configurations, the sentiment-augmented models achieve **statistically significant** improvements in predictive accuracy over price-only baselines, especially when evaluated on directional error loss functions. In other cases, improvements are positive but not statistically significant, underscoring that sentiment is not universally powerful and must be carefully integrated.

C. Economic Performance of Sentiment-Enhanced Strategies

Using the long–flat strategy defined in Section IV-F, we compare the economic value of forecasts with and without sentiment.

- For the LSTM-based strategy, incorporating sentiment yields a higher Sharpe ratio and improved cumulative returns relative to the price-only strategy, with comparable or lower drawdowns [9].
- GRU-based strategies show modest Sharpe improvements and some reduction in volatility.
- Transformer-based strategies are more volatile and exhibit mixed performance, again reflecting their sensitivity to data size and tuning.

Importantly, strategies built on sentiment-augmented forecasts tend to reduce the number of “false positive” long positions during clearly negative news cycles, suggesting that FinBERT sentiment helps the model better identify adverse conditions.

D. Robustness and Sensitivity Analyses

To test robustness, we conduct several additional analyses:

1. **Alternative sentiment constructions** – Using equal-weighted vs volume-weighted averaging and simple rolling averages vs exponential decay for the sentiment index. The main findings are robust to these choices, though heavily smoothed indices can dilute short-term effects.

2. **Alternative forecast horizons** – Extending the horizon to multi-day returns reduces the incremental value of sentiment, consistent with the view that news effects are strongest at short horizons [3], [5].
3. **Subsample analysis** – Splitting the sample into pre- and post-major market events (e.g., COVID-19 period) shows that sentiment is particularly helpful during high-volatility regimes, when news flow is dense and investors react strongly.

VI. Discussion

The TSLA case study illustrates how a structured framework, a standardized dataset, and a domain-specific sentiment model can be combined to produce more credible evidence on the role of sentiment in forecasting.

Relative to the original FNSPID study [1], this work:

- **Clarifies the sentiment construction step** by replacing ad-hoc prompting with FinBERT-based scoring [2] and transparent aggregation choices.
- **Provides explicit baseline models** and sentiment-augmente1. **Generality:** While most of our analysis has focused on TSLA, it would be useful to consider how these findings might apply to other stocks or sectors. Comparing results from different assets could help clarify whether our approach is broadly applicable or tailored to TSLA-specific behavior.
- 2. **Limitation Discussion:** Although our results suggest sentiment information can improve forecasting, it's important to recognize circumstances where it might be misleading or less useful. For instance, some market events may generate a lot of news but have little real impact on prices, or sentiment indices could be influenced by media biases. Highlighting examples or discussing these limitations adds nuance and transparency.
- 3. **Further Applications:** Beyond the academic domain, these methods could inform the development of trading strategies or risk management systems in practice. For future research, exploring how sentiment-enhanced forecasts perform during market shocks or integrating alternative data sources could be valuable next steps.
- d counterparts across multiple architectures [7], [10].
- **Applies standard forecast comparison tests** [8] and risk-adjusted performance metrics [9] to ground claims of improvement.
- **Documents a full, reproducible pipeline** that can be extended to other stocks or sectors.

At the same time, several limitations remain. Focusing only on TSLA may limit generalizability; future work should extend the analysis to multiple tickers and sectors. Our sentiment index relies solely on news articles in FNSPID; adding social media or options-implied sentiment could enrich the signal [6]. Finally, although deep learning models are flexible, their interpretability is limited; combining them with simpler models or attention-based explanations may improve understanding and trust [3], [10].

VII. Conclusion and Future Work

This paper proposes a comprehensive framework for integrating financial sentiment into forecasting models and implements it on the FNSPID dataset with a TSLA-focused case study. Using FinBERT to construct a daily sentiment index and comparing price-only and sentiment-augmented deep learning models, we find evidence that sentiment can improve both statistical and economic performance, particularly at short horizons and in volatile periods.

Our contributions are threefold: (i) a modular, reproducible pipeline for sentiment-enhanced forecasting based on a public benchmark dataset; (ii) a rigorous comparison of deep learning models with and without sentiment features, using standard statistical and economic metrics; and (iii) an explicit positioning of our approach as an extension and methodological refinement of the FNSPID study [1].

Future research directions include expanding to multi-asset portfolios, exploring alternative sentiment sources and models (e.g., multimodal transformers), performing more extensive ablation studies, and investigating reinforcement-learning-based trading policies that directly optimize risk–return objectives.

References

[1] Z. Dong, W. Han, Y. Liu, Y. Wang, and Y. Zhang, “FNSPID: Financial News Stock Price Integrated Dataset,” arXiv:2402.06698, 2024.

[2] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” arXiv:1908.10063, 2019.

[3] F. Z. Xing, E. Cambria, and R. E. Welsch, “Natural Language Based Financial Forecasting: A Survey,” Artif. Intell. Rev., vol. 50, no. 1, pp. 49–73, 2018.

[4] C. Kearney and S. Liu, “Textual Sentiment in Finance: A Survey of Methods and Models,” Int. Rev. Financ. Anal., vol. 33, pp. 171–185, 2014.

[5] P. C. Tetlock, “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” J. Finance, vol. 62, no. 3, pp. 1139–1168, 2007.

[6] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011.

[7] Y. Zhang, P. Chen, and Y. Wang, “Stock Market Prediction via Multi-source Long Short-term Memory Network,” in Proc. 2018 IEEE Int. Conf. Big Data (Big Data), pp. 2281–2290, 2018.

[8] F. X. Diebold and R. S. Mariano, “Comparing Predictive Accuracy,” J. Bus. Econ. Stat., vol. 13, no. 3, pp. 253–263, 1995.

[9] W. F. Sharpe, “Mutual Fund Performance,” J. Bus., vol. 39, no. 1, pp. 119–138, 1966.

[10] L. Yang, Y. Li, and E. Cambria, “Financial Time Series Forecasting With Deep Learning: A Systematic Literature Review: 2005–2022,” IEEE Access, vol. 10, pp. 84574–84596, 2022.