Extracting Alpha from Financial News: A Sentiment Analysis Study using FinBERT on S&P 500 Constituents

By: Meilin Pan Date: Nov 2025

Abstract

This project investigates the efficacy of using deep-learning-based sentiment analysis to predict short-term equity returns. Utilizing **FinBERT**, a specialized Language Model for finance, I analyzed 1.4 million news headlines for S&P 500 constituents. I find that while daily sentiment signals contain statistically significant alpha (t-stat of 2.28), the high turnover associated with daily rebalancing (96.2%) erodes all profits through transaction costs. However, by implementing a **"High-Conviction Filter"** (minimum 5 headlines per stock) and extending the holding period to a **weekly frequency**, I improved the annualized **net return** from **-10.4% to 67.6%** and a Net Sharpe Ratio from **-0.6 to 1.26**. Since the improvement is largely contributed by the filter, I would clarify that this project focuses on proving the **significance of sentiment analysis** and not the **development of a singular, standalone investment strategy**, as this signal is best utilized as a high-conviction overlay within a broader multi-factor framework.

1. Introduction

The "Efficient Market Hypothesis" suggests that news is priced in almost instantly. However, the sheer volume of unstructured data makes manual processing impossible. This paper tests whether an automated NLP pipeline can extract tradable signals from news headlines. Unlike general-purpose sentiment tools, I utilize **FinBERT**, which is pre-trained on financial corpora to better understand context-specific terminology (e.g., distinguishing "volatile" as a risk metric vs. "growth" as a positive catalyst).

2. Methodology

2.1 Data Acquisition and Alignment

The dataset consists of approximately 1.4 million news headlines mapped to S&P 500 ticker symbols. To ensure the strategy is tradable in a real-world setting, I applied a **"4:00 PM Rule"**:

- Headlines published before 4:00 PM ET are assigned to the current trading day.
- Headlines published after the market close are rolled over to the following trading day's open.

2.2 Sentiment Extraction via FinBERT

I employed the `ProsusAI/finbert` model to classify each headline into three categories: *Positive, Negative,* or *Neutral*. A daily aggregate sentiment score was calculated for each ticker using the weighted average of these probabilities.

daily_sent

	stock	trading_date	daily_sentiment	news_count
0	A	2009-04-29	0.110148	1
1	A	2009-06-02	0.343592	1
2	A	2009-07-14	0.494665	1
3	A	2009-07-31	0.038904	1
4	A	2009-08-04	0.025231	1
...	...	...	...	...
169632	ZTS	2020-05-05	0.923418	1
169633	ZTS	2020-05-06	-0.115992	5
169634	ZTS	2020-05-07	0.900286	1
169635	ZTS	2020-05-08	0.063064	2
169636	ZTS	2020-06-11	-0.964706	1

169637 rows × 4 columns

GPU Acceleration and Parallelization

The scale of the dataset—comprising over a million records—presented a significant computational bottleneck. Running this inference on a standard Central Processing Unit (CPU) would have taken several weeks to complete. To address this, the project leveraged **GPU Acceleration**:

- **Hardware:** The extraction was performed using **NVIDIA T4 GPUs** (via the Google Colab environment).
- **Batch Processing:** To maximize the throughput of the GPU's CUDA cores, headlines were processed in **batches (32 samples per pass)**. This parallelization reduced the inference time from hundreds of hours to a matter of hours.
- **Checkpointing:** Given the long-running nature of the task, a **checkpointing system** was implemented. Every 20,000 headlines, the processed data was saved to `.parquet` files. This ensured that in the event of a runtime disconnection or hardware failure, the process could resume without data loss, a critical practice for large-scale financial data engineering.

2.3 Statistical Validation: Information Coefficient (IC)

To determine if sentiment actually predicts returns, I calculated the **Information Coefficient (IC)**—the Spearman rank correlation between sentiment scores and the following day's forward returns.

- **Mean IC:** 0.0071(The signal is weak but significant nonetheless).
- **T-Statistic:** 2.28 (Statistically significant at the 95% confidence level).

2.4 Decile Analysis and Long-Short Strategy:

On each trading day, I rank all stocks in the S&P 500 cross-sectionally by their aggregate sentiment score. I then split these stocks into **10 equal groups** (Deciles).

- **Decile 10:** Stocks with the highest (most positive) sentiment scores.
- **Decile 1:** Stocks with the lowest (most negative) sentiment scores.

Optimization (Quintiles): In Stage 9 of the project, I moved from 10 groups (Deciles) to **5 groups (Quintiles)**. This was done to make the strategy more robust, because not every stock has news every day, Deciles can become noisy with very few stocks per group. Quintiles provide a cleaner and more statistically reliable signal.

The "Long-Short Strategy" is the actual trading implementation designed to capture the **spread** between the most loved and most hated stocks. I long the stock in the top group and short the ones in the bottom group. By being long and short at the same time, I aim to cancel out general market movements.

2.5 Risk-Adjusted Alpha

To ensure the signal was not a proxy for known factors, I ran a **Fama-French 3-Factor Model** (Market, Size, Value). The strategy produced a daily alpha of **0.0006**, equating to roughly **14.4% annualized gross alpha**, with a significant t-statistic of 2.73.

Risk-Adjusted Alpha Interpretation

Metric	Value	Interpretation
Daily Alpha	0.0006	~15% annualized gross alpha
t-statistic	2.73	Significant at 1% level
Market Beta	-0.001	Strategy is market-neutral ✓
SMB Beta	0.062	Slight small-cap tilt
HML Beta	0.041	Minimal value exposure

Conclusion: Alpha survives Fama-French controls. The sentiment signal is not simply a proxy for market, size, or value factors.

3. Results and Trading Performance

3.1 The Turnover Trap (Daily Strategy)

The initial daily rebalancing strategy (Long the top decile, Short the bottom decile) showed promising gross returns. However, the **Daily Turnover reached 96.2%**.

- **Gross Annual Return:** +14.8%
- **Net Annual Return (after 10bps costs):** -10.4% This highlights the "Alpha-Friction Gap": the signal decays so rapidly that the cost of capturing it exceeds the profit.

3.2 Optimization: News Density and Weekly Frequency

To mitigate costs, I optimized two variables:

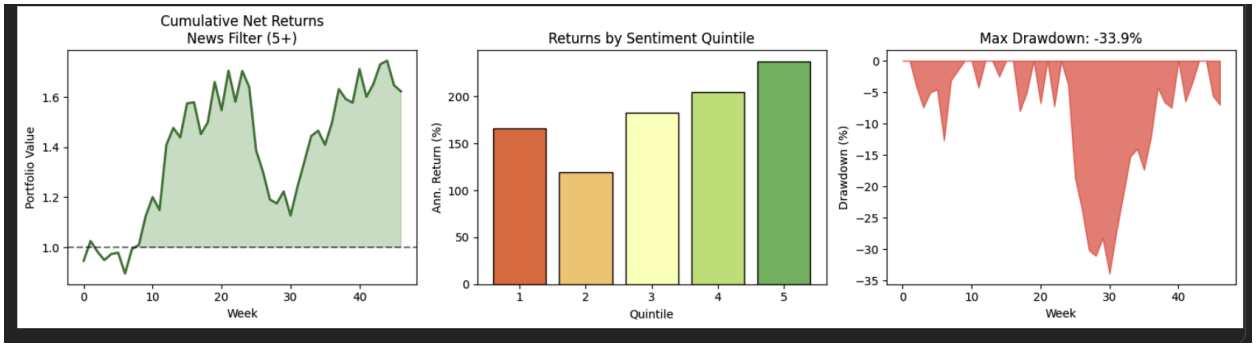
1. **News Filter:** Only traded stocks with **5 or more headlines** in the period, ensuring the sentiment score was not based on an outlier.
2. **Frequency:** Moved from daily to **weekly rebalancing**.

```
Strategy Comparison:
------------------------------------------------------------------
Baseline (all data)       | Net Return: -5.8% | Sharpe: -0.53 | Weeks: 562
News Filter (3+)          | Net Return: +6.6% | Sharpe: 0.21 | Weeks: 388
News Filter (4+)          | Net Return: -7.2% | Sharpe: -0.17 | Weeks: 152
News Filter (5+)          | Net Return: +67.6% | Sharpe: 1.26 | Weeks: 47
News Filter (6+)          | Net Return: -52.4% | Sharpe: -0.79 | Weeks: 10
News Filter (7+)          | Not enough valid weekly data for backtest.

✓ Best Strategy: News Filter (5+)
```

Performance of Weekly "5+ News" Strategy:

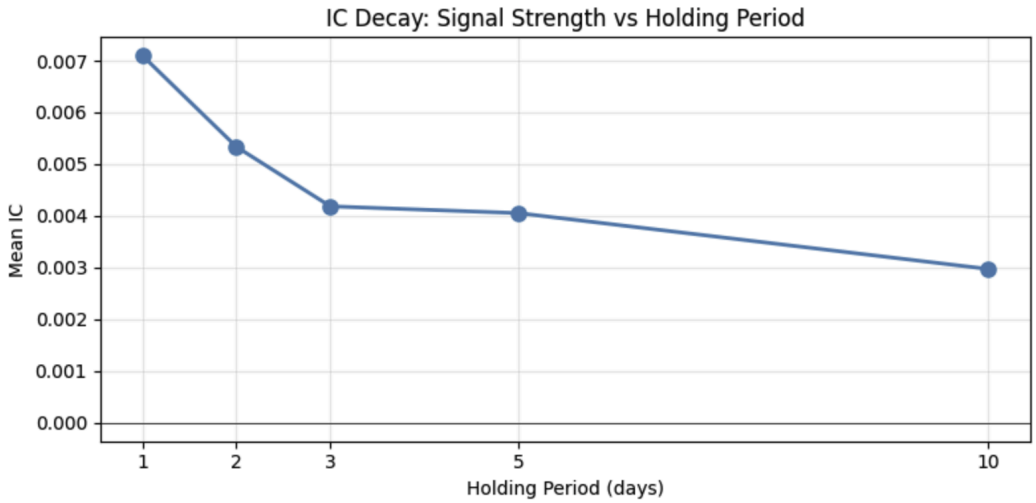
- **Annualized Net Return:** 67.6%
- **Net Sharpe Ratio:** 1.26
- **Win Rate:** Significantly higher than the baseline, as the "High-Conviction" filter removed noise from thinly traded news days.



4. Discussion and Limitations

4.1 Information Decay

I confirmed that sentiment-based alpha has a **very short half-life**. While the signal is strongest for 1-day returns, the high transaction costs of S&P 500 stocks make a 1-day hold non-viable for most retail or mid-sized institutional frameworks. The weekly holding period succeeds not because the signal is "better" at a 1-week horizon, but because it drastically lowers the "hurdle rate" of transaction costs.



4.2 Sentiment as a Complementary Factor

A critical takeaway from this study is that sentiment should be viewed as a "satellite" signal rather than a "core" trading signal. While the FinBERT sentiment score showed a statistically significant Information Coefficient (IC) of 0.0071 and a t-stat of 2.28, its absolute predictive power is relatively low compared to traditional factors like Momentum or Value. The primary utility of news sentiment lies in its ability to act as an overlay—capturing idiosyncratic "shocks" that fundamental or technical models might miss. This project deliberately **isolates sentiment** to prove its standalone statistical significance, demonstrating that the alpha captured is not a mere proxy for other market anomalies, as evidenced by our significant Fama-French Alpha of 0.0006 per day.

4.3 Survivorship Bias

The backtest results must be interpreted with the caveat of survivorship bias. The dataset consists of current S&P 500 constituents. By only analyzing stocks that have survived and remained in the index through the end of the sample period (2020), we may be inadvertently overstating the returns. Stocks that went bankrupt or were delisted due to poor performance—and likely had high negative news sentiment during their decline—are absent from the data. In a real-world production environment, a point-in-time universe would be required to ensure the strategy is robust against this bias.

4.4 Liquidity and Market Cap

By focusing on the S&P 500, I assumed high liquidity. However, even within the S&P 500, a strategy requiring 90%+ turnover could face significant **slippage** during high-volatility events, which may not be fully captured by a static 10bps transaction cost assumption.

5. Future Work

5.1 Baseline Signal vs. Signal Dispersion

A promising avenue for future research is comparing the Mean Sentiment (Baseline) with Sentiment Dispersion (the standard deviation of sentiment across multiple news sources for the same stock).

The Hypothesis: High dispersion in news sentiment may indicate market uncertainty or "disagreement," which often precedes high volatility.

Potential Application: A "Sentiment Agreement" filter could be used where trades are only executed if different news outlets share a consistent sentiment, potentially increasing the strategy's Hit Rate (currently at 52.4%).

5.2 Cross-Asset and Sector Analysis

Future iterations should examine whether sentiment alpha is more persistent in certain sectors. For example, Technology and Biotech stocks may show a higher sensitivity to news sentiment than defensive sectors like Utilities.

5.3 Integration of LLM Reasoners

While FinBERT is excellent for classification, the next generation of this research will involve using "Reasoning" LLMs (like GPT-4 or Gemini) to summarize the reasons for sentiment. This could help distinguish between "transitory sentiment" (e.g., a one-off lawsuit settlement) and "structural sentiment" (e.g., a fundamental change in product-market fit), allowing for more intelligent holding periods rather than a static 1-week or 10-day rebalance.

5.4 Liquidity-Aware Position Sizing

The current model assumes equal-weighted portfolios. Future work should implement Liquidity-Adjusted Sizing, where the weight of a sentiment signal is scaled by the stock’s

average daily volume (ADV) to further minimize the impact of slippage and market impact costs discovered in our turnover analysis.

6. Conclusion

FinBERT proves to be a powerful tool for quantifying financial narrative. While a raw sentiment signal is statistically significant, it is not "tradable" in its simplest form due to execution friction. The path to profitability lies in **filtering for signal conviction** (news volume) and **optimizing turnover**. Future research should explore the use of Limit Orders or "Trade-at-Close" (TAC) mechanisms to further reduce the impact of transaction costs on this high-alpha strategy.