Predicting Earnings Surprises from Financial News

Sven Stephan Thie

Dissertation written under the supervision of Professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, March 14th, 2026.

Abstract

This dissertation investigates whether news sentiment, extracted via dictionary methods or modern LLMs, provides incremental predictive power for S&P 500 corporate earnings surprises. Using 70,133 firm-quarter observations (2000–2024), we compare two sentiment sources (Harvard IV-4 dictionary, Mistral LLM) across seven ML architectures. No sentiment configuration significantly outperforms the controls-only baseline to predict SUE (best Ensemble R^2: 5.47% controls-only), establishing a comprehensive null result. Within this null, three findings stand out: (1) out-of-sample R^2 varies substantially across sub-periods, yet sentiment never outperforms controls in any regime; (2) SHAP analysis attributes 64–68% of feature importance to lagged SUE, with combined sentiment contributing only 6–8%; and (3) a long/short strategy based on predicted SUE quartiles yields uniformly negative and statistically insignificant Fama-French three-factor alphas, indicating that predictability does not translate into economic value. These findings indicate that for large-cap firms, news sentiment introduces estimation noise rather than incremental signal, suggesting that the information content of financial news is already captured by the structured financial variables it ultimately reflects.

Title: Predicting Earnings Surprises from Financial News

Author: Sven Stephan Thie

Keywords: Supervised Learning; Financial News Sentiment; Standardized Unexpected Earnings

Resumo

Esta dissertação investiga se o sentimento de notícias, extraído via métodos de dicionário ou LLMs modernos, fornece poder preditivo incremental para surpresas de lucros corporativos do S&P 500. Usando 70.133 observações firma-trimestre (2000–2024), comparamos duas fontes de sentimento (dicionário Harvard IV-4, LLM Mistral) através de sete arquiteturas ML. Nenhuma configuração de sentimento supera significativamente a baseline apenas com controles na previsão do SUE (melhor R^2 Ensemble: 5,47%), estabelecendo um resultado nulo abrangente. Dentro deste resultado nulo, três achados destacam-se: (1) o R^2 fora da amostra varia substancialmente entre subperíodos, mas o sentimento nunca supera os controles em nenhum regime; (2) a análise SHAP atribui 64–68% da importância das variáveis ao SUE defasado, com o sentimento combinado a contribuir apenas 6–8%; e (3) uma estratégia long/short baseada em quartis do SUE previsto gera alfas Fama-French de três fatores uniformemente negativos e estatisticamente insignificantes, indicando que a previsibilidade não se traduz em valor económico. Estes resultados indicam que, para empresas de grande capitalização, o sentimento de notícias introduz ruído de estimação em vez de sinal incremental, sugerindo que o conteúdo informacional das notícias financeiras já se encontra capturado pelas variáveis financeiras estruturadas que estas refletem.

Título: Previsão de Surpresas de Lucros a partir de Notícias Financeiras
Autor: Sven Stephan Thie
Palavras-chave: Aprendizagem Supervisionada; Sentimento de Notícias Financeiras; Lucros Inesperados Padronizados

Contents

List of Tables

List of Figures

1 Introduction

Earnings surprises, the difference between reported and expected earnings, move markets. They trigger immediate price reactions, drive post-earnings-announcement drift, and feed into analyst revisions and portfolio rebalancing. Being able to predict these surprises, even partially, matters for both asset pricing theory and investment practice. A landmark contribution in this pursuit is Tetlock et al. (2008), who demonstrated that simple lexical measures (specifically, the fraction of negative words in financial news) possess significant predictive power for firm earnings beyond traditional quantitative controls. Using the Harvard IV-4 psychosocial dictionary applied to Wall Street Journal and Dow Jones News Service articles about S&P 500 firms, they established that media content contains valuable information about firm fundamentals not captured by conventional accounting variables. However, their analysis relied on 1980–2004 data, a period that predates several fundamental changes in how information flows through financial markets. The adoption of Regulation Fair Disclosure in 2000, the proliferation of algorithmic trading, the acceleration of digital news dissemination, and the rise of sophisticated natural language processing (NLP) have collectively transformed both the information environment and the tools available for analyzing it. These developments raise a natural question: do Tetlock et al.'s dictionary-based findings remain valid in the modern period, and can contemporary NLP methods, which have evolved from simple word counting to contextualized language understanding through transformer architectures, extract more predictive information from financial text?

To address this question, we formulate the following research problem:

> *Does news sentiment, extracted via dictionary methods (Harvard IV-4) or a modern LLM (Mistral), provide incremental predictive power for S&P 500 corporate earnings, and how does the choice of sentiment extraction method and model architecture mediate this predictive value?*

We investigate this question using 70,133 firm-quarter observations from 2000 to 2024. We begin with a Tetlock Validation, applying the original methodology to a different news source (New York Times vs. Wall Street Journal and Dow Jones News Service) on a different time range (2000–2010 vs. 1980–2004) via in-sample OLS regression to establish whether dictionary sentiment retains predictive power in the modern period. We then extend the analysis to the full sample in a SUE Prediction Comparison, using expanding window cross-validation to test whether two sentiment sources (Harvard IV-4 dictionary and Mistral LLM) provide incremental predictive value over a controls-only benchmark. This comparison spans seven machine learning architectures: OLS, Ridge, Lasso, Elastic Net, Random Forest, XGBoost, and a validation-weighted Ensemble. Finally, our Ablation and Economic Significance analysis applies SHAP decomposition to quantify sentiment contributions relative to financial controls and

evaluates trading strategies to test whether any predictive power translates into economic value.

The central finding is a null: no sentiment configuration significantly outperforms the controls-only baseline, with Diebold-Mariano tests confirming the absence of statistically significant differences. The best-performing specification, a controls-only Ensemble, achieves an out-of-sample R^2 of 5.47%, while adding sentiment features consistently fails to improve upon this benchmark. The simplest explanation is that sentiment is redundant with financial controls: the information embedded in news articles is already reflected in the structured variables (notably lagged earnings surprises) that controls capture directly. The null is nonetheless informative. Predictive performance varies dramatically across sub-periods (R^2 from -1.5% to 11.4%), yet sentiment fails to outperform controls in any regime. SHAP decomposition shows that lagged SUE accounts for 64–68% of feature importance while combined sentiment contributes only 6–8%. And even the best-performing model's long-short trading strategy based on predicted SUE produces uniformly negative and statistically insignificant Fama-French three-factor alphas, confirming that statistical predictability does not yield economic value.

These findings make several contributions to the financial text analysis literature. We provide the first expanding window validation of Tetlock et al. (2008), showing that in-sample dictionary sentiment significance does not survive out-of-sample evaluation. We also establish that neither LLM-based (Mistral) nor dictionary-based (Harvard IV-4) sentiment provides incremental value over the controls-only baseline. Beyond the null itself, we demonstrate the importance of baseline transparency – controls-only models, often omitted in NLP-finance research, consistently outperform sentiment-augmented models, which means studies without proper baselines risk overstating NLP contributions. Finally, we provide possible explanations for why sentiment fails, identifying regime dependence, feature importance hierarchy, and that the signal does not convert into economic value. Limitations of the study are discussed in Section 6.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on sentiment analysis, earnings prediction, and market efficiency. Section 3 describes the data sources and variable construction. Section 4 details the methodology. Section 5 presents the empirical results, and Section 6 discusses their implications and limitations.

2 Literature Review

2.1 Dictionary-Based Sentiment Analysis

The foundation of this research rests on Tetlock (2007), who demonstrated that media pessimism, measured via the Harvard IV-4 dictionary applied to Wall Street Journal columns, predicts downward pressure on aggregate market prices and elevated trading volume. Tetlock et al. (2008) extended this to the firm level, establishing that simple lexical measures of media content contain valuable information about firm fundamentals not captured by conventional accounting variables.

For sentiment measurement, Tetlock et al. (2008) employed the Harvard IV-4 psychosocial dictionary to quantify sentiment in financial news stories. They focused specifically on the fraction of negative words in Wall Street Journal and Dow Jones News Service articles about S&P 500 firms during the 30 to 3 trading days prior to earnings announcements. We adopt this $[RDQ-30, RDQ-3]$ trading day window, hereafter the "Tetlock window," for measuring pre-announcement sentiment.

The study employed two measures of earnings surprise as dependent variables:

1. *Standardized Unexpected Earnings (SUE)*: The difference between actual and expected earnings (using a seasonal random walk benchmark), standardized by historical volatility
2. *Standardized Analyst Forecast Error (SAFE)*: The median analyst forecast error divided by earnings volatility

Using linear regression with quarterly clustered standard errors on data from 1980 to 2004, Tetlock et al. (2008) found that negative word fractions significantly predict low firm earnings (negative coefficient: -0.0637, $t = -4.69$). Moreover, stock prices briefly underreact to information embedded in negative words. The relationship holds after controlling for lagged earnings, firm size, book-to-market ratio, trading volume, past returns, analyst forecast revisions, and forecast dispersion.

While groundbreaking, several aspects of the Tetlock et al. methodology warrant scrutiny. First, the in-sample estimation approach may overstate true predictive power, as parameters are optimized on the same data used for evaluation. Second, the economic magnitude of the effect, while statistically significant, translates to modest trading profits after accounting for transaction costs. Third, the relationship between negative words and earnings may be partially mechanical: journalists may simply be reporting on firms already experiencing difficulties, meaning news sentiment proxies for publicly available information rather than providing incremental insight. These concerns motivate our validation framework with rigorous out-of-sample testing.

Beyond these general concerns, the Tetlock et al. methodology faces several specific limitations that motivate our analytical approach:

A first limitation concerns *context*: dictionary approaches treat words as independent units, missing critical contextual information. For example, "not good" would be classified as containing one positive word, when the phrase is clearly negative. Similarly, negations ("not profitable"), intensifiers ("extremely poor"), and conditional statements ("if earnings decline") are not properly captured.

A second limitation is *domain misalignment*: the Harvard IV-4 dictionary was developed for psychosocial research, not financial analysis. As Loughran and McDonald (2011) demonstrate, many words classified as negative in general contexts have neutral or even positive meanings in finance. For instance, "liability" is classified as negative in Harvard IV-4 but is a standard accounting term without inherent negative connotation. Similarly, "capital" and "debt" may be classified inappropriately for financial contexts.

Third, the *linear regression framework* assumes additive, linear relationships between sentiment and earnings. This cannot capture complex non-linear patterns, threshold dynamics where sentiment only matters above certain levels, or time-varying relationships where the sentiment-earnings link changes over time.

Finally, the 1980–2004 study period predates major changes in financial markets and information dissemination, including the adoption of Regulation Fair Disclosure (2000), the proliferation of electronic trading, and the rise of real-time news dissemination, all fundamental changes in how information flows through markets.

Loughran and McDonald (2011) address the domain misalignment problem by developing a finance-specific sentiment dictionary. Their comprehensive analysis of 10-K filings reveals that nearly three-quarters of words classified as negative by the Harvard IV-4 dictionary are not negative in financial contexts. The Loughran-McDonald dictionary provides several improvements: financial specificity, where words are classified based on their meaning in financial documents rather than general usage; multi-dimensional sentiment, extending beyond simple positive/negative classification to include categories for uncertainty, litigious language, and constraining language; and extensive coverage, containing 2,337 negative words specifically relevant to financial contexts. Loughran and McDonald (2011) demonstrate that using finance-specific dictionaries substantially improves the measurement of tone in financial disclosures and its relationship with market outcomes.

Subsequent work has explored statistical alternatives to pre-defined dictionaries. Li (2010) applied naïve Bayesian classification to identify forward-looking statements in corporate filings, demonstrating that data-driven word selection can outperform fixed dictionaries for predicting future performance. Ke et al. (2020) take this further by using word embeddings and neural net-

works to construct "sentiment" directly from text without relying on any dictionary, finding that their learned representations predict returns beyond traditional word-count measures. These studies represent a methodological progression from fixed dictionaries to learned representations, a progression that our LLM-based approach extends to the frontier of general-purpose language models.

While the Loughran-McDonald and Ke et al. approaches motivate domain-specific sentiment measures, we use Harvard IV-4 as our dictionary baseline to enable direct comparison with the original Tetlock et al. (2008) methodology.

2.2 Machine Learning for Earnings Prediction

A parallel literature has developed applying machine learning methods directly to earnings prediction, providing important context for evaluating whether sophisticated NLP can improve upon simpler approaches. This stream of research suggests that while ML can enhance predictive performance, the gains often derive from better exploitation of structured financial data rather than unstructured text.

The landmark study by Gu et al. (2020) provides the methodological foundation for applying machine learning to financial prediction, systematically comparing neural networks, tree-based methods, and linear models for stock return forecasting. Their finding that gradient boosted trees and random forests capture economically meaningful non-linear interactions among predictors directly motivates our model architecture choices. More recently, Hansen and Siggaard (2024) apply double machine learning to explain PEAD, finding that traditional variables (momentum, liquidity, and limits to arbitrage) account for the bulk of the drift, mirroring our SHAP finding that lagged SUE dominates feature importance.

Ball and Ghysels (2018) demonstrate that automated MIDAS regressions combining high-frequency macroeconomic and stock return data with quarterly accounting variables can match or exceed analyst forecast accuracy. Critically, combining automated and analyst forecasts consistently outperforms either approach alone, with macroeconomic indicators receiving the largest predictive weights. This finding highlights that even without machine learning, structured quantitative data captures much of the predictable variation in earnings, reinforcing the possibility that unstructured text contributes only marginally when financial controls are properly specified.

Chen et al. (2022) provide direct evidence on ML for earnings prediction using random forests and stochastic gradient boosting on nearly 14,000 predictors derived from XBRL filings. They demonstrate that ML models generate significant out-of-sample improvements over analyst forecasts, particularly for firms with complex financial structures. However, the gains are concentrated in earnings components (operating income, net income, EPS) and footnote dis-

closures, precisely the type of structured accounting features that our controls already capture through lagged SUE, forecast revision, and Fama-French abnormal returns. For S&P 500 firms, which represent the most intensively covered segment of the market, any incremental information from text may be rapidly impounded into consensus expectations.

van Binsbergen et al. (2023) provide complementary evidence by comparing machine learning forecasts to analyst forecasts across the term structure of earnings expectations. They find that simple linear models fail out-of-sample, particularly after the 2000s, while their random forest benchmark remains effective by capturing non-linear predictor relationships. Crucially, analyst forecasts are not redundant: they contain private information that the ML model incorporates as a key input. At longer horizons, ML reveals that analysts exhibit systematic upward bias that increases with the forecast horizon and predicts negative cross-sectional returns. Their finding that past earnings serve as the dominant predictor in the random forest aligns with our result that lagged SUE captures the majority of predictable variation in next-quarter earnings.

A key insight emerging from this literature is the importance of proper validation methodology. Studies using in-sample or simple holdout evaluation often report larger improvements than those employing expanding window cross-validation that mimics realistic forecasting scenarios. This methodological concern directly motivates our approach, which advances from in-sample validation (Tetlock Validation) to rigorous out-of-sample testing (SUE Prediction Comparison).

2.3 Earnings Surprises and Market Efficiency

The predictability of earnings surprises is fundamentally linked to market efficiency. Fama (1970) distinguished three progressively stronger efficiency hypotheses, ranging from prices that incorporate only past trading data (weak form), through prices that fully impound all public disclosures (semi-strong form), to prices that also reflect privately held information (strong form). Our research engages with this framework on two levels: first, we ask whether publicly available news sentiment helps predict future earnings surprises (an accounting question about the earnings generating process), and second, we test whether any such predictive power translates into economic value through trading strategies (a market efficiency question). The semi-strong form is directly relevant to the second question.

Ball and Brown (1968) first documented that earnings announcements contain value-relevant information not fully anticipated by the market. Bernard and Thomas (1989) extended this work by documenting post-earnings-announcement drift (PEAD), the tendency for stocks with positive earnings surprises to continue outperforming, and stocks with negative surprises to continue underperforming, for several months following the announcement. PEAD represents one of the most robust anomalies in empirical finance and suggests that markets do not immediately incorporate earnings information. Engelberg and Parsons (2011) provide causal evidence that media

coverage directly affects trading behavior around earnings announcements, using geographic variation in local newspaper coverage to establish that media exposure, not merely information release, drives investor responses. Meursault et al. (2023) demonstrate that textual features from earnings announcements explain a significant portion of the drift beyond numerical predictors, motivating our examination of whether pre-announcement news sentiment similarly contains incremental information.

However, PEAD has attenuated in recent decades. Hong et al. (2000) showed that information diffuses slowly for smaller firms with low analyst coverage, implying that PEAD should be weakest for well-covered large-cap stocks. Hendershott et al. (2011) demonstrate that algorithmic trading improves price discovery, while institutional quantitative strategies and regulatory changes, particularly Regulation Fair Disclosure in 2000, have further reduced information asymmetries. For S&P 500 firms specifically, these efficiency gains are particularly pronounced: large-cap stocks attract extensive analyst coverage, high trading volumes, and sophisticated institutional attention. This suggests that even if sentiment contains genuine information about future earnings, translating that into economic value may be difficult for well-covered firms, a prediction our empirical analysis directly tests.

A related strand of literature examines whether aggregated social signals contain predictive information. Cookson et al. (2024) examine social media attention and sentiment across Twitter, StockTwits, and Seeking Alpha, finding that attention is highly correlated across platforms while sentiment is largely idiosyncratic. Importantly, the two signals carry opposite implications for short-horizon returns: higher sentiment is associated with next-day price increases, while elevated attention is associated with next-day declines. However, their setting differs from ours in two key respects: they focus on daily social media signals rather than news article sentiment, and they predict short-horizon returns rather than quarterly earnings.

The distinction between statistical predictability and economic exploitability is crucial. Even if sentiment statistically predicts earnings surprises, the relationship may be too weak or too rapidly incorporated to generate meaningful trading profits after transaction costs. Our Ablation analysis specifically addresses this question by evaluating cumulative returns from sentiment-based trading strategies.

2.4 Modern NLP in Finance

The past decade has witnessed significant advances in NLP through transformer-based architectures and large language models (LLMs). Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), which fundamentally changed how machines understand language by processing text bidirectionally and learning contextual representations.

More recently, instruction-tuned LLMs such as GPT-4 (OpenAI, 2023) and open-source alter-

natives like Mistral (Mistral AI, 2025) have shown strong performance in understanding and classifying text sentiment. These models offer several advantages over dictionary methods:

- Contextual Understanding: LLMs process entire sentences and paragraphs, properly interpreting negations ("no longer competitive"), conditionals ("if margins contract"), and complex linguistic structures that dictionary methods miss.
- Domain Adaptation: Through careful prompting, LLMs can be instructed to evaluate sentiment from a financial analyst's perspective, focusing on implications for firm fundamentals.
- Nuanced Classification: LLMs can provide graduated sentiment scores (e.g., 5-point scales) rather than simple positive/negative word counts.
- Zero-Shot Capability: Unlike fine-tuned models, LLMs can be applied to financial text without domain-specific training data.

Recent studies have begun evaluating LLMs directly for financial prediction tasks, with results that present an optimistic but nuanced picture. Lopez-Lira and Tang (2025) demonstrate that ChatGPT-based sentiment scores from news headlines can predict next-day stock returns, with the effect concentrated in smaller, less-covered stocks. Their finding that LLM sentiment outperforms traditional dictionaries for return prediction appears to contradict our results. However, critical differences in research design explain this divergence: Lopez-Lira and Tang predict short-horizon *returns* (next-day) using headline-level sentiment, whereas we predict quarterly *earnings surprises* using article-level sentiment aggregated over 28-day windows. The short-horizon return setting captures price momentum effects that may dissipate at longer horizons, and the quarterly aggregation in our study necessarily smooths away the event-day signals that drive their results.

Bybee et al. (2024) demonstrate that news text, analyzed via topic modeling across 800,000 Wall Street Journal articles, explains approximately 25% of aggregate stock return variation, confirming that news contains genuine economic information. However, their aggregate return prediction operates at a fundamentally different level than our firm-specific quarterly earnings task, where the signal must survive cross-sectional aggregation and temporal smoothing. These studies collectively suggest that LLM sentiment may contain genuine short-lived information content that our quarterly framework cannot capture. This interpretation is consistent with our finding that in-sample sentiment significance (Tetlock Validation) does not translate to out-of-sample quarterly predictive value (SUE Prediction Comparison); the signal may exist but decay rapidly in efficient large-cap markets.

An emerging methodological consideration for LLM-based financial analysis is *prompt sensitivity*, the degree to which sentiment classifications depend on specific prompt formulation.

Recent benchmarking confirms that zero-shot LLMs can outperform domain-specific models like FinBERT (Huang et al., 2023) on financial sentiment classification, though prompt design materially affects accuracy: prompts adopting a financial analyst perspective yield superior classifications compared to generic sentiment instructions. Zhuo et al. (2024) show that prompt robustness varies systematically with model scale, with larger models exhibiting greater stability across prompt variations, while few-shot examples primarily reduce variance rather than improve mean performance. Comprehensive benchmarking by Xie et al. (2024) across 42 financial datasets confirms that LLMs excel at sentiment extraction relative to traditional approaches, but struggle with complex financial reasoning tasks such as forecasting. These findings inform our methodological choice of a standardized Financial Analyst persona prompt for zero-shot classification (Appendix A.4.1). While our consistent prompt design ensures comparability across sentiment sources, the literature suggests that alternative prompt formulations could yield meaningfully different sentiment distributions, a limitation that should temper interpretation of LLM sentiment performance relative to dictionary methods.

A critical methodological concern for LLM-based financial research is *data contamination*. Models like Mistral are trained on vast internet corpora that include historical news archives, financial commentary, and retrospective analysis. When applied to historical news articles, these models may implicitly encode knowledge of subsequent outcomes. For instance, an article about a company in 2010 might be interpreted differently by an LLM that "knows" (from training data) that the company later experienced difficulties. This contamination poses a fundamental challenge for our research design. Even with carefully constructed prompts instructing the model to evaluate sentiment "as of the article date" without hindsight, the model's underlying representations may be contaminated by post-article information. The New York Times archive we use is publicly available and likely included in LLM training corpora, making this concern particularly acute. We attempt to mitigate contamination through prompt engineering that emphasizes contemporary interpretation and by benchmarking against dictionary methods that cannot be contaminated (as they apply fixed word lists without learned representations). However, we acknowledge that complete elimination of contamination bias is infeasible with current LLM architectures, and this limitation should temper interpretation of any apparent LLM advantages.

2.5 Research Gap

Despite extensive literature on both sentiment analysis and earnings prediction, several gaps remain:

First, most studies using Tetlock's methodology rely on historical data. Contemporary validation on recent data (2000–2024) is limited, leaving open questions about whether original findings generalize to modern markets with fundamentally different information environments.

Second, while LLMs have been applied to various financial text classification tasks, systematic comparison of LLM-based sentiment extraction against traditional dictionary methods for earnings prediction with proper out-of-sample validation has not been thoroughly explored.

Third, many studies use simple train-test splits that may overstate predictive performance. Expanding window cross-validation that mimics realistic forecasting scenarios is needed to provide reliable performance estimates.

This research addresses these gaps through a systematic framework comparing Mistral sentiment extraction against Harvard IV-4 dictionary methods across multiple feature configurations using expanding window cross-validation, with explicit controls-only baselines to establish the marginal value of sentiment information.

3 Data

3.1 Overview and Data Collection

This study employs a comprehensive dataset spanning 2000 to 2024, integrating multiple data sources to construct a rich information environment for earnings prediction. The data collection process focuses exclusively on S&P 500 firms to ensure data quality and relevance to institutional investors, resulting in 70,133 firm-quarter observations covering 992 unique companies across 100 quarters. The dataset is derived from 2,259,290 New York Times articles and extensive financial databases from Wharton Research Data Services (WRDS). A complete visual overview of the data preprocessing and model training pipeline is presented in Figure 4 in the Appendix.

3.2 Data Sources

The dataset integrates three primary sources. Financial news is sourced from the New York Times Archive API (2,259,290 articles, 2000–2024), selected for its comprehensive coverage, high editorial standards, and consistent digital archive across the study period. Articles are classified into company-specific and global/macro news using a two-stage classification pipeline described in Section 3.3. Articles are aggregated at the firm-quarter level using the Tetlock window.

Financial data comes from three WRDS databases: Compustat provides quarterly earnings data (943,437 raw firm-quarters, filtered to 70,133 S&P 500 observations); I/B/E/S provides analyst forecast data (forecast dispersion and revision); and CRSP provides daily stock returns for Fama-French three-factor abnormal return computation (using factor data from Kenneth French's Data Library).

Sentiment extraction employs two approaches: (1) Mistral Small 3.2 (24B), a 24-billion parameter instruction-tuned model from the Mistral family (Mistral AI, 2025), classifying article sentiment on a 5-point scale using a Financial Analyst perspective prompt; and (2) Harvard IV-4 Dictionary, the traditional word-counting approach from Tetlock et al. (2008), computing negative word fractions ($\text{neg\_ratio} = \frac{\text{\# negative words}}{\text{total words}}$). Full prompt text and processing details are provided in Appendix A.4.1.

3.3 Article Classification

Classifying 2.3 million New York Times articles into business-relevant categories requires a scalable yet accurate pipeline. We implement a two-stage classification approach: Stage 1 fil-

ters articles for business relevance (distinguishing financial and economic news from sports, entertainment, and lifestyle content), and Stage 2 categorizes relevant articles as either company-specific (mentioning identifiable S&P 500 firms) or global/macro (general market and economic news without firm-specific focus).

For Stage 1, each article's headline and lead paragraph are evaluated for financial relevance. For Stage 2, relevant articles are further classified based on whether they discuss specific companies or broader economic conditions. This distinction is central to our feature engineering: company-specific articles contribute to firm-level sentiment, while global articles capture the macroeconomic information environment shared across all firms.

We selected GPT-4o-mini single-perspective classification for production use based on systematic evaluation against embedding-based alternatives and lexical baselines on 200 manually labeled articles. The single-perspective approach, using only the Financial Analyst persona rather than five expert personas with majority voting, achieved equivalent accuracy to the multi-perspective variant (McNemar's test $p = 0.453$) at 80% lower cost. Stage 1 relevance accuracy is 85.0% and Stage 2 company-specific accuracy is 87.0%. Articles are processed with temperature=0.1 for consistent classification. For company-specific articles, firm-article matching links articles to S&P 500 PERMNOs via GPT-4o-mini entity recognition. Full classification validation results, confusion matrices, and the complete classification prompt are provided in Appendix A.3.

3.4 Sample Construction and Data Quality

The sample construction process involves multiple stages of merging and filtering to ensure data quality:

Table 1: Sample Construction: S&P 500 Dataset

Stage	Observations	Percentage
Raw Compustat firm-quarters (all firms)	943,437	—
After date filter (2000–2024)	851,249	90.2%
After SUE filter (non-missing target)	746,139	79.1%
After PERMNO filter (CRSP match)	592,914	62.8%
After S&P 500 filter	70,135	7.4%
After sentiment filter (final sample)	70,133	7.4%
Tetlock Validation (2000–2010):		
Initial training period	32,819	—
SUE Prediction Comparison (2000–2024):		
Expanding window CV (full sample)	70,133	100.0%
Unique companies (by PERMNO)	992	—

Note: Percentages relative to raw Compustat universe (943,437). S&P 500 membership matched against 1,087 unique PERMNOs from master file. Sentiment coverage: 4,687 firm-quarters (6.7%) with company-specific news; remainder imputed with median. All control variables imputed with missing indicators for variables with >10% missingness.

Infinite SUE values (2,342 observations, 2.85%) arising from zero variance in rolling standardization were removed. SUE and lagged_SUE are winsorized at the 1st and 99th percentiles. The most substantial missing data occurs in Fama-French abnormal returns (25–30%), reflecting insufficient trading history for newly listed firms. Missingness reflects systematic features of financial markets rather than random patterns. We handle missing values through training-set mean imputation with binary missing indicators, preserving missingness information as features; full imputation methodology is described in Appendix 6.

3.5 Variable Construction

Our primary dependent variable follows the seasonal random walk model of Ball and Brown (1968) and Bernard and Thomas (1989):

$$SUE_{it} = \frac{(EPS_{it} - EPS_{it-4}) - \mu_{UE,i}}{\sigma_{UE,i}} \tag{1}$$

where EPS_{it} is diluted EPS for firm i in quarter t, EPS_{it-4} is the seasonal benchmark, and $\mu_{UE,i}$ and $\sigma_{UE,i}$ are rolling mean and standard deviation over 8 prior quarters (minimum 4 required). SUE is winsorized at the 1st and 99th percentiles. ADF tests confirm stationarity (ADF=-50.47, $p < 0.001$), and firm-level autocorrelation is near zero ($\rho = 0.036$), supporting SUE as a valid surprise measure.

Sentiment is measured within the Tetlock window. For each firm-quarter, we separately average

the sentiment of all global and all company-specific articles, then combine these two averages using article counts as weights:

$$s_{it}^{combined} = \frac{N_{it}^{global} s_{it}^{global} + N_{it}^{company} s_{it}^{company}}{N_{it}^{global} + N_{it}^{company}}.$$

In practice, this is equivalent to taking a simple average across all articles regardless of type. For dictionary methods, sentiment equals the fraction of negative words in an article; for LLMs, it is a score on a 5-point scale (see Appendix A.4.1).

We also construct two additional feature sets. The Delta configuration measures how sentiment changes from one quarter to the next ($\Delta s_t^{combined} = s_t^{combined} - s_{t-1}^{combined}$), capturing whether the tone of news coverage is improving or worsening. The Attention configuration uses the log number of articles in the window, $\log(N_{it}^{combined})$, as a proxy for investor attention, together with a binary indicator for whether any company-specific articles appeared.

Following Tetlock et al. (2008), nine control variables capture earnings momentum (lagged SUE), firm characteristics (log market cap, log book-to-market, log share turnover), past returns (Fama-French long-term alpha, pre-announcement and announcement CARs based on the Fama and French (1993) three-factor model), and analyst expectations (forecast dispersion and revision). Complete definitions are in Appendix Table 6.

3.6 Descriptive Statistics and Correlations

Descriptive statistics and correlation matrices for the Tetlock Validation sample (N=32,819) are provided in Appendix Tables 9 and 10. The most relevant pattern for interpreting the results is that the correlation between current and lagged SUE ($\rho = 0.35$) is an order of magnitude larger than any sentiment-SUE correlation ($|\rho| < 0.07$), foreshadowing lagged SUE's dominance in multivariate models.

4 Methodology

4.1 Overview and Research Design

This section details the technical implementation of the analytical framework introduced in the Introduction. The complete data preprocessing and model training pipeline is visualized in Figure 4 in the Appendix.

4.2 Tetlock Validation (2000-2010)

The Tetlock Validation establishes a rigorous baseline by applying Tetlock et al. (2008)'s methodological framework using in-sample analysis on the 2000–2010 period. While we follow Tetlock et al.'s methodological framework, our analysis constitutes a validation and extension rather than a strict replication. We employ the same regression specification and sentiment measurement approach but apply it to: (1) a different news source (New York Times vs. Wall Street Journal and Dow Jones News Service) and (2) a different time period (2000–2010 vs. 1980–2004). These differences preclude direct coefficient comparisons but allow us to test whether the fundamental sentiment-earnings relationship generalizes to modern markets with different information environments. This validation serves one critical purpose:

- Validation: Confirms whether the sentiment-earnings relationship persists with New York Times data (vs. original Wall Street Journal and Dow Jones News Service)

The in-sample approach mirrors Tetlock's original methodology, enabling assessment of whether the sentiment-earnings relationship generalizes to our setting.

We estimate the following OLS regression model exactly as specified in Tetlock et al. (2008):

$$\begin{aligned} SUE_{it} = &\alpha + \beta_1 \text{combined\_neg\_fraction}_{it} + \beta_2 \text{lagged\_SUE}_{it-1} + \beta_3 \text{forecast\_dispersion}_{it} \\ &+ \beta_4 \text{forecast\_revision}_{it} + \beta_5 \log(\text{market\_cap})_{it} \\ &+ \beta_6 \log(\text{book\_to\_market})_{it} + \beta_7 \log(\text{share\_turnover})_{it} \\ &+ \beta_8 \text{FFAlpha\_longterm}_{it} + \beta_9 \text{FFCAR\_preannouncement}_{it} \\ &+ \beta_{10} \text{FFCAR\_announcement}_{it} + \varepsilon_{it} \end{aligned} \tag{2}$$

where SUE_{it} is the standardized unexpected earnings for firm i in quarter t and combined_neg_fraction$_{it}$ is the Harvard IV-4 negative word fraction averaged across global and company-specific articles. Control variables include earnings momentum, firm characteristics,

past returns, and analyst forecasts. The error term ε_{it} is clustered by quarter to account for time-series correlation.

OLS with Clustered Standard Errors: We use ordinary least squares with Newey and West (1987) heteroskedasticity and autocorrelation consistent (HAC) standard errors, clustered by quarter-year. Clustering accounts for cross-sectional correlation in earnings surprises (common shocks affect multiple firms) and time-series correlation within firms (earnings persistence not fully captured by controls).

Winsorization: SUE and lagged_SUE are winsorized at the 1st and 99th percentiles to reduce leverage of extreme outliers in the target and its primary predictor. Other control variables, bounded by construction (e.g., log-transformed) or already well-behaved, are left unwinsorized.

Complete-Case Analysis: We use listwise deletion for missing control variables to follow Tetlock's approach. This results in $N = 32{,}819$ observations from the 2000–2010 period for S&P 500 firms with complete data.

The 2000–2010 period serves as the validation window for several reasons. First, it encompasses Tetlock's original study period end (2004), enabling comparison. Second, it post-dates Regulation Fair Disclosure (October 2000), representing the modern information environment. Third, it provides sufficient observations for stable parameter estimation. Finally, it serves as the initial training period for the SUE Prediction Comparison's expanding window validation.

The Tetlock Validation tests three hypotheses. First, H1a (Sentiment Sign): the sentiment coefficient should be negative ($\beta_1 < 0$), indicating that negative sentiment predicts lower earnings surprises, consistent with Tetlock et al. (2008). Second, H1b (Statistical Significance): β_1 should be statistically significant at the 5% level ($|t| > 2.0$). Third, H1c (Model Fit): the model should achieve similar explanatory power (Adjusted R^2) to the original study, confirming that the methodology generalizes to modern data.

4.3 SUE Prediction Comparison (2000-2024)

The SUE Prediction Comparison extends the analysis to the full 25-year period (2000–2024) using an expanding window cross-validation framework to evaluate whether sentiment provides incremental predictive value over a controls-only benchmark. Key objectives:

1. Test whether sentiment, from either an LLM (Mistral) or a traditional dictionary (Harvard IV-4), provides incremental predictive value over financial controls alone
2. Evaluate seven model architectures across four feature configurations
3. Test temporal stability through rolling out-of-sample predictions

4. Identify the best-performing model combination for economic significance analysis

We employ an expanding window approach that preserves temporal order and prevents look-ahead bias:

1. Initial Training Period: 2000–2010 (same period as the Tetlock Validation)
2. Validation Window: Rolling 8-quarter (2-year) validation period
3. Test Window: 1-month ahead predictions for final evaluation
4. Expansion: Training window grows by one month at each iteration

This design ensures that models are always evaluated on truly out-of-sample data, providing realistic estimates of predictive performance.

We evaluate seven model architectures spanning linear, regularized, tree-based, and ensemble methods.

Linear Models: *OLS* (Ordinary Least Squares) minimizes the sum of squared residuals without any penalty term, serving as the unregularized baseline. *Ridge* regression adds an L2 penalty ($\lambda \|\beta\|_2^2$) that shrinks coefficients toward zero without eliminating them, reducing variance from multicollinearity ($\lambda \in \{0.1, 1, 10, 100, 1000\}$). *Lasso* regression imposes an L1 penalty ($\lambda \|\beta\|_1$) that drives irrelevant coefficients exactly to zero, performing implicit feature selection ($\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$). *ElasticNet* combines both L1 and L2 penalties, balancing Lasso's sparsity with Ridge's grouping of correlated predictors ($\lambda \in \{0.001, 0.01, 0.1, 1\}$, l1_ratio $\in \{0.1, 0.5, 0.9\}$).

Tree-based Models: *Random Forest* constructs an ensemble of independently grown decision trees on bootstrap samples and averages their predictions, reducing variance through decorrelation ($n\_estimators \in \{50, 100, 200\}$, $max\_depth \in \{5, 10, 20, None\}$). *XGBoost* (eXtreme Gradient Boosting) sequentially fits decision trees to the residuals of the current ensemble, optimizing a regularized objective function that controls model complexity ($learning\_rate \in \{0.01, 0.03, 0.05\}$, $max\_depth \in \{2, 3\}$, $n\_estimators \in \{50, 100, 150\}$, with strong L1/L2 regularization). Full hyperparameter grids are in Appendix Tables 11–13.

Ensemble: The *Ensemble* combines the two tree-based models (Random Forest and XGBoost) into a single prediction using *validation-weighted averaging*, where each base model's weight is proportional to its validation-set R^2:

$$\hat{y}_i^{ens} = \frac{\sum_{m=1}^{2} w_m \cdot \hat{y}_i^{(m)}}{\sum_{m=1}^{2} w_m}, \quad w_m = \max(R^2_{val,m}, 0) \tag{3}$$

We evaluate four feature configurations: (1) Controls Only (9 Tetlock controls as baseline), and for each of the two sentiment sources (Mistral, HIV Dictionary): (2) Sentiment (controls + combined sentiment), (3) Delta (controls + quarter-over-quarter sentiment change), and (4) Attention (controls + log article counts as investor attention proxies, without sentiment scores).

With 5 sentiment-augmented configurations (2 per sentiment source plus source-independent Attention) $\times$ 7 architectures, plus 7 controls-only models, we evaluate 42 total model specifications. The controls-only configuration serves as the primary benchmark: any sentiment-augmented model must outperform this baseline to demonstrate incremental predictive value beyond standard financial variables.

Within each expanding window, nested cross-validation selects hyperparameters maximizing validation R^2, ensuring out-of-sample evaluation is uncontaminated. The primary performance metric is out-of-sample R^2 computed across rolling test windows.

4.4 Ablation and Economic Significance

We use SHAP (SHapley Additive exPlanations) feature importance (Shapley, 1953; Lundberg and Lee, 2017) to quantify the marginal contribution of each feature to predictive performance. SHAP values are computed from the Ensemble's two tree-based components (RandomForest and XGBoost) using TreeSHAP, then combined as a validation-R^2-weighted average. Linear models are excluded because kernel SHAP is computationally prohibitive at this scale and tree-based models dominate Ensemble weight, making the TreeSHAP-based decomposition a close approximation of overall Ensemble feature attribution.

To assess economic significance, we construct quartile-sorted portfolios at each of 48 test quarters (2013Q1–2024Q4). Each quarter, firms are sorted by predicted SUE into quartiles; the long-short strategy buys Q4 (highest predicted SUE) and sells Q1 (lowest). Equal-weighted daily returns are collected within an event window of $[-1,+1]$ trading days around each firm's earnings announcement date (RDQ), with $[-1,+2]$ and $[-1,+3]$ windows as sensitivity checks. Calendar-time portfolio returns are computed as $R_{LS,t} = R_{long,t} - R_{short,t}$. To assess whether returns compensate for systematic risk, we regress daily long-short returns on the Fama and French (1993) three-factor model with Newey-West standard errors. Transaction costs of 10 basis points one-way are applied to each leg entry and exit (40 bps total per event).

5 Results

5.1 Tetlock Validation (2000-2010)

Table 2 presents OLS regression coefficients from our validation of the Tetlock et al. (2008) specification on 2000–2010 S&P 500 data:

Table 2: OLS Regression Coefficients: Tetlock Validation (2000-2010)

Variable	Coef	t-stat
const	0.516	1.79
combined_neg_fraction	−3.160	−2.07**
lagged_SUE	0.339	19.20***
log_market_cap	−0.020	−2.34**
log_book_to_market	−0.024	−0.71
log_share_turnover	0.006	0.31
FFCAR_announcement	0.334	1.11
FFCAR_preannouncement	0.359	3.55***
FFAlpha_longterm	0.236	5.57***
forecast_dispersion	−0.064	−4.36***
forecast_revision	20.87	7.70***
Model Statistics:		
N	32,819	
N Clusters	49	
Adj. R^2	0.131	

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Standard errors clustered by quarter (2000–2010).

The combined negative fraction exhibits a significant negative coefficient (−3.160, t=−2.07), consistent in sign with Tetlock et al. (2008)'s finding that negative sentiment predicts lower earnings surprises (−0.064, t=−4.69). The raw coefficient is roughly 50 times larger than Tetlock's because our sentiment variable, averaged across hundreds of articles per firm-quarter, has an extremely compressed distribution ($\sigma = 0.011$, range 0.090–0.151). Tetlock's measure aggregates fewer and more targeted WSJ/Dow Jones articles per firm-quarter, retaining greater cross-sectional dispersion. OLS mechanically inflates the coefficient when the predictor has narrow dispersion. In standardized terms, a one-standard-deviation increase in combined_neg_fraction corresponds to only a 0.035-standard-deviation decrease in SUE, a negligible economic effect. The dominant predictor is lagged_SUE (0.339, t=19.20), with a standardized effect of 0.335, nearly ten times larger than sentiment, indicating substantial earnings persistence. Among the other controls, log_market_cap (−0.020, t=−2.34) indicates that larger firms have lower surprise magnitudes. FFCAR_preannouncement (0.359, t=3.55) and FFAlpha_longterm (0.236, t=5.57) capture momentum effects. Forecast_revision (20.87, t=7.70) shows that analyst updates con-

tain predictive information. Log_share_turnover loses significance (t=0.31), possibly reflecting changes in market microstructure post-2000.

Our validation achieves comparable Adjusted R^2 (0.131 vs. Tetlock's 0.119), demonstrating that the fundamental sentiment-earnings relationship persists despite different data sources, time periods, and sample composition. However, the economically trivial standardized sentiment effect suggests that statistical significance alone overstates the practical relevance of news tone, a theme the SUE Prediction Comparison addresses next.

5.2 SUE Prediction Comparison (2000-2024)

The SUE Prediction Comparison evaluates two sentiment extraction approaches (Mistral, HIV Dictionary) across seven ML architectures using expanding window cross-validation. Table 3 presents the main results.

Table 3: SUE Prediction Comparison: Average Test R² Across Test Periods by Sentiment Model and ML Architecture (%)

Configuration	OLS	Ridge	Lasso	ENet	RF	XGB	Ensemble
Baseline:							
Controls Only	1.24	1.26	1.38	1.74	5.14	5.30	5.47
HIV Dictionary + Controls:							
Sentiment	0.96	0.98	1.20	1.58	4.94	4.90	5.14
Delta	1.27	1.28	1.39	1.75	5.11	5.18	5.40
Attention (Article Count Only):							
Attention	1.29	1.30	1.49	1.85	4.95	5.07	5.18
Mistral + Controls:							
Sentiment	1.21	1.23	1.37	1.73	4.96	4.64	5.16
Delta	1.24	1.26	1.40	1.76	5.14	5.30	5.45

Note: Test R² from expanding window cross-validation (2000-2010 initial training, 8-quarter validation, 1-month test). N≈49,000 training, ≈5,300 validation, ≈660 test per window. RF = RandomForest; XGB = XGBoost; ENet = ElasticNet. Ensemble uses validation-weighted averaging of Random Forest and XGBoost.

The controls-only Ensemble achieves the best test R² (5.47%), outperforming all sentiment-augmented configurations. The validation-weighted Ensemble also outperforms both XGBoost (5.30%) and RandomForest (5.14%) for controls-only, with this advantage extending across most configurations. Adding Mistral sentiment does not help: Mistral sentiment+controls achieves only 5.16% (Ensemble), slightly worse than controls alone, and performs comparably to the HIV dictionary (5.14%). Across the board, tree-based models (RandomForest, XGBoost, Ensemble) outperform linear models regardless of configuration. Delta features (sen-

timent change over time) are largely neutral, with Mistral Delta achieving the best sentiment-augmented result (5.45% Ensemble).

Figure 1 displays out-of-sample R^2 across expanding windows using a 4-quarter moving average to smooth quarterly noise. The rolling analysis reveals high volatility in predictive performance across time periods, with no systematic advantage for any sentiment model. The controls-only baseline maintains competitive performance throughout, achieving the highest average R^2 of 5.47%. A notable structural break appears around 2018, where all models experience substantial performance degradation during the 2018–2021 period before recovering in 2022–2024.

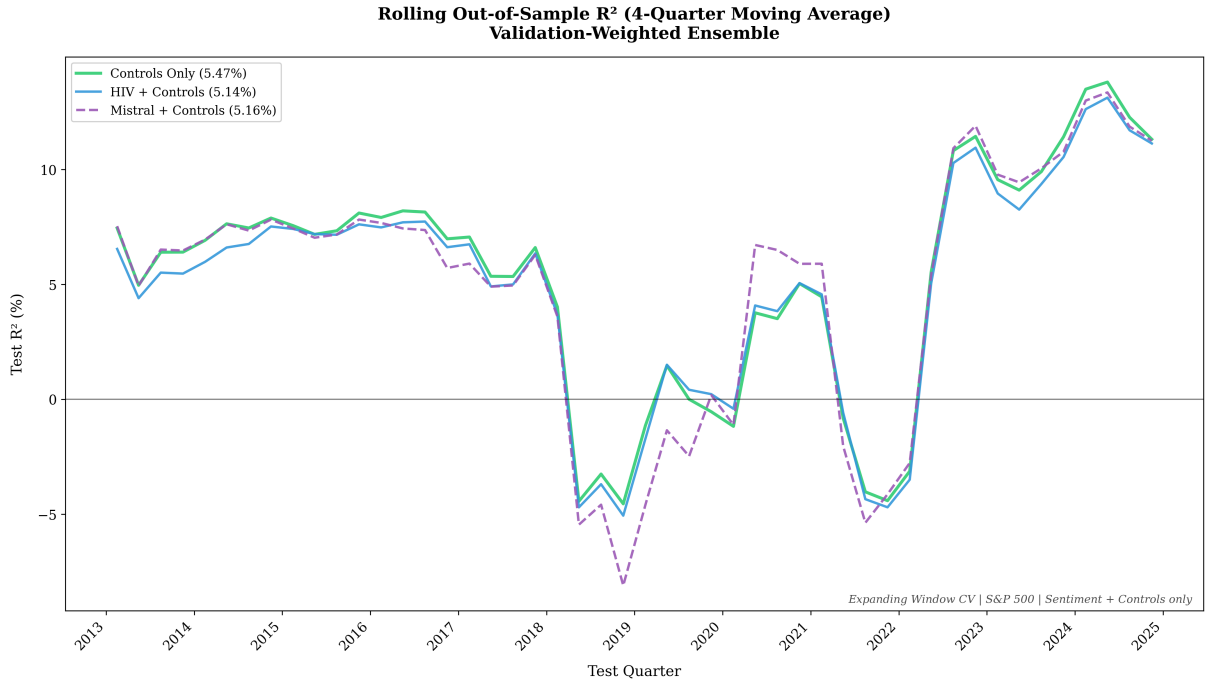

Figure 1: Rolling Out-of-Sample Test R^2 (4-Quarter Moving Average) for Ensemble Models. The controls-only model (solid green) achieves the highest average R^2 (5.47%), with sentiment-augmented models showing no consistent improvement.

To formally assess whether performance differences are statistically significant, we conduct Diebold-Mariano tests comparing each sentiment configuration against the controls-only baseline across all 48 expanding windows (see Appendix A.9 for full results). The tests reveal that no sentiment configuration significantly outperforms controls-only. In fact, HIV Sentiment+Controls (DM=-3.33, p=0.001) and Attention+Controls (DM=-2.09, p=0.037) perform *significantly worse* than the baseline. The best-performing configuration, Mistral Delta+Controls, shows a negligible R^2 difference ($<0.1\%$) that is not statistically distinguishable from zero (DM=-0.41, p=0.682). Furthermore, Diebold-Mariano tests comparing Mistral against HIV dictionary sentiment show no significant differences (Appendix Table 16), confirming that even LLM-based sentiment does not significantly outperform the traditional

dictionary approach.

The rolling R^2 patterns in Figure 1 suggest substantial time variation in predictive performance. To examine this, we divide the 48 test quarters into three sub-periods and report average Ensemble R^2 descriptively (Table 4). Sub-period Diebold-Mariano tests are not conducted due to insufficient quarters per period for reliable inference.

Table 4: Structural Break Analysis: Average SUE Prediction R^2 (%) by Sub-Period

Source	Config	Full	Pre-2018	2018–2021	Post-2022
Controls	Controls Only	**5.47**	7.20	−1.12	**11.39**
HIV	Sentiment	5.14	6.71	−1.12	10.87
HIV	Delta	5.40	7.10	−1.20	11.35
–	Attention	5.18	7.28	−1.29	10.32
Mistral	Sentiment	5.16	6.82	−1.54	11.31
Mistral	Delta	**5.45**	7.16	−1.16	**11.43**

Note: Average out-of-sample R^2 (%) from expanding window cross-validation. All configurations use the validation-weighted Ensemble. Pre-2018: 2013Q1–2017Q4 (20 quarters); 2018–2021: 2018Q1–2021Q4 (16 quarters); Post-2022: 2022Q1–2024Q4 (12 quarters). Bold indicates best-performing configuration per column.

The sub-period results reveal distinct regimes. The pre-2018 period exhibits stable predictability, with Attention marginally outperforming controls (7.28% vs. 7.20% R^2). The 2018–2021 period produces negative R^2 for all configurations; the degradation begins in 2018, two years before the COVID-19 pandemic, suggesting that macroeconomic regime shifts gradually eroded historical earnings relationships, with the pandemic intensifying this breakdown. The post-2022 period shows the highest R^2 across all configurations (10.32–11.43%), with Mistral Delta slightly exceeding controls (11.43% vs. 11.39%), reflecting both a return to predictable earnings patterns and the expanding training window's incorporation of crisis-era data. Despite these marginal advantages in individual sub-periods, no sentiment configuration achieves a statistically significant improvement over controls on average, confirming that the sub-period patterns primarily reflect macroeconomic regime shifts rather than consistent sentiment informativeness. For context, Ensemble in-sample R^2 ranges from 14.55% (controls-only) to 14.83% (Mistral Sentiment), highlighting a substantial overfitting gap relative to out-of-sample performance (5.14–5.47%). The complete R^2 breakdown by sub-period for all configurations is provided in Appendix Table 17; we discuss the implications in Section 6.

The disconnect between the Tetlock Validation's in-sample significance ($t = -2.07$) and the SUE Prediction Comparison's out-of-sample null motivates a deeper investigation into which features actually drive the models' predictions. The Ablation addresses this through SHAP-based feature decomposition and economic significance testing.

5.3 Ablation and Economic Significance

Figure 2 presents SHAP-based feature importance from the Ensemble model (validation-weighted average of RandomForest and XGBoost contributions), quantifying each feature's contribution to predictions. The analysis reveals a striking hierarchy: lagged SUE dominates with 64–68% of total feature importance in sentiment-augmented models, confirming that past earnings surprises are by far the strongest predictor of future surprises. Combined news sentiment contributes 6–8% of importance, meaningful but far less than lagged SUE, with Mistral's combined sentiment (7.6%) exceeding the HIV dictionary (6.2%). Delta (quarter-over-quarter change) features contribute less (2.6–3.4%), consistent with their marginal R^2 improvement. Financial controls collectively exceed sentiment features, with FFAlpha_longterm, forecast_revision, and FFCAR_preannouncement each providing meaningful contributions.

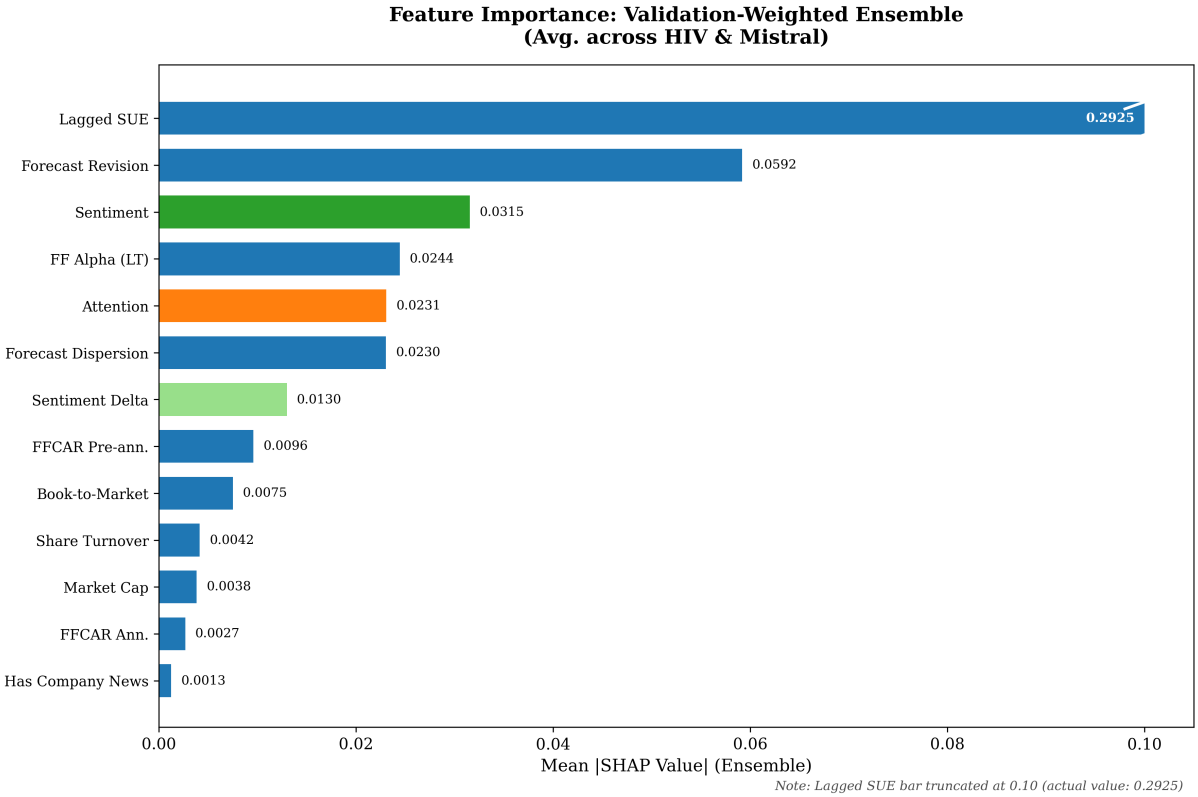

Figure 2: SHAP Feature Importance: Sentiment Models with Controls (Ensemble). Lagged SUE dominates prediction (64–68% importance for sentiment models, 68% for controls-only), followed by combined sentiment (6–8% for Sentiment configurations, 2.6–3.4% for Delta configurations).

Decomposing the Ensemble SHAP values into constituent model contributions reveals instructive differences between RandomForest and XGBoost. RandomForest concentrates substantially more importance on lagged SUE compared to XGBoost, reflecting fundamental algorith-

mic differences. RandomForest's bagging procedure creates trees that independently discover the dominant lagged SUE signal. XGBoost's sequential boosting, by contrast, progressively focuses residual attention on secondary features including sentiment. These algorithmic differences have practical consequences. RandomForest's lower baseline sentiment allocation means it is more sensitive to changes in sentiment signal strength across sub-periods, while XGBoost's higher baseline allocation leaves less room for adjustment. The Ensemble's intermediate SHAP values represent a more balanced feature attribution, averaging RF's conservative and XGBoost's liberal sentiment weighting.

Table 5 presents Fama and French (1993) three-factor risk-adjusted returns for the $[-1,+1]$ event window across sentiment configurations using the validation-weighted Ensemble. For comparability, a random-sort null baseline, averaging 500 permutations of random quartile assignment, is included to confirm that the trading pipeline does not generate spurious alpha. All long-short strategies produce negative risk-adjusted alphas, and no alpha is statistically significant at conventional levels.

Table 5: Fama-French 3-Factor Risk-Adjusted Returns (Ensemble, $[-1,+1]$ Window)

Source	Config	α (%/yr)	t-stat	β_{Mkt}	β_{SMB}	β_{HML}	R^2
Baseline	*Random Sort (Null)*	−0.04	−0.00	0.00	−0.01	0.00	0.002
Controls	Controls Only	−14.53	−0.55	−0.11	−0.31	−0.28	0.006
HIV	Sentiment	−2.50	−0.09	−0.10	−0.34	−0.31	0.007
HIV	Delta	−12.42	−0.46	−0.13	−0.28	−0.29	0.006
Mistral	Sentiment	−18.22	−0.67	−0.05	−0.23	−0.30	0.005
Mistral	Delta	−21.43	−0.80	−0.12	−0.35	−0.33	0.008
–	Attention	−1.48	−0.06	−0.16	−0.33	−0.27	0.007

Note: Fama-French 3-factor regression on daily calendar-time long-short portfolio returns. Newey-West standard errors. Ensemble uses validation-weighted averaging of Random Forest and XGBoost. α expressed as annualized percentage ($\alpha_{\text{daily}} \times 252$). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. *Random Sort (Null)* averages over 500 permutations of random quartile assignment; near-zero alpha confirms the null baseline. Sample: 2013Q1–2024Q4 (48 quarters). Results for $[-1,+2]$ and $[-1,+3]$ windows are qualitatively similar (Appendix Table 18).

The trading results are uniformly negative. All Fama and French (1993) three-factor alphas are negative across all six configurations, with no alpha statistically significant at conventional levels. Annualized alphas range from -1.48% (Attention + Controls, $t = -0.06$) to -21.43% (Mistral Sentiment Delta, $t = -0.80$), with the Controls-only baseline at -14.53% ($t = -0.55$). No sentiment configuration improves upon controls. Results for the $[-1,+2]$ and $[-1,+3]$ windows are qualitatively identical (Appendix Table 18). The near-zero R^2 (0.005–0.008) indicates that systematic risk factors explain virtually none of the long-short return variation, consistent with idiosyncratic earnings news dominating in short event windows. Figure 3 illustrates the

cumulative long-short return paths for all models.

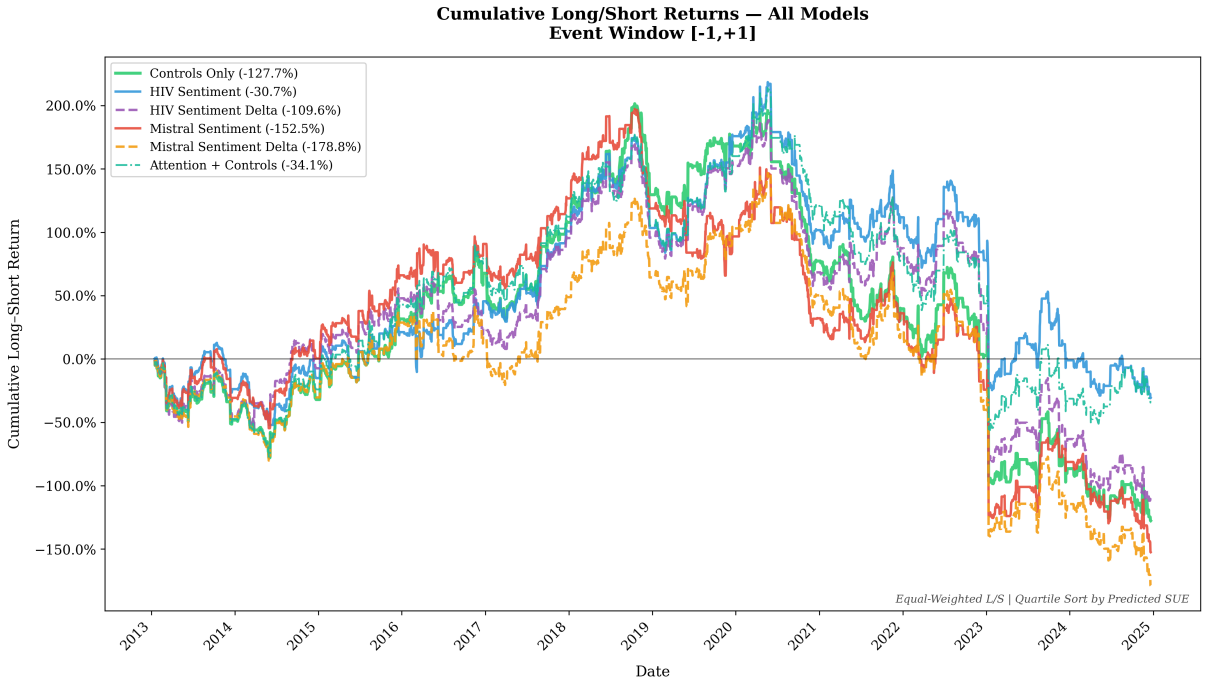

Figure 3: Cumulative Long-Short Returns by Model ($[-1, +1]$ Event Window). All models exhibit negative cumulative returns over the 48-quarter test period (2013Q1–2024Q4), with no sentiment configuration outperforming the controls-only baseline.

6 Discussion

Financial controls consistently dominate sentiment-augmented models across all architectures and time periods. At quarterly horizons, news sentiment functions as noise rather than signal for S&P 500 earnings prediction. The information journalists convey about earnings is already captured by lagged SUE and the other financial controls, making sentiment redundant. Worse, adding weakly correlated features ($|\rho| < 0.07$ with SUE) forces models to estimate extra parameters on mostly random variation, which degrades out-of-sample performance. Sentiment-augmented models perform equal to or worse than controls-only across all 42 specifications (Table 3), confirming that sentiment adds no new information and the extra degrees of freedom hurt generalization.

The statistical evidence reinforces this interpretation. Lagged SUE accounts for 64–68% of Ensemble SHAP importance (Figure 2), consistent with its pairwise correlation with current SUE ($\rho = 0.35$), an order of magnitude larger than any sentiment measure. This reflects the earnings autocorrelation documented by Bernard and Thomas (1989): approximately two-thirds of the predictable variation in next-quarter surprises is captured by the previous quarter's surprise alone. Any incremental news signal must compete with, and fails to improve upon, this powerful baseline.

This dominance of controls is consistent with van Binsbergen et al. (2023), Chen et al. (2022), and So (2013), who show that ML models for earnings prediction derive value from known financial regularities rather than novel signals. Our results extend this to the text domain: tree-based models achieve meaningful R^2 by capturing non-linear interactions among financial controls, not by extracting value from sentiment. Notably, combined sentiment receives 6–8% of SHAP importance yet does not improve out-of-sample R^2. This apparent paradox has a simple explanation: during training, models occasionally use sentiment to make predictions, but when sentiment is removed, the models use controls to make the same predictions just as accurately. Sentiment is a substitute for controls, not an independent source of information. The attributed importance thus reflects in-sample model behavior, not genuine predictive contribution.

The null result for LLM-based sentiment, where Mistral performs comparably to the HIV dictionary but neither outperforms controls (Table 3), contradicts optimistic assessments such as Lopez-Lira and Tang (2025), who report that ChatGPT sentiment predicts next-day stock returns. However, those studies examine daily or weekly return prediction, a fundamentally different task from quarterly earnings forecasting. At daily horizons, sentiment may capture short-lived attention effects (Hirshleifer et al., 2009) that dissipate before our quarterly aggregation window closes. The discrepancy likely reflects *horizon dependence* of sentiment informativeness rather than contradictory evidence. Several additional factors may explain the LLM's failure to outperform dictionaries. LLMs trained on internet-scale corpora optimize for

general language understanding, not for extracting market-relevant information from financial text. As a result, the model may respond to linguistic features (writing style, article length, narrative complexity) that are orthogonal to earnings prediction. Furthermore, our quarterly aggregation smooths away any event-day sentiment signals that LLMs might detect, so the information content that justifies the computational cost of LLM inference may reside at horizons our framework cannot capture. Studies applying BERT-based models (Huang et al., 2023) have similarly reported improvements in sentiment labeling accuracy, but classification accuracy and earnings forecasting value are distinct objectives; superior contextual understanding does not necessarily translate to superior prediction when the fundamental relationship between text and quarterly outcomes is weak. Our null result provides a necessary counterweight to publication bias, establishing boundary conditions for when NLP approaches add value.

The sub-period analysis (Table 4) reveals that R^2 variation across regimes is driven by changes in earnings persistence, not sentiment informativeness, consistent with Garcia (2013). Because lagged SUE accounts for 64–68% of model importance, when macroeconomic disruptions break the link between last quarter's earnings surprise and this quarter's, as during 2018–2021, the entire prediction framework loses its dominant input. This is a property of the earnings generating process, not of market prices or sentiment. Sentiment-augmented models track the same trajectory as controls-only in every sub-period, and during the 2018–2021 breakdown they degraded *more* than controls-only, confirming that sentiment amplifies forecast errors rather than compensating for weakened earnings persistence. The in-sample significance of dictionary sentiment in the Tetlock Validation ($t = -2.07$) alongside its out-of-sample failure illustrates the broader pattern: in-sample relationships that do not generalize under proper temporal validation.

Our trading strategy results constitute evidence for semi-strong market efficiency (Fama, 1970). Despite meaningful out-of-sample R^2, Fama and French (1993) three-factor alphas are uniformly negative and statistically insignificant across all models and event windows (Table 5; Appendix Table 18). The predictable component of earnings, driven by lagged SUE, is publicly available and already reflected in stock prices (Welch and Goyal, 2008). The Ensemble achieves the highest R^2 (5.47%) yet produces a negative annualized alpha of -14.53% ($t = -0.55$), as model averaging compresses the forecast distribution and weakens the tail discrimination that quartile sorting requires.

It is worth distinguishing two separate mechanisms behind these results, which the literature often conflates. Sentiment's inability to improve earnings predictions reflects the noise argument above: news is redundant with financial controls. This is a property of the earnings generating process itself, where persistence arises from business momentum, accounting conservatism, and analyst learning. The uniformly negative trading alphas, by contrast, constitute direct evidence for semi-strong efficiency: even the predictable component of earnings (driven by lagged

SUE) is publicly known and already impounded in stock prices (Fama, 1970). This distinction matters because in-sample relationships that fail out-of-sample are better characterized as non-robust rather than "arbitraged away," since arbitrage operates on prices, not on earnings. Statistical predictive performance (R^2) and portfolio profitability also require fundamentally different conditions. R^2 measures average forecast precision across the cross-section, while long-short returns depend on precision at the tails where the strategy concentrates.

Several limitations constrain the interpretation and generalizability of these findings. A first concern is data contamination. Mistral was trained on internet-scale text corpora that almost certainly include historical news archives, including the New York Times. Despite our careful prompt design instructing the model to evaluate sentiment as it would have appeared at publication and to ignore hindsight knowledge, we cannot guarantee that the model does not encode information about what occurred after articles were published. For instance, a 2008 article expressing uncertainty about a company might receive different sentiment treatment because the model has learned that the financial crisis subsequently occurred. This contamination could bias sentiment estimates in either direction and represents an inherent limitation of using pre-trained LLMs for temporal prediction tasks.

Additionally, we employ a single prompt design, the Financial Analyst persona, for all LLM sentiment extraction. While this prompt was carefully constructed based on prompt engineering best practices, no ablation study was conducted across prompt variations. Different prompt framings could yield materially different sentiment scores and potentially different forecasting results.

Our analysis focuses on S&P 500 constituents, but index composition changes over time. Firms that exit the index due to bankruptcy, merger, or declining market capitalization may have systematically different sentiment-earnings dynamics than surviving firms. If distressed firms exhibit stronger sentiment effects as negative news coverage intensifies before exit, our sample may understate the true predictive relationship. Moreover, quarterly sentiment aggregation necessarily sacrifices granular information about sentiment dynamics. Event-day sentiment spikes are smoothed into quarterly averages; higher-frequency prediction horizons might reveal sentiment effects invisible at the quarterly level. Finally, relying exclusively on the New York Times constrains the information environment; alternative sources such as earnings call transcripts, SEC filings, or social media may contain different signals. Our LLM-based article classification pipeline achieves 85–87% accuracy, meaning 13–15% of articles are misclassified, with errors propagating through the aggregation pipeline and potentially diluting sentiment signals. Beyond these individual concerns, our results are conditioned on specific design choices (a single aggregation window of 28 trading days, a single prediction horizon of one quarter ahead, and a fixed set of nine controls) whose interactions were not systematically explored.

For practitioners developing earnings forecasting systems, these results deliver a clear message:

for S&P 500 and similarly efficient large-cap markets, the marginal cost of NLP infrastructure (API fees, prompt engineering, article classification pipelines) is unlikely to be justified by prediction improvements that are effectively zero across 42 specifications. The prediction-to-profit gap has direct practical consequences: quarterly rebalancing implies high portfolio turnover, and even the best long-short configurations produce negative risk-adjusted alphas after transaction costs of 10 basis points per leg. Resources would be better directed toward improving control variable quality (more timely analyst forecast data, refined earnings momentum measures) or targeting less efficient market segments where sentiment may retain informational value. For researchers, our expanding window methodology with controls-only baselines, SHAP decomposition, and Diebold-Mariano tests provides a template for rigorous sentiment evaluation that avoids the in-sample overfitting common in the literature. For LLM applications specifically, our findings suggest that off-the-shelf models without domain-specific fine-tuning may not outperform simpler dictionary alternatives for quarterly earnings prediction.

7 Conclusion

This dissertation investigated whether news sentiment, extracted via dictionary methods (Harvard IV-4) or modern LLMs (Mistral), provides incremental predictive power for S&P 500 earnings surprises beyond financial controls. Across 42 model specifications spanning seven architectures, the answer is consistently negative: no sentiment configuration significantly outperforms the controls-only baseline. The simplest explanation is that news sentiment operates as noise rather than signal for quarterly earnings prediction of large-cap firms; the information journalists convey is already captured by the structured financial variables (notably lagged SUE) that controls encode directly.

The null result is nonetheless informative. Predictive performance varies dramatically across sub-periods because earnings persistence itself fluctuates with macroeconomic conditions, yet sentiment never outperforms controls in any regime. SHAP decomposition reveals a clear hierarchy: combined sentiment receives 6–8% of feature importance yet does not improve R^2, reflecting in-sample attribution redistribution rather than genuine predictive contribution. And even where statistical predictability exists, it does not translate into economic value: the best R^2 (5.47%) yields uniformly negative Fama-French three-factor alphas (all statistically insignificant) across models and event windows, confirming market efficiency for S&P 500 firms.

These findings contribute to the literature in several ways. We provide the first expanding window validation of Tetlock et al. (2008), showing that in-sample dictionary sentiment significance ($t = -2.07$) does not survive out-of-sample evaluation. We establish that Mistral does not outperform the Harvard IV-4 dictionary for earnings prediction. The controls-only baseline demonstrates that sentiment's apparent value in prior work may reflect inadequate benchmarking.

The limitations discussed in Section 6 constrain the generalizability of these findings and motivate several directions for future research. The sub-period analysis suggests investigating whether sentiment informativeness co-varies systematically with macroeconomic regimes. Less covered markets, such as small-cap stocks and emerging markets, may exhibit slower information processing where sentiment signals persist long enough to be exploitable. Domain-specific LLM fine-tuning, which our off-the-shelf Mistral approach forgoes, could improve the extraction of earnings-relevant information from text. Higher-frequency prediction horizons might reveal sentiment effects that quarterly aggregation obscures, and alternative text sources, such as earnings call transcripts, SEC filings, and social media, may contain sentiment signals absent from general news coverage. The frontier of NLP-finance research lies not in demonstrating that sentiment predicts large-cap earnings, but in identifying the conditions (market segments, time regimes, and information types) where it can generate economically meaningful value.

Word Count: 8,418 words.

References

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2), 159–178.

Ball, R. T., & Ghysels, E. (2018). Automated earnings forecasts: Beat analysts or combine and conquer? *Management Science*, 64(10), 4936–4952. https://doi.org/10.1287/mnsc.2017.2864

Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27, 1–36.

Chen, X., et al. (2022). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research*. https://doi.org/10.1111/1475-679x.12429

Cookson, J., et al. (2024). The social signal. *Journal of Financial Economics*. https://doi.org/10.1016/j.jfineco.2024.103870

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1), 1–33. https://doi.org/10.1111/j.1540-6261.2010.01624.x

Hong, H., Lim, T., & Stein, J. C. (2000). Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance*, 55(1), 265–295. https://doi.org/10.1111/0022-1082.00206

Ke, Z. T., Kelly, B. T., & Xiu, D. (2020). Predicting returns with text data. Working Paper, National Bureau of Economic Research. https://doi.org/10.3386/w26186

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, 79(5), 3105–3147. https://doi.org/10.1111/jofi.13377

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679x.2010.00382.x

Lopez-Lira, A., & Tang, Y. (2025). Can ChatGPT forecast stock price movements? Return predictability and large language models. Working Paper, University of Florida.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

van Binsbergen, J. H., Han, X., & Lopez-Lira, A. (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of Financial Studies*, 36(6), 2361–2396. https://doi.org/10.1093/rfs/hhac085

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307–317.

So, E. C. (2013). A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, 108(3), 615–640. https://doi.org/10.1016/j.jfineco.2013.02.002

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.

OpenAI. (2023). GPT-4 Technical Report. Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

Mistral AI. (2025). Mistral Small 3.2 (24B). Technical Report. https://mistral.ai/news/mistral-small-3-1/

Welch, I., & Goyal, A. (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4), 1455–1508. https://doi.org/10.1093/rfs/hhm014

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1950–1976.

Xie, Q., Han, W., Lai, Y., Peng, M., & Huang, J. (2024). FinBen: A Holistic Financial Benchmark for Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 37.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Meursault, V., Liang, P. J., Routledge, B. R., & Scanlon, M. M. (2023). PEAD.txt: Post-earnings-announcement drift using text. *Journal of Financial and Quantitative Analysis*, 58(6), 2299–2326. https://doi.org/10.1017/S0022109022001181

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267–1300. https://doi.org/10.1111/jofi.12027

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67–97. https://doi.org/10.1111/j.1540-6261.2010.01626.x

Hirshleifer, D., Lim, S. S., & Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5), 2289–2325. https://doi.org/10.1111/j.1540-6261.2009.01501.x

Hansen, J. H., & Siggaard, M. (2024). Double machine learning: Explaining the post-earnings announcement drift. *Journal of Financial and Quantitative Analysis*, 59(3), 1003–1030. https://doi.org/10.1017/S0022109023000534

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708. https://doi.org/10.2307/1913610

A Appendix

A.1 Variable Definitions and Construction

Table 6: Complete Variable Definitions

Variable	Definition
Target Variable:	
SUE	Standardized Unexpected Earnings: $SUE_t = \frac{EPS_t - E[EPS_t]}{\sigma(EPS)}$. Winsorized at 1%/99%.
Control Variables (9 features):	
Lagged SUE	One-quarter lagged SUE (SUE_{t-1}). Controls for earnings momentum.
Log(Market Cap)	Natural log of market capitalization. Controls for firm size.
Log(B/M)	Natural log of book-to-market ratio. Controls for value vs. growth.
Log(Turnover)	Natural log of average daily share turnover. Controls for liquidity.
FFCAR Announcement	Fama-French abnormal return on trading day -2 relative to RDQ. Captures immediate pre-announcement price discovery.
FFCAR Pre-ann.	Cumulative FF abnormal returns from day -30 to -3. Pre-announcement drift.
FFAlpha Long-term	Cumulative FF abnormal returns from day -252 to -31. Long-term momentum.
Forecast Dispersion	Standard deviation of analyst EPS forecasts. Analyst disagreement.
Forecast Revision	Change in median forecast over 90 days: $\Delta forecast/price$.
Sentiment Variables (per model: HIV, Mistral):	
Combined Sentiment	Average of global and company-specific sentiment. HIV: negative word fraction; Mistral: 1-5 scale.
Delta Configuration (sentiment change):	
Combined Sent. Delta	Change in combined sentiment vs. prior quarter: $\Delta s_t^{combined} = s_t^{combined} - s_{t-1}^{combined}$.
Attention Configuration (news volume):	
Log(N Combined Articles)	Natural log of total number of news articles in Tetlock window.
has_company_news	Binary indicator: 1 if any company-specific article exists in the Tetlock window.

Note: FF abnormal returns use rolling 252-day regression on FF3 factors. Missing controls are mean-imputed with binary indicators.

A.2 Data Preprocessing Pipeline

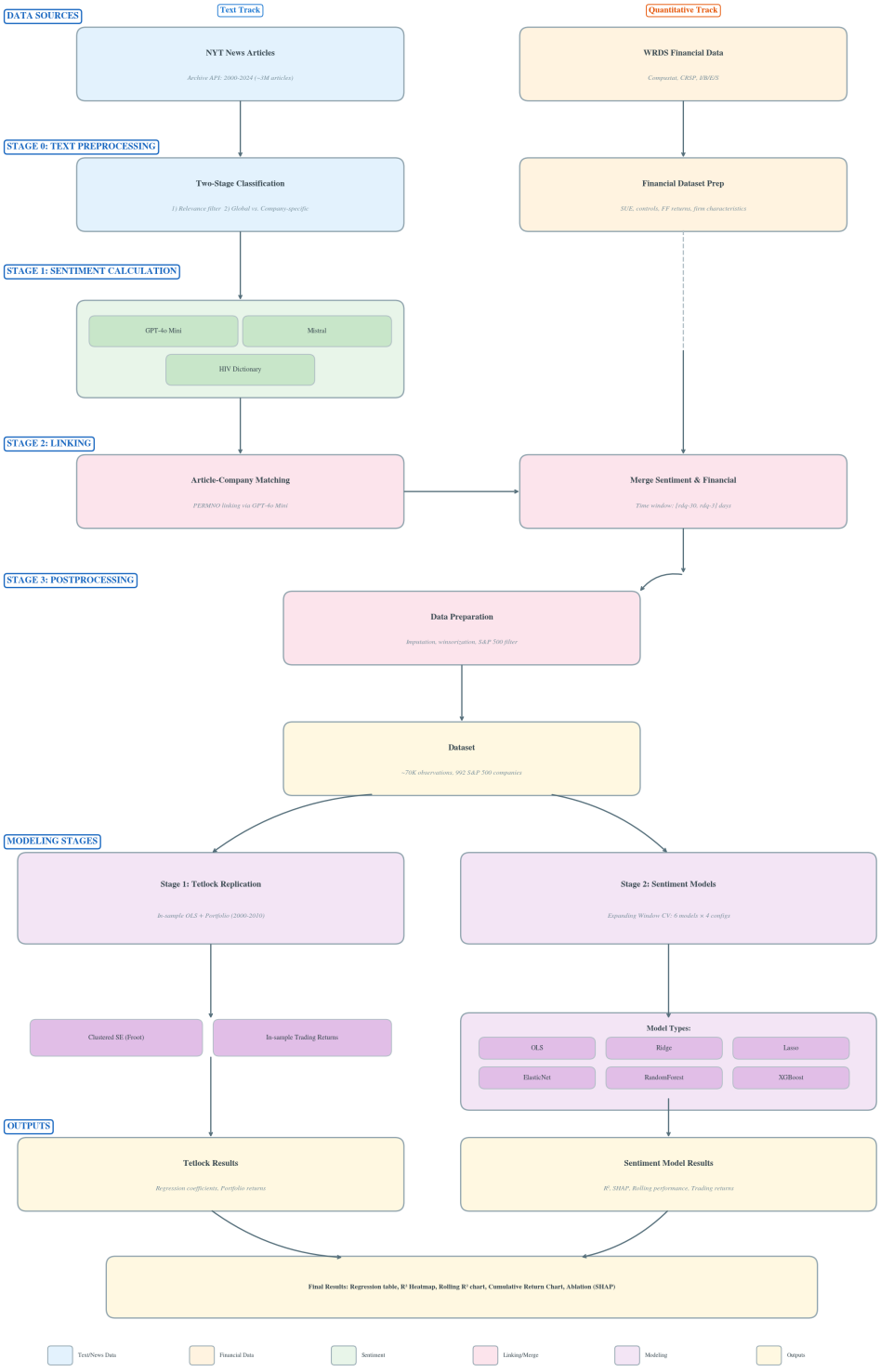

Figure 4: Complete Data Preprocessing and Model Training Pipeline

A.3 Article Classification Pipeline

A.3.1 LLM Classification Prompt

```
You are a financial news classifier. Analyze the given news article and
classify it.

Your task:
1. First determine if this article is RELEVANT (about business, economics,
finance, markets, companies) or IRRELEVANT (sports, entertainment, politics
without economic focus, lifestyle, etc.)
2. If RELEVANT, determine if it's COMPANY-SPECIFIC (focuses on a specific
company or companies) or GLOBAL (general market/economic news)

Respond ONLY with a JSON object in this exact format:
{"is_relevant": 0 or 1,
"is_company_specific": 0 or 1,
"confidence": 0.0-1.0,
"reasoning": "brief explanation"}

Rules:
- is_relevant=1 for business/economic/financial news, 0 otherwise
- is_company_specific=1 only if the article focuses on
specific company(ies), 0 for general market/economic news
- If is_relevant=0, then is_company_specific must be 0
- confidence should reflect your certainty (0.5-1.0 range)
```

A.3.2 Validation Results

Table 7: Classification Model Validation Results (N=200)

Model	Stage 1 Acc.	Stage 2 Acc.[a]	Category
LLM-based classifiers:			
GPT-4o-mini Multi-Perspective	86.5%	**88.9%**	LLM
GPT-4o-mini Single-Perspective	85.0%	87.0%	LLM
Mistral Small 3.2 (24B)	83.5%	85.2%	LLM
Embedding-based:			
OpenAI-emb	**85.0%**	81.5%	Embedding
OpenAI + BM25	82.5%	79.6%	Hybrid
BGE + BM25	80.5%	85.2%	Hybrid
FinBERT + BM25	51.0%	79.6%	Hybrid
Lexical baseline:			
BM25-only	31.5%	40.7%	Lexical

[a] Stage 2 accuracy computed on 54 truly relevant articles only.

Table 8: McNemar's Test: Pairwise Classification Comparisons

Model 1	Model 2	p (Stage 1)	p (Stage 2)	Sig.
Top-tier pairwise (no significant differences):				
GPT-4o Multi-Persp.	GPT-4o-mini	0.453	1.000	
GPT-4o-mini	Mistral	0.549	1.000	
GPT-4o-mini	OpenAI-emb	1.000	0.607	
GPT-4o-mini	Gemma	0.774	0.500	
Mistral	OpenAI-emb	0.868	0.803	
Top-tier vs. baselines (significant differences):				
GPT-4o-mini	FinBERT	<0.001	0.481	***
GPT-4o-mini	BM25-only	<0.001	<0.001	***
Mistral	FinBERT	<0.001	0.648	***
Mistral	BM25-only	<0.001	<0.001	***
OpenAI-emb	FinBERT	<0.001	0.017	**[b]
OpenAI-emb	BM25-only	<0.001	<0.001	***

Note: McNemar's test (McNemar, 1947) on 200 labeled articles (Stage 1) and 54 relevant articles (Stage 2). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Exact binomial p-values used when $n_{\text{discordant}} < 25$.
[b] Stage 2 only.

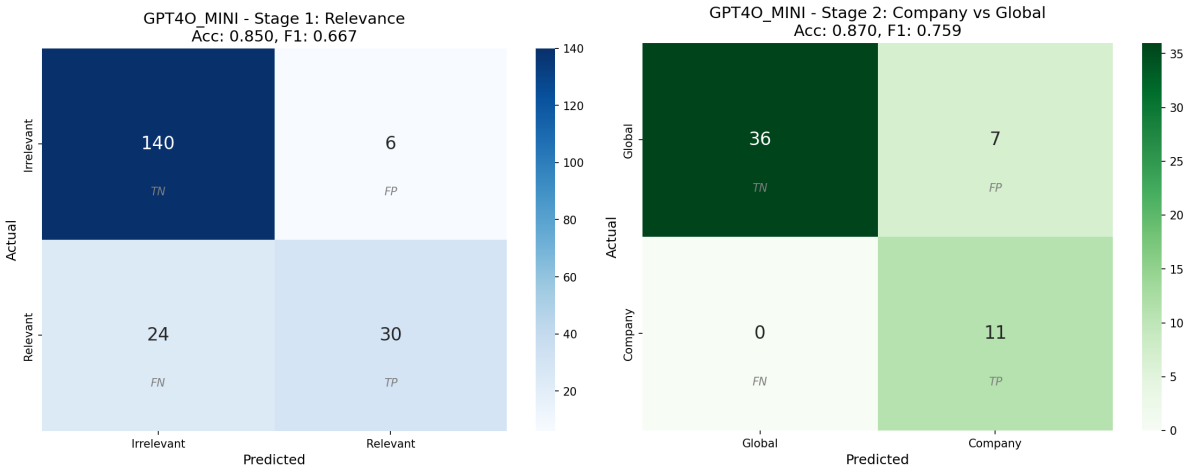

Figure 5: GPT-4o-mini confusion matrices. Left: Stage 1 (85.0% accuracy). Right: Stage 2 (87.0% accuracy on 54 relevant articles).

A.4 Sentiment Extraction and Aggregation

A.4.1 LLM Sentiment Extraction Prompt

The following system prompt is used for Mistral sentiment extraction. The prompt was designed to elicit earnings-focused sentiment assessment while explicitly preventing hindsight bias:

```
You are a SENIOR FINANCIAL ANALYST at a major investment bank.
```

```
PURPOSE: You are helping to build a dataset for academic
research that investigates the predictive power of news
articles on Standardized Unexpected Earnings (SUE) for
S&P 500 companies.  Your accurate sentiment classification
is critical for the validity of this research.

YOUR EXPERTISE: You specialize in analyzing news for investment decisions,
identifying market-moving information, and predicting how news affects
corporate earnings.

TASK: For each article, assign a sentiment score from 1-5 based on its
likely impact on corporate earnings.

SCORING GUIDE:
1 = Very Negative:  News strongly suggests earnings will MISS expectations
- Fraud, major lawsuits, product recalls, executive scandals
- Severe revenue decline, bankruptcy risk, major customer loss
2 = Negative:  News suggests earnings may DISAPPOINT
- Layoffs, missed targets, increased competition
- Regulatory issues, supply chain problems, cost overruns
3 = Neutral:  News has UNCLEAR or BALANCED impact on earnings
- Routine announcements, mixed signals
- Industry-wide trends with unclear company impact
4 = Positive:  News suggests earnings may EXCEED expectations
- New contracts, expansion plans, cost cutting success
- Market share gains, positive analyst coverage
5 = Very Positive:  News strongly suggests earnings BEAT
- Breakthrough products, major acquisitions
- Exceptional demand, strong guidance upgrade

IMPORTANT:
- Focus on EARNINGS IMPACT, not general market sentiment
- For company-specific news:  assess direct earnings impact
- For global/macro news:  assess how it affects corporate earnings broadly
- Be consistent in your ratings across similar news types

CRITICAL - AVOID HINDSIGHT BIAS:
- Evaluate each article ONLY based on information contained in the text
itself
- Ignore any knowledge you have about what actually
happened to the companies after publication
- Imagine you are reading this article on its publication date with NO
knowledge of future events
- Your rating should reflect what a well-informed analyst would predict AT
THAT TIME

Return ONLY a JSON array with objects containing:
- "id":  article index (0-indexed)
- "sentiment":  score from 1-5
- "confidence":  your confidence in the rating (0.5-1.0)
```

A.4.2 Firm-Quarter Aggregation

$$s_{it}^{combined} = \frac{N_{it}^{global} s_{it}^{global} + N_{it}^{company} s_{it}^{company}}{N_{it}^{global} + N_{it}^{company}} \tag{4}$$

where N_{it}^{type} is the number of articles of each type, and s_{it}^{global} and $s_{it}^{company}$ are each computed

as equal-weighted averages of article-level scores within the Tetlock window:

$$s_{it}^{type} = \frac{1}{|\mathscr{A}_{it}^{type}|} \sum_{j \in \mathscr{A}_{it}^{type}} s_j \tag{5}$$

A.5 Data Summary Statistics

Table 9: Descriptive Statistics for All Variables (N = 32,819)

Variable	Mean	Std	Min	P25	Median	P75	Max
Target Variable:							
SUE	−0.006	1.015	−2.340	−0.662	0.031	0.697	2.149
Earnings Momentum:							
lagged_SUE	−0.011	1.004	−2.336	−0.647	0.000	0.675	2.147
Sentiment Measures:							
combined_neg_fraction	0.121	0.011	0.090	0.113	0.121	0.129	0.151
combined_sentiment_mistral	2.894	0.086	2.629	2.833	2.895	2.940	3.161
delta_combined_neg_fraction	0.001	0.010	−0.039	−0.004	0.001	0.006	0.037
delta_combined_sentiment_mistral	−0.004	0.062	−0.227	−0.042	−0.007	0.034	0.216
Attention:							
log_n_combined_articles	7.282	0.228	3.401	7.146	7.212	7.401	7.915
has_company_news	0.094	0.292	0.000	0.000	0.000	0.000	1.000
Control Variables:							
log_market_cap	9.109	0.858	0.237	9.311	9.311	9.311	13.131
log_book_to_market	−1.041	0.484	−5.884	−1.095	−1.095	−1.095	4.787
log_share_turnover	7.502	0.651	1.470	7.239	7.557	7.647	11.689
FFAlpha_longterm	0.053	0.230	−2.478	0.006	0.032	0.098	3.324
FFCAR_preannouncement	0.006	0.080	−0.931	−0.009	0.003	0.020	3.395
FFCAR_announcement	0.000	0.028	−0.935	−0.004	0.000	0.005	0.675
forecast_dispersion	0.451	1.269	0.000	0.000	0.330	0.646	184.027
forecast_revision	−0.001	0.042	−3.368	0.000	0.000	0.000	1.000

Note: Summary statistics for S&P 500 sample (N = 32,819 firm-quarters, 0% missing). Combined sentiment variables aggregate global and company-specific news using article-count-weighted averaging. SUE and lagged_SUE winsorized at 1st and 99th percentiles; other variables unwinsorized.

A.6 Correlation Matrix

Table 10: Correlation Matrix of Key Variables (N = 32,819)

	SUE	HIV	Mst	ΔHIV	ΔMst	Att	LSue	MC	B/M	TO	Pre	FFα	Rev
SUE	1.00	−0.05	0.06	0.00	0.03	−0.03	**0.35**	−0.01	−0.03	0.00	0.03	0.08	0.03
HIV	−0.05	1.00	**−0.63**	**0.42**	**−0.23**	**0.49**	−0.04	**−0.17**	**0.10**	**0.20**	−0.01	0.00	−0.01
Mst	0.06	**−0.63**	1.00	−0.08	**0.27**	**−0.25**	0.06	0.06	−0.07	−0.08	0.01	−0.01	0.02
ΔHIV	0.00	**0.42**	−0.08	1.00	**−0.54**	−0.05	0.02	0.03	−0.04	−0.02	0.00	0.00	0.01
ΔMst	0.03	**−0.23**	**0.27**	**−0.54**	1.00	0.09	0.00	−0.06	0.07	0.06	0.01	0.01	0.01
Att	−0.03	**0.49**	**−0.25**	−0.05	0.09	1.00	−0.06	**−0.25**	**0.20**	**0.23**	0.00	0.01	−0.01
LSue	**0.35**	−0.04	0.06	0.02	0.00	−0.06	1.00	0.01	−0.04	−0.01	0.00	0.07	0.03
MC	−0.01	**−0.17**	0.06	0.03	−0.06	**−0.25**	0.01	1.00	**−0.24**	**−0.18**	−0.03	−0.03	0.00
B/M	−0.03	**0.10**	−0.07	−0.04	0.07	**0.20**	−0.04	**−0.24**	1.00	0.09	0.02	−0.01	−0.03
TO	0.00	**0.20**	−0.08	−0.02	0.06	**0.23**	−0.01	**−0.18**	0.09	1.00	0.03	0.03	−0.03
Pre	0.03	−0.01	0.01	0.00	0.01	0.00	0.00	−0.03	0.02	0.03	1.00	0.03	0.01
FFα	0.08	0.00	−0.01	0.00	0.01	0.01	0.07	−0.03	−0.01	0.03	0.03	1.00	0.05
Rev	0.03	−0.01	0.02	0.01	0.01	−0.01	0.03	0.00	−0.03	−0.03	0.01	0.05	1.00

Note: Correlations computed on initial training sample (N=32,819). Bold indicates $|r| \geq 0.10$. HIV = combined_neg_fraction; Mst = combined_sentiment_mistral; ΔHIV/ΔMst = quarter-over-quarter sentiment change; Att = log(n_combined_articles); LSue = lagged SUE; MC = log market cap; B/M = log book-to-market; TO = log share turnover; Pre = FFCAR pre-announcement; FFα = FF alpha long-term; Rev = forecast revision. FFCAR announcement, forecast dispersion, and has_company_news are omitted (all $|r| < 0.05$ with SUE and other variables). The strongest predictor of SUE is lagged SUE (0.35). Sentiment measures show weak correlations with SUE ($|r| \leq 0.06$). HIV and Mistral are negatively correlated (-0.63) as expected since higher HIV indicates more negative words while higher Mistral indicates more positive sentiment. Sentiment deltas are strongly negatively correlated (ΔHIV–ΔMst: -0.54).

A.7 Model Hyperparameters

A.7.1 XGBoost Hyperparameter Search Space

Table 11: XGBoost Hyperparameter Search Space (SUE Prediction Comparison)

Hyperparameter	Values Tested	Description
`learning_rate`	[0.01, 0.03, 0.05]	Step size shrinkage (learning rate)
`max_depth`	[2, 3]	Maximum tree depth (reduced for generalization)
`n_estimators`	[50, 100, 150]	Number of boosting rounds
`subsample`	[0.6, 0.7, 0.8]	Row sampling ratio per tree
`colsample_bytree`	[0.5, 0.6, 0.7]	Column sampling ratio per tree
`reg_alpha`	[1.0, 10.0, 50.0]	L1 regularization term
`reg_lambda`	[10.0, 50.0, 100.0]	L2 regularization term
`min_child_weight`	[5, 10, 20]	Minimum sum of instance weight in child

Note: RandomizedSearchCV with 20 iterations and 5-fold cross-validation. GPU acceleration via `tree_method='gpu_hist'` on NVIDIA A100. Strong L1/L2 regularization used to prevent overfitting on financial time series. Early stopping implemented with 20-round patience. Best parameters selected based on validation R^2.

A.7.2 Linear Model Hyperparameters

Table 12: Ridge, Lasso, and ElasticNet Hyperparameter Search Space (SUE Prediction Comparison)

Model	Hyperparameter	Values Tested	Description
Ridge	`alpha`	[0.1, 1, 10, 100, 1000]	L2 regularization strength
Lasso	`alpha`	[0.001, 0.01, 0.1, 1, 10]	L1 regularization strength
ElasticNet	`alpha`	[0.001, 0.01, 0.1, 1]	Overall regularization strength
	`l1_ratio`	[0.1, 0.5, 0.9]	L1 vs. L2 mixing parameter

Note: All linear models evaluated on validation set (2016–2021). Ridge uses pure L2 penalty, Lasso uses pure L1 penalty, ElasticNet combines both: penalty $= \alpha \times [(1 - \text{l1\_ratio}) \times \text{L2} + \text{l1\_ratio} \times \text{L1}]$. All models trained with standardized features.

A.7.3 Random Forest Hyperparameters

Table 13: Random Forest Hyperparameter Search Space (SUE Prediction Comparison)

Hyperparameter	Values Tested	Description
`n_estimators`	[50, 100, 200]	Number of trees in forest
`max_depth`	[5, 10, 20, None]	Maximum tree depth (None = unlimited)
`min_samples_split`	2 (default)	Minimum samples to split internal node
`min_samples_leaf`	1 (default)	Minimum samples required at leaf node
`max_features`	`sqrt` (default)	Number of features per split: $\sqrt{p}$

Note: Grid search over *n_estimators* and *max_depth* parameters: $3 \times 4 = 12$ combinations. Default parameters used for other hyperparameters. Random state fixed at 42 for reproducibility. Bootstrap sampling enabled (default).

A.8 SUE Prediction Comparison: Detailed Results

Table 14: Complete Test R² Results by Sentiment Model and Configuration (%)

Source	Config	OLS	Ridge	Lasso	ElasticNet	RF	XGBoost	Ensemble
Baseline:								
Controls	Controls Only	1.24	1.26	1.38	1.74	5.14	5.30	**5.47**
HIV Dictionary + Controls:								
HIV	Sentiment	0.96	0.98	1.20	1.58	4.94	4.90	5.14
HIV	Delta	1.27	1.28	1.39	1.75	5.11	5.18	5.40
Attention (Article Count Only):								
–	Attention	1.29	1.30	1.49	1.85	4.95	5.07	5.18
Mistral + Controls:								
Mistral	Sentiment	1.21	1.23	1.37	1.73	4.96	4.64	5.16
Mistral	Delta	1.24	1.26	1.40	1.76	5.14	5.30	**5.45**

Note: Test R² from expanding window cross-validation. RF = RandomForest. "ctrl" = with 9 Tetlock control variables. Ensemble uses validation-weighted averaging of Random Forest and XGBoost.

A.9 Diebold-Mariano Test Results

Table 15: Diebold-Mariano Test: Sentiment+Controls vs Controls-Only (Ensemble)

Source	Config	R^2 Controls	R^2 Sentiment	DM stat	p-value	Sig.
HIV	Sentiment	5.47%	5.14%	−3.33	**0.001**	***
Mistral	Sentiment	5.47%	5.16%	−0.82	0.414	
HIV	Delta	5.47%	5.40%	−1.36	0.175	
Mistral	Delta	5.47%	5.45%	−0.41	0.682	
–	Attention	5.47%	5.18%	−2.09	**0.037**	**

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. DM test conducted over 48 expanding windows. Negative DM statistic indicates controls-only outperforms. Ensemble uses validation-weighted averaging of Random Forest and XGBoost. No sentiment configuration significantly outperforms controls-only.

Table 16: Diebold-Mariano Tests: Mistral vs. HIV Dictionary (Ensemble)

Config	Model	DM stat	p-value	Winner
Sentiment	Ensemble	0.04	0.970	No sig. diff.
Delta	Ensemble	0.83	0.407	No sig. diff.

Note: Diebold-Mariano (1995) test comparing Mistral sentiment against HIV dictionary sentiment. HAC-robust standard errors. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Results shown for validation-weighted Ensemble. Full sample: 2013Q1–2024Q4 (48 quarters).

A.10 R^2 Breakdown by Sub-Period

Table 17: In-Sample and Out-of-Sample R^2 by Sub-Period (Ensemble, Sentiment + Controls)

		In-Sample	Out-of-Sample			
Source	Config	Full	Full	Pre-2018	2018–2021	Post-2022
Controls	Controls Only	14.55%	**5.47%**	7.20%	−1.12%	**11.39%**
HIV	Sentiment	14.75%	5.14%	6.71%	−1.12%	10.87%
HIV	Delta	14.64%	5.40%	7.10%	−1.20%	11.35%
–	Attention	14.82%	5.18%	7.28%	−1.29%	10.32%
Mistral	Sentiment	14.83%	5.16%	6.82%	−1.54%	11.31%
Mistral	Delta	14.62%	**5.45%**	7.16%	−1.16%	**11.43%**

Note: In-Sample R^2 is the average training R^2 across all expanding windows. Out-of-sample columns show mean test R^2 (%) from validation-weighted Ensemble. Pre-2018: 2013Q1–2017Q4 (20 quarters); 2018–2021: 2018Q1–2021Q4 (16 quarters); Post-2022: 2022Q1–2024Q4 (12 quarters). Bold indicates best-performing configuration per column. All models collapse to negative R^2 in the 2018–2021 regime. The gap between in-sample (∼14.6–14.8%) and out-of-sample (∼5.1–5.5%) R^2 highlights substantial overfitting across all configurations.

A.11 Trading Returns and Risk-Adjusted Performance

Table 18: Fama-French 3-Factor Risk-Adjusted Returns: Extended Event Windows (Ensemble, Long-Short)

Window	Source	Config	α (%/yr)	t-stat	β_{Mkt}	β_{SMB}	β_{HML}	R^2
$[-1,+2]$	Controls	Controls Only	-14.65	-0.61	-0.04	-0.45	-0.30	0.010
$[-1,+2]$	HIV	Sentiment	-9.27	-0.38	-0.05	-0.45	-0.33	0.011
$[-1,+2]$	HIV	Delta	-22.10	-0.88	-0.08	-0.40	-0.31	0.009
$[-1,+2]$	Mistral	Sentiment	-16.81	-0.69	-0.02	-0.40	-0.32	0.009
$[-1,+2]$	Mistral	Delta	-21.72	-0.91	-0.05	-0.46	-0.35	0.011
$[-1,+2]$	–	Attention	-15.72	-0.65	-0.07	-0.46	-0.33	0.011
$[-1,+3]$	Controls	Controls Only	-11.46	-0.60	0.02	-0.43	-0.41	0.014
$[-1,+3]$	HIV	Sentiment	-7.15	-0.36	0.00	-0.45	-0.42	0.015
$[-1,+3]$	HIV	Delta	-21.75	-1.08	-0.02	-0.39	-0.42	0.013
$[-1,+3]$	Mistral	Sentiment	-15.47	-0.81	0.03	-0.38	-0.39	0.013
$[-1,+3]$	Mistral	Delta	-17.21	-0.89	0.02	-0.44	-0.43	0.015
$[-1,+3]$	–	Attention	-12.86	-0.66	-0.00	-0.43	-0.43	0.015

Note: Fama-French 3-factor regression on daily calendar-time long-short portfolio returns. Newey-West standard errors. Ensemble uses validation-weighted averaging of Random Forest and XGBoost. α expressed as annualized percentage ($\alpha_{\text{daily}} \times 252$). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. No alpha is statistically significant at the 10% level. Sample: 2013Q1–2024Q4 (48 quarters, $\sim$2,100–2,400 trading days depending on window). The $[-1,+1]$ window results are reported in Table 5.