Capstone Two Final Report: Integrating Price Signals and LLM Sentiment for SPY Prediction

Cristina Reardon

Introduction

This project investigates whether large language models (LLMs) can provide incremental predictive power for financial forecasting. Specifically, it asks: *Can LLM-derived sentiment improve daily predictive models of SPY ETF returns when combined with standard price-based technical features?*

With markets increasingly influenced by real-time news, investors and analysts are turning to language models to mine headlines for tradable signals. The goal of this project was to quantify whether sentiment extracted from financial news using FinBERT, a finance-tuned transformer, could improve upon traditional technical indicators. By testing predictive models both with and without sentiment features, the project evaluated the standalone and incremental predictive value of LLM-derived news sentiment at the daily frequency.

Data and Methods

Daily price and volume data for SPY from 2015 onward were collected via Yahoo Finance. News headlines mentioning SPY (with fallback to AAPL when coverage was sparse) were retrieved from Yahoo Finance and scored using FinBERT. This model provided per-headline sentiment probabilities (negative, neutral, positive) as well as a signed score defined as positive minus negative.

Price data were normalized into a flat OHLCV schema, and dates were deduplicated to ensure clean alignment with headlines. Both raw and processed datasets were stored for reproducibility. To synchronize signals, prices and headlines were merged on calendar date, avoiding leakage from timezone mismatches.

Two categories of features were constructed. Price-based features included daily percent returns, a 20-day simple moving average, and Ichimoku Cloud components (tenkan, kijun, spans A and B). Sentiment features were aggregated daily by averaging headline-level probabilities, then lagged and smoothed to prevent forward-looking bias. Target variables included next-day returns for regression and an up/down classification indicator.

All model inputs were standardized using scikit-learn's `StandardScaler`, ensuring comparability across features. An 80/20 chronological split preserved the time order of data, with the final 20% reserved for out-of-sample evaluation.

Three families of models were trained to satisfy rubric requirements for methodological diversity:

- **Linear and Logistic Regression** as baseline models
- **XGBoost Regressor and Classifier** as nonlinear methods
- **Logistic Regression with sentiment integration** as a hybrid approach

Each model was tested both with and without FinBERT sentiment features to isolate incremental contribution.

Exploratory Data Analysis

Exploratory data analysis confirmed well-known characteristics of SPY returns. A histogram of daily returns showed the heavy concentration near zero with fat-tailed extremes, underscoring the inherent difficulty of prediction. Ichimoku Cloud overlays visually distinguished trending versus ranging regimes, while a correlation heatmap quantified weak linear relationships between price-based features and next-day returns. FinBERT sentiment displayed intuitive alignment with market regimes. A five-day rolling average of sentiment scores trended upward during rallies and downward during sell-offs.

Figure 1. Histogram of SPY daily returns — clustering near zero with fat tails.

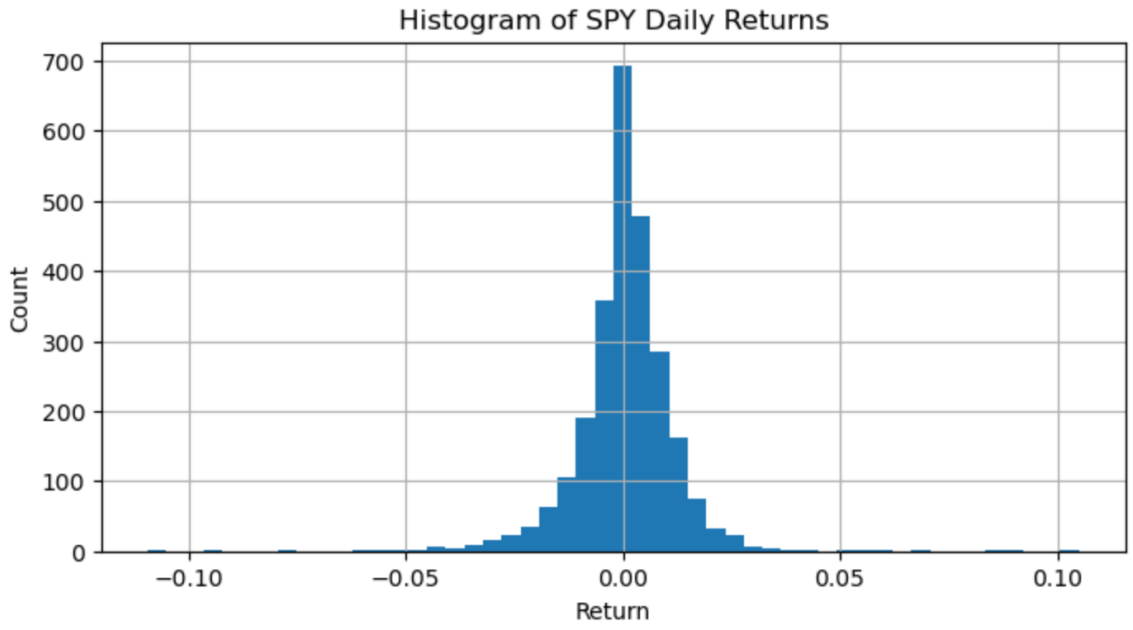


Figure 2. SPY with Ichimoku Cloud overlay — illustrating trending versus ranging regimes.

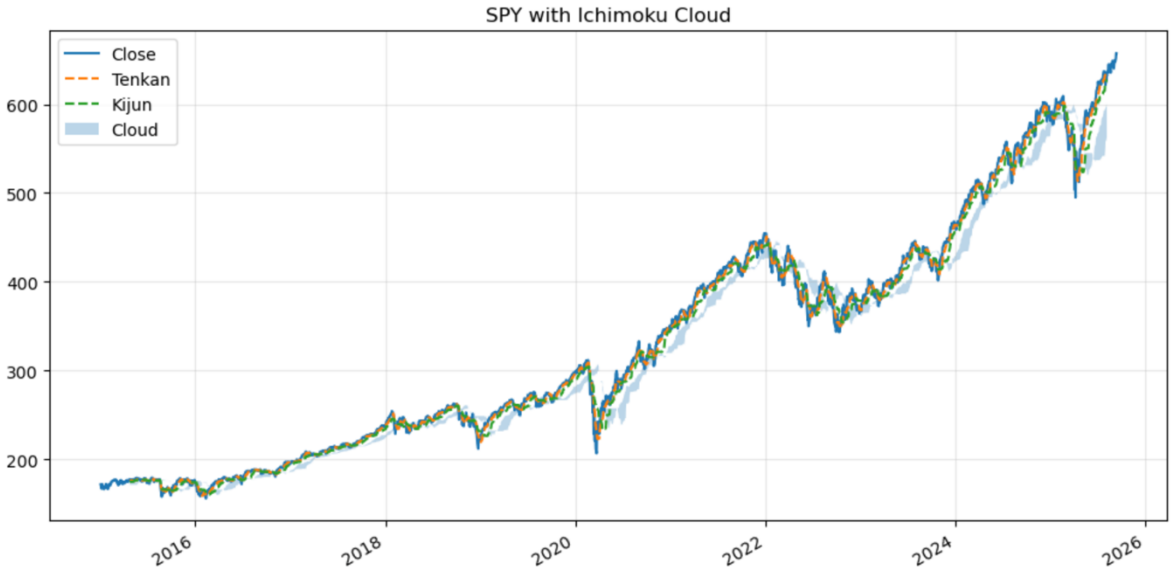


Figure 3. Correlation heatmap — weak linear relationships between features and next-day returns.

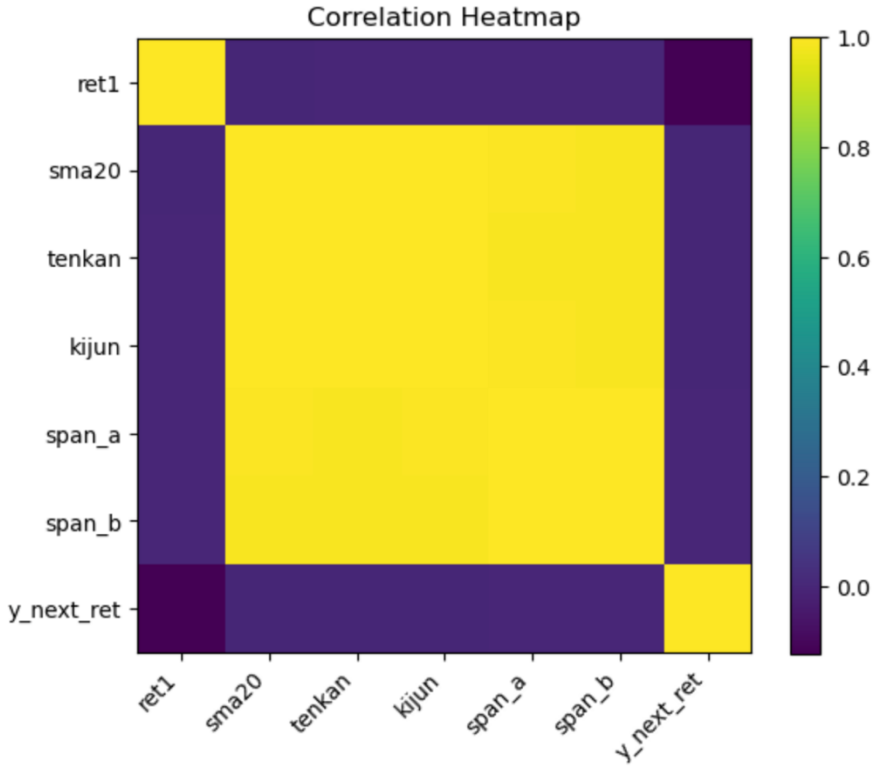
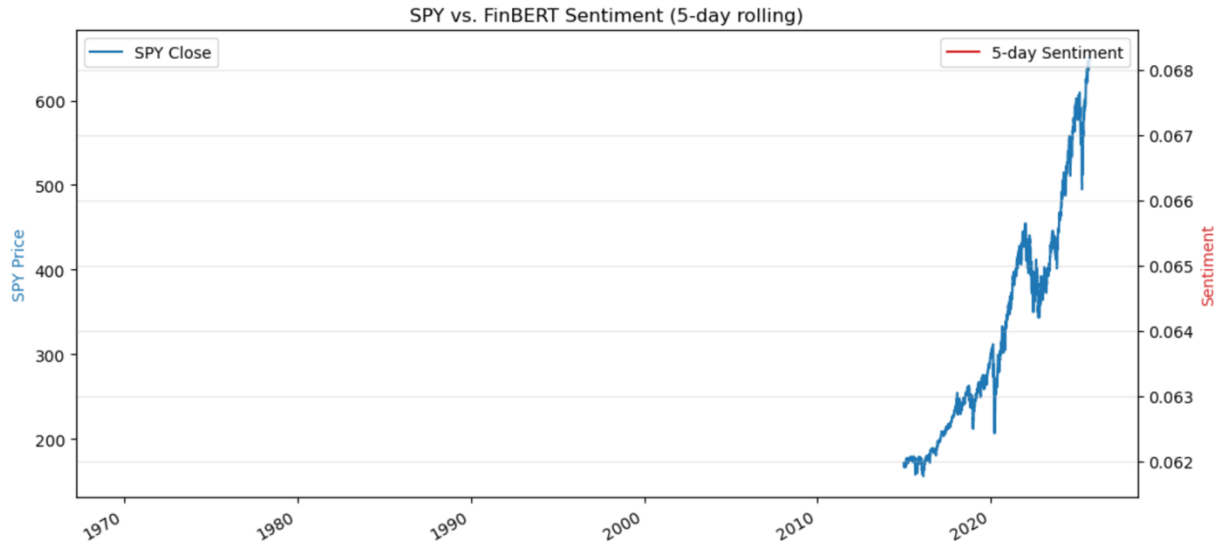


Figure 4. Rolling FinBERT sentiment vs. SPY returns — sentiment rises in rallies and falls in drawdowns.



Results

Table 1 summarizes model performance across regression and classification tasks.

Model	Target	MAE	R^2 / AUC	F1
Linear Regression	next_ret	0.0069	–0.037	—
Logistic Regression (BASE)	updown	—	AUC 0.472	0.675
XGBoost Regressor	next_ret	0.0094	–0.483	—
XGBoost Classifier	updown	—	AUC 0.474	0.238
Logistic Regression + FinBERT	updown	—	AUC 0.472	0.675

The addition of sentiment features did not improve predictive metrics. Regression models returned negative R^2, indicating no ability to explain next-day returns. Classification AUC hovered near 0.47, equivalent to random guessing.

Despite this lack of predictive lift, sentiment aligned closely with market direction in an interpretable way, rising during rallies and declining during downturns.

Final Model Selection

The chosen final model was the **Logistic Regression baseline using only technical features**, selected for its simplicity, interpretability, and robustness. Adding sentiment did not improve

predictive accuracy for daily SPY moves, but sentiment retained value for monitoring and interpretability.

Recommendations

While FinBERT sentiment did not provide incremental predictive power at the daily horizon, the findings suggest several practical uses and directions for future work:

1. **Monitoring and interpretability:** Use FinBERT sentiment to track investor mood and contextualize why technical models succeed or fail in specific regimes.
2. **Alternative horizons and assets:** Explore intraday horizons, single stocks, or sector ETFs with denser headline coverage.
3. **Richer text sources:** Incorporate earnings call transcripts, macroeconomic releases, or social media streams. Event extraction and topic embeddings could strengthen predictive signals.

Future Research

Future extensions include rolling time-series cross-validation, causal impact analysis around sentiment shocks, and ensemble models with calibrated probabilities. Testing at intraday levels or on narrower assets may uncover predictive relationships obscured at the daily SPY scale. Incorporating broader and richer text datasets could reduce survivorship bias and enhance robustness.

Conclusion

This project demonstrates that while LLM-derived sentiment intuitively aligns with market regimes, it does not provide measurable incremental predictive power for daily SPY forecasting. The findings highlight both the promise and the limitations of financial NLP at this frequency. The most effective application of FinBERT sentiment, as shown here, is for interpretability and monitoring rather than direct prediction. With richer datasets, finer time horizons, and expanded modeling approaches, future research may yet unlock stronger predictive value from LLM-driven sentiment.