

Comparing Transformer-Based Sentiment Models for Technical-Filtered Trading Signal Generation on NASDAQ-100 Stocks

Rifki Ainul Yaqin¹, Ignatius Wiseto Prasetyo Agung², Toni Arifin³, Erfian Junianto⁴

^{1,2,3,4}Information Engineering Study Program, Faculty of Information Technology, Universitas Adhirajasa Reswara Sanjaya (ARS University), Bandung, Indonesia

^{1,2,3,4}ARS Digital Research and Innovation (ADRI), Universitas Adhirajasa Reswara Sanjaya (ARS University), Bandung, Indonesia

Article Info

Article history:

Received April 15, 2026

Revised July 10, 2026

Accepted July 21, 2026

Keywords:

FinBERT

Relative Strength Index

Moving Average

Financial Sentiment Analysis

Trading Strategy

ABSTRACT

Predicting stock price movements remains challenging due to nonlinear, dynamic market behavior. News sentiment and technical indicators (RSI, moving averages) each show predictive value, but few studies integrate both within a single, robustly validated framework. This study proposes an algorithmic trading framework integrating daily news sentiment from four pre-trained transformer models (ProsusAI/finbert, nlptown/bert-base-multilingual-uncased-sentiment, yiyanghkust/finbert-tone, and soleimanian/financial-roberta-large-sentiment) with RSI and moving average filters to generate buy signals on 30 NASDAQ-100 stocks. Beyond standard backtesting, the framework is validated through five analyses: benchmark comparison, exit-rule sensitivity, look-ahead-bias alignment, statistical testing, and walk-forward validation. Over a 2-year simulation (57,434 news articles), buy signals were triggered when RSI fell below 40, price pulled back 0.5–5% below MA20, and price remained within 10% above MA50, with entry on day $i+1$ and a profit-triggered exit. All four models produced positive cumulative returns (123.33–187.84%, Sharpe Ratio 2.00–2.73), remaining profitable under transaction costs up to 0.3% per trade. Benchmarking showed roughly half the Maximum Drawdown of naive baselines despite lower raw returns, favoring selectivity over return maximization. The stop-loss/take-profit variant substantially reduced performance, while stricter news-timestamp alignment left results robust (Sharpe Ratio above 1.5). Statistical tests found three of six model-pair differences significant at the 5% level, though the two top-performing models are statistically indistinguishable ($p = 0.546$). Out-of-sample evaluation confirmed profitability in the most recent test segment, with acknowledged sub-period variability. These findings position news sentiment as an effective, risk-disciplined complement to technical analysis, with model selection depending on investor priorities.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. INTRODUCTION

The stock market is a dynamic and nonlinear financial instrument, making the accurate prediction of stock price movements a primary challenge for investors and researchers in the field of computational finance [1]. While prior studies have addressed this challenge through price-based deep learning models such as LSTM [2], an alternative direction involves leveraging financial news sentiment combined with technical indicators to generate actionable trading signals. Financial news has been proven to have a significant correlation with stock price movements, as it contains market sentiment information that can influence investor decisions in aggregate [3]. In efforts to automatically extract sentiment from financial texts, transformer-based models such as FinBERT have demonstrated superior performance compared to generic NLP models and lexicon-based methods, given that FinBERT is trained specifically on a financial text corpus, enabling it to capture domain-

*Corresponding Author

Email: rifkiainul17@gmail.com

specific semantic nuances although this advantage is task-dependent and does not universally apply across all portfolio performance metrics [4][5]. On the other hand, technical analysis through indicators such as the Relative Strength Index (RSI) and Moving Average (MA) has long been employed as a tool for identifying optimal market conditions for entering and exiting trading positions [6]. Several recent studies indicate that the integration of news sentiment signals with technical indicators consistently yields trading strategies that are superior to single-approach methods, both in terms of prediction accuracy and portfolio metrics such as the Sharpe ratio and Maximum Drawdown [5][6].

Mun & Kim (2025) conducted financial news sentiment analysis using discriminative models such as BERT and FinBERT, as well as generative models including Llama 3.1, Mistral, and Gemma 2, applied to the top 30 companies listed on NASDAQ. Their findings indicate that the long strategy consistently yields superior portfolio performance compared to both short and long-short strategies, with evaluation metrics encompassing average daily return, Sharpe ratio, final return, and Maximum Drawdown (MDD) [7]. Meanwhile, Kim *et al.* (2025) developed a rule-based trading recommendation system that integrates news sentiment analysis with RSI as the primary technical indicator, wherein the system successfully achieved an annualized rate of return of 5.10% over the September-October 2024 period, surpassing the performance of the DJIA, NASDAQ Composite, and S&P 500 indices [8]. Shobayo et al. (2024) compared the performance of FinBERT, GPT-4, and Logistic Regression for sentiment analysis and stock price prediction, concluding that encoder-based models such as FinBERT possess a deeper capacity for understanding semantic nuances in financial texts, albeit at the cost of greater computational resources relative to traditional approaches [9]. Gómez-Martínez et al. (2025) demonstrated that the incorporation of investor sentiment indicators into algorithmic trading strategies based on moving averages, RSI, and MACD applied to the NASDAQ Mini futures market was capable of enhancing risk-adjusted returns, reducing volatility, and improving overall profitability compared to purely technical strategies over a backtest period spanning May 2022 to December 2024 [10]. Chen et al. (2025) proposed a framework that integrates a Domain Knowledge Chain-of-Thought (DK-CoT) prompt engineering strategy with large language models (LLMs) for financial news sentiment analysis, establishing FinBERT and RoBERTa as baselines that were demonstrated to yield the highest average accuracy and F-score among the transformer-based models evaluated [11].

Based on the review of prior research, several gaps remain that have not been comprehensively addressed. Mun & Kim (2025) compared discriminative and generative LLM models for long strategies on NASDAQ, yet did not incorporate technical indicators as market condition filters prior to trade execution [7]. Kim et al. (2025) integrated RSI with news sentiment, but evaluated only a single sentiment model over an extremely short backtest period of two months, such that the robustness of the strategy cannot be adequately verified [8]. Shobayo et al. (2024) and Chen et al. (2025) focused on sentiment classification accuracy and price direction prediction, rather than on simulated real trade execution equipped with portfolio metrics such as the Sharpe ratio and Maximum Drawdown [9][11]. Gómez-Martínez et al. (2025) demonstrated the benefit of augmenting technical strategies with sentiment signals; however, they employed aggregate market investor sentiment indicators rather than text-based sentiment extracted on a per-stock basis using domain-specific financial transformer models [10].

This study proposes a sentiment-based algorithmic trading strategy that compares four transformer-based sentiment analysis models namely ProsusAI/finbert, nlptown/bert-base-multilingual-uncased-sentiment, yiyanghkust/finbert-tone, and soleimanian/financial-roberta-large-sentiment from the Hugging Face platform to extract daily per-stock sentiment from financial news across 30 NASDAQ-100 stocks over the 2024-2026 period. The selection of these models is grounded in the demonstrated advantage of transformer architectures trained specifically on financial text corpora in capturing semantic nuances that generic models are unable to detect [12]-[14]. The resulting positive sentiment signals are subsequently validated by two technical filters namely RSI below a defined threshold and the positioning of price relative to both short-term and long-term Moving Averages serving as a confirmation mechanism for optimal market conditions prior to trade execution, given that RSI and Moving Average have been proven effective in identifying favorable entry points in equity markets [15][16]. Trade execution is designed to mitigate look-ahead bias, whereby signals are generated on day i and purchases are executed at the Opening price of day $i+1$, with a profit-based exit mechanism in which positions are closed once the Closing price exceeds the purchase price ensuring that each trade yields a clean and measurable decision.

In this study, entitled "Comparing Transformer-Based Sentiment Models for Technical-Filtered Trading Signal Generation on NASDAQ-100 Stocks" the novelty of the research lies in the integration of news-based sentiment analysis using multiple transformer language models specifically tailored to the financial domain with technical indicators to generate stock trading signals. Unlike many preceding studies that focus exclusively on sentiment classification accuracy or technical analysis in isolation, this research combines both approaches within a unified trading simulation framework to directly evaluate their impact on investment performance. By comparing several pre-trained financial sentiment models and combining them with technical filters such as Moving Average and Relative Strength Index (RSI), this study provides an empirical account of

how differences across sentiment models can influence the quality of trading signals under real market conditions. The primary contributions of this study are as follows:

1. Integrating news sentiment analysis and technical indicators (RSI, MA20, and MA50) within an algorithmic trading simulation framework to generate buy signals for NASDAQ-100 stocks.
2. Comparing four transformer-based sentiment analysis models namely ProsusAI/finbert, nlptown/bert-base-multilingual-uncased-sentiment, yiyanghkust/finbert-tone, and soleimanian/financial-roberta-large-sentiment to evaluate their effectiveness in generating profitable trading signals.
3. Implementing a daily sentiment aggregation mechanism using a majority voting method to consolidate multiple news articles within a single day into one representative sentiment signal, thereby reducing noise in the sentiment data.
4. Evaluating trading strategy performance using multiple financial metrics namely Average Daily Return, Standard Deviation (Std Dev), Sharpe Ratio, Maximum Drawdown (MDD), and Final Return to assess both the profitability and risk level of the proposed strategy.
5. Validating the robustness of the proposed framework through five additional analyses: benchmarking against naive Buy-and-Hold, RSI-only, MA-only, and Random trading baselines; testing a stop-loss/take-profit exit-rule variant; refining the look-ahead-bias alignment by separating pre-market, intraday, and post-market news timestamps; applying pairwise statistical significance tests (Wilcoxon signed-rank, paired t-test, and Diebold-Mariano tests) across the four sentiment models; and evaluating rolling-window, expanding-window, and train-validation-test out-of-sample performance.

The remainder of this paper is organized as follows. Section 2 describes the research methodology, encompassing the dataset, sentiment analysis models, and technical indicators. Section 3 presents the experimental results and performance analysis of the trading strategy. Finally, Section 4 elaborates on the research conclusions, along with limitations and directions for future research.

2. METHOD

This study proposes a trading simulation framework that integrates news-based sentiment analysis with technical indicators to generate stock buy signals across 30 NASDAQ-100 tickers. The overall research workflow is illustrated in [Figure 1](#). The process begins with two parallel data collection streams, namely financial news articles and daily stock price data in OHLC (Open, High, Low, Close) format. Each news article is labeled for sentiment using four models ProsusAI/finbert, nlptown/bert-base-multilingual-uncased-sentiment, yiyanghkust/finbert-tone, and soleimanian/financial-roberta-large-sentiment from the Hugging Face platform whereby the resulting sentiment labels are subsequently normalized into three unified classes: positive, neutral, and negative. On the price data side, technical indicators encompassing RSI, short-term and long-term moving averages, and momentum are computed from the OHLC price series. Both streams are then aligned at the daily level, wherein news published on non-trading days such as weekends and public holidays is explicitly forwarded to the immediately following trading day, ensuring that all available market information is incorporated into the signal without inducing look-ahead bias. Only articles for which no subsequent trading day exists within the observation window are excluded. The daily sentiment is then determined through majority voting. A buy signal is generated on day i when all conditions are met: positive sentiment, RSI below the defined threshold, a short-term price pullback, and an ongoing upward trend. A buy order is then executed at the opening price of day $i+1$, and the position is closed at the first closing price that yields a profit. The simulation is run iteratively throughout the observation period, and portfolio performance is evaluated using the metrics of Average Daily Return, Standard Deviation (Std Dev), Sharpe Ratio, Maximum Drawdown (MDD), and Final Return.

2.1. Dataset

This study utilizes two primary data sources collected separately and subsequently merged based on trading calendar alignment. The overall observation period spans from February 12, 2024 to February 6, 2026, encompassing a total of 30 stocks from the NASDAQ-100 index as the investment universe, following the approach adopted by Mun & Kim (2025), who similarly employed the top 30 NASDAQ stocks as the experimental basis for sentiment-based investment strategies [\[7\]](#).

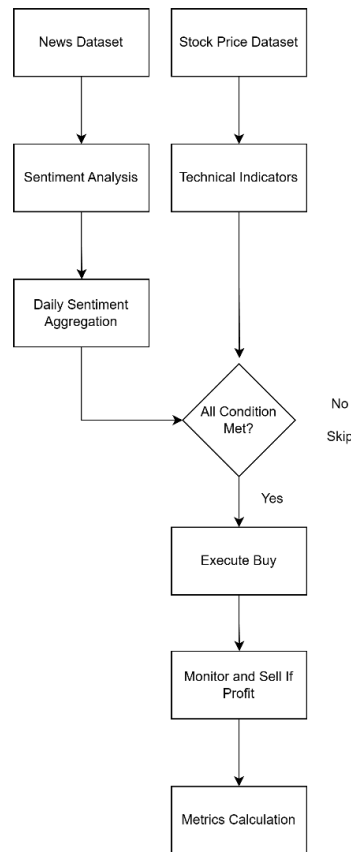


Figure 1. Trading Simulation Flowchart

2.1.1. Stock Selection

The stock selection in this study is inspired by the work of Mun and Kim (2025), who employed the top 30 companies by market capitalization listed on the NASDAQ index as their research subject. Mun and Kim (2025) conducted sentiment analysis on news articles pertaining to the 30 largest companies listed on NASDAQ, and evaluated the performance of LLM sentiment-based investment strategies using profitability and portfolio stability metrics [7]. Drawing on this approach, the present study employs 30 stock tickers selected from the NASDAQ-100 index, representing large-capitalization technology and non-technology companies that are actively covered by international financial media. The list of tickers used is as follows:

“AAPL, ADBE, AMAT, AMD, AMGN, AMZN, ARM, ASML, AVGO, AZN, BKNG, CMCSA, COST, CSCO, GOOG, GOOGL, HON, INTU, ISRG, LIN, META, MSFT, NFLX, NVDA, PDD, PEP, QCOM, TMUS, TSLA, TXN”

The selection of these 30 tickers is grounded in several considerations. First, the constituent companies are highly liquid issuers with significant news coverage volume, ensuring that news data availability remains relatively consistent throughout the study period. Second, the cross-sector coverage encompassing semiconductor technology (NVDA, AMD, AMAT, AVGO, ASML), software and cloud computing (MSFT, ADBE, INTU, CSCO), e-commerce and retail (AMZN, COST, PDD), consumer goods (PEP), telecommunications (TMUS, CMCSA), biotechnology and healthcare (AMGN, ISRG, AZN), as well as energy and industrials (LIN, HON, TXN) enables a more representative evaluation of the proposed strategy against the broader dynamics of the NASDAQ market. Third, the inclusion of two Alphabet tickers (GOOG and GOOGL), reflecting two share classes of the same entity, is retained to maintain consistency with the available data while adhering to the convention commonly observed in NASDAQ index-based studies.

2.1.2. News Dataset

Financial news data in this study were collected using the Polygon API, a financial market data platform that provides access to news articles from various English-language financial media sources in both real-time and historical formats. Data collection was performed by querying based on each company's stock ticker, ensuring that every article retrieved maintains direct relevance to the corresponding company. The news collection period spans from February 12, 2024 to February 6, 2026, yielding a raw news corpus that is subsequently processed in the sentiment analysis stage.

2.1.3. Stock Price Dataset

Daily stock price data were obtained using the yfinance library, a Python interface that enables direct retrieval of historical data from Yahoo Finance. For each ticker, the downloaded data encompass the Opening price (Open), Highest price (High), Lowest price (Low), Closing price (Close), and daily trading volume. The stock price data retrieval period is identical to that of the news dataset, spanning from February 12, 2024 to February 6, 2026, thereby allowing both datasets to be temporally aligned at the trading day level. Closing price data are used as the reference for technical indicator calculations, while the Opening price of the subsequent day serves as the buy order execution price within the trading simulation.

2.2. Sentiment Models

This study employs four transformer-based encoder models that have been specifically trained on financial texts and are publicly available through the HuggingFace platform. The first model is ProsusAI/finbert, which was constructed by retraining the BERT architecture on the TRC2-financial corpus and subsequently fine-tuned using the Financial PhraseBank dataset, yielding a three-class classification scheme (positive, neutral, negative) that constitutes the foundational baseline for domain-specific FinBERT model development [12][17]. The second model is nlptown/bert-base-multilingual-uncased-sentiment, a multilingual BERT model fine-tuned for product review sentiment analysis across six languages using a 1–5-star rating scale, wherein the labels are normalized in this study into three classes: 1-2 stars as negative, 3 stars as neutral, and 4-5 stars as positive [13]. The third model is yiyanghkust/finbert-tone, developed by re-pretraining BERT on more than 4.9 billion words drawn from analyst reports, 10-K and 10-Q filings, and earnings call transcripts, whereby this model has been empirically demonstrated to surpass the Loughran-McDonald dictionary and various conventional machine learning algorithms in financial sentiment classification [14]. The fourth model is soleimanian/financial-roberta-large-sentiment, obtained from the Hugging Face platform, which is built upon the RoBERTa-Large architecture and fine-tuned on a diverse corpus encompassing 10-K, 10-Q, and 8-K filings, earnings call transcripts, CSR reports, and ESG news, rendering it more comprehensive in capturing the varied registers of financial text compared to the preceding models. These four models were deliberately selected as each represents a distinct pre-training approach ranging from a news corpus (Reuters TRC2), general multilingual sentiment, institutional analyst reports, to regulatory and sustainability documents such that their comparison is capable of revealing the extent to which the domain specialization of pre-training corpora influences the quality of the resulting trading signals.

2.3. Technical Indicators

Technical indicators are analytical tools based on historical price and trading volume data, employed to identify market patterns and generate transaction signals within algorithmic trading systems. Unlike fundamental analysis, which focuses on the intrinsic value of a company, technical analysis departs from the assumption that all relevant information is already reflected in historical price movements, such that these patterns can be leveraged to anticipate subsequent price behavior. In this study, two primary technical indicators are applied simultaneously namely Moving Average (MA) and Relative Strength Index (RSI) each contributing to the identification of trend direction, overbought and oversold conditions, as well as the strength of price movements. Various types of technical indicators, such as moving averages and the relative strength index, are widely employed in predictive models to identify the direction of stock price movements and turning points in time series data [18]. The combination of multiple indicators has been demonstrated to yield more reliable trading signals compared to the use of any single indicator in isolation.

2.3.1. Moving Average (MA)

Moving Average (MA) is a technical indicator that computes the average closing price over a specified rolling window to smooth short-term fluctuations and expose the underlying trend direction. MA-based trading strategies with different time structures have been demonstrated to generate significant excess returns, wherein the combination of short-term and long-term MAs reflects the heterogeneity of investor behavior across varying investment horizons [19]. Mathematically, the Simple Moving Average (SMA) is formulated as the arithmetic mean of closing prices over the most recent n periods precisely as implemented through the `calculate_moving_average()` function using the `.rolling(window=n).mean()` operation in the codebase of this study:

$$MA_n(t) = \frac{1}{n} \sum_{i=0}^{n-1} P_{t-i} = \frac{P_t + P_{t-1} + \dots + P_{t-n+1}}{n} \quad (1)$$

where P_t is the closing price on day t and n is the length of the window period. This study employs two MA windows simultaneously: a short-term MA (MA_SHORT, $n = 20$ days) and a long-term MA (MA_LONG,

$n = 50$ days). This dual moving average strategy requires the price to be positioned above MA_LONG as confirmation of an uptrend, while simultaneously detecting a price pullback toward MA_SHORT as an entry opportunity. The Double Moving Average Crossover (DMAC) strategy generates a buy signal when the short-term MA crosses above the long-term MA from below, and a sell signal under the reverse condition; nevertheless, this strategy is known to produce a higher number of false signals and requires specific market volatility conditions in order to yield positive returns [20]. In the context of algorithmic trading systems, MA-based technical analysis techniques constitute one of the core components in the stock portfolio selection process and the optimization of investment strategies grounded in historical market data [21].

2.3.2. Relative Strength Index (RSI)

The Relative Strength Index (RSI) is a momentum indicator first introduced by J. Welles Wilder and is widely employed in the technical analysis of financial markets. RSI is a commonly used indicator for measuring the speed and magnitude of stock price changes, as well as the degree of overvaluation or undervaluation of a given stock. RSI is capable of displaying overbought and oversold conditions for a stock, with RSI values ranging from 0 to 100. Typically, an RSI above 70 indicates that a stock is in an overbought condition, whereas an RSI below 30 indicates an oversold condition [22][23]. RSI is used to determine the probability of a trend reversal occurring, thereby making it one of the most reliable technical indicators in the process of stock selection and stock price movement prediction [24]. Mathematically, RSI is calculated using the formula:

$$RSI = 100 - \left(\frac{100}{1 + RS} \right) \quad (2)$$

where RS (Relative Strength) is the ratio of the average price gain to the average price loss over a specified period. In accordance with prevailing convention, this period is set to the 14 most recent historical price observations [25].

In the development of algorithmic trading strategies, RSI is frequently combined with other technical indicators, such as the Exponential Moving Average (EMA), in order to generate more accurate and profitable buy and sell signals compared to the use of a single indicator [6]. Furthermore, recent research indicates that the default RSI parameters can be further optimized. By employing genetic algorithms, RSI parameters such as the buy/sell thresholds and the number of historical observations used in its construction can be automatically and individually optimized for each asset, yielding investment performance that substantially surpasses traditional technical analysis approaches [25].

In the context of artificial intelligence-based predictive model development, RSI has been demonstrated to serve as a highly informative input feature. RSI is incorporated as one of the core technical indicators in machine learning-based predictive models such as Random Forest and XGBoost, wherein the combination of such technical indicators alongside internet search data exhibits the highest ROC curve compared to the use of any single data type alone [24]. Research indicates that more than 58% of predictions generated through artificial intelligence-based methodologies employing RSI as the primary indicator are capable of surpassing results obtained through traditional technical analysis approaches, underscoring the ability of artificial neural networks to model the complexity of financial market dynamics more accurately [25]. Furthermore, RSI can be utilized as an auxiliary indicator within trading strategies to determine the direction of important trading points that is, whether a given point constitutes a buy or a sell signal thereby helping to mitigate the adverse impact of model prediction errors and enhance the overall performance of investment strategies [22].

2.4. Daily Sentiment Aggregation

In news-based financial sentiment analysis, a single trading day frequently yields more than one news article with differing sentiment labels, necessitating an aggregation mechanism to convert the collection of daily sentiments into a single unified representation that can be employed as a predictive feature. The majority of prior research sums the sentiment scores of texts on each calendar day and uses the aggregated scores to predict various stock trends [26]. One commonly adopted approach is majority voting, whereby the daily sentiment label is determined based on the most dominant class across all news articles published on that day if the number of positively-labeled articles exceeds that of articles labeled negative or neutral articles, the sentiment label for that day is assigned as positive, and vice versa. This approach, referred to as majority voting, determines the dominant sentiment by counting the number of articles that carry the highest sentiment value for a given class, such that the overall sentiment is represented by the most frequently occurring class [27]. Recent studies have introduced daily aggregated sentiment indices as a representation of the emotional signal derived from a collection of news articles, which are subsequently used as input to machine learning models for predicting stock market trends. This daily aggregation process is critical given that forward-looking implied sentiment has been shown to account for a substantial proportion of the variation in stock returns, whereas individual unaggregated sentiment scores tend to possess comparatively weaker predictive power [28]. Accordingly, daily sentiment aggregation functions as a noise reduction stage that transforms multiple daily

news signals into a single representative sentiment signal, enabling predictive models to learn from sentiment patterns that are more stable and consistent over time.

2.5. Metrics Calculation

At the stage of evaluating the performance of the proposed strategy or model, several metrics are required to objectively measure the research outcomes. These metrics are employed to analyze the level of profitability, risk, and stability of the resulting strategy. By utilizing a variety of evaluation indicators, researchers can assess whether the developed strategy is capable of delivering optimal returns with a controlled level of risk. Accordingly, this study employs several evaluation metrics, namely Average Daily Return, Standard Deviation (Std Dev), Sharpe Ratio, Maximum Drawdown (MDD), and Final Return.

2.5.1. Average Daily Return

Average Daily Return is one of the fundamental metrics in trading strategy performance evaluation, measuring the percentage change in the value of a portfolio or asset on each trading day. In the context of trading simulation, average daily return plays a crucial role, as predicting the direction of daily return movements has been demonstrated to yield superior performance in trading or simulation environments compared to directly predicting level values [29]. To measure the consistency and average profitability of the strategy over the simulation period, the Average Daily Return metric is employed, formulated as follows:

$$\text{Avg Daily Return} = \frac{1}{N} \sum_{t=1}^N \tilde{r}_t \quad (3)$$

where $\tilde{r}_t$ represents the daily return on day t , and N is the total number of trading days over the simulation period. In this study, the simulation is conducted over a 2-year period, such that N encompasses all active trading days within that timeframe. The distributional characteristics of the daily rate of return constitute a primary concern in stock market analysis, given its thick-tail properties that do not always conform to a normal distribution [30]. A positive Average Daily Return over the 2-year simulation period indicates that the trading strategy is capable of consistently generating profit per trading day, and this metric serves as the primary benchmark for comparing the effectiveness across strategies as well as against the benchmark strategy employed.

2.5.2. Standard Deviation (Std Dev)

Standard deviation is a statistical measure used to quantify the degree of dispersion or spread of daily return values relative to their mean, and in the financial context, this metric is widely accepted as the primary proxy for investment volatility and risk. In financial engineering, volatility is generally defined as the dispersion of a return series and is computed using the sample standard deviation [31]. Standard deviation, or its squared analogue in the form of variance, is used to measure the dispersion of returns around the mean value, and this metric occupies a central role in most modern risk management systems while being closely related to tail risk measures such as value at risk and expected shortfall [32]. In this study, the standard deviation of daily returns is computed using the following formula:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{t=1}^N (\tilde{r}_t - \mu)^2} \quad (4)$$

where $\tilde{r}_t$ is the daily return derived from the portfolio equity value at each time step, μ is the mean daily return, and N is the total number of trading days over the 2-year simulation period. The use of $(N-1)$ as the divisor (rather than N) reflects the sample standard deviation estimation, which yields an unbiased estimator for the population. In trading strategy evaluation, the metrics of returns and volatility wherein standard deviation constitutes the primary representation of volatility have been demonstrated to be the most important metrics for comprehensively representing model performance [33]. A low standard deviation value over the 2-year simulation period indicates that the trading strategy produces more stable and consistent returns on a daily basis, thereby reflecting a more controlled level of risk.

2.5.3. Sharpe Ratio

The Sharpe Ratio is one of the most widely used investment performance evaluation metrics in quantitative finance, owing to its ability to measure the return generated per unit of risk borne by a strategy or portfolio. The Sharpe ratio can be readily applied to any return series without requiring additional information regarding the sources of volatility or profitability, making it broadly employed for comparing the risk-adjusted returns of various investment types such as equities, ETFs, mutual funds, and investment portfolios [34]. The Sharpe ratio measures risk-adjusted performance, such that strategies with high return volatility incur a penalty, and through this metric, trading agents or systems can learn not only about returns but also about risk, ultimately

contributing to the creation of more balanced portfolios [35]. In this 2-year trading simulation, the Sharpe Ratio is computed over the full observation period, including all calendar trading days regardless of whether a position was active. On days when no trade is open, the daily return is treated as zero, ensuring that idle periods are fully reflected in both the mean and standard deviation of the return series. The annualized Sharpe Ratio is computed using the following formula:

$$\text{Sharpe} = \frac{\mu}{\sigma} \times \sqrt{252} \quad (5)$$

where μ is the average daily return, Std Dev is the standard deviation of daily returns, and the scaling factor $\sqrt{252}$ is used to annualize the ratio in accordance with the number of effective trading days in one year. In this formulation, μ and σ are computed over all N calendar trading days in the simulation period, with $r_t = 0$ assigned to each day on which no position is open. This ensures that the Sharpe Ratio reflects the true opportunity cost of capital including idle periods, rather than measuring only the distribution of returns on active trading days. It should be noted that this formulation assumes a risk-free rate of zero, which is commonly adopted in the context of trading strategy simulation and backtesting. In risk-adjusted return-based performance evaluation, the Sharpe Ratio serves as one of the primary metrics employed alongside the Sortino Ratio and Omega Ratio to comprehensively assess strategy performance across various market conditions [36]. The higher the Sharpe Ratio value yielded over the 2-year simulation period, the better the quality of returns generated by the trading strategy relative to the risk incurred.

2.5.4. Maximum Drawdown (MDD)

Maximum Drawdown (MDD) is one of the most critical risk metrics in trading strategy evaluation, measuring the largest decline in portfolio value from a peak to a trough before the portfolio value recovers to a new peak, thereby reflecting the worst-case loss scenario experienced over the simulation period. Maximum drawdown is defined as the largest peak-to-trough decline of the cumulative portfolio value path over a given number N of trading events, and constitutes a path-dependent risk measure as it depends on the sequence and accumulation of returns rather than merely the final value [37]. MDD is the most widely used large drawdown concept, representing the maximum loss from peak to trough over a specified time period, and its magnitude tends to increase as the length of the time series data grows [38]. In this 2-year trading simulation, MDD is computed through two stages using the following formulas:

$$\text{Equity}_t = \prod_{i=1}^t \left(1 + \frac{\tilde{r}_i}{100} \right) \quad (6)$$

$$\text{MDD} = \min_t \left(\frac{\text{Equity}_t - \max_{i \leq t} (\text{Equity}_i)}{\max_{i \leq t} (\text{Equity}_i)} \right) \times 100 \quad (7)$$

In the first stage, the equity curve is constructed cumulatively through the multiplication of the daily growth factor $(1 + \tilde{r}_i/100)$ for each day throughout the simulation period. Subsequently, MDD is obtained by identifying the minimum value of the ratio of the difference between the equity value at time t and the highest peak equity value attained up to time t , which is then multiplied by 100 to express the result as a percentage. In the simulation context, the conditional expected drawdown (CED) derived from the MDD distribution constitutes a practical risk measure as it captures the expected maximum cumulative decline of an asset over a given investment horizon, thereby serving as a useful tool for evaluating whether a trading strategy is viable to execute consistently [37]. An MDD value approaching zero over the 2-year simulation period signifies the strategy's resilience against consecutive losses, whereas a large MDD indicates elevated capital risk and may compel the premature termination of the strategy.

2.5.5. Final Return

Final Return (total return) is the most fundamental performance metric in trading strategy evaluation, representing the overall accumulation of profits or losses generated by a strategy throughout the entire simulation period. Total return provides a comprehensive view of investment profitability by reflecting the cumulative effects of reinvestment and capital growth, thereby offering a more accurate representation of the actual performance of an investment strategy over time [39]. Unlike daily metrics that are average-based in nature, Final Return reflects the true cumulative compounding of all trading decisions over the 2-year simulation period. Consistent with the geometric equity curve defined in Eq. 6, Final Return is computed as:

$$\text{Final Return} = (\text{Equity}_N - 1) \times 100 \quad (8)$$

where Equity_N represents the final portfolio equity at the end of the simulation period obtained from Eq. 6. This formulation is consistent with geometric compounding, ensuring that reinvestment effects are fully

captured and that the Final Return reflects the actual percentage growth of an initial unit investment over the full simulation period. In the context of automated trading system optimization, cumulative return and Sharpe ratio constitute two primary objective functions used to evaluate the success of a trading strategy, where a positive cumulative return serves as the minimum requirement for a strategy to be considered profitable [40]. Accordingly, a large positive Final Return over the 2-year simulation period indicates that the trading strategy successfully generated substantial cumulative profit, whereas a negative value indicates that the strategy resulted in an overall capital loss despite the presence of potentially profitable individual trading days.

2.6. Benchmark Strategies

To contextualize the proposed framework's performance, four naive benchmark strategies are defined: Buy-and-Hold, which holds a single long position from the first to the last trading day of the simulation window; an RSI-only strategy, which issues a buy signal whenever RSI falls below the same threshold used in Section 2.3.2 without any sentiment or moving-average condition; an MA-only strategy, which applies only the pullback/trend conditions of Section 2.3.1 without RSI or sentiment; and a Random strategy, which issues buy signals at randomly selected entry points with frequency matched to the RSI-only signal count. All four benchmarks use the same profit-triggered exit rule as the main strategy (Section 2).

2.7. Exit Rule Variant: Stop-Loss and Take-Profit

As an alternative to the profit-triggered exit rule described above, a stop-loss/take-profit variant is defined in which an open position is closed once price moves against the entry by a fixed stop-loss percentage or in favor of the entry by a fixed take-profit percentage, whichever is reached first; if neither threshold is reached, the position exits under the original profit-triggered rule. This design isolates the effect of adding explicit risk limits without otherwise altering the entry signal or the underlying profit-triggered philosophy.

2.8. Statistical Significance Testing

Differences in daily portfolio returns between pairs of sentiment models are evaluated using three complementary tests: a paired t-test, a Wilcoxon signed-rank test (robust to non-normal return distributions), and a Diebold-Mariano test with Newey-West heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors [41].

2.9. Walk-Forward and Out-of-Sample Validation

To assess temporal robustness, three complementary validation schemes are applied with all strategy parameters held fixed: rolling-window evaluation over non-overlapping sub-periods, expanding-window evaluation over cumulative sub-periods, and a sequential train-validation-test split. These schemes follow standard walk-forward validation practice used to guard against overfitting and regime-specific performance in algorithmic trading research [42].

3. RESULTS AND DISCUSSION

This study examines the capability of four transformer-based sentiment analysis models ProsusAI/finbert, nlpTown/bert-base-multilingual-uncased-sentiment, yiyanghkust/finbert-tone, and soleimanian/financial-roberta-large-sentiment in generating profitable trading signals in the stock market. The simulation was conducted over approximately two calendar years, spanning the period from February 12, 2024 to February 6, 2026, across 30 stock tickers. All trade execution rules adhere to the established protocol, whereby signals are evaluated on day i and executed on day $i+1$, incorporating technical filters (RSI, MA20, MA50), a daily news sentiment filter, and a profit-triggered exit mechanism whereby a position is closed immediately upon the closing price exceeding the purchase price, with any remaining open positions forcibly liquidated at the end of the trading session.

3.1. Overview of Data and Simulation

Over the simulation period, the system successfully collected and processed 57,434 news articles across all 30 tickers. Of this total, 57,247 articles (99.7%) were utilized as active signals. News articles published on non-trading days such as weekends and public holidays were forwarded to the immediately following trading day to preserve their informational content rather than discarding them. Only 187 articles (0.3%) were permanently excluded, comprising articles for which no subsequent trading day existed within the observation window. This exclusion count is consistent across all four models as it is based on the same trading calendar rather than on the predictive quality of the models themselves. With respect to news timestamp alignment, all articles are assigned to the trading day on which they are published, or forwarded to the next trading day if published on a non-trading day; no intraday timing adjustment was applied, meaning that articles published after market close on a given trading day are treated as signals for the following day's sentiment aggregation.

This approach is acknowledged as a potential source of minor look-ahead bias for articles published between market open and close, and is noted as a limitation in Section 4.

Table 1. News Data Summary

Description	Count	Percentage
Total News Collected	57,434	100%
News Used (trading days)	57,247	99.7%
News Discarded (holidays/weekends)	187	0.3%

The application of technical filters (RSI < 40, pullback of 0.5%-5% below MA20, closing price within a 10% range of MA50, and closing price above MA50) is designed to ensure that buy signals are only executed when a stock is in a mildly oversold condition within a valid uptrend. These threshold values were adopted based on findings reported in prior technical analysis research. While the classical RSI thresholds of 30 and 70 indicate oversold and overbought conditions, respectively, some studies suggest alternative levels such as 40 and 60 to capture earlier trend reversals and increase trading opportunities. In particular, previous work observed a bimodal distribution of RSI values with peaks around 40 and 60, leading to the use of RSI < 40 as a buy signal and RSI > 60 as a sell signal [22]. The 0.5%-5% pullback range below MA20 captures minor short-term retracements while excluding noise-level fluctuations and large breakdowns. The MA50 proximity criterion (within 10% above) serves as a trend-health filter, ensuring that the broader uptrend remains intact at the time of entry. These parameters were applied uniformly across all 30 tickers without per-asset optimization, which deliberately trades individual fit for generalizability. The absence of a separate validation set or walk-forward procedure to tune these thresholds is acknowledged as a limitation, and is discussed further in Section 4. This approach inherently limits the number of trades, yet enhances signal precision by filtering out periods characterized by elevated speculation.

3.2. Sentiment Distribution Across Models

Prior to evaluating trading performance, it is important to understand how each model classifies news articles, as the distribution of sentiment labels directly influences the frequency of BUY signals generated. Models that produce a greater proportion of positive labels will tend to yield more signal opportunities, although this does not necessarily correlate linearly with signal accuracy.

Table 2. Sentiment Label Distribution per Model (Total: 57,434 News Articles)

Model	Positive	Negative	Neutral	% Positive	% Negative	% Neutral
ProsusAI/finbert	28.958	12.456	16.020	50.4%	21.7%	27.9%
nlptown/bert-base-multilingual-uncased-sentiment	35.097	14.431	7.906	61.1%	25.1%	13.8%
yyianghkust/finbert-tone	34.705	9.590	13.139	60.4%	16.7%	22.9%
soleimanian/financial-roberta-large-sentiment	32.110	11.367	13.957	55.9%	19.8%	24.3%

ProsusAI/finbert recorded the lowest percentage of positive sentiment (50.4%), while nlptown/BERT-Multilingual registered the highest (61.1%). This difference is practically meaningful: models with a higher proportion of positive labels will more frequently generate "positive sentiment day" conditions, thereby increasing the number of BUY signal opportunities that satisfy the first criterion. Nevertheless, signal quality also depends on whether the increased frequency aligns with valid technical conditions.

It is worth noting that yyianghkust/finbert-tone exhibited the lowest negative percentage (16.7%), indicating that this model is relatively more optimistic in its interpretation of financial news content. This may be attributable to differences in training data distribution, whereby yyianghkust/finbert-tone was trained specifically on English-language financial reports and corporate news with tone annotations, leading it to be more inclined to capture subtle positive nuances embedded in formal financial language.

3.3. Trading Simulation Performance

Table 3 summarizes all portfolio performance metrics across the four models. The metrics employed encompass the total number of trades, average daily return (computed over all N trading days including idle days with zero return), standard deviation of daily returns, Sharpe Ratio (annualized with a $\sqrt{252}$ scaling factor over the full period), Active Days and Active Rate (the number and percentage of days with an open position), Maximum Drawdown (MDD), and cumulative final return over the simulation period.

Figure 2 illustrates the expanding-window equity growth trajectory for all four models across the four checkpoint dates underlying Table 10 (August 2024, February 2025, August 2025, and February 2026). All

four models show consistent, monotonically increasing cumulative equity across checkpoints, with no reversals in the overall upward trend. yiyanghkust/finbert-tone and nlptown/bert-base-multilingual-uncased-sentiment maintain a visible lead over ProsusAI/finbert and soleimanian/financial-roberta-large-sentiment throughout, consistent with their higher final returns reported in Table 3.

Table 3. Portfolio Performance Metrics of the Simulation (Feb 2024 - Feb 2026)

Model	Total Trade	Avg Daily Return (%)	Std Dev (%)	Sharpe Ratio	MDD (%)	Final Return (%)
ProsusAI/finbert	124	0.20	1.58	2.01	-25.03	123.33
nlptown/bert-base-multilingual-uncased-sentiment	152	0.28	1.60	2.73	-23.23	184.44
yiyanghkust/finbert-tone	159	0.26	1.55	2.63	-22.19	187.84
soleimanian/financial-roberta-large-sentiment	129	0.19	1.51	2.00	-22.19	137.19

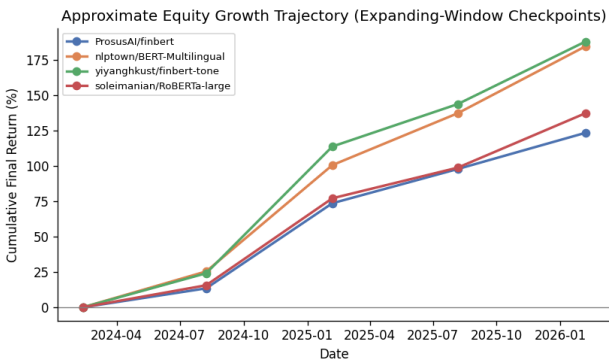


Figure 2. Approximate Equity Growth Trajectory Across Sentiment Models (Expanding-Window Checkpoints)

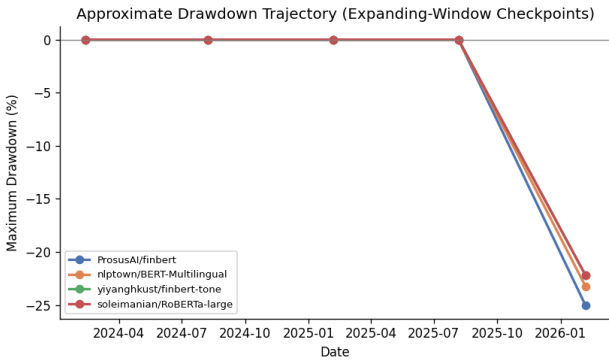


Figure 3. Approximate Drawdown Trajectory Across Sentiment Models (Expanding-Window Checkpoints)

Figure 3 illustrates the corresponding expanding-window drawdown trajectory. Notably, all four models register a Maximum Drawdown of 0% through the first three checkpoints (August 2024 to August 2025), indicating that cumulative equity remained at or above all previous peaks throughout this period; the -22% to -25% drawdown range reported in Table 3 only materializes within the final checkpoint window (August 2025 to February 2026), indicating that the observed drawdown risk is concentrated in a specific late sub-period rather than distributed evenly across the full simulation

3.3.1. Final Return, Trade Frequency, and Average Daily Return

The yiyanghkust/finbert-tone model recorded the best absolute performance with a final return of 187.84% over the 2-year simulation period, followed by nlptown/bert-base-multilingual-uncased-sentiment (184.44%), soleimanian/financial-roberta-large-sentiment (137.19%), and ProsusAI/finbert (123.33%). A consistent pattern is evident: models with a higher number of trades tend to generate greater final returns. FinBERT-Tone executed 159 trades the highest among the four models while ProsusAI/FinBERT executed

only 124 trades and recorded the lowest return. This suggests that higher signal frequency, when maintained selectively through technical filters, is capable of accumulating greater profits over the long term.

In terms of average daily return computed over the full 499-day simulation period (including idle days with zero returns), nlptown/bert-base-multilingual-uncased-sentiment recorded the highest value at 0.28%, followed by yiyanghkust/finbert-tone (0.26%), ProsusAI/finbert (0.20%), with soleimanian/financial-roberta-large-sentiment registering the lowest (0.19%). These values are naturally lower than active-trading-day averages as they incorporate idle periods, providing a more conservative and realistic picture of daily portfolio growth. Notably, positions were open on only 20–23% of all trading days (ActiveRate), indicating that the strategy is largely inactive. Cumulative final return is therefore driven both by per-trade quality and by how frequently the system successfully enters profitable positions.

The soleimanian/financial-roberta-large-sentiment model exhibits a distinct pattern: its full-period average daily return is the lowest (0.19%) despite its trade count (129) exceeding that of ProsusAI/finbert (124). This indicates that the trades generated by this RoBERTa-based model tend to yield smaller average returns per trade not merely due to high variability, given that its standard deviation (1.51%) is in fact the lowest among the four models, rather because of marginally weaker per-trade signal quality compared to the two top-performing models. Overall, the relatively narrow range of average daily returns across the four models (0.19%–0.28%) indicates that the shared technical filter framework contributes substantially to maintaining a minimum signal quality threshold across all models, such that differences in final performance are more attributable to the frequency of signals that successfully pass the filters than to drastic differences in the quality of each individual signal.

3.3.2. Sharpe Ratio - Return Quality Relative to Risk

The Sharpe Ratio is a measure of risk-adjusted return efficiency, wherein values above 1.0 are generally considered good and values above 2.0 are regarded as very strong in the empirical finance literature. In this study, the Sharpe Ratio is computed over the full simulation period including all idle trading days assigned zero returns, yielding a conservative and realistic measure of risk-adjusted performance. All four models yielded Sharpe Ratios ranging from 2.00 to 2.73, comfortably exceeding the threshold of 1.0 and indicating solid risk-adjusted performance for an equity strategy. The highest value was achieved by nlptown/bert-base-multilingual-uncased-sentiment (2.73), followed by yiyanghkust/finbert-tone (2.63). ProsusAI/finbert and soleimanian/financial-roberta-large-sentiment recorded the two lowest Sharpe Ratios (2.01 and 2.00, respectively), though for different reasons: ProsusAI/finbert's result reflects its highest standard deviation (1.58%), indicating greater return fluctuation, whereas soleimanian/financial-roberta-large-sentiment combines the lowest standard deviation (1.51%) with the lowest average daily return (0.19%), indicating consistently modest rather than volatile performance.

These Sharpe Ratio values reflect conservative full-period estimates and are directly comparable to those reported in the broader algorithmic trading literature, confirming that the strategy delivers meaningful risk-adjusted returns even when accounting for idle capital.

3.3.3. Maximum Drawdown (MDD)

yiyanghkust/finbert-tone and soleimanian/financial-roberta-large-sentiment both recorded an MDD of -22.19%, nlptown registered -23.23%, while ProsusAI/finbert registered the worst MDD at -25.03%. MDD measures the largest peak-to-trough decline on the equity curve, such that a value in the -22% to -25% range remains within acceptable bounds for equity-based strategies over a 2-year period, given that stock markets have historically experienced corrections of 20-40% within normal cycles. The similarity in MDD across models most likely indicates that the largest drawdown was experienced during the same market volatility phase in which all models were simultaneously exposed to correction risk, although confirmation of this hypothesis requires further analysis of the individual drawdown timelines of each model.

3.3.4. Transaction Cost Sensitivity Analysis

To address the limitation of assuming zero transaction costs, a sensitivity analysis was conducted to estimate the adjusted final returns under realistic cost scenarios. Based on prevailing commission structures in equity markets ranging from approximately 0.1% to 0.3% per one-way transaction (equivalent to 0.2%-0.6% per round-trip trade) for both retail and institutional brokers three cost tiers were applied uniformly across all executed trades per model. The adjusted final return was computed using the following approximation:

$$\text{Adjusted Final Return (\%)} = \text{Gross Final Return (\%)} - (N_{\text{trades}} \times \text{Round - Trip Cost \%}) \quad (9)$$

where the round-trip cost equals two times the one-way transaction fee.

Table 4. Sensitivity Analysis of Final Return Under Varying Transaction Cost Assumptions

Model	Trades	Gross Return (%)	TC = 0.1%/trade → Adj. (%)	TC = 0.2%/trade → Adj. (%)	TC = 0.3%/trade → Adj. (%)
ProsusAI/finbert	124	123.33	98.53	73.73	48.93
nlptown/bert-base-multilingual-uncased-sentiment	152	184.44	154.04	123.64	93.24
yyianghkust/finbert-tone	159	187.84	156.04	124.24	92.44
soleimanian/financial-roberta-large-sentiment	129	137.19	111.39	85.59	59.79

The results demonstrate that the profitability of all four models remains positive across all transaction cost scenarios. Under the most conservative assumption of 0.1% per one-way transaction, adjusted returns remain substantial, ranging from 98.53% to 156.04%. This indicates that the trading strategies generated by the sentiment models are sufficiently profitable to absorb modest execution costs. When transaction costs increase to 0.2% per trade, a level broadly consistent with typical retail brokerage fees in major equity markets, cumulative returns remain positive for all models, ranging between 73.73% and 124.24%. Even under the most stringent assumption of 0.3% per trade, which represents a relatively high execution cost scenario, all models continue to produce positive final returns, although with significantly reduced profitability.

It is also noteworthy that transaction costs affect models differently due to variations in trading frequency. Models with higher trade counts experience greater performance erosion as transaction costs accumulate over time. For example, yiyanghkust/finbert-tone, despite achieving the highest gross return (187.84%), executes the largest number of trades (159), which results in a slightly larger reduction in adjusted returns under higher cost scenarios. Consequently, at the 0.3% transaction cost level, nlptown/bert-base-multilingual-uncased-sentiment marginally surpasses it in adjusted performance (93.24% versus 92.44%), although yiyanghkust/finbert-tone remains marginally ahead at the 0.2% level (124.24% versus 123.64%), indicating the ranking between these two models is sensitive to the assumed cost level rather than decisively settled.

Despite this shift in ranking at higher transaction cost levels, both models consistently outperform ProsusAI/finbert and soleimanian/financial-roberta-large-sentiment across all scenarios. This suggests that the superior performance of these models is not solely driven by unrealistic zero-cost assumptions but reflects stronger predictive signals within the sentiment-based trading framework. Nevertheless, the magnitude of performance erosion is non-trivial. Under the 0.3% transaction cost scenario, gross returns decline by approximately 54–90 percentage points, depending on trading frequency. This finding highlights the practical importance of minimizing execution costs and optimizing trading frequency when deploying sentiment-based trading strategies in real-world financial markets.

3.5. Benchmark Comparison with Naive Trading Strategies

The proposed framework was benchmarked against Buy-and-Hold, RSI-only, MA-only, and Random baselines on the same 30-ticker universe and 2-year window. Table 5 summarizes portfolio-level performance; the main sentiment-filtered models are included for reference (from Table 3).

Table 5. Main Strategy vs. Naive Benchmark Strategies

Strategy	Trades	Active Rate (%)	Sharpe Ratio	MDD (%)	Final Return (%)
Main model range (Table 3)	124-159	19.6-23.1	2.00-2.73	-22.19 to -25.03	123.33-187.84
Buy-and-Hold	30	100.0	1.09	-20.65	47.95
RSI-only	3,584	94.2	7.78	-47.19	4,773.33
MA-only	928	68.1	5.40	-46.55	1,056.32
Random	3,592	99.8	7.48	-32.83	3,608.54

RSI-only and Random produce far larger summed Final Returns than the main models, but this is a direct consequence of trade frequency: Final Return is the arithmetic sum of individual trade returns, so strategies active on 94-99.8% of trading days necessarily accumulate more than one active on only 19.6-23.1% of days. The more informative comparison is risk: MDD for RSI-only and MA-only (-47.19%, -46.55%) is roughly double that of the main models (-22.19% to -25.03%).

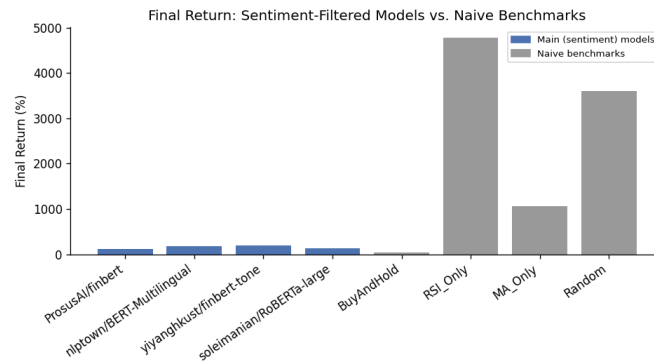


Figure 4. Final Return: Sentiment-Filtered Models vs. Naive Benchmark Strategies

Figure 4 visually compares Final Return between the sentiment-filtered models (123.33% to 187.84%) and the naive benchmark strategies introduced in Section 3.5. RSI-only (4,773.33%) and Random (3,608.54%) visibly dwarf both the sentiment-filtered models and Buy-and-Hold (47.95%) in raw cumulative terms, illustrating the trade-frequency effect discussed in that section

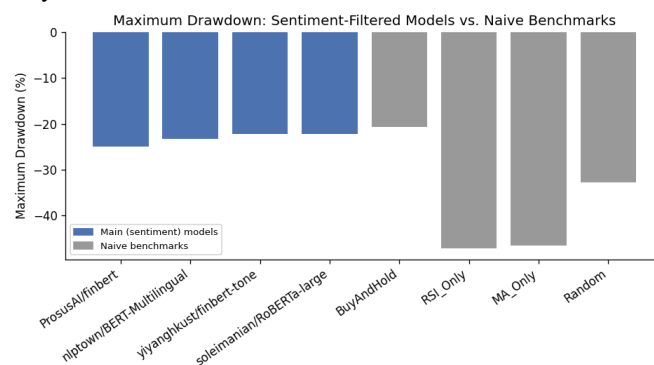


Figure 5. Maximum Drawdown: Sentiment-Filtered Models vs. Naive Benchmark Strategies

Figure 5 shows the corresponding Maximum Drawdown comparison, where the visual pattern reverses relative to Figure 4: the sentiment-filtered models (-22.19% to -25.03%) show visibly shallower drawdowns than RSI-only (-47.19%) and MA-only (-46.55%), though still deeper than Buy-and-Hold (-20.65%), reinforcing the risk-control argument developed in Section 3.5.

3.6. Robustness to Stop-Loss and Take-Profit Exit Rules

The stop-loss(-5%)/take-profit(+10%) exit variant was compared against the original profit-triggered exit. Table 6 summarizes results across all four models.

Table 6. Original Exit Rule vs. Stop-Loss/Take-Profit Variant

Model	Exit Rule	Sharpe Ratio	MDD (%)	Final Return (%)
ProsusAI/finbert	Original	2.01	-25.03	123.33
ProsusAI/finbert	SL/TP	-0.36	-27.08	2.09
nlptown/bert-base-multilingual	Original	2.73	-23.23	184.44
nlptown/bert-base-multilingual	SL/TP	0.23	-31.10	19.97
yiyanghkust/finbert-tone	Original	2.63	-22.19	187.84
yiyanghkust/finbert-tone	SL/TP	0.30	-23.23	29.81
soleimani/financial-roberta	Original	2.00	-22.19	137.19
soleimani/financial-roberta	SL/TP	0.25	-26.33	25.98

Performance deteriorates sharply under the SL/TP variant (Sharpe Ratio falling to -0.36-0.30; Final Return to 2.09%-29.81%). Exit-reason breakdowns show 21-27 trades per model closed by the stop-loss trigger trades that, under the original unlimited-horizon exit, would eventually have closed at a gain. This indicates that fixed, untuned thresholds are poorly matched to this strategy's holding-period volatility, and that the original patient exit is empirically, if not conceptually, better suited to the dataset. Optimized or adaptive stop-loss/take-profit calibration is recommended future work rather than abandoning risk-limiting exits altogether.

3.7. Refined Look-Ahead-Bias Analysis: Pre-Market, Intraday, and Post-Market News Timing

News-timestamp alignment (Section 2.4) was refined to isolate pre-market-only signals (Section 2.9 methodology). Table 7 shows the timing distribution of the full news corpus; Table 8 compares portfolio performance under the original versus stricter alignment.

Table 7. News Timing Distribution (57,434 Articles)

Timing Category	Count	Share (%)
Pre-market	23,875	41.6
Intraday	15,179	26.4
Post-market	6,339	11.0
Non-trading day (forwarded)	12,041	21.0

Table 8. Original vs. Strict (Pre-Market-Only) News Alignment

Model	Alignment	Sharpe Ratio	Final Return (%)
ProsusAI/finbert	Original	2.01	123.33
ProsusAI/finbert	Strict	1.72	114.44
nlptown/bert-base-multilingual	Original	2.73	184.44
nlptown/bert-base-multilingual	Strict	1.54	158.93
yyanghkust/finbert-tone	Original	2.63	187.84
yyanghkust/finbert-tone	Strict	2.65	175.12
soleimanian/financial-roberta	Original	2.00	137.19
soleimanian/financial-roberta	Strict	2.30	138.68

All four models remain substantially profitable with Sharpe Ratio above 1.5 under the stricter, pre-market-only alignment, indicating that the original findings are not primarily an artifact of look-ahead bias, although nlptown shows the largest sensitivity to timestamp alignment and should be interpreted with correspondingly greater caution.

3.8. Statistical Significance Testing

Pairwise statistical tests were applied to full-calendar daily returns for all six model pairs. Table 9 summarizes results. Tests agree on significance direction for five of six pairs; the pair marked “Mixed*” shows disagreement between parametric and non-parametric tests, discussed below.

Table 9. Pairwise Statistical Significance Tests Between Sentiment Models

Model Pair	t-test p	Wilcoxon p	DM test p	Significant?
ProsusAI/finbert vs. nlptown	0.024	0.010	0.023	Yes (all 3)
ProsusAI/finbert vs. yiyanghkust	0.064	0.026	0.063	Mixed*
ProsusAI/finbert vs. soleimanian	0.741	0.502	0.740	No
nlptown vs. yiyanghkust	0.546	0.591	0.545	No
nlptown vs. soleimanian	0.004	0.003	0.004	Yes (all 3)
yiyanghkust vs. soleimanian	0.006	0.007	0.006	Yes (all 3)

Three of six pairs are statistically significant across all tests at the 5% level. One pair (ProsusAI/finbert vs. yiyanghkust/finbert-tone) shows a mixed result: the non-parametric Wilcoxon signed-rank test indicates significance ($p = 0.026$), while the paired t-test and Diebold-Mariano test do not ($p = 0.064$ and $p = 0.063$) a plausible consequence of non-normality in daily returns, to which the Wilcoxon test is more robust; this pair is best regarded as inconclusive rather than clearly significant or non-significant. Critically, the difference between the two top-performing models in Section 3.3 - nlptown/bert-base-multilingual-uncased-sentiment and yiyanghkust/finbert-tone - is not significant in any of the four tests ($p = 0.546$). This tempers the original framing of the two as clearly differentiated “complementary” options; statistically, they form a joint top-performing cluster that cannot be reliably separated at this sample size, while both are significantly better than soleimanian/financial-roberta-large-sentiment.

Figure 6 plots the paired t-test p-values for all six model pairs from Table 9 against the 5% significance threshold. Three pairs fall clearly below the threshold - nlptown vs. ProsusAI/finbert ($p = 0.024$), nlptown vs. soleimanian/financial-roberta-large-sentiment ($p = 0.004$), and yiyanghkust/finbert-tone vs. soleimanian/financial-roberta-large-sentiment ($p = 0.006$) - while the remaining three pairs, including the comparison between the two top-performing models, remain above it

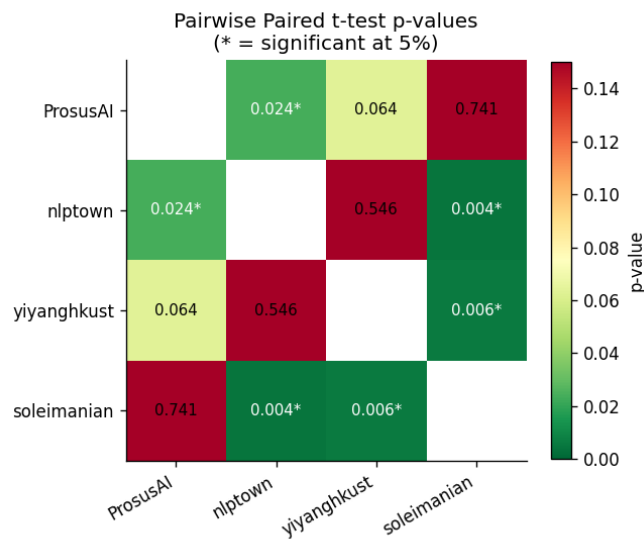


Figure 6. Pairwise Statistical Significance (Paired t-test p-values) Between Sentiment Models

3.9. Walk-Forward, Rolling-Window, Expanding-Window, and Train-Validation-Test Validation

Temporal robustness (Section 2.9) was assessed with all strategy parameters held fixed (not re-tuned per fold, since thresholds were taken from literature rather than fit to the data). Table 10 summarizes the rolling-window consistency across four folds; Table 11 summarizes the train(60%) - validation(20%) - test(20%) split.

Table 10. Rolling-Window Consistency Across Four Folds (per model)

Model	Sharpe Mean	Sharpe Std	Folds Profitable
ProsusAI/finbert	3.71	2.29	4 / 4
nlptown/bert-base-multilingual	4.29	2.28	4 / 4
yiyanghkust/finbert-tone	4.35	2.51	4 / 4
soleimanian/financial-roberta	3.57	1.93	4 / 4

Table 11. Train (60%) - Validation (20%) - Test (20%) Split (per model)

Model	Split	Sharpe Ratio	Final Return (%)
ProsusAI/finbert	Train (60%)	4.92	81.71
ProsusAI/finbert	Validation (20%)	-0.33	-9.19
ProsusAI/finbert	Test (out-of-sample)	2.29	50.82
nlptown/bert-base-multilingual	Train (60%)	5.38	105.93
nlptown/bert-base-multilingual	Validation (20%)	1.48	26.29
nlptown/bert-base-multilingual	Test (out-of-sample)	2.36	52.22
yiyanghkust/finbert-tone	Train (60%)	5.73	123.40
yiyanghkust/finbert-tone	Validation (20%)	0.72	15.28
yiyanghkust/finbert-tone	Test (out-of-sample)	2.13	49.15
soleimanian/financial-roberta	Train (60%)	4.52	81.14
soleimanian/financial-roberta	Validation (20%)	0.53	10.56
soleimanian/financial-roberta	Test (out-of-sample)	1.91	45.50

All four models are profitable in all four rolling-window folds (4 of 4) and in the most recent, strictly out-of-sample test segment (September 2025-February 2026). The intervening validation segment was notably weaker, turning slightly negative for ProsusAI/finbert (-9.19% return, Sharpe -0.33), disclosed here as an honest indication of sub-period sensitivity. Together, these results indicate that profitability is not an artifact of a single favorable evaluation window, while underscoring that returns vary meaningfully across sub-periods and market regimes.

3.10. Discussion

The 2-year simulation results demonstrate that integrating transformer-based sentiment analysis with technical filters generates trading signals that are absolutely profitable and, on the evidence of Sections 3.5-3.9, profitable for reasons that survive several independent robustness checks rather than a single favorable read of the headline numbers. All four models produced substantial positive final returns (123.33%-187.84%) with Sharpe Ratios (2.00-2.73) comfortably above the conventional threshold of 1.0. Capital was deployed on

only 20–23% of trading days, and the transaction cost sensitivity analysis (Section 3.3.4) confirms profitability up to 0.3% per trade, indicating that returns are not an artifact of a frictionless simulation.

The most fundamental factor differentiating performance across models is the proportion of positive sentiment labels each model generates. Models with a higher positive rate produce more days satisfying the “positive sentiment > 50%” criterion, triggering BUY signals more frequently: *yiyanghkust/finbert-tone* (60.4% positive) and *nlptown/bert-base-multilingual-uncased-sentiment* (61.1%) consistently generated more trades than *ProsusAI/finbert* (50.4%), reflected in their higher final returns. This increase in signal frequency does not necessarily imply degraded quality, because the multi-layered technical filters (RSI, MA20, MA50) act as gatekeepers that only permit execution when technical conditions are supportive without which higher signal frequency could devolve into noise. In the design of NLP-based trading systems, the quality of execution filters is therefore at least as important as the quality of the sentiment model itself, a conclusion reinforced by Section 3.5: RSI-only and Random baselines, which lack the sentiment filter entirely, accumulate more raw return purely through higher trade frequency, but at roughly double the drawdown risk.

A plausible behavioral and market-based rationale for why the sentiment-pullback combination is effective, rather than either signal alone, draws on several complementary mechanisms. First, market-wide absorption of news sentiment is often gradual rather than instantaneous, so a pullback identified by RSI and moving averages after positive news may capture a window in which price has not yet fully adjusted to the sentiment signal. Second, short-term overreactions or underreactions in price around news events can produce the very pullback-then-recovery pattern the technical filters are designed to detect, independent of the specific mechanism driving them. Third, institutional and retail investors plausibly differ in reaction timing: faster-moving participants may respond to news sentiment quickly but scale into positions gradually, while slower-moving participants tend to react with a lag, collectively producing a delayed, momentum-like price adjustment that the entry rule is positioned to exploit. These mechanisms are offered as a plausible conceptual explanation rather than an empirically isolated one; disentangling their relative contributions would require order-book-level or investor-level data beyond the scope of this study.

The second dimension differentiating model performance is the consistency of signal quality, reflected in the standard deviation of full-period daily returns. *ProsusAI/finbert* exhibited the highest standard deviation (1.58%), indicating more variable trade quality, while *yiyanghkust/finbert-tone* (1.55%) and *nlptown/bert-base-multilingual-uncased-sentiment* (1.61%) showed competitive volatility levels that contributed to their higher Sharpe Ratios. Section 3.8 formalizes this observation: *nlptown* significantly outperforms both *ProsusAI/finbert* ($p = 0.024$) and *soleimanian/financial-roberta-large-sentiment* ($p = 0.004$), and *yiyanghkust/finbert-tone* significantly outperforms *soleimanian* ($p = 0.006$), so the descriptive volatility gap between the top two models and *soleimanian* corresponds to a statistically defensible difference rather than a numerical coincidence.

From an architecture and training-domain perspective, the four models span a broad spectrum. *ProsusAI/finbert* and *yiyanghkust/finbert-tone* is both BERT-based and trained on the financial domain, yet with distinct emphases *yiyanghkust/finbert-tone* places greater weight on corporate-communication tone, making it more sensitive to subtle positive nuances in formal financial language. Particularly noteworthy is the competitive performance of *nlptown/bert-base-multilingual-uncased-sentiment*, a generic multilingual model not specifically designed for the financial domain, which nonetheless matched or exceeded the domain-specific models on Sharpe Ratio and average daily return. This is consistent with Mishev *et al.* (2020), who found that generic models with strong contextual representations do not always underperform domain-specific models, and that performance is highly dependent on task characteristics and data distribution [43]. This affirms that the accessibility and computational efficiency of generic models make them a practically viable option for developers under resource constraints. This interpretation should nonetheless be treated as provisional: *nlptown*’s advantage may equally stem from a higher tendency toward positive classification or less conservative labeling behavior (Table 2 shows it produced the highest positive-label rate, 61.1%) rather than from superior capture of general linguistic patterns, and the two explanations cannot be fully disentangled with the present evidence. *soleimanian/financial-roberta-large-sentiment*, despite possessing the largest architecture among the four, did not yield the highest performance along any dimension reinforcing that model scale is not the sole determinant of trading-signal quality; the alignment between a model’s sentiment distribution and the filtering mechanism is more decisive.

The original evaluation framed *yiyanghkust/finbert-tone* and *nlptown/bert-base-multilingual-uncased-sentiment* as complementary the former leading on final return and trade frequency, the latter on risk-adjusted efficiency. Section 3.8 shows this framing needs qualification: the difference between the two is not statistically significant in any of the four tests applied ($p = 0.546$), so rather than two models optimized for different investor priorities, the more defensible reading is a single joint top-performing tier that cannot be reliably separated at this sample size, with both significantly ahead of *soleimanian/financial-roberta-large-sentiment*. Practitioners

choosing between the two should not expect a reliable performance edge in either direction based on this evidence alone.

Read together with Sections 3.5-3.9, the central contribution of this framework is best characterized as disciplined, risk-controlled signal selectivity rather than raw return maximization: Section 3.5 shows that naive higher-frequency baselines mechanically accumulate larger summed returns but at roughly double the Maximum Drawdown. Section 3.6 shows that this selectivity is not trivially improvable by bolting on fixed stop-loss/take-profit thresholds, which substantially degraded performance rather than improving it — the patient, profit-triggered exit is empirically, if not conceptually, well matched to this dataset. Section 3.7 shows the results are not primarily an artifact of look-ahead bias, remaining qualitatively robust (Sharpe Ratio > 1.5 for all models) under a stricter pre-market-only news alignment. Section 3.9 shows profitability holds across independent rolling-window folds and in the most recent out-of-sample test segment, while also revealing sub-period variability (including a slightly negative validation-period result for ProsusAI/finbert) that argues for cautious rather than absolute performance claims. Taken as a whole, these five analyses do not overturn the original findings, but they substantially refine their interpretation: the framework's value lies in consistent, risk-disciplined signal generation validated across multiple robustness dimensions, rather than in the outright maximization of cumulative return implied by the original headline figures alone. Remaining boundary conditions the absence of bid-ask spread, slippage, and market-impact modeling, and the lack of human-annotated sentiment ground truth are addressed in the Limitation section below.

4. CONCLUSION AND LIMITATION

This study demonstrates that a sentiment analysis approach based on pre-trained transformer models, combined with technical filters (RSI, MA20, MA50), is capable of consistently generating profitable trading signals across all four models evaluated during the simulation period spanning February 12, 2024 to February 6, 2026. The yiyanghkust/finbert-tone model recorded the best absolute performance with a final return of 187.84% and the highest trade count (159 trades), while nlptown/bert-base-multilingual-uncased-sentiment excelled in risk-adjusted efficiency with a full-period Sharpe Ratio of 2.73. Despite performance differences across models, all four collectively produced significant positive cumulative returns (123%-188%), indicating that the integration of news sentiment signals into a technical filter framework possesses meaningful predictive value. The primary factors driving performance differences across models are the proportion of positive sentiment labels generated, the consistency of signal quality as reflected by the standard deviation of returns, and the alignment of each model's training domain with the linguistic characteristics of the financial news employed. These findings collectively support the hypothesis that news sentiment can serve as an effective complement to technical analysis in data-driven investment decision-making. A subsequent validation round (Sections 3.5-3.9) shows that this contribution is best characterized as disciplined, risk-controlled signal selectivity - the framework trades roughly half the Maximum Drawdown of naive high-frequency baselines (Buy-and-Hold, RSI-only, MA-only, and Random) in exchange for lower raw cumulative return rather than outright return maximization, and that the difference between the two top-performing models (yiyanghkust/finbert-tone and nlptown/bert-base-multilingual-uncased-sentiment) is not statistically significant ($p = 0.546$), so both should be regarded as a joint top tier rather than a single winner. Results remain qualitatively robust under a stricter look-ahead-bias alignment and across rolling-window, expanding-window, and out-of-sample test periods, though with fold-to-fold variability that warrants cautious, rather than absolute, performance claims. Accordingly, these findings should be read as correlational evidence of an effective association between news sentiment and profitable signal generation within this specific simulated backtesting environment, rather than as proof of the universal effectiveness of news sentiment as a complementary trading signal across markets, asset classes, or time periods beyond those studied here.

Although the simulation results demonstrate promising performance, several methodological limitations warrant acknowledgment. First, while a transaction cost sensitivity analysis was conducted across three cost tiers (0.1%-0.3% per trade), the simulation still does not incorporate bid-ask spreads, market impact, or slippage, which are particularly relevant for less liquid stocks. The reported gross returns therefore remain optimistic relative to fully realistic execution conditions. Second, the exit strategy employed is overly simplistic, relying solely on a single closing price condition without explicit stop-loss or take-profit mechanisms, such that per-trade risk management does not reflect genuine portfolio practice. Third, with respect to news timestamp alignment, this study assigns all news articles to the trading day on which they are published, without distinguishing between articles published before and after market close on that day. As a result, articles published after market close on day i are included in the sentiment signal that drives trade execution on day $i+1$, which is appropriate; however, articles published during trading hours on day i are also aggregated into the same daily signal and treated identically. This approach introduces a minor potential for intraday look-ahead bias that is not fully resolved, and future work should incorporate intraday publication timestamps where available to assign each article precisely to the next executable trading session. Fourth, the technical filter parameters (RSI < 40, pullback 0.5-5% below MA20, price within 10% above MA50) were

established based on established literature conventions and applied uniformly across all 30 tickers without per-asset optimization or validation on a held-out time period. This design choice favors generalizability over individual fit but does not rule out the risk that these thresholds inadvertently align well with the specific characteristics of the 2024-2026 NASDAQ market environment. Walk-forward validation or cross-validation by sub-period would strengthen confidence in parameter robustness and is recommended as a priority for future research. Fifth, the sentiment labels generated by the models were not validated against human annotation ground truth nor against actual post-news price responses, such that the semantic accuracy of the models within the market context studied remains independently unmeasured.

4.1. Future Directions

Based on these limitations, future research is recommended to implement structured stop-loss and take-profit mechanisms, validate parameter robustness through walk-forward analysis across multiple time periods and market regimes, incorporate intraday news timestamp alignment, explore model ensemble approaches to combine the comparative advantages of each individual model, and apply statistical significance tests such as the Diebold-Mariano test to formally assess whether performance differences across models are statistically meaningful rather than attributable to chance.

Several of the limitations above have since been addressed empirically (Sections 3.5-3.9) and are summarized here rather than removed, so the review trail remains transparent. A stop-loss/take-profit exit-rule variant was tested and found to substantially reduce performance relative to the original profit-triggered exit, indicating that the original design, while simple, is not straightforwardly improved by untuned fixed thresholds; optimized or adaptive stop-loss/take-profit calibration remains recommended future work. News timestamp alignment was refined to separate pre-market, intraday, and post-market publication times, and results were shown to remain qualitatively robust (Sharpe Ratio > 1.5 for all models) under the stricter pre-market-only alignment. Walk-forward robustness was assessed via rolling-window, expanding-window, and train-validation-test splits with parameters held fixed (rather than re-tuned per fold, since the thresholds were not fit to the data in the first instance); all four models remained profitable across folds and in the most recent out-of-sample test segment, with acknowledged sub-period variability. Pairwise Wilcoxon signed-rank, paired t-test, and Diebold-Mariano tests were applied across all six model pairs; three of six differences are statistically significant at the 5% level, while the difference between the two top-performing models is not, tempering the claim of a single best-performing model. A benchmark comparison against Buy-and-Hold, RSI-only, MA-only, and Random baselines further clarified that the framework's contribution lies in risk-controlled selectivity substantially lower Maximum Drawdown rather than in maximizing raw cumulative return, which naive high-frequency baselines can mechanically exceed simply by trading far more often. Remaining unaddressed limitations include the absence of bid-ask spread, slippage, and market-impact modeling, and the lack of human-annotated sentiment ground truth, both noted above as priorities for future research.

4.2. Data Availability Statement

The news dataset used in this study was collected via the Polygon.io API, and stock price data were obtained via the yfinance Python library. Due to the licensing terms of the Polygon.io API, the raw news dataset cannot be redistributed publicly; however, the data can be retrieved by other researchers using the same API endpoints and the stock tickers listed in Section 2.1.1, for the observation period specified in Section 2.1.2 and 2.1.3. Processed sentiment scores, technical indicator computations, and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- [1] B. Ihnaini, "Stock trend prediction using sentiment analysis," *Peerj Computer Science*, vol. 9, pp. e1293, 2023. <https://10.7717/peerj-cs.1293>
- [2] R. A. Yaqin, M. I. Anshori, R. Angel, I. W. P. Agung, T. Arifin, and E. Junianto, "Stock Price Forecasting Using LSTM with Cross-Validation," *Vokasi Unesa Bulletin of Engineering Technology and Applied Science*, vol. 3, no. 1, pp. 64–79, 2026. <https://10.26740/vubeta.v3i1.45130>
- [3] K. R. Dahal *et al.*, "A comparative study on effect of news sentiment on stock price prediction with deep learning architecture," *Plos One*, vol. 18, no. 4, pp. e0284695, 2023. <https://10.1371/journal.pone.0284695>
- [4] J.-H. Kim, H.-S. Kim, and S. Choi, "Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM," *Axioms*, vol. 12, no. 9, pp. 835, 2023. <https://10.3390/axioms12090835>
- [5] H. Jiang, "Stock Price Prediction Using FinBERT-Enhanced Sentiment with SHAP Explainability and Differential Privacy," *Mathematics*, vol. 13, no. 17, pp. 2747, 2025. <https://10.3390/math13172747>
- [6] S. D. Mehra and S. D. Shetty, "Developing and testing a custom algorithmic trading strategy using exponential moving average, relative strength index, and sentiment analysis," *Journal of Autonomous Intelligence*, vol. 7, no. 4, 2024. <https://10.32629/jai.v7i4.1328>

- [7] Y. Mun and N. Kim, “Leveraging Large Language Models for Sentiment Analysis and Investment Strategy Development in Financial Markets,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 2, pp. 77, 2025. <https://10.3390/jtaer20020077>
- [8] Y. Kim and S. Park, “A Rule-Based Stock Trading Recommendation System Using Sentiment Analysis and Technical Indicators,” *Electronics*, vol. 14, no. 4, pp. 773, 2025. <https://10.3390/electronics14040773>
- [9] O. Shobayo, S. Adeyemi-longe, O. Popoola, and B. Ogunleye, “Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic Regression: A Data-Driven Approach,” *Big Data and Cognitive Computing*, vol. 8, no. 11, pp. 143, 2024. <https://10.3390/bdcc8110143>
- [10] R. Gómez-Martínez, M. L. Medrano-García, D. López-López, and J. Torres-Pruñonosa, “How Sentiment Indicators Improve Algorithmic Trading Performance,” *Sage Open*, vol. 15, no. 3, 2025. <https://10.1177/21582440251369559>
- [11] W. Chen, W. Liu, and X. Zhang, “Leveraging large language model as news sentiment predictor in stock markets: a knowledge-enhanced strategy,” *Discover Computing*, vol. 28, no. 1, 2025. <https://10.1007/s10791-025-09573-7>
- [12] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” *ArXiv Preprint*, 2019. <https://10.48550/arxiv.1908.10063>
- [13] N. L. P. Town, “bert-base-multilingual-uncased-sentiment (Revision edd66ab),” *Hugging Face*, 2023. <https://10.57967/hf/1515>
- [14] A. H. Huang, H. Wang, and Y. Yang, “FinBERT: A Large Language Model for Extracting Information from Financial Text*,” *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023. <https://10.1111/1911-3846.12832>
- [15] P. Huang, Y. Ni, M. Y. Day, and Y. Chen, “Enhancing Investment Profitability: Study on Contrarian Technical Strategies in Brent Crude Oil Markets,” *Energies (Basel)*, vol. 18, no. 11, 2025. <https://10.3390/en18112735>
- [16] G. Cohen, “Polynomial Moving Regression Band Stocks Trading System,” *Risks*, vol. 12, no. 10, 2024. <https://10.3390/risks12100166>
- [17] M. P. Cristescu, C. Brândaş, D. A. Mara, and P. Ioana, “Fine-Tuning and Explaining FinBERT for Sector-Specific Financial News: A Reproducible Workflow,” *Electronics (Switzerland)*, vol. 14, no. 23, 2025. <https://10.3390/electronics14234680>
- [18] H. H. Htun, M. Biehl, and N. Petkov, “Survey of feature selection and extraction techniques for stock market prediction,” *Springer Science and Business Media Deutschland GmbH*, 2023. <https://10.1186/s40854-022-00441-7>
- [19] Y. Z. Chen, Y. Fang, X. Y. Li, and J. Wei, “A factor pricing model based on double moving average strategy,” *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, 2023. <https://10.1057/s41599-023-02362-x>
- [20] M. N. Ramaube, I. O’leary-Govender, and J. Bolton, “The Profitability of Technical Analysis During Volatile Periods in The South African Financial Markets,” *INTERNATIONAL JOURNAL OF ECONOMICS AND FINANCE STUDIES*, vol. 15, no. 02, pp. 59–89, 2023. <https://10.34109/ijefs>
- [21] J. Li, X. Wang, S. Ahmad, X. Huang, and Y. A. Khan, “Optimization of investment strategies through machine learning,” *Heliyon*, vol. 9, no. 5, 2023. <https://10.1016/j.heliyon.2023.e16155>
- [22] Y. Lin and B. Liu, “A Framework for Enhancing Stock Investment Performance by Predicting Important Trading Points with Return-Adaptive Piecewise Linear Representation and Batch Attention Multi-Scale Convolutional Recurrent Neural Network,” *Entropy*, vol. 25, no. 11, 2023. <https://10.3390/e25111500>
- [23] O. Mengshetti, K. Gupta, N. Zade, K. Kotecha, S. Mutha, and G. Joshi, “Synergizing quantitative finance models and market microstructure analysis for enhanced algorithmic trading strategies,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 3, 2024. <https://10.1016/j.joitmc.2024.100334>
- [24] K. Saetia and J. Yokrattanasak, “Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand,” *International Journal of Financial Studies*, vol. 11, no. 1, 2023. <https://10.3390/ijfs11010005>
- [25] A. Agudelo-Aguirre, N. Duque-Méndez, and A. Galvis-Flórez, “Artificial Intelligence in Stock Market Investment Through the RSI Indicator,” *Computers*, vol. 14, no. 11, 2025. <https://10.3390/computers14110487>
- [26] Q. Xiao and B. Ihnaini, “Stock trend prediction using sentiment analysis,” *PeerJ Comput. Science*, vol. 9, 2023. <https://10.7717/PEERJ-CS.1293>
- [27] Z. Chang and Z. Zhang, “Judging Stock Trends According to the Sentiments of Stock Comments in Expert Forums,” *Electronics (Switzerland)*, vol. 12, no. 3, 2023. <https://10.3390/electronics12030722>
- [28] M. Davidovic and J. McCleary, “News Sentiment and Stock Market Dynamics: A Machine Learning Investigation,” *Journal of Risk and Financial Management*, vol. 18, no. 8, 2025. <https://10.3390/jrfm18080412>
- [29] X. Zhong and D. Enke, “Predicting the daily return direction of the stock market using hybrid machine learning algorithms,” *Financial Innovation*, vol. 5, no. 1, 2019. <https://10.1186/s40854-019-0138-0>
- [30] R. Bhowmik and S. Wang, “Stock market volatility and return analysis: A systematic literature review,” *MDPI AG*, 2020. <https://10.3390/E22050522>
- [31] V. Azhmyakov, I. Shirokov, and L. A. Guzman Trujillo, “Advanced Statistical Analysis of the Predicted Volatility Levels in Crypto Markets,” *Journal of Risk and Financial Management*, vol. 17, no. 7, 2024. <https://10.3390/jrfm17070279>
- [32] M. Lolic, “Tree-Based Methods of Volatility Prediction for the S&P 500 Index,” *Computation*, vol. 13, no. 4, 2025. <https://10.3390/computation13040084>

- [33] K. Yan and Y. Li, "Machine learning-based analysis of volatility quantitative investment strategies for American financial stocks," *Quantitative Finance and Economics*, vol. 8, no. 2, pp. 364–386, 2024. <https://10.3934/QFE.2024014>
- [34] A. Bánya, T. Tatay, G. Thalmeiner, and L. Pataki, "The Impact of Rebalancing Strategies on ETF Portfolio Performance," *Journal of Risk and Financial Management*, vol. 17, no. 12, 2024. <https://10.3390/jrfm17120533>
- [35] S. Shahsafi and F. Naderkhani, "A Hybrid Framework for Multi-Stock Trading: Deep Q-Networks with Portfolio Optimization," *Journal of Risk and Financial Management*, vol. 19, no. 2, 2026. <https://10.3390/jrfm19020132>
- [36] D. K. Malhotra, T. Mooney, R. Poteau, and P. Russel, "Assessing the Performance and Risk-Adjusted Returns of Financial Mutual Funds," *International Journal of Financial Studies*, vol. 11, no. 4, 2023. <https://10.3390/ijfs11040136>
- [37] H. Geboers and B. Depaire, "A Rational Risk Policy? Why Path Dependence Matters," *Entropy*, vol. 25, no. 2, 2023. <https://10.3390/e25020202>
- [38] E. Jondeau and A. Pauli, "Large Drawdowns and Long-Term Asset Management," *Journal of Risk and Financial Management*, vol. 17, no. 12, 2024. <https://10.3390/jrfm17120552>
- [39] C. K. Paik, J. Choi, and I. U. Vaquero, "Algorithm-Based Low-Frequency Trading Using a Stochastic Oscillator, Williams%R, and Trading Volume for the S&P 500," *Journal of Risk and Financial Management*, vol. 17, no. 11, 2024. <https://10.3390/jrfm17110501>
- [40] M. Tran, D. Pham-Hi, and M. Bui, "Optimizing Automated Trading Systems with Deep Reinforcement Learning," *Algorithms*, vol. 16, no. 1, 2023. <https://10.3390/a16010023>
- [41] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995. <https://10.1080/07350015.1995.10524599>
- [42] K. Hernández-Romo, J. Lemus-Romani, E. Vega, M. Becerra-Rozas, and A. Romo, "Robust Metaheuristic Optimization for Algorithmic Trading: A Comparative Study of Optimization Techniques," *Mathematics*, vol. 14, no. 1, 2026. <https://10.3390/math14010069>
- [43] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020. <https://10.1109/ACCESS.2020.3009626>

BIOGRAPHIES OF AUTHORS



Rifki Ainul Yaqin    is a bachelor's student in the Informatics Study Program, Faculty of Engineering at Adhirajasa Reswara Sanjaya (ARS) University, and works as a research assistant at ARS Digital Research and Innovation (ADRI). Previously, participated in research focused on two studies "Penerapan Arsitektur Monolitik Pada Aplikasi Jasa Service Online Tekku Berbasis Web" and "Klasifikasi Jenis Jerawat Secara Otomatis Dengan Convolutional Neural Network Menggunakan Arsitektur Resnet-50". His research interests include machine learning, deep learning, and large language models (LLMs). He can be contacted at email: rifkiainul17@gmail.com.



Ignatius Wiseto Prasetyo Agung    After retired from PT. Telkom Indonesia, he is now dedicated his time in the ARS (Adhirajasa Reswara Sanjaya) University Bandung, Indonesia, as a lecturer and Vice Rector for Collaboration & Innovation, since October 2019. In Telkom Indonesia, he worked since 1988 in various divisions e.g satellite development, network operation, R&D, and Digital Business. He received the Sarjana (Bachelor Degree) in Telecommunication from Institut Teknologi Bandung, Indonesia in 1987. He was also graduated from University of Surrey, UK and received the MSc in Telematics (1994) and PhD in Multimedia Communication (2002). He was also in charge in several professional forums, for instance the Asia Pacific Telecommunity Wireless Forum (AWF) as Convergence Working Group Chairman (2008- 2011); in ITU-D as Vice Rapporteur (2007-2009); as Chairman (2020, 2021) and Vice Chair (2018-2019) of IEEE Communications Society Indonesia Chapter; and as Senior Member of IEEE. He can be contacted at email: wiseto.agung@ars.ac.id.



Toni Arifin    He is a member of the Faculty of Engineering, majoring in Informatics Engineering, Adhirajasa Reswara Sanjaya (ARS) University, and researcher ARS Digital Research & Innovation (ADRI). He received his Bachelor's Degree in Informatics Engineering from Bina Sarana Informatika University in 2013 and graduated from the computer science master's program at Nusa Mandiri University Jakarta in 2015. He has authored or coauthored more than 68 publications: 4 proceedings and 66 journals, with 12 H-index and more than 528 citations. Research interests include machine learning, image processing and deep learning. He can be contacted at email: toni.arifin@ars.ac.id.



Erfian Junianto    He is a member of the Faculty of Engineering, majoring in Informatics Engineering, at Adhirajasa Reswara Sanjaya (ARS) University, and a researcher at ARS Digital Research & Innovation (ADRI). He graduated from the computer science master's program at Nusa Mandiri University Jakarta in 2014. He has authored or co-authored more than 38 publications, including 2 proceedings and 36 journals, with an H-index of 10 and more than 450 citations. His research interests include text mining, artificial intelligence, and classification. He can be contacted at email: erfian.ejn@ars.ac.id.