

Sentiment-Augmented Financial Prediction: A Data Mining Approach to Market Signal Extraction

May 8, 2026

Max Bassler
bassler@uark.edu
University of Arkansas
Fayetteville, AR, USA

Daniel Draper
dedraper@uark.edu
University of Arkansas
Fayetteville, AR, USA

ABSTRACT

The objective of this study was to determine whether incorporating NLP-derived sentiment features extracted from financial text into models trained on financial-statement and market-derived features meaningfully improved short-term stock direction predictions. To accomplish this, a three-layer feature pipeline was constructed, the performance of each feature layer in classifying twenty-one-day forward excess return direction was obtained for five different supervised learning classifier models, and pairwise differences in performance across feature layers for each of the candidate models were evaluated by means of paired t-test to determine whether adding a specific feature layer into the feature space significantly improved predictions.

Maybe add concluding sentence about best model prediction findings, then combine all into one paragraph

I. INTRODUCTION

Financial markets are highly complex systems characterized by a high degree of uncertainty and influenced by a wide range of factors, such as firm-specific performance, investor behavior, macroeconomic conditions, sector trends, and market expectations. Because of this complexity and uncertainty, predicting future stock price movement remains a challenging task in both academic research and industry practice. Traditional quantitative models have historically relied on structured data such as accounting ratios, historical returns, volatility, and volume; however, market participants also respond to unstructured information, including news articles, corporate filings, analyst commentary, and event disclosures. This presents a unique opportunity to test whether sentiment extracted from unstructured financial text can add meaningful predictive signal beyond traditional structured financial and market data.

The present study examines this question within a controlled supervised learning framework. A three-layer feature space architecture was developed in which each layer additively introduces a distinct category of information: Layer 1 contains financial-statement ratios derived from SEC filings; Layer 2 augments the feature space with market-behavior features such as momentum, volatility, and volume signals; and Layer 3 further augments the feature space with sentiment features generated by applying the FinBERT financial language model to event text from Capital IQ Key Developments and SEC filings. The objective of layering features in this manner is to isolate the marginal predictive contribution of each information category in order to determine whether incorporating an information category significantly improves predictions rather than to optimize a single end-to-end predictor.

A central technical concern throughout the project was the prevention of temporal information leakage. In any supervised learning task involving sequential, time-series market data, the validity of model performance estimates

depends critically on whether features used at prediction time are restricted to information that would have only been available at that time in practice. Although this appears to be a stock prediction problem, the core data mining challenges were the treatment of look-ahead bias, the alignment of quarterly fundamentals with daily market data, and the use of purged expanding-window cross-validation.

The remainder of the report is organized as follows. Section II describes the methodology, including the feature layer architecture, the selected model cohort, the validation and evaluation strategies. Section III details the initialization of the modeling pipeline, including data collection, preprocessing, and feature engineering procedures. Section IV reports the evaluation results across each (feature layers, candidate model) combination and presents the paired t-test analysis of pairwise differences in predictive power. Section V discusses the practical implications of the findings, Section VI enumerates the limitations of the present implementation, Section VII outlines directions for future work, and Section VIII concludes.

II. METHODOLOGY

A. Feature Layer Architecture

The project used an additive feature design. The feature space was decomposed into three additive layers, each corresponding to a distinct category of predictive information. The motivation for this decomposition is the isolation of the marginal contribution of each category to model performance.

1. Layer 1: Financial-Statement Features

Layer 1 establishes the fundamental baseline of the feature space and consists of financial statement features derived from SEC quarterly filings. The feature set includes measures of liquidity, solvency, profitability, efficiency, growth, and earnings quality, which are presented in the table that follows.

Feature Category	Feature Pool
Liquidity	Current Ratio, Quick Ratio, Cash Ratio
Solvency	Debt-to-Assets, Debt-to-Equity, Long-Term Debt Ratio
Profitability	Operating Margin, Net Margin, Return-on-Assets, Return-on-Equity
Efficiency	Asset Turnover, Inventory Turnover
Growth	Revenue Growth, Earnings Growth
Earnings	CFO-to-Net-Income, Accruals Ratio, Free Cash Flow Features

2. Layer 2: Market-Derived Features

Layer 2 augments Layer 1 with point-in-time market behavior features measured prior to the prediction reference date to test whether this inclusion of price and volume context improves predictions beyond the fundamental baseline feature space contained in Layer 1. The feature set includes measures of momentum, volatility, risk proxies, volume, and relative strength, which are presented in the table that follows.

Feature Category	Feature Pool
Momentum	Recent Returns, Relative Returns, Return Shocks
Volatility	Realized Volatility, Volatility Ratios, Return Z-Scores
Risk Proxies	Beta-to-Sector, Drawdown, Overnight Gap
Volume	Volume Ratio, Log Volume, Abnormal Volume
Relative Strength	Stock Return Relative to Sector or Market Benchmark

3. Layer 3: Sentiment-Derived Features

Layer 3 augments Layer 2 with FinBERT-scored text-derived sentiment features to test whether this inclusion of quantified, unstructured information improves predictions beyond the financial-and-market feature space contained in Layer 2. The feature set includes measures of sentiment level, sentiment change, attention, relative sentiment, and coverage, which are presented in the table that follows.

Feature Category	Feature Pool
Sentiment Level	Mean Sentiment Score, Positive, Negative, and Neutral Probabilities
Sentiment Change	Sentiment Momentum, Sentiment Delta
Attention	News Counts, Text Chunk Counts
Relative Sentiment	Within-Sector Adjusted Sentiment
Coverage	Sentiment Availability Across Event Windows

B. Model Cohort

Five supervised learning classifiers were selected for the statistical evaluation, spanning the range of model families relevant to tabular financial classification, which are presented in the table that follows.

Model	Role
Logistic Regression	Regression Baseline
Random Forest	Bagging Baseline
Gradient Boosting	Boosting Baseline
XGBoost	Regularized Boosting
CatBoost	Ordered Boosting

Logistic Regression was selected as the linear, interpretable baseline against which non-linear models could be compared. Random Forest was selected as the bagging-based ensemble, providing a variance-reduction benchmark robust to feature scaling and outliers. Histogram Gradient Boosting was selected as the scikit-learn-native boosting baseline. XGBoost and CatBoost were selected as the leading regularized gradient-boosting implementations, with XGBoost serving as the primary advanced model and CatBoost providing an ordered-boosting comparator that is generally considered less prone to target leakage on small panels.

Hyperparameters for each model were tuned using a Bayesian optimization procedure over the cross-validation folds defined in the following Section II.C, with model-specific search spaces covering tree depth, learning rate, subsample fractions, and L1/L2 regularization strengths for the boosted models, and inverse regularization strength and penalty type for logistic regression.

C. Evaluation Strategy

Each (feature layer, model) combination was evaluated under a unified time-series-split cross-validation consisting of purged expanding-window cross-validation with a five-fold structure, a five-trading-day embargo between training and validation segments, and a final 2024 calendar-year holdout that was reserved entirely for out-of-sample reporting. Performance was assessed using AUC-ROC as the primary metric, F1-score as the secondary metric, and log loss as the calibration metric. Confusion matrices and SHAP-based feature attributions were also generated to support interpretation.

The marginal contribution of each feature layer was assessed by comparing the holdout AUC obtained when each layer was added against the previous layer under the same model. The pairwise differences between candidate models under the locked Layer 3 configuration were assessed using paired t-tests across the five cross-validation folds, with two-sided p-values reported alongside Cohen's d_z effect sizes.

III. INITIALIZATION

A. Data Collection

The sample universe for evaluation consisted of thirty-four tickers within the Consumer Staples sector, designated as Universe V1. Constituent tickers included large consumer-staples names such as WMT, COST, PG, KO, PEP, KR, TGT, CL, KMB, HSY, GIS, KHC, and TSN. The decision to restrict the universe to a single sector was made to control for cross-sector heterogeneity in the structural relationships between fundamentals, market behavior, and sentiment, and to permit the evaluation of within-sector relative sentiment as a feature without confounding from cross-sector dispersion. Generalization beyond Consumer Staples was treated as a future-work item rather than as a claim of the present study.

Data were drawn from four primary sources, which are summarized in the table that follows.

Data Type	Source	Use Case
Fundamental Ratios	SEC Filings (via <i>edgartools</i> Python Library)	Layer 1
Adjusted Price & Volume	Yahoo Finance (via <i>yfinance</i> Python Library)	Layer 2
Event Text	Capital IQ Key Developments	Layer 3 (Primary)
Filing Text	SEC 10-K & 10-Q Filings (via <i>edgartools</i>)	Layer 3 (Secondary)

The fundamentals source was selected because the SEC EDGAR system provides the authoritative public record of regulated financial reporting, and the *edgartools* Python library standardizes the underlying XBRL-tagged accounts into a consistent set of approximately ninety-five accounting concepts that permit cross-company ratio comparison. The market data source was selected for because of Yahoo Finance’s complete OHLCV (Open, High, Low, Close, and Volume) coverage of the sample universe across the relevant historical window. The Capital IQ Key Developments feed was selected as the primary text source because it provided substantially better temporal coverage for the Consumer Staples universe than the sparser external-news datasets considered in initial pipeline design, and the SEC filing text was retained as a secondary source to capture forward-looking management disclosures that complement the externally reported event text.

The historical window spanned the period over which all three data sources had concurrent availability, with the year 2024 reserved as the final holdout window for out-of-sample evaluation and reporting.

B. Data Preprocessing

The preprocessing pipeline was designed to enforce point-in-time correctness and robust handling of distributional irregularities common in financial features. Missing market values were forward-filled rather than back-filled to prevent the introduction of future information into the feature stream. Missing financial ratios were imputed using sector medians computed only on the training fold of each cross-validation split, which prevented the leakage of validation-period distributional information into the training procedure. Extreme outliers in financial ratios – which can arise from near-zero denominators in the underlying ratio computations – were winsorized at the first and ninety-ninth percentiles of the training-fold distribution. Feature scaling was performed using the *RobustScaler*, which centers on the median and scales by the interquartile range, and is therefore less sensitive to the heavy tails present in financial features than the more conventional *StandardScaler*.

Categorical encoding for sector and industry indicators was performed within the cross-validation fold to prevent target leakage through the encoding step. All features were lagged by at least one trading day relative to the prediction reference date to ensure that no feature value reflected information that would not have been observable in practice.

The structural integrity of the resulting panel was verified by a comprehensive data layer audit, which evaluated 103 individual integrity and alignment checks. The audit returned the outcomes summarized in the table that follows.

Audit Status	Count
PASS	98
REVIEW	2
INFO	2
LIMITED	1
FAIL	0

The absence of FAIL outcomes was treated as a precondition for trustworthy downstream modeling, and the small number of REVIEW, INFO, and LIMITED outcomes were inspected and judged not to materially affect the modeling pipeline that consumed the panel.

The classification target was constructed as the sign of the twenty-one-day forward excess return relative to a sector-adjusted threshold. For each event date t and each ticker i , the forward excess return was computed as the cumulative log return of ticker i from $t+1$ to $t+21$ minus the cumulative log return of the relevant sector benchmark over the same window. The label was set to one when this excess return exceeded the threshold and to zero otherwise. The use of a sector-adjusted excess-return label rather than a raw forward-return label was motivated by the within-sector character of the universe: because all thirty-four tickers belonged to Consumer Staples, an unadjusted classifier would have been dominated by sector-wide co-movement rather than the cross-sectional discrimination that was the actual object of interest.

C. Feature Engineering

Each of the three feature layers was implemented as a distinct engineering pipeline whose output was joined into the final modeling panel by the (ticker, date) key. The implementation details specific to each layer are summarized in the subsections that follow.

1. Layer 1 Implementation

Layer 1 features were constructed from quarterly 10-Q and annual 10-K filings retrieved through the *edgartools* Python library. A fundamentals integrity rebuild was performed prior to modeling. The rebuild replaced an earlier feature-construction pass that exhibited a substantial number of ratio sanity flags arising from arithmetic edge cases, sign inconsistencies, and unit mismatches across filer-specific reporting conventions. The reduction in flagged observations after the integrity rebuild is summarized in the table that follows.

Universe	Metric	Before	After
Universe V2	Feature Ratio Sanity Flags	10,514	181
Universe V2	Panel Ratio Sanity Flags	8,260	158
Universe V2	Amount Ratio Sanity Flags	2,374	3
Universe V1	Feature Ratio Sanity Flags	2,966	63
Universe V1	Panel Ratio Sanity Flags	2,402	53
Universe V1	Amount Ratio Sanity Flags	806	0

The order-of-magnitude reduction in flagged observations across all categories was treated as a precondition for trustworthy downstream modeling, since the presence of structurally invalid ratios would

otherwise have propagated through scaling, fold splitting, and model fitting, and would have distorted both performance metrics and feature-attribution analyses.

2. Layer 2 Implementation

Layer 2 features were constructed from adjusted OHLCV (Open, High, Low, Close, and Volume) data retrieved through the *yfinance* Python library. The momentum features were computed as multi-window relative returns over five-day, ten-day, twenty-one-day, and sixty-three-day windows, together with relative-strength measures computed against both the relevant sector ETF and the broad market benchmark. Realized volatility was computed over rolling twenty-one-day and sixty-three-day windows, and the volatility ratio of short-window to long-window realized volatility was retained as an indicator of regime change. Volume features were computed as the ratio of current volume to the trailing twenty-day average volume, log-transformed volume, and the residual abnormal volume from a regression of log volume on calendar effects. Distributional features included rolling skewness and kurtosis of returns, used as proxies for tail risk and asymmetry in the recent return distribution. Return shock indicators were constructed to flag observations where the most recent absolute return exceeded a multi-sigma threshold relative to the trailing distribution.

3. Layer 3 Implementation

Layer 3 features were generated by applying FinBERT (ProsusAI/finbert) – a BERT-based transformer model pre-trained on a large financial corpus and fine-tuned for three-class sentiment classification on the Financial PhraseBank – to the event text of the project. FinBERT was selected over generic sentiment tools (such as VADER or TextBlob) because it is specifically calibrated to financial language, in which terms such as “depreciation,” “exposure,” and “leverage” carry domain-specific connotations that diverge from their everyday meanings.

Two text sources fed the FinBERT inference pipeline. The primary source was Capital IQ Key Developments, a curated event-text feed covering corporate disclosures, M&A activity, regulatory filings, and analyst-relevant company news. The secondary source was the Management Discussion & Analysis section of SEC 10-K and 10-Q filings, retrieved through the *edgertools* Python library. Coverage diagnostics for the Capital IQ source are summarized in the table that follows.

Coverage Window	Overall Coverage	2024 Holdout Coverage
7 Days	60.11%	55.47%
30 Days	88.18%	86.13%
63 Days	93.77%	92.70%

The within-sector relative sentiment feature, which played a central role in the empirical results reported in Section IV, was constructed by subtracting the contemporaneous sector-wide mean sentiment from the ticker-level sentiment. The motivation for this construction was the observation that absolute sentiment levels can be confounded by sector-wide news cycles, while sector-relative sentiment captures the idiosyncratic component of textual signal that should, in principle, be informative for cross-sectional ranking within the universe.

A panel integrity check was performed to verify that the sentiment join did not duplicate or drop event rows. The check is summarized in the table that follows.

Panel Check	Value
Base Panel Row Count	1,108
Sentiment Enriched Row Count	1,108

Row Count Difference	0
Duplicate Event Rows	0

The zero row-count delta confirmed that the sentiment join was structurally clean.

IV. RESULTS

This section reports the empirical results of the modeling pipeline. The performance evaluation in Section IV.A compares the layered feature configurations and the candidate models on the 2024 holdout, and the statistical analysis in Section IV.B formalizes the pairwise differences between candidate models using .

A. Performance Evaluation

1. Active Quarterly Benchmark – Layer 1 + Layer 2 Configuration

The locked benchmark performance of the three primary models under the Layer 1+2 (financial statement plus market) configuration on the 2024 holdout is summarized in the table that follows.

Model	CV AUROC Mean	CV AUROC Std	Worst Fold AUROC	2024 Holdout AUROC	Holdout Log Loss	Holdout F1
Logistic Regression	0.5022	0.0571	0.4198	0.4992	0.7330	0.5490
Random Forest	0.5093	0.0711	0.3732	0.4487	0.7080	0.5439
XGBoost	0.5127	0.0827	0.3522	0.5365	0.7448	0.6018

Under the Layer 1+2 configuration, all three models produced holdout AUROC values close to chance, with XGBoost the best of the three at 0.5365. The relatively wide cross-validation AUROC standard deviations and worst-fold AUROC values below 0.40 indicated substantial fold-to-fold variability, consistent with regime sensitivity in the underlying market features.

2. Capital IQ Layered Sentiment Ladder

The layered sentiment ladder is summarized in the table that follows. Each row represents an additive feature configuration tested under the same validation contract, and the selected model for each rung is the model that achieved the highest mean cross-validated AUROC for that feature set.

Rung	Feature Set	Selected Model	Mean CV AUROC	Worst Fold AUROC	2024 Holdout AUROC	Holdout Log Loss	Holdout F1
A	Quarterly Core	XGBoost	0.4878	0.3872	0.5501	0.7289	0.5421
B	Quarterly core + market	XGBoost	0.5304	0.4083	0.5020	0.7776	0.5357
C	Core + market + raw sentiment	Logistic Regression	0.4927	0.4012	0.5958	0.6939	0.4792
D	Core + market + within-sector	Logistic Regression	0.4989	0.4203	0.6038	0.7115	0.3765

	relative sentiment						
--	-----------------------	--	--	--	--	--	--

The strongest configuration was Rung D, in which within-sector relative sentiment features were appended to the financial-and-market feature space. Rung D achieved a 2024 holdout AUROC of 0.6038, an improvement of 0.1018 over the financial-and-market benchmark of Rung B and 0.0537 over the fundamentals-only benchmark of Rung A. Rung D also achieved the second-lowest holdout log loss of the four rungs, at 0.7115. The interpretation of these results is consistent with the project's central hypothesis: sentiment information, when properly aligned and demeaned by sector, provides a meaningful marginal contribution to short-horizon directional prediction. The cross-validated AUROC for Rung D, at 0.4989, did not improve in parallel with the holdout AUROC, which is treated as evidence that the holdout improvement should be characterized as promising rather than decisive.

- 3. Stability of the Sentiment Lift Across Pseudo-Holdout Years
- 4. SHAP-Based Feature Attribution

B. Statistical Analysis

Write Here (If Necessary)

- 1. Test Design
Write Here (If Necessary)
- 2. AUROC Comparisons
Write Here (If Necessary)
- 3. F1-Score Comparisons
Write Here (If Necessary)
- 4. Log Loss Comparisons
Write Here (If Necessary)
- 5. Summary of Statistical Findings
Write Here (If Necessary)

Write Here (If Necessary)

V. DISCUSSION

Write Here (If Necessary)

VI. LIMITATIONS

Write Here (If Necessary)

VII. FUTURE IMPROVEMENTS

Write Here (If Necessary)

VIII. CONCLUSION

Write Here (If Necessary)