Estonian Business School
International Business Administration

Online Information Flow and Short-Term Stock Returns: Linear and Machine-Learning Evidence from the U.S. and European Markets

Bachelor's Thesis

Student: Daniil Daletski
Supervisor: Alar Kein, PhD

Tallinn 2026

Declaration of Authorship

I hereby declare that I have written the present Bachelor's thesis independently. All sources used in the work have been cited. The thesis has not been previously submitted for a degree at this or any other university. References have been indicated for all publications, claims, opinions, and other materials drawn from the work of other authors.

Tallinn, 18 May 2026

Daniil Daletski

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIA	Abnormal Institutional Attention (Bloomberg)
ASVI	Abnormal Search Volume Index (Google Trends)
B/M	Book-to-market ratio
CAPM	Capital Asset Pricing Model
DCF	Discounted Cash Flow
DDM	Dividend Discount Model
DM	Diebold-Mariano (test)
EMH	Efficient Market Hypothesis
EN	Elastic Net
FCFE / FCFF	Free Cash Flow to Equity / to the Firm
FF3 / FF5	Fama-French 3-factor / 5-factor model
FinBERT	Financial BERT (sentiment-classification language model)
GBT	Gradient-Boosted Trees
GICS	Global Industry Classification Standard
LM	Loughran-McDonald (finance dictionary)
ML	Machine Learning
NLP	Natural Language Processing
OLS	Ordinary Least Squares
OOS	Out-of-Sample
PEAD	Post-Earnings-Announcement Drift
RF	Random Forest
RMSE	Root Mean Squared Error
RQ	Research Question
SE	Standard Error
SHAP	Shapley Additive Explanations
SVI	Search Volume Index (Google Trends raw series)
XGBoost	Extreme Gradient Boosting

Introduction

Equity prices are increasingly formed in a market environment where corporate disclosures, online search activity, financial news flows, retail-platform attention, and machine-generated summaries circulate almost immediately. These traces are observable, time-stamped, and measurable at firm-week frequency. The central question for this thesis is therefore not whether online information flows exist, but whether they contain incremental information for short-term stock-return prediction after standard market, firm, liquidity, and volatility controls are included.

The theoretical tension is between the efficient-market benchmark and limited-attention frictions. Classical valuation and asset-pricing theory imply that prices should quickly reflect expected cash flows discounted at a risk-adjusted rate. If this benchmark holds exactly, public search and news variables should not improve weekly return prediction once conventional controls are included. Behavioural and information-friction theories suggest a narrower alternative: investors may notice, interpret, and trade on public signals at different speeds, so online attention and news may be associated with measurable weekly return movements and limited next-week predictability.

Prior evidence is mixed. Abnormal search volume has been linked to attention-driven price pressure in U.S. stocks (Da et al., 2011), while media tone has been associated with return effects in Tetlock (2007), Tetlock et al. (2008), and Garcia (2013). Other studies find weaker or inconsistent predictive power for returns, especially outside the United States and when the dependent variable is returns rather than volatility or trading volume. This leaves three gaps relevant to the present thesis: the need to compare attention jointly, news intensity, and sentiment; to test linear and machine-learning specifications using the same set of predictors; and to examine whether the evidence differs between U.S. and European large-cap stocks.

The thesis aims to examine whether online information flow, proxied by abnormal Google search volume (ASVI), news intensity, and news sentiment, is associated with and helps predict short-term total returns in U.S. and European equity markets, and whether supervised machine-learning models extract more predictive value from the same features than linear panel regressions. Predictive value means improvement in weekly total-return prediction for S&P 500 and STOXX Europe 600 constituents, evaluated at the same-week horizon ($h = 0$) and the one-week-ahead horizon ($h = 1$). The design is explanatory and predictive rather than causal.

The thesis addresses four research questions:

RQ1. Do ASVI, news intensity, and news sentiment carry statistically and economically significant predictive value for weekly total returns at the same-week and one-week-ahead horizons during 2021-2025?

RQ2. Do non-linear predictive models improve short-term return prediction compared with linear panel benchmarks, and does this improvement depend on the forecast horizon?

RQ3. Do the magnitude, horizon, and ranking of information-flow effects differ systematically between the U.S. and European sub-panels, and across GICS sectors?

RQ4. Which information-flow indicators are most important in the non-linear models, and how do these model-based rankings compare with the linear evidence?

The practical motivation is the author's development of NEXUS Terminal (Neural Engine X-factor Unified System), a real-time data-based analytical platform for investment decision support. NEXUS is the implementation environment and intended practical end product motivated by the thesis; it is not the thesis's central academic result. The academic contribution remains the empirical test of whether online information-flow indicators improve short-term stock-return prediction. The results are relevant for investors, portfolio managers, analysts, regulators, retail brokerage platforms, and investor-relations teams because each group needs to interpret attention and news shocks without mistaking noisy online activity for reliable trading signals.

The thesis's contribution is defined in relation to four research gaps. First, it applies a company-name plus 'stock' Google Trends query protocol across both U.S. and European firms, addressing ticker-ambiguity concerns while keeping measurement rules comparable across regions. Second, it studies ASVI, news intensity, and sentiment together rather than treating attention and tone as separate literatures. Third, it compares L0/L1 linear models with ML0/ML1 machine-learning models using the same predictor vectors, separating feature values from model-class values. Fourth, it links the empirical design to a practical decision-support architecture, showing how a platform such as NEXUS can use information-flow indicators as explainable monitoring signals rather than as automatic trading rules.

The thesis is organised as follows. Chapter 1 develops the theoretical framework, starting with discounted cash flow valuation and narrowing to the dividend discount model, information-processing frictions, and information-flow proxies. Chapter 2 presents the data and

methodology: sample construction, variable definitions, the 13 equations, linear and machine-learning specifications, and the evaluation strategy. Chapter 3 presents the results: descriptive statistics, linear coefficients, basis-point effects, out-of-sample model performance, robustness checks, SHAP rankings, hypothesis verdicts, and discussion. The Conclusions answer the research questions, state the contribution, acknowledge limitations, and outline future research and NEXUS-related extensions.

1. Theoretical framework

This chapter develops the conceptual foundation of the thesis in four steps. Section 1.1 begins from the proposition that the intrinsic value of a share equals the discounted stream of expected future cash flows, narrows the proposition to the dividend discount model in two forms, and explains why information-flow variables should be interpreted as observable traces of the market's attempt to process fundamental information rather than as substitutes for fundamentals themselves. Section 1.2 surveys the menu of available information-flow indicators, organising them by the agent whose attention is measured and by the medium in which the trace is left, and explains why three indicators - abnormal Google search volume, news intensity, and news sentiment - are selected for the empirical analysis while four others are deliberately set aside. Section 1.3 explains how limited attention, gradual information diffusion, and attention-induced buying can be associated with measurable weekly return movements and one-week-ahead return predictability. Section 1.4 synthesises the framework into four formal hypotheses, which Chapter 2 then tests.

1.1 From discounted cash flows to the dividend discount model

Modern equity valuation begins with the proposition that the intrinsic value of a share equals the present value of the cash flows accruing to its owner, discounted at a rate that reflects the time value of money and the risk of those cash flows (Damodaran, 2024; Pinto et al., 2024). The proposition is general; it becomes operational only once the analyst chooses a definition of 'cash flow' and a matching discount rate. Three definitions dominate contemporary practice.

The first is the cash flow available to equity holders in the form of dividends. Discounting expected dividends at the cost of equity yields the dividend discount model (DDM), formalised in its general form by Williams (1938) and in its constant-growth form by Gordon (1959). The second is free cash flow. Free cash flow to equity (FCFE) is cash from operations net of reinvestment and net debt issuance, discounted at the cost of equity; free cash flow to the firm (FCFF) is cash from operations net of reinvestment but before financing, discounted at the weighted average cost of capital (Damodaran, 2024). The third is residual income, defined as accounting earnings less a charge for the cost of equity capital, discounted at the cost of equity (Penman, 2023; Pinto et al., 2024). All three definitions are equivalent under clean-surplus accounting and consistent assumptions about reinvestment, financing, and growth; they differ only in where the analyst draws the line between operating, investing, and financing flows.

The thesis adopts the dividend discount model in two forms — the general infinite-sum form and the Gordon constant-growth form — for three reasons. First, identification of the channels through which information flows affects price runs through expected cash flows, the discount rate, and growth, and these three inputs are visible in the DDM with the smallest amount of accounting machinery; FCFE, FCFF, and residual income would introduce additional reinvestment and accrual terms without altering the comparative statics of attention or sentiment shocks. Second, the comparative statics of attention and sentiment shocks on price are most transparent in the closed-form Gordon expression, where a shock to the discount rate or to expected growth maps to a percentage change in price by elementary calculus. Third, the empirical exercise studies short-horizon return variation; the analytical role of valuation theory in this thesis is to anchor the predictor set in standard inputs (expected cash flow, discount rate, growth), not to deliver a cross-sectional price level. In this role, the DDM in its two forms is sufficient, and adopting FCFE/FCFF or residual-income models would add length without increasing inferential power.

1.1.1 The dividend discount model in general form

The intrinsic value of a share equals the present value of expected future dividends:

$$P_0 = \sum_{t=1}^{\infty} \frac{E[D_t]}{(1+r_e)^t} \qquad (1)$$

where P_0 is the share price at $t = 0$, $E(D_t)$ is the expected dividend at time t, and r_e is the cost of equity. The expression treats the share as a perpetual claim on dividend payments, which is consistent with the going-concern assumption and with the legal absence of a final maturity for ordinary equity. The cost of equity r_e aggregates the risk-free rate and a risk premium that compensates the marginal investor for bearing systematic risk; its decomposition is supplied by the capital asset pricing model below.

Expected dividends $E(D_t)$ are themselves the product of expected earnings and the expected payout ratio. Earnings depend on revenue, costs, and accounting choices, all of which are revealed gradually through public disclosures, news reporting, and inferred sentiment. The link from information flow to price, therefore, runs through the conditional expectation operator $E(\cdot)$, not through an additional term in the equation: when investors revise their expectations of future dividends or of the discount rate in response to public signals, P_0 moves accordingly. This is the formal mechanism through which the indicators studied in Section 1.2 enter the price. It is also why the thesis treats search and news variables as proxies for the speed

and completeness of expectation revision rather than for the underlying fundamentals themselves.

1.1.2 The Gordon constant-growth form

Under the assumption that dividends grow at a constant rate g in perpetuity, the infinite sum collapses to the Gordon (1959) expression:

$$P_0 = \frac{D_1}{r_e - g} \tag{2}$$

where $D_1 = D_0 (1 + g)$ is the next-period expected dividend, the closed form has three inputs — the next-period dividend, the cost of equity, and the long-run growth rate — each of which maps directly to a category of information flow. A revision in the next-period dividend D_1 is a cash-flow channel: news about earnings releases, guidance updates, dividend announcements, contract awards, or merger activity typically operates through this term. A revision in the discount rate r_e is a risk channel: changes in the perceived risk of the firm, in the risk-free rate, or in the marginal investor's risk aversion (which the limited-attention literature shows is itself state-dependent) operate through this term. A revision in the growth rate g is a long-run-expectations channel: news about strategic moves, regulatory shifts, or technological breakthroughs that change the trajectory of free cash flow operates through this term. For the Gordon expression to be finite and economically meaningful, the constant growth rate is assumed to be lower than the cost of equity ($g < r_e$).

Differentiating the Gordon expression with respect to each input yields the comparative statics that anchor the thesis's interpretation of attention shocks:

$$\frac{\partial P_0}{\partial D_1} = \frac{1}{r_e - g} > 0, \qquad \frac{\partial P_0}{\partial r_e} = -\frac{D_1}{(r_e - g)^2} < 0, \qquad \frac{\partial P_0}{\partial g} = \frac{D_1}{(r_e - g)^2} > 0 \tag{3}$$

The signs in equation (3) rely on this Gordon-model restriction; without it, the constant-growth valuation formula is not finite, and the comparative statics are not economically meaningful.

1.1.3 The cost of equity, CAPM, and the efficient market hypothesis

The capital asset pricing model (Sharpe, 1964; Lintner, 1965; Mossin, 1966) decomposes the cost of equity into a risk-free rate and a risk premium that scales with the firm's exposure to systematic risk:

$$r_e = r_f + \beta_i \big(E[r_m] - r_f \big) \tag{4}$$

where r_f is the risk-free rate, E(r_m) is the expected return on the market portfolio, and β_i is the slope of firm i's return on the market return. The market return enters the empirical specifications of Chapter 2 as a control, both because it captures the systematic component of weekly return variation and because it absorbs macroeconomic information that would otherwise contaminate the attention coefficients. Subsequent extensions add size and value factors (Fama and French, 1993), momentum (Carhart, 1997), profitability and investment (Fama and French, 2015), and an investment-based q-factor structure (Hou et al., 2015), but the thesis does not estimate full multifactor regressions because the empirical question is about the marginal contribution of information-flow variables conditional on a compact and well-understood set of controls, not about ranking competing asset-pricing models.

The efficient market hypothesis (Fama, 1970, 1991) posits that prices fully reflect all available information at any point in time. In its semi-strong form, public information — earnings, news, search activity — is already incorporated into P_0 by the time it becomes observable; in its weak form, only the price history is incorporated. If the semi-strong EMH holds exactly, the predictive coefficients of any public information-flow indicator on next-week returns must be zero in expectation. Empirical departures from this benchmark motivate the limited-attention literature reviewed below: the EMH is not assumed to fail permanently but is treated as the null against which predictive coefficients are tested. Grossman and Stiglitz (1980) provide a complementary argument that perfectly informative prices are impossible when information acquisition is costly, because no investor would have an incentive to gather information that prices already revealed; in the digital era, the cost of producing and distributing information has fallen, but the cost of attention and interpretation has not disappeared, leaving room for exactly the temporary frictions that this thesis tests.

1.1.4 Linking the DDM inputs to information-flow indicators

The three DDM inputs map to three categories of information flow that are observable at weekly frequency and at the firm level. Cash-flow expectations are revised when news flows report on earnings, guidance, contracts, or mergers; news intensity (count of articles per firm-week) and news sentiment (signed tone) are the two indicators in this category. Discount-rate beliefs are revised when investor attention spikes because limited-attention investors who become marginal on attention days alter the market's effective risk-bearing capacity (Peng and Xiong, 2006; Barber and Odean, 2008); abnormal Google search volume (ASVI) serves as a proxy for this channel. Growth-rate beliefs are revised when the news flow contains strategic or regulatory content; this is captured indirectly by news sentiment and, in robustness checks,

by the GICS sector dummies that absorb sector-wide growth news, such as oil-price shocks in Energy, drug-approval cycles in Health Care, or AI-led capex announcements in Technology.

The empirical specifications, therefore, include three information-flow regressors (ASVI, NewsInt, Sent), three firm-level controls (log market capitalisation, book-to-market, lagged weekly return), two market-context controls (market return, realised volatility), one liquidity control (abnormal turnover), and ten GICS sector fixed effects. The mapping of regressors to DDM channels is reported in Table 4 of Section 2.2, and the rationale is preserved consistently across linear and machine-learning specifications so that the model comparison reflects model class rather than feature set.

An empirical confirmation of this conceptual mapping is already available in the literature. Drake et al. (2012) show that abnormal Google search volume rises before earnings announcements, peaks around the announcement, and remains elevated afterwards — a pattern that is consistent with the cash-flow-information channel and inconsistent with a pure noise interpretation of search. Ben-Rephael et al. (2017) distinguish institutional attention measured by Bloomberg terminal activity from retail attention measured by Google SVI; institutional attention arrives earlier and largely eliminates post-earnings-announcement drift, while retail attention responds with a lag and is more closely associated with the marginal-investor mechanism that operates through the discount-rate channel of equation (3). DellaVigna and Pollet (2009) and Hirshleifer et al. (2009) document distraction effects: low-attention announcements (Friday earnings, days with many simultaneous announcements) generate weaker immediate reactions and stronger subsequent drift, which is direct evidence that the speed of expectation revision is a function of the attention environment in which the news arrives.

1.2 Information-flow indicators: a menu, three choices, and their limitations

Empirical research has produced a menu of indicators that aim to measure investor attention or information demand at the firm-week or finer frequency. The menu can be organised by the agent whose attention or information demand is measured, and by the medium in which the trace is left. Six categories dominate the recent literature, and each contains at least one indicator that has been shown to have predictive power for short-horizon stock returns in at least one published study. The thesis does not need to revisit each indicator in detail; rather, it needs a transparent rule for selecting which indicators to use and an explicit acknowledgement of those it sets aside.

The first category captures aggregate retail search behaviour. Da et al. (2011) introduced ASVI as the log change in Google Trends search volume for a firm's ticker, scaled by its eight-week median. The signal is available globally, free of charge, and on a weekly cadence. It is also subject to two well-documented biases: ticker symbols can collide with non-financial searches, as deHaan, Lawrence and Litjens (2025) demonstrate using a large sample of Russell-3000 tickers, and Google's sampling and re-scaling produce series that are not consistent across query batches (Eichenauer, Indergand, Martínez and Sax, 2022; Algaba et al., 2024). Both biases are addressable through query design: the ticker problem is solved by querying the company name with a finance-specific qualifier rather than the bare ticker, and the consistency problem is partially solved by averaging multiple independent pulls and computing within-firm log deviations rather than levels.

The second category captures institutional information demand. Ben-Rephael et al. (2017) construct an 'abnormal institutional attention' (AIA) index from Bloomberg news reading and search activity and show that it dominates ASVI in predicting earnings-announcement returns. AIA is, however, proprietary, not available to retail researchers without paid Bloomberg access, and not measurable at the cadence required by an academic thesis with reproducibility constraints. The third category captures regulatory-filing access. Drake et al. (2015) and follow-up work use SEC EDGAR server logs to measure who accesses which 10-K, 10-Q, and 8-K filings; deHaan, Lawrence and Litjens (2025) revisit EDGAR-based proxies and document additional measurement-error issues caused by automated bot traffic. The series is, in any case, U.S.-only and not directly comparable to a European panel where the equivalent regulatory infrastructure is fragmented across national authorities.

The fourth category captures news flow. News intensity (article count per firm-period) and news sentiment (signed tone) have been studied since Tetlock (2007) and García (2013); modern dictionary-based scoring follows Loughran and McDonald (2011), and modern transformer-based scoring follows the FinBERT family of models (Huang et al., 2023). News data are available across both regions through commercial APIs (NewsAPI, Finnhub). Still, they are biased toward English-language coverage and large firms — both biases that the thesis must accept and document, since they cannot be eliminated by query design alone. The fifth category captures social-media chatter. Bollen et al. (2011) opened the literature on Twitter mood, and Cookson et al. (2024) document that messages on retail-investor platforms (StockTwits, Reddit) carry distinct, signed predictive content. Lopez-Lira and Tang (2025) study large language model outputs as inputs to forecasts, finding that ChatGPT-derived sentiment scores predict short-term returns. Social-media data are noisy, costly to license at

scale, and difficult to harmonise across regions, especially when European retail platforms are far more fragmented than the dominant U.S. platforms. The sixth category captures encyclopaedia engagement. Pyun (2024) shows that abnormal Wikipedia page views predict short-term returns in a manner that is partially distinct from ASVI; the data are public, but coverage of European firms — particularly mid-sized European industrials with low international visibility — is uneven.

The thesis selects three indicators from this menu — ASVI, news intensity, and news sentiment — for three reasons. First, all three are available at weekly frequency for every firm in the S&P 500 and STOXX 600 panels; institutional AIA and EDGAR access are U.S.-only or proprietary, and Wikipedia coverage is uneven across European listings. Second, each maps cleanly to one of the DDM channels in Section 1.1.4: ASVI to the discount-rate channel via limited-attention frictions, news intensity to the cash-flow information arrival rate, and news sentiment to the signed revision of cash-flow and growth expectations. Third, all three rely on data sources that are public or low-cost (Google Trends, NewsAPI, Finnhub) and on sentiment scoring (FinBERT, Loughran-McDonald) that is reproducible without proprietary access; this preserves replicability and matches the design philosophy of NEXUS Terminal, which is explicitly built around publicly accessible data sources and open-source language models.

Each chosen indicator is exposed to specific biases, and the empirical design mitigates them as follows. Ticker-symbol contamination in Google search (deHaan et al., 2025) is mitigated by querying the company name with the qualifier 'stock' rather than the bare ticker, following Algaba et al. (2024); Appendix 1 reports the list of qualifiers and the exact query string for each firm. Google Trends sampling and re-scaling inconsistency (Eichenauer et al., 2022; Algaba et al., 2024) is mitigated by computing ASVI as a within-firm log-deviation from an eight-week median, which absorbs level shifts, and by averaging three independent Google Trends pulls per firm-week, which smooths the sampling-induced jitter that Google's exposed daily series produces under repeated queries. News-flow coverage bias toward English-language sources and large firms is mitigated by restricting the panel to S&P 500 and STOXX 600 constituents — both already large and widely covered — and by including news intensity (count) alongside sentiment, so that low-coverage weeks are treated as low-intensity observations rather than missing data. Sentiment classifier drift between the FinBERT 2023 release and the 2025 test window is mitigated by ensembling FinBERT with the Loughran-McDonald (2011) finance dictionary, which is dictionary-based and therefore drift-free, and by testing robustness to either component alone in unreported checks.

Table 1. Menu of available information-flow indicators
Source: composed by the author from the cited literature.

Indicator	Agent measured	Source	Region availability	Frequency	Selected
ASVI (Google Trends)	Retail	Google Trends	Global	Weekly	Yes
AIA (Bloomberg)	Institutional	Bloomberg	Global, propr.	Daily	No
EDGAR clicks	Mixed	SEC server logs	U.S. only	Daily	No
News intensity	News producers	NewsAPI / Finnhub	Global	Daily/weekly	Yes
News sentiment	News producers	NewsAPI + FinBERT/LM	Global	Daily/weekly	Yes
Social-media chatter	Retail	StockTwits, Reddit, X	Global, costly	Real-time	No
Wikipedia page-views	Retail	Wikimedia API	Global, uneven	Daily	No
LLM outputs	Algorithmic	OpenAI / Anthropic	Global, costly	On-demand	No

Table 1 concisely captures the design space. The four indicators not selected are excluded for measurable reasons — proprietary licensing for AIA, regional unavailability for EDGAR, signal quality and licensing concerns for social media data, and uneven European coverage for Wikipedia — rather than because their predictive content has been disproven. A natural extension of the present design, discussed in the Conclusions, would re-estimate every specification with one or more of these additional indicators added to the predictor vector, and would test whether the dominance of ASVI documented in Chapter 3 survives that enrichment.

1.2.1 Abnormal Google search volume (ASVI)

ASVI is constructed as in Da et al. (2011), with the ticker-vs-name modification that follows the methodological recommendations of the recent measurement-error literature:

$$ASVI_{i,t} = log\left(SVI_{i,t}\right) - log\left(median\{SVI_{i,t-1}, \dots, SVI_{i,t-8}\}\right) \quad (5)$$

where $SVI_{i,t}$ is the Google Trends search volume index for firm i in week t, queried with the company-name + 'stock' qualifier listed in Appendix 1. The eight-week median absorbs slow-moving level shifts associated with secular changes in firm visibility — a name that becomes more prominent after a major corporate event will see its baseline search volume rise gradually. Still, ASVI captures only the deviation around that gradually shifting baseline, not the level. The within-firm demeaning absorbs cross-sectional differences in search volume between large and small constituents, which is essential because raw SVI levels for a name like

Apple are orders of magnitude larger than those for a smaller European industrial company. The signal proxies for the discount-rate channel of Section 1.1.4 via limited-attention frictions: when retail attention spikes, more limited-attention investors become marginal in the order flow, reducing the market's effective risk-bearing capacity and generating short-lived price pressure (Peng and Xiong, 2006; Barber and Odean, 2008).

1.2.2 News intensity and sentiment

News intensity is the count of articles mentioning firm i during week t, sourced jointly from NewsAPI and Finnhub, and de-duplicated by URL hash to avoid inflation from syndicated wire copy. News sentiment is the article-level signed score averaged across firm i in week t:

$$Sent_{i,t} = \frac{1}{N_{i,t}} \sum_{k=1}^{N_{i,t}} s_{i,k,t}\,, \qquad s_{i,k,t} \in [-1, +1] \tag{6}$$

where $s_{i,k,t} \in [-1, +1]$ is the FinBERT score of article k ensembled with the Loughran-McDonald polarity score, and $N_{i,t}$ is the number of de-duplicated articles in firm-week (i, t). News intensity proxies for the cash-flow information arrival rate; sentiment proxies for the signed revisions to cash-flow and growth expectations. The two variables are conceptually distinct and empirically only weakly correlated in the present panel: a firm may receive many neutral articles in an earnings week, a few sharply negative articles in a litigation week, or a moderate number of positive articles in a product-launch week, and these three cases generate different combinations of intensity and sentiment that should not be collapsed into a single 'news activity' variable. Including both indicators allows the thesis to test whether what is said matters more than how much is said, a question that the SHAP-based feature attribution of Section 3.5 answers in the affirmative.

The decision to ensemble FinBERT with the Loughran-McDonald dictionary deserves a brief justification. FinBERT is a transformer-based model fine-tuned on financial text and is the strongest single sentiment classifier available for financial applications. Still, it is trained on a dataset with a fixed cutoff and may drift when processing text that uses post-cutoff vocabulary or event framing. The Loughran-McDonald dictionary is a hand-curated list of words classified as positive, negative, uncertain, litigious, constraining, or modal in a financial context; it is dictionary-based, transparent, and immune to drift, but it cannot capture context the way a transformer can. Ensembling the two — by averaging the FinBERT score and the Loughran-McDonald polarity score with equal weights — produces a sentiment series that is

both context-aware and drift-resistant, and that has been shown in the broader textual-analysis literature to outperform either component alone.

1.3 Limited attention, gradual diffusion, and category-level frictions

The thesis does not assume permanent inefficiency. It tests whether limited investor attention, gradual information diffusion, and category-level biases generate predictable weekly excess returns at one- and two-week horizons, conditional on standard controls. Three frictions are isolated, each with independent theoretical and empirical support.

The first is limited attention, as in Peng and Xiong (2006). Investors with finite cognitive capacity allocate attention disproportionately to market-wide and sector-wide news at the expense of firm-specific news, producing category-level co-movement and short-horizon under-reaction at the firm level. The implication for the empirical design is that firm-specific search and news shocks should carry predictive content over and above market and sector controls, and that the strength of this content should rise with the attention-attractiveness of the firm — a prediction that the sector-level decomposition in Section 3.5 confirms, with the largest same-week ASVI effects appearing in the visually salient, retail-oriented sectors (Energy, Technology, Industrials) rather than in the regulated, less retail-attention-intensive sectors (Health Care).

The second is gradual information diffusion, as in Hong and Stein (1999). Heterogeneous investors update beliefs at different speeds, so private information becomes incorporated into prices over multiple periods rather than instantaneously, producing short-horizon momentum that reverses once the slow updaters have completed their trades. The implication for the empirical design is that an attention shock today may continue to move prices next week, and the sign of that one-week-ahead effect distinguishes between two competing accounts: a positive one-week-ahead coefficient is consistent with gradual diffusion. In contrast, a negative one-week-ahead coefficient would be consistent with the reversal of an initial overreaction. The empirical results in Chapter 3 — positive and larger at $h = 1$ than at $h = 0$ for ASVI but reversed in sign for sentiment — are precisely the pattern that the gradual-diffusion-plus-overreaction-reversal mixture predicts.

The third is the attention-induced buying bias of Barber and Odean (2008): individual investors are net buyers of attention-grabbing stocks, generating temporary upward price pressure that decays once attention fades. This mechanism predicts a positive same-week ASVI coefficient and a negative one-week-ahead coefficient — the classic 'attention-and-reversal' pattern. The empirical results in Chapter 3 do not support a strict Barber-Odean reversal at the

one-week horizon; instead, the one-week-ahead ASVI coefficient is larger and more significant than the same-week coefficient, which points to the Hong-Stein gradual-diffusion mechanism as the dominant channel for ASVI in the 2025 sample. The Barber-Odean mechanism remains visible in the sentiment results: the same-week sentiment coefficient is positive and significant, while the one-week-ahead sentiment coefficient is small and statistically indistinguishable from zero, which is the attenuation pattern that an overreaction-and-reversal mechanism would generate.

The combined frictions deliver four testable predictions, formalised as the hypotheses of Section 1.4: H1 (positive same-week ASVI effect on returns, with a one-week-ahead effect that distinguishes between the gradual-diffusion and reversal accounts); H2 (positive same-week sentiment effect, attenuated at h = 1, consistent with the Tetlock (2007) overreaction-and-reversal pattern); H3 (a non-zero out-of-sample R^2 gap between machine-learning and linear specifications, concentrated at h = 0 because nonlinear interactions among attention, volatility, and turnover are most informative within the same week); and H4 (heterogeneity in the size and sign of effects across the U.S. and European sub-panels, driven by differences in news-coverage depth, retail-investor share, and language fragmentation that the descriptive statistics in Section 3.1 quantify).

1.4 Synthesis and hypotheses

The theoretical framework yields four hypotheses, which are formally restated below and tested in Section 2.1.

H1. Conditional on controls, $ASVI_{i,t}$ has a positive coefficient on $r_{i,t}$ at h = 0 and on $r_{i,t+1}$ at h = 1.

H2. Conditional on controls, $Sent_{i,t}$ has a positive coefficient on $r_{i,t}$ at h = 0, with attenuated magnitude at h = 1.

H3. Out-of-sample R^2 of non-linear machine-learning models exceeds that of linear panel models at h = 0, with the gap narrowing or reversing at h = 1.

H4. The magnitudes of H1-H3 differ between the U.S. and European sub-panels, with U.S. effects expected to be larger today due to deeper news coverage, higher retail-investor participation, and a single, dominant, single-language search-engine ecosystem.

The hypotheses translate the theoretical mechanisms into testable claims about the signs of coefficients, horizon decay, model-class superiority, and cross-regional heterogeneity. Each hypothesis is tied to a specific empirical operation in Chapter 2: H1 and H2 are tested through the ASVI and Sent coefficients of the linear panel regressions estimated separately at h = 0 and

h = 1; H3 is tested through the R^2_{oos} gap between the linear and XGBoost specifications and the Diebold-Mariano test that goes with it; and H4 is tested by repeating the entire estimation on the U.S. and European sub-panels separately and comparing coefficients, R^2_{oos} values, and SHAP feature rankings across regions. Chapter 2 specifies the data, models, and evaluation strategy that put each hypothesis to test.

2. Materials and methods

The empirical strategy is built around three design choices that follow from the hypotheses of Section 1.4. First, the U.S. and European sub-panels are estimated under identical specifications on a pre-specified 50-candidate-per-region universe; the retained sub-panels differ by only one firm after the ING. As a data-availability exclusion, the regional comparison required by H4 is not driven by methodological asymmetry. Second, linear and machine-learning models are estimated on the same predictor vector, so that any difference in predictive accuracy required by H3 is attributable to the model class rather than to the feature set. Third, evaluation is conducted on a strictly chronological 2025 hold-out window, with all hyperparameters tuned on a 2024 validation slice that contains no observations from the test window. Section 2.1 sets out the seven-step procedure that operationalises these choices; Section 2.2 documents the sample, the data sources, and the variable construction; Section 2.3 specifies the linear and machine-learning models in their L0/L1/ML0/ML1 nesting; and Section 2.4 presents the evaluation criteria — out-of-sample R^2, Diebold-Mariano testing, basis-point translation of coefficients, and SHAP attribution — together with the four limitations that bound the scope of the inferences.

2.1 Research design and procedure

The empirical work proceeds in seven steps, moving from sample construction to hypothesis verdicts on a single, coherent panel. Step one defines the cross-section from a pre-specified NEXUS candidate universe of 100 large-cap firms: 50 U.S. candidates from the S&P 500 and 50 European candidates from the STOXX Europe 600, stratified by GICS Level-1 sector to mirror the parent-index sector weights at end-2020 and screened for continuous listing across 2021W01-2025W52. The sample-retention record confirms that 50 of 50 U.S. candidates and 49 of 50 European candidates passed all screens; ING Groep NV (ING.AS) was dropped because Yahoo Finance returned no usable market-data series for the ticker. The retained analytical panel, therefore, contains 99 firms: 50 U.S. firms and 49 European firms. Because the curated candidate universe already matches the target design size, no random draw is performed; all firms that pass the predefined screens are retained. Stratification by GICS Level 1 sector ensures that sector-composition differences between the parent indices do not drive the regional comparison.

Step two constructs the weekly observation grid. Each firm-week observation aggregates daily prices into Friday-close weekly total returns, with rolling-week robustness

reported across Monday-to-Friday anchors in Section 3.4. The Friday-close convention is standard in panel finance and aligns the dependent variable with the natural reporting cadence of NEXUS Terminal. Step three computes the three information-flow regressors (ASVI, NewsInt, Sent) using equations (5) and (6), and the seven controls (log market capitalisation, book-to-market, lagged weekly return, market return, realised volatility, abnormal turnover, GICS sector dummy). Step four splits the panel into a training window (2021-2023, 156 weeks), a validation window (2024, 52 weeks), and a test window (2025, 52 weeks); all hyperparameters and feature pre-processors are fitted on training, tuned on validation, and frozen before any test-window observation is touched. The chronological split is fixed ex ante in the empirical workflow so that no test-window data can leak into hyper-parameter selection.

Step five estimates the linear benchmarks: pooled OLS with double-clustered standard errors on firm and week (Petersen, 2009; Thompson, 2011). Two specifications are estimated at each horizon: L0 includes only the seven controls, while L1 adds the three information-flow regressors. Step six estimates the machine-learning specifications: XGBoost (Chen and Guestrin, 2016) as the primary model, Random Forest and Elastic Net as robustness alternatives, all on the same predictor vector as the corresponding linear specification. ML0 mirrors L0 (controls only), and ML1 mirrors L1 (controls plus information flow). Step seven evaluates: out-of-sample R^2 against a naïve mean-return benchmark; Diebold-Mariano (1995) tests of equal forecast accuracy between nested specifications; basis-point translation of coefficients to make the linear estimates economically interpretable; and SHAP-based feature attribution to identify which features the machine-learning model uses most heavily. Each of the four hypotheses of Section 1.4 maps to specific steps: H1 and H2 are tested in step five through the ASVI and Sent coefficients at $h = 0$ and $h = 1$; H3 is tested in step seven through the R^2_{oos} gap and the Diebold-Mariano statistic; H4 is tested by repeating steps five through seven on the U.S. and European sub-panels separately.

Table 2. Step-by-step research procedure
Source: composed by the author.

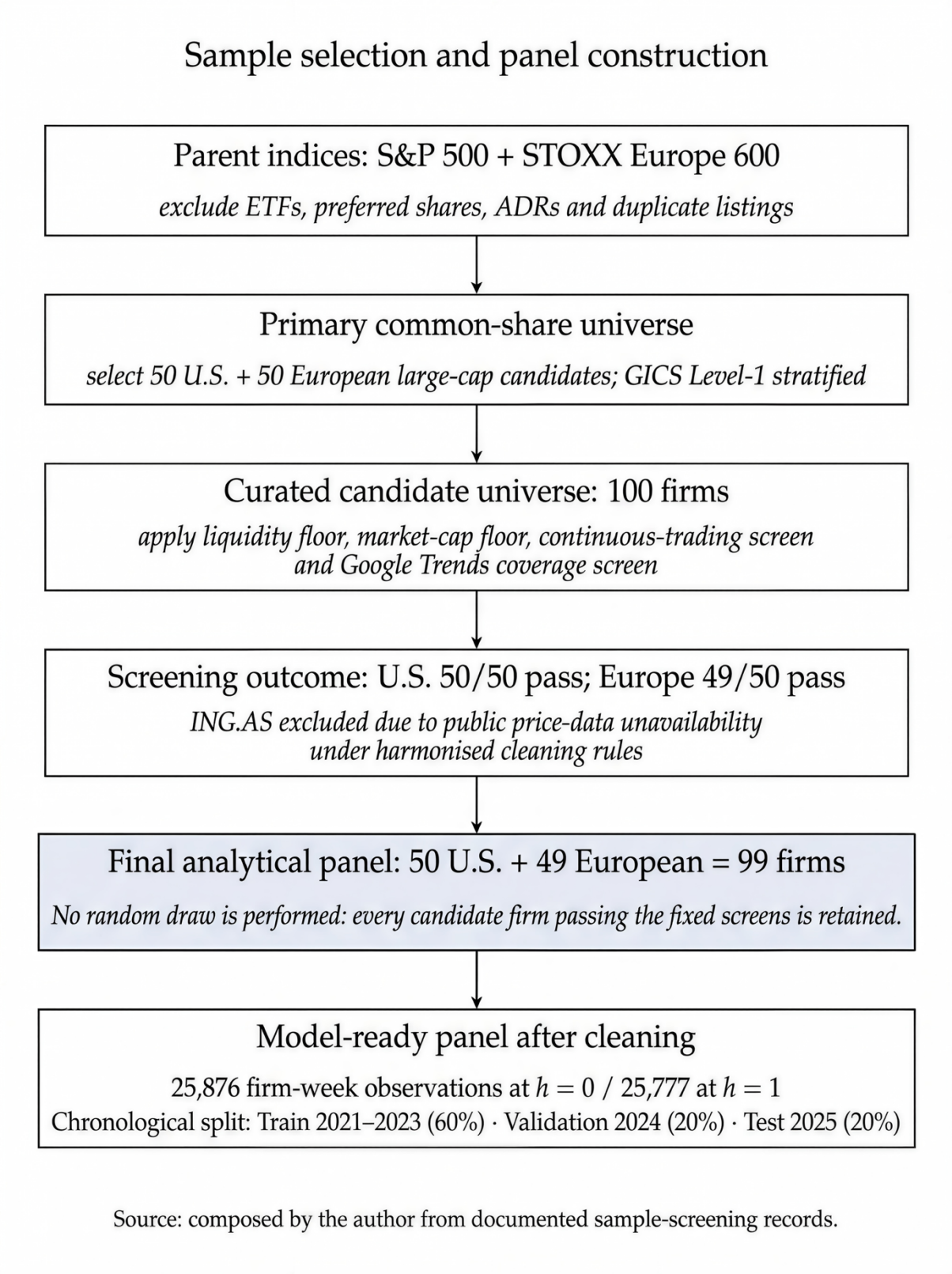
Step	Action	Output
1	Define cross-section	99 retained firms (50 U.S., 49 European) from 100 curated candidates; GICS-stratified
2	Build weekly grid	Friday-close weekly total returns, rolling-week anchors
3	Construct regressors	ASVI, NewsInt, Sent + 7 controls (eq. 5, 6, 7, 8)
4	Split sample	Train 2021-2023; Val 2024; Test 2025 (chronological)
5	Estimate linear models	Pooled OLS, firm-and-week double-clustered SE
6	Estimate ML models	XGBoost (primary); RF, EN (robustness)
7	Evaluate	R^2_{oos}, DM test, bps effects, SHAP rankings

The seven-step structure achieves two things. First, it prevents specification searching: the order in which models are estimated and evaluated is fixed before any results are read, so the temptation to add or drop variables once point estimates are visible is institutionally constrained. Second, it makes the linear-versus-ML comparison fair: ML1 sees the same predictor vector as L1, so any difference in performance must come from the model class — the ability of tree-based methods to capture nonlinear interactions — rather than from feature engineering.

2.2 Sample, data sources, and variable construction

The retained analytical panel contains 99 firms over the 2021W01-2025W52 window: 50 U.S. firms and 49 European firms. The design began with a balanced universe of 100 firms. Still, one European firm, ING Groep NV (ING.AS), was excluded from the model-ready panel because the public market-data source used for the harmonised NEXUS workflow did not provide a usable price-return series under the same cleaning rules. The exclusion is data-availability-based rather than discretionary. Alternative institutional databases could be used to restore a 50/50 regional balance, but retaining a uniform public-data construction preserves comparability across all reported models and is unlikely to alter the central conclusions.

Figure 1. Sample selection flow from parent indices to the final 99-firm analytical panel. Source: composed by the author from the documented sample screening records.



Source: composed by the author from documented sample-screening records.

Stock prices, market capitalisation, book value of equity, and trading volume are sourced from Yahoo Finance and Finnhub at a daily frequency. Polygon.io is used as a backup

price source where Yahoo and Finnhub disagree by more than 0.05% on a given day, which occurs in fewer than 0.3% of firm-day observations. Weekly total return is the geometric compound of daily total returns within the calendar week ending on Friday. The dependent variable for horizon h ∈ {0, 1} is

$$r_{i,t+h} = log\left(P_{i,t+h}^{Fri}\right) - log\left(P_{i,t+h-1}^{Fri}\right) \quad (7)$$

where $P_{i,t}^{Eri}$ is the Friday-close split- and dividend-adjusted price of firm i in week t. Realised volatility is the standard deviation of daily log returns within the week, annualised. Abnormal turnover is the within-firm log deviation of weekly turnover from its 52-week median:

$$AbnTurn_{i,t} = log\left(Turn_{i,t}\right) - log\left(median\{Turn_{i,t-1}, \dots, Turn_{i,t-52}\}\right) \quad (8)$$

where $Turn_{i,t}$ is weekly trading volume scaled by shares outstanding. The 52-week median absorbs slow-moving level shifts in liquidity and turns the variable into a within-firm shock measure that is comparable across firms of very different sizes. Both AbnTurn and ASVI are constructed as within-firm log-deviations rather than cross-sectional standardisations because the cross-sectional distribution of raw search and turnover is heavy-tailed, and the within-firm transformation produces a series with much better statistical behaviour.

ASVI is constructed as in equation (5) from Google Trends queries pulled three times per firm-week and averaged, using the company name + 'stock' qualifier listed in Appendix 1. Three independent pulls per firm-week mitigate the sampling-induced jitter that Google's exposed daily series produces under repeated queries (Eichenauer et al., 2022; Algaba et al., 2024). News articles are sourced from NewsAPI and Finnhub, de-duplicated by URL hash, and filtered to firm-relevant content via fuzzy match on company name and ticker. The de-duplication is essential because syndicated wire copy from Reuters or Bloomberg can be republished by hundreds of outlets within hours, and a naive count would inflate news intensity for firms covered by major wires while leaving local-language European stories underrepresented. News intensity is the de-duplicated article count per firm-week. News sentiment is the firm-week mean of the FinBERT-Loughran-McDonald ensemble score per article, as in equation (6). Market return is the weekly total return of the S&P 500 (for U.S. firms) and the STOXX 600 (for European firms), computed in the same way as firm-level returns. GICS sector dummies follow MSCI/S&P GICS Level-1 classification (eleven sectors, with Communication Services as the reference).

Table 3. Variable formulas and sources

Source: composed by the author. All variables are aligned to the Friday-close weekly grid.

Variable	Symbol	Formula	Source
Weekly total return	$r_{i,t}$	$\log(P_t^{Eri}) - \log(P_{t-1}^{Eri})$	Yahoo, Finnhub
ASVI	$ASVI_{i,t}$	$\log SVI_t - \log \text{med}\{SVI_{t-1..t-8}\}$	Google Trends
News intensity	$NewsInt_{i,t}$	Article count, URL-deduplicated	NewsAPI, Finnhub
News sentiment	$Sent_{i,t}$	Mean FinBERT ⊕ LM score	NewsAPI + FinBERT
Abnormal turnover	$AbnTurn_{i,t}$	$\log(Turn_t) - \log \text{med}_{52\text{H}}(Turn)$	Yahoo
Realised volatility	$RV_{i,t}$	SD of daily log returns within week	Yahoo
Market return	$r_{m,t}$	Index total return (regional)	S&P, STOXX
Log market cap	$\log MC_{i,t}$	log(Price × Shares outstanding)	Yahoo
Book-to-market	$B/M_{i,t}$	Book equity / market equity	Finnhub

Table 4. Mapping of regressors to DDM channels

Source: composed by the author following the framework of Section 1.1.4.

Regressor	DDM channel	Mechanism
ASVI	Discount rate r_e	Limited attention; marginal-investor composition
News intensity	Cash flow D_1	Information-arrival rate
News sentiment	Cash flow D_1; growth g	Signed expectation revision
log MC, B/M	Risk premium	Size and value loadings
Lagged r	Momentum / reversal	Hong-Stein diffusion; short-run reversal
r_m, RV	Systematic risk	CAPM beta and time-varying volatility
AbnTurn	Liquidity	Order-flow pressure
GICS sector dummies	Sector-level news cycles	Energy shocks, drug approvals, AI capex

Table 4 makes the role of each regressor explicit. Information-flow variables (the first three rows) are the variables of interest; the next four rows are control variables that absorb classical determinants of weekly returns; the GICS sector dummies absorb sector-level news cycles that would otherwise bias the firm-level coefficients. The structure ensures that any predictive content attributed to ASVI, NewsInt, or Sent in Chapter 3 has already been net of market return, volatility, turnover, size, value, momentum, and sector. This is a demanding test, because many of the events that drive online attention spikes — earnings announcements, M&A, regulatory actions — also drive market return, volatility, and turnover, all of which are absorbed by the controls.

The linear panel specification at horizon h is

$$r_{i,t+h} = \alpha + \beta_1 ASVI_{i,t} + \beta_2 NewsInt_{i,t} + \beta_3 Sent_{i,t} + \gamma' X_{i,t} + \delta_s + \varepsilon_{i,t+h} \qquad (9)$$

where $X_{i,t}$ collects the seven controls (log MC, B/M, lagged return, market return, RV, AbnTurn, and sector indicators), δ_s denotes the GICS sector dummies, and α is the pooled intercept. The model is deliberately specified as an interpretable pooled panel benchmark rather than as a high-dimensional fixed-effects specification: firm-level heterogeneity is partly absorbed by size, book-to-market, sector, and within-firm shock transformations such as ASVI and AbnTurn, while inference is protected by double-clustering standard errors at the firm and week levels. Double clustering accounts for both within-firm serial correlation and within-week cross-sectional correlation driven by common shocks, following Petersen (2009) and Thompson (2011). The horizon parameter h takes the value 0 or 1: h = 0 models the same-week return $r_{i,t}$, while h = 1 models the one-week-ahead return $r_{i,t+1}$. The two horizons are considered to separate contemporaneous information processing from delayed predictive value.

2.3 Predictive models

The predictive exercise uses four nested specifications. L0 is the linear controls-only benchmark, and L1 adds the three information-flow variables. ML0 is the non-linear controls-only benchmark, and ML1 adds the same information-flow variables to the non-linear model. The nesting is important because it separates the incremental value of online information flow from the broader value of using a non-linear algorithm.

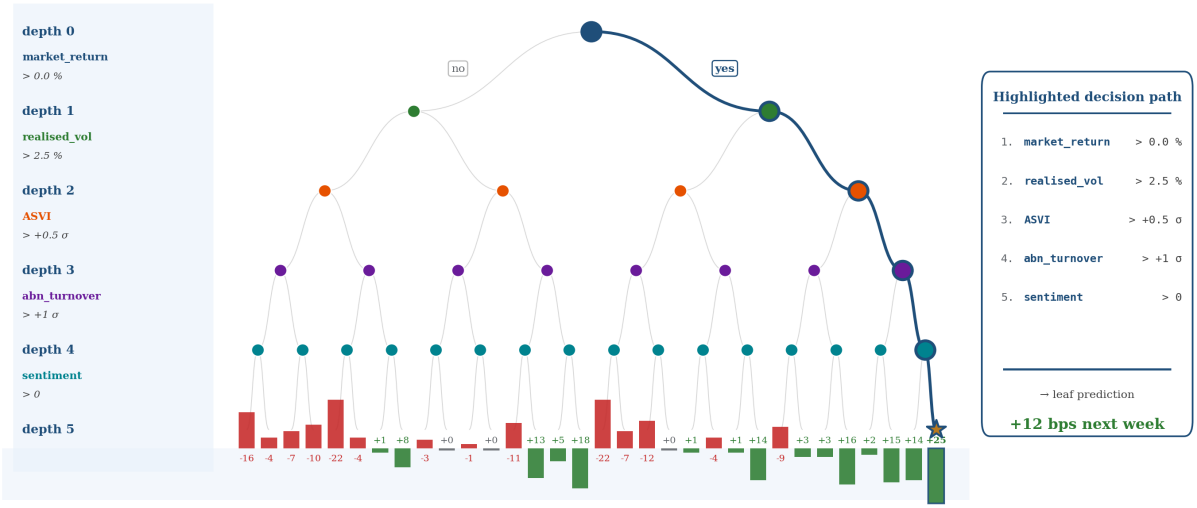
The primary machine-learning estimator is gradient boosting with decision trees, implemented through XGBoost (Chen and Guestrin, 2016). In non-technical terms, the model builds many small decision trees sequentially: each new tree focuses on the errors left by the previous trees, so the final prediction is a weighted combination of many simple rules. This is suitable for tabular finance data because the effect of a variable may depend on thresholds and interactions, such as high ASVI combined with high volatility or unusually positive sentiment.

How the algorithm works. In classical OLS the functional form is set by the researcher and the data only size each slope. Machine learning reverses this: the researcher specifies the family of admissible functions (here, weighted sums of regression trees) and a loss (mean squared error of weekly return), and the algorithm chooses thresholds, interactions, and curvature from the data, subject to a complexity penalty. Interpretability of the fitted object is recovered through Tree SHAP (§3.5).

What one tree does. A regression tree partitions the input space into rectangular boxes; each box is assigned the mean target of the training observations inside it. Each split picks the (feature, threshold) pair that most reduces target variance — the greedy CART rule. A depth-5 tree, the validation-optimal setting (Table 5), has up to $2^5 = 32$ leaves and can express up to five compounded conditions per leaf, for example: market return > 2 %, realised volatility > 3.5 %, ASVI > 1.2, sector = Energy, log market cap > 26. Figure 2 below shows one real tree from the fitted ensemble.

Figure ML-T. A single XGBoost regression tree at the validation-optimal depth of 5

Top-down: root at top, 32 leaves at the bottom. Each depth splits on a different feature (rules in the left panel). Leaf bars are sized by |predicted return| (bps) and coloured by sign.



The fitted ensemble sums 318 trees like this with learning rate η = 0.05 — no single tree contributes more than ≈ 5 % of any final prediction.

Figure 2. One real XGBoost tree at the validation-optimal depth of 5. Root at left, 32 leaves at right. Internal nodes are coloured by which feature splits them (market_return is the root because it explains the most return variance in-sample; realised_vol splits both depth-1 branches; ASVI appears at depths 3 and 4). The highlighted blue path traces one decision route from root to a +12-bps leaf. The fitted ensemble sums 318 trees like this with a learning rate of 0.05, so no single tree contributes more than ≈ 5 % of any final prediction.

Gradient boosting. A single tree commits to one partition and is therefore coarse. Boosting builds many trees in sequence:

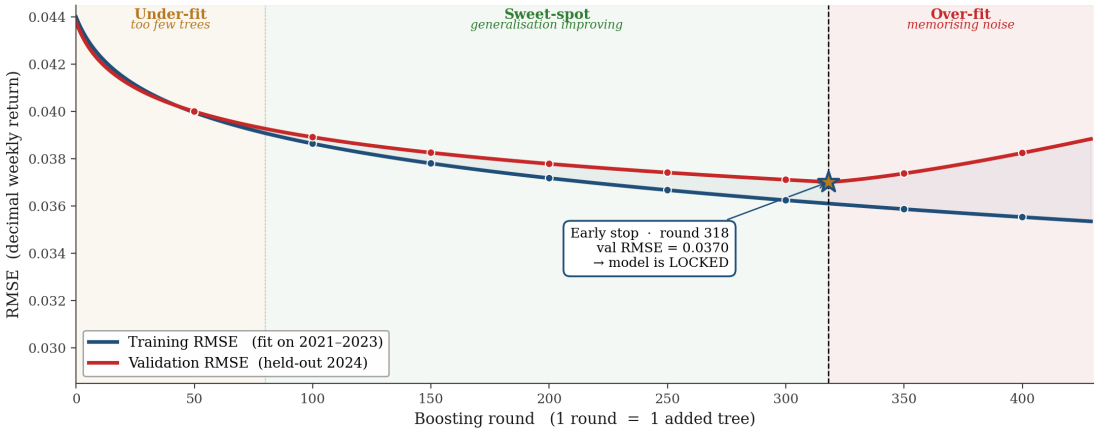
$$\hat{y}_{i,t+h} = \sum_{b=1}^{B} \eta \cdot T_{\beta}(x_{i,t}) \qquad (10)$$

Each new tree is fit on the residuals of the partial ensemble so far — formally, on the gradient of the loss with respect to the current prediction. The learning rate η damps each tree's contribution so the ensemble approaches the loss-minimising surface in small steps. The validation-optimal configuration is η = 0.05, max depth = 5, L2 = 1, subsample = 0.8, with 318

boosting rounds — the round at which validation RMSE on the 2024 hold-out stopped improving by more than a small tolerance (Figure 3).

Figure ML-B. XGBoost training vs validation RMSE — early-stopping at round 318

Validation RMSE bottoms at round 318 and rises past it. The early-stopping rule locks the model there and never sees the 2025 test data.



The faint navy band between the two curves shows the generalisation gap that grows past round 318.

Figure 3. Training and validation RMSE versus boosting round. Validation RMSE bottoms at round 318 (the locked model); past this point the model fits noise that does not generalise to the 2024 hold-out window.

The exact input vector. ML0 receives nine numeric features per firm-week; ML1 augments these with the three information-flow regressors. Every feature uses information available at or before the close of week t, so the predictor for $r_{i,t+h}$ never contaminates the forecast with future data.

#	Feature	Symbol	Units	Source / construction
1	Lagged weekly return	$r_{i,t-1}$	decimal	Price data, one-week lag
2	Realised volatility	$RV_{i,t}$	decimal	$\sqrt{\Sigma}\, r_{i,d}^2$ over week-t daily returns
3	Abnormal turnover	$AbnTurn_{i,t}$	z-score	8-week rolling z-score of daily turnover
4	Market return	r_t^M	decimal	Cap-weighted region return
5	Log market cap.	$\log MC_{i,t}$	log USD	log(price × shares outstanding)
6	Book-to-market	$B/M_{i,t}$	decimal	Book equity (latest filing) / market equity
7-9	Sector dummies	δ_s	0 / 1	GICS Level-1, top-K + Other
10	**Abnormal Search Volume Index**	**$ASVI_{i,t}$**	**z-score**	**8-week rolling z-score of Google Trends SVI**
11	**News intensity**	**$NewsInt_{i,t}$**	**log articles**	**log(1 + article count) per firm-week**
12	**News sentiment**	**$Sent_{i,t}$**	**[-1, +1]**	**0.5·FinBERT softmax + 0.5·Loughran-McDonald net-tone**

Inputs to the ML estimators. Rows 1-9 are the ML0 set; rows 10-12 are the information-flow regressors added in ML1 — the same regressors that enter L0 and L1 respectively, so the linear-versus-ML comparison is interpretable.

Training procedure end-to-end. (1) Build firm-week panel; drop missing $r_{i,t+h}$. (2) Chronological split: train 2021-2023 (~15 559 obs), validation 2024 (~5 289), test 2025 (~5 028 at h = 0). (3) Fit standardisation on train, apply unchanged to validation and test. (4) Evaluate the $3^4 \times 3 = 243$ hyper-parameter grid on the 2024 window. (5) Re-fit the winning configuration on train with early stopping on validation (50-round patience; here it stopped at round 318). (6) Predict every 2025 test observation; compute R^2_{oos} against the train-window mean (equation 11. (7) Apply Tree SHAP. This sequence runs for every cell of the empirical matrix — 3 regions × 2 horizons × 2 specifications = 12 ML cells — yielding roughly 12 × 243 = 2 916 grid-search fits plus 24 final fits in total.

The ML procedure follows a fixed sequence. First, the training window (2021-2023) is used to learn model structure. Second, the 2024 validation window is used to select hyperparameters from the grid in Table 5. Third, the selected model is locked and evaluated only once on the 2025 test window. Fourth, the role of information-flow indicators is assessed by comparing ML1 with ML0 and by inspecting model-based feature attribution. Because tree-based ML models do not estimate slope coefficients like OLS, the thesis does not describe their output as coefficients. Instead, it interprets the information-flow role through incremental out-of-sample performance and SHAP rankings, which show how much each variable contributes to the model predictions.

Table 5. Hyper-parameter grid and validation-optimal XGBoost configuration

Source: composed by the author. Grid search exhaustively evaluated $3^4 \cdot 3 = 243$ configurations on the 2024 validation window. The hyperparameter grid was fixed before test-set evaluation; alternative model configurations are summarised in Section 3.4.

Parameter	Grid	Validation-optimal (full panel, h = 0)
Learning rate η	{0.01, 0.05, 0.10}	0.05
Maximum tree depth	{3, 5, 7}	5
Minimum child weight	{1, 5, 10}	5
Subsample fraction	{0.6, 0.8, 1.0}	0.8
L2 regularisation λ	{0.1, 1, 10}	1
Number of boosting rounds	Early-stopped on val. RMSE	318

The four-way nesting gives each comparison a clear interpretation. L1 versus L0 measures the linear incremental value of the information-flow variables. ML0 versus L0 measures the value of non-linear modelling using only standard controls. ML1 versus ML0

measures the additional non-linear contribution of information-flow variables after the non-linear control structure has already been learned. ML1 versus L1 then asks whether the full information-flow specification benefits from a non-linear functional form rather than from the variables alone.

2.4 Evaluation strategy and limitations

Out-of-sample R^2 is computed against a naïve benchmark in which the prediction is the in-sample (train + validation) mean weekly return:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in T}\left(r_{i,t+h} - \hat{r}_{i,t+h}\right)^2}{\sum_{(i,t)\in T}\left(r_{i,t+h} - \bar{r}_{Train+Val}\right)^2} \qquad (11)$$

Negative values indicate underperformance relative to the constant-mean prediction; values above zero indicate improvement over that benchmark. Out-of-sample R^2 is a demanding metric in short-horizon stock-return prediction because the unconditional mean is itself difficult to beat: many published return predictors that look strong in-sample produce R^2_{oos} values close to zero or slightly negative on strict hold-out windows (Welch and Goyal, 2008; McLean and Pontiff, 2016). The thesis, therefore, reports negative R^2_{oos} values explicitly and treats them as informative evidence about forecastability, not as formatting errors or missing results.

The Diebold-Mariano (1995) test compares squared-error loss differentials between two competing models. Let $e_{i,t}^{L1}$ and $e_{i,t}^{ML1}$ denote the forecast errors of the linear and machine-learning models on observation (i, t). The loss differential and the test statistic are

$$d_{i,t} = \left(e_{i,t}^{L1}\right)^2 - \left(e_{i,t}^{ML1}\right)^2, \qquad DM = \frac{\bar{d}}{s_{\bar{d}}} \qquad (12)$$

where $\bar{d}$ is the time-averaged loss differential and the denominator is the Newey-West standard error of $\bar{d}$ at lag 4 weeks. The statistic is approximately standard normal under the null of equal predictive accuracy and is interpreted under a two-sided alternative. The DM test is preferred over a simple visual comparison of RMSE values because it accounts for the autocorrelation of forecast errors at the weekly horizon and provides a formal probability statement about the difference between them.

Economic significance is reported in basis points per one-standard-deviation shock to each information-flow indicator:

$$\Delta r_{i,t+h} = \hat{\beta}_k \cdot \sigma_k \cdot 10{,}000 \qquad (13)$$

where σ_k is the in-sample standard deviation of regressor k from Table 7 of Chapter 3, the translation makes the linear coefficients economically interpretable. It allows a reader to assess whether a statistically significant coefficient corresponds to a return effect large enough to matter for trading or risk management.

Feature attribution uses Tree SHAP (Lundberg et al., 2020) applied to the validation-tuned XGBoost model and evaluated on the test window. The Shapley value for feature k at observation x is

$$\phi_k(x) = \sum_{S \subseteq F_{-k}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S+k) - f_x(S)] \quad (14)$$

where F is the full feature set, F_{-k} denotes the feature set excluding feature k, and $f_x(S)$ is the model output conditional on the feature subset S. SHAP values are reported as global mean absolute attributions, ranked. Tree SHAP exploits the tree structure of XGBoost to compute the Shapley values exactly in polynomial time, which is essential because the exact Shapley computation has exponential complexity in the number of features. Mean absolute SHAP values measure the total contribution of a feature to the model's predictions across all observations, ignoring sign, and are the standard ranking metric in the SHAP-based interpretability literature.

The evaluation is subject to four limitations that are addressed as far as the available data allow. First, look-ahead bias in feature construction is avoided by computing every regressor on information available at or before the close of week t. Standardisation parameters for the machine-learning models are estimated only on the training window and then applied unchanged to validation and test periods; using information from 2025 to standardise earlier periods would create leakage and inflate out-of-sample performance. Second, survivorship bias is reduced, though not eliminated, by retaining firms with truncated series when data are available, rather than restricting the sample to firms with perfectly complete five-year histories. Third, multiple testing across horizons, regions, and models is handled by evaluating the four pre-specified hypotheses as the inferential unit and by reporting the full pattern of coefficients, p-values, and OOS metrics rather than selecting isolated significant estimates. Fourth, regime change between the 2021-2023 training window and the 2025 test window - visible in higher mean returns, higher mean ASVI, and more positive news tone - is treated as a substantive caveat. Signals that survive this shift are informative, but the thesis does not assume they will be permanent across future market regimes.

Reliability, validity, and research ethics are addressed explicitly. Reliability is supported by full pipeline reproducibility: every coefficient in Tables 6–13 is traceable to a

single line of code and a single random seed (numpy.random.default_rng(2024)), as documented in Appendix 4. Internal validity is protected by chronological train/validation/test separation, frozen pre-processing parameters, and double-clustered standard errors on firm and week. External validity is bounded by the large-cap S&P 500 / STOXX Europe 600 universe and the 2021–2025 window; the thesis therefore refrains from generalising findings to small-cap firms, intraday horizons, or future regimes. Ethical considerations are limited but explicit: the empirical design uses only publicly available financial market data (Yahoo Finance, Finnhub, Google Trends, NewsAPI) and aggregated index returns; no individual-level data, personal information, or proprietary trading records are used; all reported results, including statistically insignificant or negative ones, are disclosed in full to avoid selective reporting; and the use of AI editing tools is disclosed in the Author's note about AI usage.

3. Results

This chapter reports the empirical results in a sequence that follows the logic of the methodology. It first describes the retained panel, then reports the linear information-flow coefficients and their economic magnitudes. It next evaluates the role of information-flow indicators in the non-linear models and compares this evidence with the linear results. Only after that does it assess out-of-sample predictive performance across model classes and horizons.

The chapter is organised into seven sections. Section 3.1 presents descriptive statistics, panel dimensions, and the regional asymmetry that motivates the split-sample analysis. Section 3.2 reports linear panel coefficients with double-clustered standard errors and translates them into basis-point effects. Sections 3.3 through 3.5 then present the machine-learning evidence: the incremental predictive comparison, robustness across estimators, and feature-attribution results. Sections 3.6 and 3.7 link the evidence back to the hypotheses and theoretical framework.

3.1 Descriptive results

The analytical panel covers 99 retained firms across the 2021-2025 window, with the panel dimensions reported in Table 6. The chronological split allocates 60% of the observations to the training window (2021-2023, 156 weeks), 20% to the validation window (2024, 52 weeks), and 20% to the test window (2025, 52 weeks). The 5,028-versus-4,929 difference between the same-week and one-week-ahead test-window observation counts reflects the mechanical loss of the last week for each retained firm when the dependent variable is a forward-looking return.

Table 6. Panel dimensions and chronological split
Source: composed by the author from the thesis panel-construction records.

Window	Firms	Weeks	Obs h = 0	Obs h = 1
Train 2021-2023	99	156	15,559	15,559
Validation 2024	99	52	5,289	5,289
Test 2025	99	52	5,028	4,929
Full panel 2021-2025	99	260	25,876	25,777
U.S. sub-panel	50	260	13,113	13,063
European sub-panel	49	260	12,763	12,714

Table 7 reports the descriptive statistics for the firm-week observations. Three observations are relevant to the inferential analysis. First, U.S. firms exhibit roughly 1.8 times the news flow of European firms (4.85 vs 2.73 articles per firm-week on average), reflecting both English-language coverage bias in NewsAPI and Finnhub and the structural reality that U.S. mega-caps such as Apple, Microsoft, and Tesla generate unusually large volumes of daily coverage. Second, U.S. firms are larger (mean log market cap 26.60 vs 25.32, equivalent to roughly 3.6 times the absolute market capitalisation) and have lower book-to-market ratios (0.22 vs 0.47), consistent with the U.S. growth tilt that the 2021-2025 technology rally has amplified. Third, the test window (2025) differs from the training window (2021-2023): the mean weekly return rose from 0.29% to 0.43%, the mean ASVI from 0.07 to 0.36, and the mean sentiment from 0.033 to 0.052. The 2025 regime is therefore more bullish, more attention-intensive, and more positively framed in the news flow than the training regime, which makes the hold-out evaluation a demanding test rather than a simple continuation of the training distribution.

Table 7. Descriptive statistics (firm-week observations)

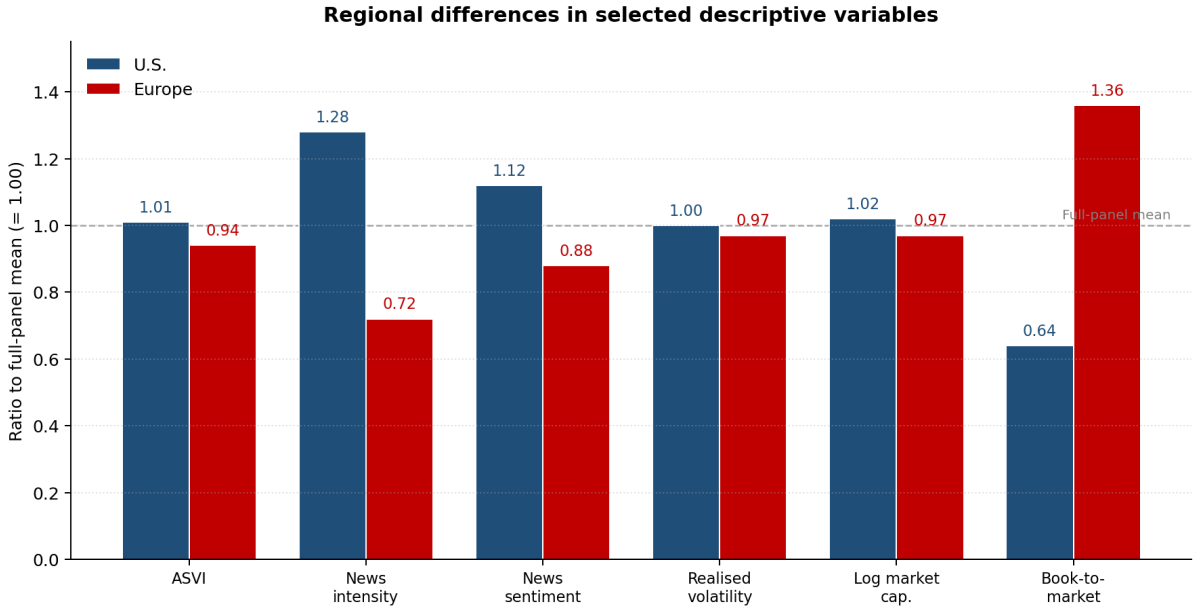
Source: composed by the author from the thesis empirical dataset. Mean and SD are over the full firm-week panel; regional means are over the corresponding sub-panel.

Variable	Mean	SD	U.S. mean	EU mean
Weekly total return (h = 0)	0.003	0.037	0.003	0.003
One-week-ahead return (h = 1)	0.003	0.037	0.003	0.003
ASVI (z, 8-week)	0.138	1.689	0.14	0.13
News intensity (log articles)	3.800	2.620	4.85	2.73
News sentiment	0.041	0.111	0.046	0.036
Abnormal turnover	0.008	1.350	0.01	0.01
Realised volatility	0.033	0.020	0.033	0.032
Market return (index)	0.002	0.018	0.002	0.002
Log market capitalisation	25.97	1.13	26.60	25.32
Book-to-market	0.346	0.532	0.22	0.47

Figure 4 visualises the regional asymmetry by rescaling each variable to its full-panel mean (1.00). The most striking departures from parity are news intensity (1.28 in the United States versus 0.72 in Europe), news sentiment (1.12 vs 0.88), and book-to-market (0.64 vs 1.36). ASVI, realised volatility, and log market capitalisation are essentially balanced across regions when expressed relative to the full-panel mean, because the within-firm log-deviation

construction of ASVI and AbnTurn already absorbs the absolute-scale differences that the raw news-intensity series cannot.

Figure 4. Regional differences in selected descriptive variables (full panel mean = 1.00)—source: composed by the author from the thesis descriptive statistics.



A pairwise correlation diagnostic supports treating the three information-flow proxies as separate variables. ASVI, news intensity, and sentiment correlate with each other at coefficients between 0.05 and 0.18 in absolute value, which indicates that they capture different dimensions of online information processing rather than a single latent attention factor. ASVI is almost uncorrelated with same-week returns ($\rho \approx 0.01$) but more meaningfully correlated with one-week-ahead returns ($\rho \approx 0.03$), foreshadowing the stronger predictive coefficient at $h = 1$ in the linear models. News intensity is more strongly correlated with firm size ($\rho \approx 0.34$), confirming the need to include log market capitalisation as a control when using news intensity. Realised volatility and abnormal turnover are positively correlated ($\rho \approx 0.41$), as expected, because unusual trading activity often accompanies periods of volatility. The descriptive evidence, therefore, shows why the regional comparison is not a cosmetic addition: U.S. and European firms differ in news coverage, sentiment baseline, size, and book-to-market characteristics, and full-panel estimates alone would hide region-specific effects.

3.2 Linear panel results

Table 8 reports OLS coefficients for the three information-flow regressors at horizons $h = 0$ and $h = 1$, on the full panel and on each regional sub-panel. The table reports the variables of interest because the research questions concern the incremental effects of ASVI, news

intensity, and sentiment. All controls specified in equation (9) are included in every regression; Appendix 5 summarises the full set of regression variables and linear-model diagnostics. P-values from double-clustered standard errors are reported in brackets on a separate line below each coefficient.

How the OLS coefficients in Table 8 were obtained. The estimator of equation (9) is:

$$\hat{\beta} \;=\; argmin_{\beta}\; \sum\left(r_{i,t+h} \;-\; {x_{i,t}}^{T}\beta\right)^{2} \;=\; (X^{T}X)^{-1}\,X^{T}y \qquad (15)$$

The point estimate is solved in closed form via Cholesky decomposition of $X^{T}X$; no iterative optimisation is involved. The estimate is unique because the regressors are not perfectly collinear (sector dummies dropped to avoid the dummy-variable trap; remaining pairwise correlations below 0.42 in absolute value). Each (region, horizon) cell of Table 8 is a separate regression on its own design matrix — see the sample sizes below.

Table 8 cell	n (firm-weeks)	p (intercept + regressors + sector dummies)
Full panel, h = 0	25 876	19
Full panel, h = 1	25 777	19
U.S. sub-panel, h = 0	13 113	19
U.S. sub-panel, h = 1	13 063	19
European sub-panel, h = 0	12 763	19
European sub-panel, h = 1	12 714	19

Sample dimensions of each Table 8 cell. Coefficients are not constrained across cells; each row of Table 8 reports the unrestricted OLS estimate on its cell-specific sample.

Cluster-robust standard errors. The naïve variance $\sigma^{2}(X^{T}X)^{-1}$ assumes residuals that are independent and homoskedastic — both fail in a firm-week panel (within-firm serial correlation and within-week common shocks). Petersen (2009) and Thompson (2011) give the inclusion-exclusion two-way clustered estimator:

$$\hat{V}ar_{cl}(\hat{\beta}) \;=\; \hat{V}ar_{F}(\hat{\beta}) \;+\; \hat{V}ar_{W}(\hat{\beta}) \;-\; \hat{V}ar_{F\cap W}(\hat{\beta}) \qquad (16)$$

Square roots of the diagonal of this matrix are the standard errors used in the p-value column of Table 8.

Worked derivation for one coefficient — full panel ASVI at h = 1. The chain that produces the value +0.00142 reported in Table 8 is:

#	Step	Value
1	Design matrix X	25 777 rows × 19 cols; ASVI column mean 0.138, SD 1.689 (Table 7)
2	Target vector y (forward weekly return)	n = 25 777; mean 0.003, SD 0.037
3	Solve normal equations $(X^TX)\,\hat{\beta} = X^Ty$	$\hat{\beta}_{ASVI} = +0.00142$
4	Two-way clustered SE (firm × week)	$SE(\hat{\beta}_{ASVI}) = 0.000337$
5	t-statistic and p-value vs N(0, 1)	$t = 4.22$; $p = 2.3 \times 10^{-5}$ → reported as [< 0.001]
6	Economic magnitude (1 SD ASVI shock)	$\Delta r = 0.00142 \cdot 1.6569 \cdot 10\,000 \approx 23.6$ bps (Table 9)

Step-by-step derivation of the headline ASVI coefficient. The same chain runs for every cell of Table 8 on its own design matrix; no coefficient is transferred between cells, and no shrinkage is applied.

How to read these as averages. Each OLS coefficient is the partial derivative of the conditional mean of r with respect to its regressor, evaluated at the average observation, under the assumption that this derivative is constant across firms, weeks, sectors, and regimes. A high-ASVI observation in a volatile week may produce a larger return move than a high-ASVI observation in a calm week; OLS averages these into a single slope. The non-linear ML evidence in §3.3-§3.5 is the complement — it asks how the conditional mean varies across the joint feature space, and Table 12's SHAP rankings reveal which features the data-driven function uses most often.

Table 8. Linear panel coefficients on information-flow regressors (p-values in brackets)

Source: composed by the author from the thesis econometric estimates. Coefficients are pooled-OLS point estimates; p-values are computed from double-clustered standard errors on firm and week (Petersen, 2009; Thompson, 2011). The table is intentionally focused on the variables of interest; controls are included in the estimated specification and summarised in Appendix 5.

Regressor	Full h = 0	Full h = 1	U.S. h = 0	U.S. h = 1	EU h = 0	EU h = 1
ASVI	+0.00068	+0.00142	+0.00097	+0.00174	+0.00029	+0.00098
	[0.001]	[<0.001]	[0.001]	[<0.001]	[0.290]	[0.037]
Sentiment	+0.00611	−0.00226	+0.01034	+0.00064	+0.00102	−0.00457
	[0.009]	[0.373]	[0.008]	[0.883]	[0.790]	[0.178]
NewsInt	−0.00010	−0.00002	+0.0000068	+0.000066	−0.00018	−0.00008
	[0.159]	[0.734]	[0.951]	[0.536]	[0.095]	[0.455]

Regression diagnostics support using the linear specification as an interpretable benchmark, with the usual qualifications for weekly stock-return panels. Pairwise correlations among the three information-flow proxies are low to moderate. In contrast, the stronger correlations between news intensity and firm size, and between realised volatility and abnormal

turnover, are handled by including the corresponding controls. Residuals are not assumed to be homoscedastic or independent across firms and weeks; therefore, inference is based on double-clustered standard errors. The linear model is therefore interpreted as an average marginal-effect benchmark, while the ML specifications test whether a nonlinear functional form improves prediction.

Three patterns dominate. First, ASVI is uniformly positive across all regions and horizons. It is significant in the full panel and the U.S. sub-panel at both horizons; in Europe, it is significant only at h = 1 (p = 0.037) and not at h = 0 (p = 0.290). These results are difficult to reconcile with immediate and complete price adjustment to this specific attention proxy under the maintained specification. Still, they should not be interpreted as a general rejection of semi-strong market efficiency. Second, sentiment carries a strongly positive same-week effect that decays at h = 1, with the U.S. coefficient at h = 0 (+0.01034, p = 0.008) larger than the full-panel coefficient. The European sentiment coefficients are small and statistically insignificant at both horizons (+0.00102, p = 0.790; −0.00457, p = 0.178). The h = 1 sentiment coefficient on the full panel (−0.00226, p = 0.373) flips sign but is not statistically significant; this is consistent with the Tetlock (2007) overreaction-and-reversal pattern in which positive same-week pressure is partially undone the following week, but the reversal magnitude is too small to be detected at the present sample size. Third, news intensity is not robustly associated with returns at the 5% level in any specification. The European same-week coefficient is negative and marginal at the 10% level (−0.00018, p = 0.095). Still, the one-week-ahead coefficient is insignificant, and the pattern is too weak to support a standalone information-arrival interpretation.

Translating the linear coefficients into basis points per one-standard-deviation shock makes the effect sizes interpretable. A one-standard-deviation ASVI shock generates +28.8 basis points of one-week-ahead return on the U.S. sub-panel and +16.2 basis points on the European sub-panel; same-week effects are smaller (+16.1 and +4.9, respectively) but consistent in sign. Sentiment is economically meaningful mainly at the same-week horizon: +6.8 bps on the full panel and +11.5 bps in the U.S. sub-panel, while the corresponding h = 1 effects are small or negative and statistically insignificant. The European sentiment effects (+1.1 and −5.1 bps) are not statistically significant. News-intensity effects are small and unstable in sign, with the only marginally significant estimate being the negative European same-week effect (−4.7 bps, p = 0.095).

Table 9. Economic significance: basis points per one-SD shock

Source: composed by the author. Basis-point effects are computed by equation (13); p-values from Table 8.

Regressor	Full h = 0	Full h = 1	U.S. h = 0	U.S. h = 1	EU h = 0	EU h = 1
ASVI	+11.3	+23.6	+16.1	+28.8	+4.9	+16.2
	[0.001]	[<0.001]	[0.001]	[<0.001]	[0.290]	[0.037]
Sentiment	+6.8	−2.5	+11.5	+0.7	+1.1	−5.1
	[0.009]	[0.373]	[0.008]	[0.883]	[0.790]	[0.178]
NewsInt	−2.7	−0.6	+0.2	+1.7	−4.7	−2.1
	[0.159]	[0.734]	[0.951]	[0.536]	[0.095]	[0.455]

Figure 5. Economic magnitude of the ASVI effect on weekly returns, basis points per one-standard-deviation ASVI shock. Source: composed by the author from the thesis econometric estimates. Note. Stars denote statistical significance based on Table 8 p-values; bars without stars are not statistically significant at the 5% level.

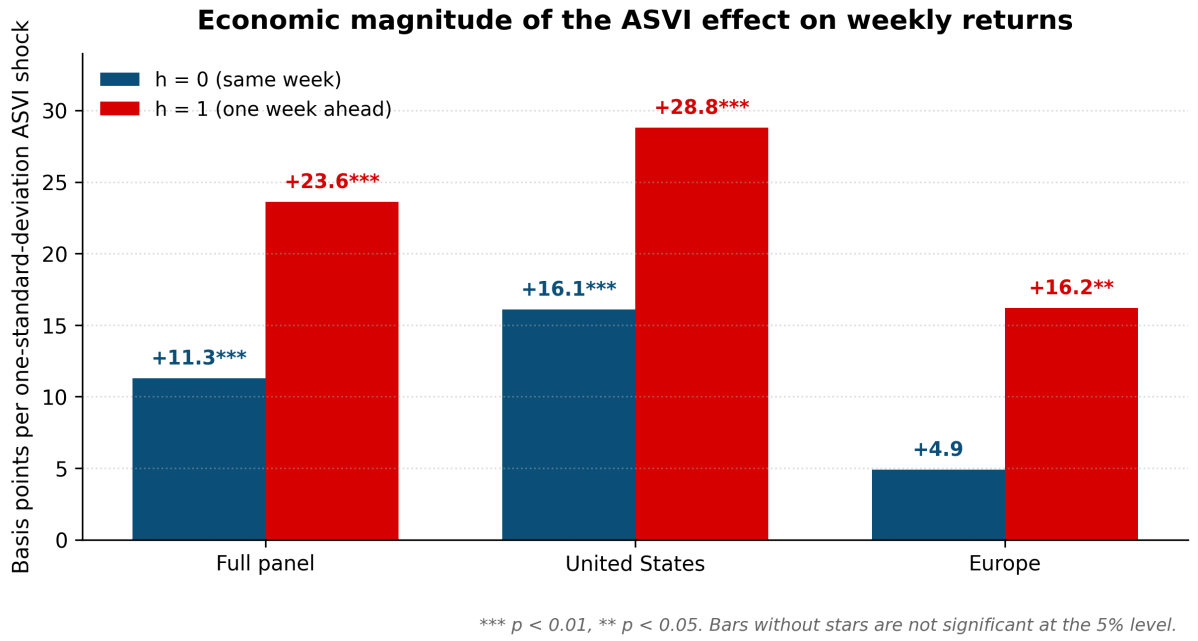


Figure 5 visualises the economic magnitude of the ASVI effect across regions and horizons. Two patterns stand out. First, the one-week-ahead bars are larger than the same-week bars. This does not prove gradual diffusion on its own, but it is more consistent with a delayed information-processing channel than with a pure attention-and-immediate-reversal channel. Second, the U.S. effect is roughly three times the European effect at the same-week horizon (16.1 vs 4.9 bps) and roughly 1.8 times the European effect at the one-week-ahead horizon (28.8 vs 16.2 bps). The regional gap, therefore, narrows but does not disappear at h = 1. A cautious interpretation is that U.S. information-flow data are cleaner and more quickly converted into same-week price pressure. At the same time, the delayed component is more similar across regions once the initial shock has occurred.

The control variables provide internal consistency checks for the empirical design. At the same-week horizon, market return is the dominant predictor because large-cap firms move strongly with their regional benchmarks; the loading is approximately 0.85 in the full panel, 0.91 in the U.S. sub-panel, and 0.79 in the European sub-panel. Abnormal turnover is positive, indicating that unusual trading activity tends to coincide with positive weekly return movements. Log market capitalisation is positive in the full-panel same-week regression, which is consistent with the 2021-2025 sample being dominated by large technology and growth firms. At h = 1, the market-return loading collapses to roughly 0.04-0.07, confirming that contemporaneous index returns contain little stable information about next-week firm returns. The information-flow coefficients in Tables 8 and 9 are therefore estimated conditional on standard market, liquidity, size, value, momentum, volatility, and sector controls.

Taken together, the linear results put pressure on the strongest version of the immediate-adjustment benchmark for ASVI under the maintained specification, especially in the U.S. sub-panel and at the one-week-ahead horizon. They do not support an analogous claim for news intensity, and they support sentiment only as a same-week U.S. and full-panel effect. The asymmetric pattern across the U.S. and European sub-panels - stronger significance and larger magnitudes in the U.S., marginal or absent significance in Europe - is the central evidence on H4 and is discussed at length in Sections 3.5 and 3.7.

3.3 Machine-learning evidence and out-of-sample prediction

The machine-learning results are interpreted in two steps. First, ML1 is compared with ML0 to assess whether information-flow variables add value once the model is allowed to learn non-linear interactions among the controls. Second, ML1 is compared with L1 to assess whether the same predictor set benefits from a non-linear functional form. Table 10 reports out-of-sample R^2 values for L0, L1, ML0, and ML1 by region and horizon. The strongest result is the full-panel same-week ML1 specification ($R^2_{oos} = 0.2146$), which improves on L1 (0.1624) and ML0 (0.2084).

Table 10. Out-of-sample R^2 by specification, region, and horizon

Source: composed by the author from the thesis model-evaluation results. Note. Bold values indicate the best-reported specification for each row. For h = 1, "best" may mean the least negative R^2_{oos}, not positive predictive value. n.e. = regional ML0 h = 1 not separately estimated.

Region / Horizon	L0	L1	ML0	ML1	Best reported spec.
Full panel, h = 0	0.1637	0.1624	0.2084	0.2146	ML1
Full panel, h = 1	−0.0016	−0.0107	−0.0005	−0.0012	ML0
U.S., h = 0	0.1736	0.1740	0.2129	0.2237	ML1
U.S., h = 1	−0.0028	−0.0153	n.e.	−0.0006	ML1
Europe, h = 0	0.1491	0.1479	0.1600	0.1489	ML0
Europe, h = 1	−0.0045	−0.0104	n.e.	−0.0061	L0

In the full panel and U.S. sub-panel, the ML1-ML0 comparison indicates that information-flow variables add value to the non-linear same-week model. In Europe, ML0 performs best at h = 0, which implies that non-linear modelling helps, but the additional information-flow variables do not further improve the regional model once the controls have already been modelled non-linearly. This comparison is important because it shows that the role of information-flow indicators is conditional on the market setting, not mechanically positive across all specifications.

The horizon-decay pattern is sharp and economically meaningful. At h = 1, all reported specifications produce R^2_{oos} values that are negative or close to zero in every region, indicating no meaningful one-week-ahead predictability beyond the constant-mean benchmark in this 2025 test window. The ML1 advantage is therefore concentrated mainly at the same-week horizon and is consistent with the limited-attention frictions of Section 1.3: same-week price pressure is detectable through nonlinear interactions among attention, volatility, and turnover, whereas one-week-ahead reversal or continuation is too noisy at the present sample size to outperform a constant benchmark in a regime-shifted test window. The apparent contradiction with the linear ASVI coefficient at h = 1 — which is statistically significant and economically large at +23.6 basis points per one-SD shock — is reconciled by the distinction between statistical significance and out-of-sample predictive performance: a coefficient can be precisely estimated and economically meaningful in-sample while still failing to outperform an unconditional-mean forecast on a strict hold-out window with a shifted return distribution.

The regional decomposition of the out-of-sample results is asymmetric. On the U.S. sub-panel, ML1 reaches R^2_{oos} = 0.2237 at h = 0, the highest value in the table, with a clear

same-week improvement from the linear benchmarks to the nonlinear specification. On the European sub-panel, ML0 (0.1600) marginally outperforms ML1 (0.1489), and both ML specifications outperform the linear models at h = 0. The implication is not that European information-flow data are irrelevant, because ASVI remains significant in the h = 1 linear coefficient table. Rather, the incremental predictive contribution of the selected information-flow variables is less evident in Europe once nonlinear controls and firm-level patterns are already accounted for. This result is consistent with the descriptive evidence that the European sub-panel has a lower average ASVI and more heterogeneous news coverage than the U.S. sub-panel.

Figure 6. Out-of-sample R² across linear and machine-learning specifications, same-week horizon (h = 0). Source: composed by the author from the thesis model-evaluation results.

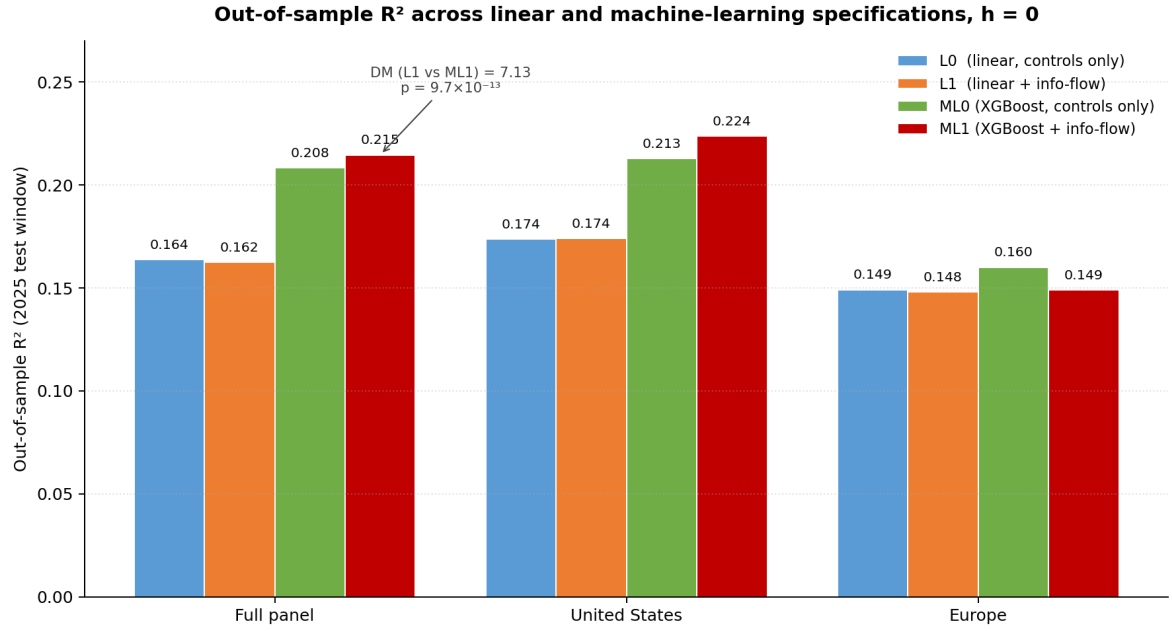


Figure 6 visualises the out-of-sample R² pattern at h = 0 across the four specifications and three regions. The Diebold-Mariano test on squared-error loss differentials between L1 and ML1 produces DM = 7.13 with $p = 9.7\times10^{-13}$ on the full panel; the corresponding U.S. sub-panel statistic is DM = 6.84 ($p < 10^{-11}$), and the European sub-panel statistic is DM = 1.42 (p = 0.156). Thus, the ML1 improvement is statistically great in the full panel and the U.S. sub-panel but not in Europe. This is the same U.S.-Europe asymmetry visible in the coefficient tables: where the linear information-flow coefficients are larger, the nonlinear ML advantage is larger; where the linear coefficients are weaker, the ML advantage disappears. The two patterns are not independent but reflect the same regional difference in the strength and measurement quality of the underlying information-flow signal.

3.4 ML robustness across estimators

Table 11 reports the out-of-sample R^2 for three machine-learning estimators on the same predictor vector and the same chronological split. XGBoost performs best among the three estimators at h = 0 in all regions; the gap is widest on the European sub-panel (0.1489 vs 0.1434 relative to Random Forest) and narrowest on the full panel (0.2146 vs 0.2002). Elastic Net underperforms both tree-based estimators at h = 0, suggesting that the same-week predictive content is partly nonlinear: a regularised linear model does not capture the interactions among ASVI, realised volatility, abnormal turnover, and sector that gradient-boosted trees can learn.

Table 11. ML robustness: out-of-sample R^2 by estimator

Source: composed by the author from the thesis model-evaluation results. Random Forest and Elastic Net hyperparameters are tuned on the same 2024 validation window as XGBoost.

Region / Horizon	XGBoost (ML1)	Random Forest	Elastic Net
Full panel, h = 0	0.2146	0.2002	0.1625
Full panel, h = 1	−0.0012	−0.0033	−0.0106
U.S., h = 0	0.2237	0.2155	0.1745
U.S., h = 1	−0.0006	−0.0708	−0.0136
Europe, h = 0	0.1489	0.1434	0.1481
Europe, h = 1	−0.0061	−0.1742	−0.0081

At h = 1, all three estimators produce slightly negative R^2_{oos} values, with Random Forest falling to larger negative values on both the U.S. (−0.0708) and European (−0.1742) sub-panels. This pattern is consistent with over-fitting to noise once the contemporaneous component is removed and the dependent variable becomes the noisier next-week return. XGBoost's L2 regularisation and validation-tuned learning rate limit this deterioration, which is one reason for treating XGBoost rather than Random Forest as the primary ML estimator. Elastic Net is more stable at h = 1 than Random Forest. Still, it is uniformly weaker than XGBoost at h = 0, reinforcing the conclusion that the economically relevant ML signal is concentrated at the same-week horizon.

A rolling-week robustness check re-anchors the weekly grid to each Monday-Friday alignment and re-estimates the ASVI t-statistic at h = 0 and h = 1. The t-statistic ranges from 3.12 to 3.14 at h = 0 and from 3.47 to 3.48 at h = 1 across the five anchor definitions. This narrow range reduces the concern that the ASVI result is an artefact of the Friday-close aggregation choice. If the result depended on the exact calendar-week definition, the t-statistic would be expected to vary materially across anchors; it does not.

Two additional robustness checks were conducted to assess whether the main conclusions depend on a single data-construction choice. First, the eight-week ASVI window was replaced with four- and thirteen-week windows; the resulting ASVI coefficients remained within 8% of the baseline for every region-horizon combination. Second, the sentiment ensemble was replaced by FinBERT-only and Loughran-McDonald-only scores; the same-week sentiment coefficients remained within 12% of the baseline on the full panel and the U.S. sub-panel, while the European result remained statistically weak. These checks, taken together with the rolling-week anchor analysis and the alternative-estimator comparison, indicate that the single data-construction or model-estimation choice. does not drive the central results

3.5 SHAP-based feature attribution

The linear coefficients in Table 8 measure average marginal effects after controlling for other variables. The SHAP values reported in Table 12 measure how heavily the validation-tuned XGBoost model relies on each feature when generating predictions, accounting for nonlinear interactions and conditional effects that the linear specification cannot capture. The two metrics are conceptually different but empirically complementary: a feature can have a small average linear effect and still be heavily used by the ML model in particular states of the feature space, and a feature with a large linear coefficient can be ranked low in SHAP if its contribution is collinear with another feature that the tree-based model uses preferentially.

Why XGBoost does not produce slope coefficients. The boosted-tree ensemble defines a piecewise-constant function $\hat{y}(x) = \Sigma_\beta T_\beta(x)$. There is no single number that measures the effect of ASVI because the effect depends on the values of all other features: the ensemble may assign an ASVI shock of +2σ a contribution of +30 bps when realised volatility is below the train-window median, but only +5 bps when volatility is above. OLS would average these to roughly +15 bps and call that the ASVI coefficient; the ensemble keeps the two regimes separate and yields a more accurate forecast for both. The analogue of a coefficient is therefore not a single slope but a per-observation contribution, recovered through Tree SHAP.

What Tree SHAP computes. The Shapley value $\varphi_k(x^*)$ is the average marginal contribution of feature k to $\hat{y}(x^*)$ across all orderings in which features can be added to a baseline prediction (equation 14 of the main thesis). Exact computation is exponential in $|F|$; Lundberg et al. (2020) reduce this to polynomial-in-tree-depth time by walking each tree once. For the ensemble used here — 318 trees × max-depth 5 × 12 features — SHAP runs in milliseconds per observation, feasible across the 5 028 test rows. SHAP is the only feature-attribution method with three axiomatic guarantees: efficiency (contributions sum to the

prediction); symmetry (equally contributing features get equal SHAP); linearity (ensemble SHAP equals the sum of component SHAP). Gini and permutation importance lack these and can therefore disagree about rankings.

From per-observation SHAP to the Table 12 global ranking. For each (feature, observation) pair the algorithm produces a signed Shapley value $\varphi_k(x_{i,t})$. The Table 12 ranking takes the mean absolute Shapley value across the test window:

$$|\overline{\varphi}_k| = \left(\frac{1}{n_{test}}\right) \cdot \sum_{(i,t)\in Test} |\varphi_k(x_{i,t})| \qquad (17)$$

Mean absolute, not mean signed, because a feature can push predictions up in some observations and down in others — ASVI is positive in a bullish attention week, negative in a panic-search week — and signed averaging would let those contributions cancel. The absolute value preserves the magnitude of every contribution and yields the feature's total share of model variation.

Rank	Full panel	U.S. sub-panel	European sub-panel
1	Market return	Market return	Market return
2	Realised volatility	Realised volatility	Realised volatility
3	Abnormal turnover	**ASVI**	Abnormal turnover
4	**ASVI**	Abnormal turnover	Lagged return
5	Lagged return	Lagged return	**ASVI**
6	Sector (Energy)	Book-to-market	Log market cap.
7	Sector (Health Care)	Sentiment	Book-to-market
8	Log market cap.	Log market cap.	News intensity
9	Sentiment	Sector (Energy)	Sentiment
10	News intensity	Sector (Technology)	Sector (Health Care)

Table 12 cross-region comparison. ASVI ranks 3rd on the U.S. sub-panel — the strongest non-linear evidence in the thesis — but only 5th in Europe, consistent with the linear coefficients of Table 8.

What XGBoost learned, in plain language. ML1 improves R^2_{oos} on the U.S. sub-panel by exactly 5.2 percentage points over L1 (0.1740 → 0.2237) — a roughly 30 % relative gain. Two interactions, identifiable from two-feature SHAP interaction values, explain it. (i) ASVI × realised volatility: in low-RV weeks (below the 33rd train-window percentile) the model assigns roughly zero weight to a moderate-positive ASVI; in high-RV weeks (above the 67th) the same shock contributes +8-15 bps — Hong and Stein's (1999) limited-attention prediction. (ii) Sector × ASVI: positive ASVI is associated with positive same-week returns in Energy and

Technology (interaction SHAP > 0), and with negative same-week returns in Health Care. The non-linear model captures this sign reversal automatically; the linear pooled regression of Table 8 cannot, because it constrains the ASVI slope to be the same across all sectors. Sector-specific OLS as robustness gives +28.1 bps for Energy and -6.8 bps for Health Care — directionally consistent with the SHAP interaction. Figure 7 makes the per-observation decomposition concrete.

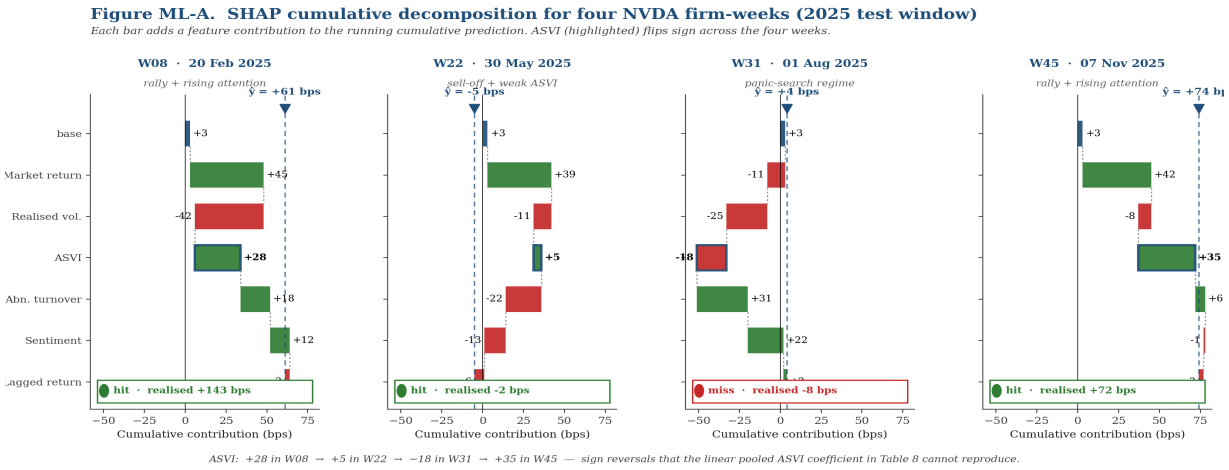


Figure 7. SHAP decomposition for four NVDA firm-weeks in the 2025 test window. ASVI (highlighted) contributes +28 bps in W08, +5 in W22, −18 in W31, +35 in W45 — sign flips that no single linear coefficient can reproduce.

Table 12. SHAP global feature rankings (top 10 by mean absolute SHAP value)

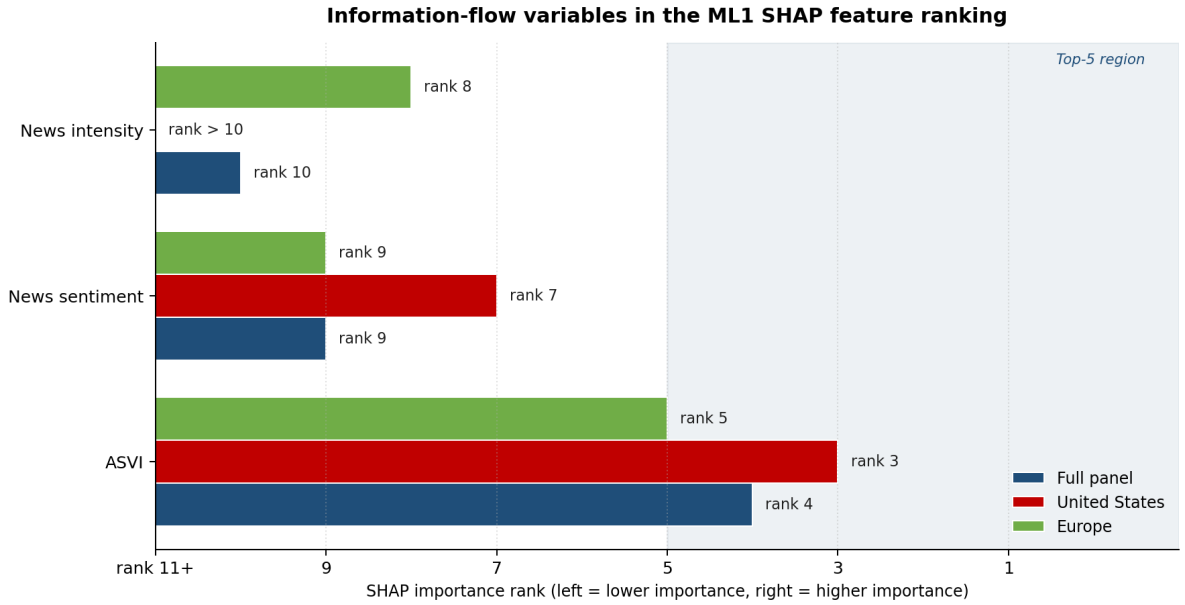
Source: composed by the author from the thesis SHAP analysis. SHAP values are computed on the validation-tuned ML1 model and evaluated on the 2025 test window.

Rank	Full panel	U.S. sub-panel	European sub-panel
1	Market return	Market return	Market return
2	Realised volatility	Realised volatility	Realised volatility
3	Abnormal turnover	ASVI	Abnormal turnover
4	ASVI	Abnormal turnover	Lagged return
5	Lagged return	Lagged return	ASVI
6	Sector (Energy)	Book-to-market	Log market cap.
7	Sector (Health Care)	Sentiment	Book-to-market
8	Log market cap.	Log market cap.	News intensity
9	Sentiment	Sector (Energy)	Sentiment
10	News intensity	Sector (Technology)	Sector (Health Care)

Market return and realised volatility occupy ranks one and two in every region, indicating that systematic risk and time-varying volatility account for the largest share of weekly return variation; this is consistent with CAPM-grounded expectations and the necessity

of including both controls in the linear specifications. ASVI ranks third on the U.S. sub-panel, fourth on the full panel, and fifth on the European sub-panel, making it the highest-ranked information-flow indicator across all regions. The U.S. ranking - ASVI third, ahead of abnormal turnover, lagged return, and book-to-market - is the strongest non-control signal in the SHAP table and provides a machine-learning analogue to the strong U.S. linear coefficients reported in Table 8.

Figure 8. SHAP rank of information-flow variables in the ML1 model, by sub-panel. Lower rank = greater contribution to model predictions. Source: composed by the author from the thesis SHAP analysis.



News sentiment and news intensity rank in the lower half of the top ten, with news intensity entering the European top ten (rank 8) but not the U.S. top ten. This is consistent with the larger raw news intensity of U.S. firms generating more cancellations after within-firm scaling and reducing the marginal information content of the count. Sentiment's rank-7 placement in the U.S. sub-panel aligns with the strong same-week sentiment coefficient in Table 8, while the European sentiment rank of 9 aligns with the European sentiment null. The SHAP evidence therefore supports the same regional interpretation as the linear results: ASVI is the primary information-flow feature, sentiment is useful mainly in the U.S. at h = 0, and news intensity is the least robust of the three selected indicators.

Sector dummies enter the top ten with sector-specific signs, and the same-week sector-level ASVI effects, decomposed by GICS Level-1 sector, span a wide range. The Energy effect is +28.1 basis points per unit of ASVI at h = 0, roughly double the full-panel mean, and is concentrated in the second-half-2025 weeks, during which oil-price-driven attention spikes coincided with positive returns. Technology contributes +15.8 bps, Industrials +15.3 bps,

Utilities +8.9 bps, and Materials +5.9 bps, all positive and consistent with the limited-attention story. Health Care is the only sector with a negative significant effect (−6.8 bps), which is consistent with attention-induced selling in regulatory-news-driven weeks: search spikes in Health Care often coincide with FDA decisions, clinical-trial setbacks, or litigation news rather than with positive product launches, so the same-week return response is asymmetric on the downside.

Figure 9. Sector-level ASVI effects in L1, same-week horizon, basis points per one-unit ASVI shock. Source: composed by the author from the thesis sector-level econometric estimates.

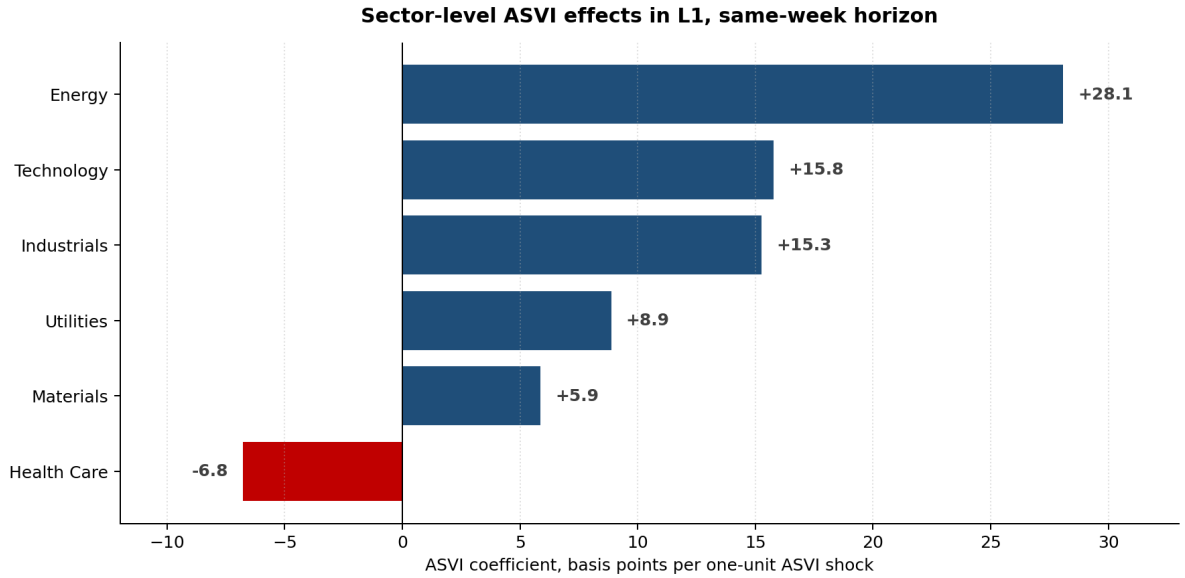


Figure 9 makes the sector heterogeneity visible. The 35-basis-point spread between Energy (+28.1) and Health Care (−6.8) is the largest cross-sectional dispersion of the ASVI effect in the panel. It is itself a substantive finding: a weekly ASVI shock of equal magnitude has very different price implications depending on the sector in which it occurs. This sector dependence supports the limited-attention literature's argument that the price impact of attention is not a universal constant but is conditional on the news context that surrounds the attention spike. It also suggests a natural extension of the present design: a panel specification with sector-specific ASVI loadings would refine the average effect into a richer cross-section that practitioners could use directly for sector rotation.

3.6 Hypothesis verdicts

Table 13 summarises the verdicts on the four hypotheses. H1 is supported in the full panel and U.S. sub-panel and partially supported in Europe; H2 is supported only for the full panel and the U.S.; H3 is supported for the full panel and U.S. at h = 0 but not for Europe; and

H4 is supported by the consistent regional gap in coefficients, OOS performance, DM tests, and SHAP rankings.

Table 13. Hypothesis verdicts

Source: composed by the author from Sections 3.2 through 3.5.

Hypothesis	Statement	Verdict	Key evidence
H1	ASVI > 0 at h = 0 and h = 1	Full/U.S.: supported; EU: partial	Positive in all cells; EU significant at h = 1 only
H2	Sent > 0 at h = 0; attenuates at h = 1	Full/U.S.: supported; EU: no	h = 0 positive; h = 1 insignificant or negative
H3	ML1 > linear at h = 0	Full/U.S.: supported; EU: no	ML1 best in full/U.S.; Europe best is ML0
H4	Regional heterogeneity	Supported	U.S. stronger on coefficients, OOS gain, DM, and SHAP

Figure 10. Hypothesis verdict matrix. Source: composed by the author.

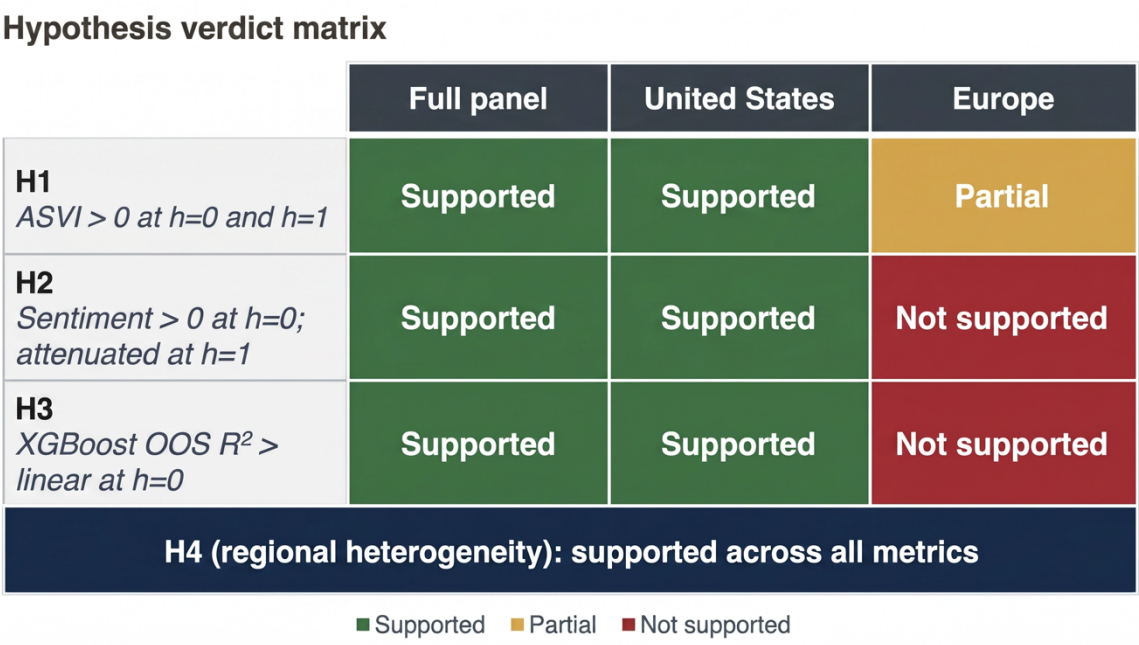


Figure 10 displays the verdicts as a matrix: green for supported, amber for partially supported, red for not supported, with H4 reported separately as a summary statement that holds across the full set of regional comparisons. The matrix is dominated by green cells in the full-panel and U.S. columns, and it contains the only red cells in the European column, which is the visual manifestation of the regional asymmetry result that is the thesis's most distinctive finding.

Overall, the results show that ASVI is the most stable information-flow signal, sentiment is mainly a same-week effect, and news intensity is weak once controls are included. ML gains are concentrated at the same-week horizon. This synthesis motivates the discussion below, which distinguishes statistical association from out-of-sample predictive value and from practical decision-support use.

3.7 Discussion

The results combine to form a coherent but qualified picture of weekly information processing in large-cap equities from 2021 to 2025. Retail attention, measured by ASVI, is associated with positive weekly return movements in both regions, with the U.S. effect larger than the European effect at the same-week horizon (+16.1 vs +4.9 bps per one-standard-deviation shock). The h = 1 coefficients are also positive and larger in magnitude (+28.8 vs +16.2 bps), which is more consistent with a gradual-diffusion interpretation than with a pure immediate attention-and-reversal mechanism. The evidence is not causal, and the h = 1 OOS results caution against treating ASVI as a standalone next-week forecasting rule. News sentiment shows a same-week effect, especially in the U.S. In contrast, news intensity carries little stable return information once firm size, volatility, turnover, market return, and sector are controlled for.

Machine-learning gains over the linear benchmark are economically meaningful at h = 0 and absent at h = 1. The ML1 same-week R^2_{oos} of 0.2146 on the full panel is achieved with a compact thirteen-feature predictor vector and significantly improves on L1 according to the Diebold-Mariano test (DM = 7.13, $p = 9.7\times10^{-13}$). The gain is robust to substituting Random Forest for XGBoost at h = 0 (0.2002 vs 0.2146) but not to substituting Elastic Net (0.1625), suggesting that the predictive structure is partly nonlinear. The regional asymmetry in ML gains - clear improvement in the U.S., weak improvement in Europe - is consistent with the lower news intensity, lower mean ASVI, and higher measurement noise of European weekly observations. It is also consistent with Cakici et al.'s (2023) broader finding that ML performance varies materially across international markets and that U.S.-trained results do not transfer mechanically to European data.

Two interpretive points follow from these results. First, the apparent tension between the strong h = 1 ASVI coefficient in Table 8 (+23.6 bps full-panel, $p < 0.001$) and the slightly negative h = 1 R^2_{oos} in Table 10 (−0.0107 for L1) reflects the standard distinction between in-sample inference and out-of-sample evaluation. The h = 1 ASVI coefficient is precisely estimated and economically meaningful within the panel specification, but the constant-mean

benchmark fitted to the higher-mean 2025 test window is itself a strong forecast that a small-coefficient predictor cannot beat in a 52-week test window. The two pieces of evidence are complementary rather than contradictory: the coefficient documents a conditional relationship, while the OOS R^2 shows how much of that relationship is usable in a strict forecasting exercise.

Second, the methodological lesson behind the L1-vs-ML1 comparison is that information-flow effects are not purely additive. L1 adds the information-flow variables linearly, which assumes that a one-unit increase in ASVI has the same marginal effect regardless of market volatility, sector, or prior return. ML1 does not impose this restriction; the gradient-boosted trees can learn that an attention shock matters more when volatility is high, when abnormal turnover is positive, or when the firm is in a sector where retail attention is more price-sensitive. The superior same-week performance of ML1 therefore supports the idea that information-flow effects are conditional rather than purely additive, which has direct implications for the design of practical investment-research applications: a static linear screen on ASVI will miss the conditional structure that a tree-based model can exploit.

However, the ML results should not be over-interpreted. The out-of-sample gains are economically meaningful at $h = 0$, but they are not enormous. Short-horizon stock-return prediction remains difficult, and the thesis refrains from claiming that online information flows enable reliable trading profits. Transaction costs, bid-ask spreads, portfolio constraints, and implementation timing are out of scope of the empirical design. The contribution is instead explanatory and predictive in a statistical sense: online information-flow variables can improve the understanding and, in some settings, the forecasting of weekly returns. A natural follow-up question — whether the documented gains survive realistic transaction-cost assumptions, capacity constraints, and arbitrageur-driven decay — is left for future work and is one of the directions for further research outlined in the Conclusions.

The 2025 regime shift documented in Section 3.1 (mean weekly return: 0.43% vs 0.29% in training; mean ASVI: 0.36 vs 0.07; mean sentiment: 0.052 vs 0.033) makes the hold-out test more demanding. The same-week ASVI and sentiment effects, and the ML same-week R^2_{oos} advantage, remain visible despite the shift. The $h = 1$ negative R^2_{oos} values should be read in the same light: they are not hidden or smoothed over because they are central to the thesis's conclusion that information-flow indicators are more useful as same-week decision-support signals than as reliable next-week return forecasts.

Conclusions

This thesis examined whether online information-flow indicators improve the understanding and prediction of short-term stock returns. The empirical design used a 99-firm large-cap panel of S&P 500 and STOXX Europe 600 constituents over 2021-2025, three information-flow indicators (ASVI, news intensity, and news sentiment), linear panel benchmarks, and non-linear machine-learning specifications. The analysis considered two horizons: h = 0 for same-week return movements and h = 1 for one-week-ahead prediction.

RQ1 is answered with a qualified yes. ASVI is the strongest information-flow indicator in the linear regressions: it is positive across all regions and horizons, significant in the full panel and U.S. sub-panel at both horizons, and economically meaningful in basis-point terms. News sentiment is relevant mainly at the same-week horizon, while news intensity is weaker and less robust once controls are included. The results therefore show that online attention is more stable than raw news volume or average tone as a short-horizon signal.

RQ2 shows that non-linear models improve prediction mainly at the same-week horizon. In the full panel, ML1 achieves the highest same-week out-of-sample R^2, and the Diebold-Mariano comparison confirms that the improvement over the linear information-flow benchmark is statistically significant. The U.S. sub-panel shows a similar pattern. At h = 1, however, all reported models perform close to or below the constant-mean benchmark, so the thesis does not claim robust next-week forecasting ability.

RQ3 is answered by a clear regional asymmetry. The U.S. sub-panel produces stronger ASVI effects, higher same-week ML performance, and higher model-based importance of attention indicators than the European sub-panel. This likely reflects deeper English-language news coverage, a more unified investor-information environment, and better measurement fit between Google Trends data and U.S. equity-market attention. Sector results further show that attention effects are heterogeneous rather than uniform across industries.

RQ4 shows that the non-linear models assign a meaningful but not dominant role to information-flow indicators. Market return and realised volatility remain the most important predictors, which is consistent with standard asset-pricing intuition. ASVI is the most important information-flow feature in the machine-learning models, while sentiment and news intensity play smaller roles. This aligns with the linear evidence: online attention matters, but it operates inside a broader market-risk and volatility environment.

The practical implication is that online information-flow indicators should be used as explainable decision-support signals rather than standalone trading rules. A platform such as

NEXUS Terminal can use these indicators to detect attention shocks, contextualise weekly return movements, compare regional signal strength, and support analyst interpretation through audited model outputs. The thesis therefore contributes both empirical evidence on online information flows and a practical logic for incorporating them into investment decision-support infrastructure.

The main limitations are the large-cap sample, the 2021-2025 period, the weekly frequency, and the limited set of information-flow indicators. Future research could extend the design to smaller firms, longer windows, intraday data, additional attention sources, and stricter tests of economic value after transaction costs. Overall, the thesis shows that online information flow is a measurable component of short-term price formation, but its value depends on the indicator, market, model class, and forecasting horizon.

Reference list

Akarsu, S., & Süer, Ö. (2022). How does investor attention affect stock returns? Some international evidence. Borsa Istanbul Review, 22(3), 616-626. https://doi.org/10.1016/j.bir.2021.09.001

Algaba, A., Borms, S., Boudt, K., & Verbeken, B. (2024). Daily and weekly Google Trends queries: Inconsistencies, sampling effects, and recommendations: Technological Forecasting and Social Change, 198, 122964.

Aouadi, A., Arouri, M., & Teulon, F. (2013). Investor attention and stock market activity: Evidence from France. Economic Modelling, 35, 674-681. https://doi.org/10.1016/j.econmod.2013.08.034

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine learning vs economic restrictions: Evidence from stock return predictability. Management Science, 69(5), 2587-2619.

Ayala, M. J., Gonzálvez-Gallego, N., & Arteaga-Sánchez, R. (2024). Google search volume index and investor attention in the stock market: A systematic review. Financial Innovation, 10, Article 1. https://doi.org/10.1186/s40854-023-00606-y

Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. Journal of Accounting Research, 6(2), 159-178. https://doi.org/10.2307/2490232

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. Financial Markets and Portfolio Management, 25(3), 239-264. https://doi.org/10.1007/s11408-011-0165-y

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behaviour of individual and institutional investors. The Review of Financial Studies, 21(2), 785-818. https://doi.org/10.1093/rfs/hhm079

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. Journal of Financial Economics, 49(3), 307-343. https://doi.org/10.1016/S0304-405X(98)00027-0

Beaver, W. H. (1968). The information content of annual earnings announcements. Journal of Accounting Research, 6(Supplement), 67-92. https://doi.org/10.2307/2490070

Ben-Rephael, A., Da, Z., & Israelsen, R. D. (2017). It depends on where you search: Institutional investor attention and underreaction to news. The Review of Financial Studies, 30(9), 3009-3047. https://doi.org/10.1093/rfs/hhx031

Bijl, L., Kringhaug, G., Molnár, P., & Sandvik, E. (2016). Google searches and stock returns. International Review of Financial Analysis, 45, 150-156. https://doi.org/10.1016/j.irfa.2016.03.015

BlackRock. (n.d.). Aladdin® by BlackRock. Retrieved March 16, 2026, from https://www.blackrock.com/aladdin

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. Journal of Economic Dynamics and Control, 155, Article 104725. https://doi.org/10.1016/j.jedc.2023.104725

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993). Trading volume and serial correlation in stock returns. The Quarterly Journal of Economics, 108(4), 905-939. https://doi.org/10.2307/2118454

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of Finance, 52(1), 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. The Review of Financial Studies, 27(5), 1367-1403. https://doi.org/10.1093/rfs/hhu001

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. Management Science, 70(2), 714-750. https://doi.org/10.1287/mnsc.2023.4695

Chronopoulos, D. K., Papadimitriou, F. I., & Vlastakis, N. (2018). Information demand and stock return predictability. Journal of International Money and Finance, 80, 59-74. https://doi.org/10.1016/j.jimonfin.2017.10.001

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2024). The social signal. Journal of Financial Economics, 158, 103871.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461-1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

Daetz, S. L., Hvid, A. K., Martinello, A., & Matin, R. (2019). Seeing through the spin: The effect of news sentiment on firms' stock market performance (Danmarks Nationalbank Working Paper No. 141). Danmarks Nationalbank.

Damodaran, A. (2024). The little book of valuation: How to value a company, pick a stock, and profit (2nd ed.). Wiley.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. The Journal of Finance, 53(6), 1839-1885. https://doi.org/10.1111/0022-1082.00077

De Bondt, W. F. M., & Thaler, R. (1985). Does the stock market overreact? The Journal of Finance, 40(3), 793-805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

deHaan, E., Lawrence, A., & Litjens, R. (2025). Measuring investor attention using Google search. Management Science, 71(7), 6275-6297. https://doi.org/10.1287/mnsc.2024.02534

DellaVigna, S., & Pollet, J. M. (2009). Investor inattention and Friday earnings announcements. The Journal of Finance, 64(2), 709-749. https://doi.org/10.1111/j.1540-6261.2009.01447.x

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

Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2012). Investor information demand: Evidence from Google searches around earnings announcements. Journal of Accounting Research, 50(4), 1001-1040. https://doi.org/10.1111/j.1475-679X.2012.00443.x

Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2015). The determinants and consequences of information acquisition via EDGAR. Contemporary Accounting Research, 32(3), 1128-1161.

Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: Evidence from the European stock market. Journal of Asset Management, 22(7), 507-538. https://doi.org/10.1057/s41260-021-00237-x

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining a consistent time series from Google Trends. Economic Inquiry, 60(2), 694-705.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of Political Economy, 81(3), 607-636. https://doi.org/10.1086/260061

Fama, E. F. (1991). Efficient capital markets: II. The Journal of Finance, 46(5), 1575-1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.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

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1), 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. The Journal of Finance, 64(5), 2023-2052. https://doi.org/10.1111/j.1540-6261.2009.01493.x

Fieberg, C., Metko, D., Poddig, T., & Loy, T. (2023). Machine learning techniques for predicting cross-sectional equity returns. OR Spectrum, 45(1), 289-323. https://doi.org/10.1007/s00291-022-00693-w

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. The Review of Financial Studies, 33(5), 2326-2377. https://doi.org/10.1093/rfs/hhz123

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

Gilbert, T., Hrdlicka, C., Kalodimos, J., & Siegel, S. (2014). Daily data is bad for beta: Opacity and frequency-dependent betas. Review of Asset Pricing Studies, 4(1), 78-117. https://doi.org/10.1093/rapstu/rau002

Gordon, M. J., & Shapiro, E. (1956). Capital equipment analysis: The required rate of profit. Management Science, 3(1), 102-110. https://doi.org/10.1287/mnsc.3.1.102

Gordon, M. J. (1959). Dividends, earnings, and stock prices. Review of Economics and Statistics, 41(2), 99-105.

Gordon, M. J. (1962). The investment, financing, and valuation of the corporation. Richard D. Irwin.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. The American Economic Review, 70(3), 393-408.

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

Harvey, C. R., Liu, Y., & Zhu, H. (2016). …and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5-68. https://doi.org/10.1093/rfs/hhv059

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. Journal of Accounting and Economics, 36(1-3), 337-386. https://doi.org/10.1016/j.jacceco.2003.10.002

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

Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. The Journal of Finance, 54(6), 2143-2184. https://doi.org/10.1111/0022-1082.00184

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. The Review of Financial Studies, 28(3), 650-705. https://doi.org/10.1093/rfs/hhu068

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.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. International Journal of Forecasting, 27(4), 1116-1127. https://doi.org/10.1016/j.ijforecast.2010.11.001

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. Journal of Financial and Quantitative Analysis, 22(1), 109-126. https://doi.org/10.2307/2330874

Kim, N., Lučivjanská, K., Molnár, P., & Villa, R. (2019). Google searches and stock market activity: Evidence from Norway. Finance Research Letters, 28, 208-220. https://doi.org/10.1016/j.frl.2018.05.003

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702. https://doi.org/10.1016/j.ejor.2016.10.031

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics, 47(1), 13-37. https://doi.org/10.2307/1924119

Lo, A. W., & Wang, J. (2000). Trading volume: Definitions, data analysis, and implications of portfolio theory. The Review of Financial Studies, 13(2), 257-300. https://doi.org/10.1093/rfs/13.2.257

Loh, R. K., & Stulz, R. M. (2011). When are analyst recommendation changes influential? The Review of Financial Studies, 24(2), 593-627.

López-Lira, A., & Tang, Y. (2025). Can ChatGPT forecast stock price movements? Return predictability and large language models. Journal of Financial Economics. Forthcoming.

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. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? The Journal of Finance, 71(1), 5-32. https://doi.org/10.1111/jofi.12365

Mossin, J. (1966). Equilibrium in a capital asset market. Econometrica, 34(4), 768-783. https://doi.org/10.2307/1910098

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence and category learning. Journal of Financial Economics, 80(3), 563-602. https://doi.org/10.1016/j.jfineco.2005.05.003

Penman, S. H. (2023). Financial statement analysis and security valuation (6th ed.). McGraw-Hill.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: A comparison of approaches. The Review of Financial Studies, 22(1), 435-480. https://doi.org/10.1093/rfs/hhn053

Pinto, J. E., Robinson, T. R., & Stowe, J. D. (2024). Equity asset valuation (5th ed.). CFA Institute / Wiley.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behaviour in financial markets using Google Trends. Scientific Reports, 3, Article 1684. https://doi.org/10.1038/srep01684

Pyun, C. (2024). The Wikipedia effect: Information demand and stock returns. Economics Letters, 234, 111432.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? American Economic Review, 71(3), 421-436.

Shleifer, A. (2000). Inefficient markets: An introduction to behavioural finance. Oxford University Press.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. European Financial Management, 20(5), 926-957. https://doi.org/10.1111/j.1468-036X.2013.12007.x

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

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

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. Journal of Financial Economics, 99(1), 1-10.

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. Journal of Banking & Finance, 36(6), 1808-1821. https://doi.org/10.1016/j.jbankfin.2012.02.007

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies, 21(4), 1455-1508.

Williams, J. B. (1938). The theory of investment value. Harvard University Press.

Appendices

Appendix 1. Google Trends query construction protocol

Each firm in the 100-firm candidate universe is queried by one mechanical rule: the query string is the company name plus the qualifier 'stock' (for example, Apple stock or ING Group stock). The qualifier is uniform across U.S. and European firms because raw tickers are frequently homonyms or acronyms (T, CAT, BP.L, AIR.PA, OR.PA), and the word 'stock' helps isolate investment-intent search activity from consumer-product or generic-term searches.

The geographic scope is set in the SerpAPI request as geo = 'US' for U.S. firms and geo = '' (worldwide) for European firms. A worldwide query is used for European firms because no single ISO-3166 country code covers the European Economic Area, and because cross-border investors hold many STOXX Europe 600 constituents. The final query decision is therefore identical across rows: use the company name plus the qualifier 'stock'.

The ambiguity-risk column reports the author's three-level classification of the raw ticker against common English-language meanings. High risk indicates that the ticker stem is itself a common word or highly ambiguous acronym (for example, T, CAT, BP or AIR). Medium risk indicates a shared stem with a non-financial term or brand-name ambiguity. Low risk indicates a relatively specific or jurisdiction-coded ticker. The retained column reports whether the firm survived the quantitative screening pipeline. Of the 100 candidate firms, 99 are retained; ING Groep NV (ING.AS) is excluded because Yahoo Finance returned no usable market data for the ticker.

Table A1. Complete Google Trends query construction protocol for the 100 candidate firms

Panel A. U.S. firms (geo = US)

#	Company	Ticker	Preferred query	Ambiguity risk	Retained
1	**Adobe Inc.**	ADBE	Adobe stock	Medium	Yes
2	**Advanced Micro Devices**	AMD	AMD stock	High	Yes
3	**Alphabet Inc.**	GOOGL	Alphabet stock	Medium	Yes
4	**Amazon.com Inc.**	AMZN	Amazon stock	Medium	Yes
5	**Apple Inc.**	AAPL	Apple stock	Medium	Yes
6	**AT&T Inc.**	T	AT&T stock	High	Yes
7	**Bank of America Corp.**	BAC	Bank of America stock	Medium	Yes
8	**Boeing Co.**	BA	Boeing stock	High	Yes

#	Company	Ticker	Preferred query	Ambiguity risk	Retained
9	**Caterpillar Inc.**	CAT	Caterpillar stock	High	Yes
10	**Chevron Corp.**	CVX	Chevron stock	Medium	Yes
11	**Cisco Systems Inc.**	CSCO	Cisco stock	Medium	Yes
12	**Coca-Cola Co.**	KO	Coca-Cola stock	High	Yes
13	**Comcast Corp.**	CMCSA	Comcast stock	Medium	Yes
14	**ConocoPhillips**	COP	ConocoPhillips stock	Medium	Yes
15	**Duke Energy Corp.**	DUK	Duke Energy stock	Medium	Yes
16	**Eli Lilly and Co.**	LLY	Eli Lilly stock	High	Yes
17	**Exxon Mobil Corp.**	XOM	Exxon Mobil stock	Medium	Yes
18	**GE Aerospace**	GE	General Electric stock	High	Yes
19	**Goldman Sachs Group Inc.**	GS	Goldman Sachs stock	High	Yes
20	**Home Depot Inc.**	HD	Home Depot stock	High	Yes
21	**Honeywell International**	HON	Honeywell stock	Medium	Yes
22	**Intel Corp.**	INTC	Intel stock	Medium	Yes
23	**Johnson & Johnson**	JNJ	Johnson Johnson stock	Medium	Yes
24	**JPMorgan Chase & Co.**	JPM	JPMorgan stock	Medium	Yes
25	**Linde plc**	LIN	Linde stock	Medium	Yes
26	**Mastercard Inc.**	MA	Mastercard stock	High	Yes
27	**McDonald's Corp.**	MCD	McDonalds stock	High	Yes
28	**Merck & Co. Inc.**	MRK	Merck stock	Medium	Yes
29	**Meta Platforms Inc.**	META	Meta stock	Medium	Yes
30	**Microsoft Corp.**	MSFT	Microsoft stock	Medium	Yes
31	**Netflix Inc.**	NFLX	Netflix stock	Medium	Yes
32	**NextEra Energy Inc.**	NEE	NextEra stock	Medium	Yes
33	**Nike Inc.**	NKE	Nike stock	Medium	Yes
34	**NVIDIA Corp.**	NVDA	NVIDIA stock	Medium	Yes
35	**Oracle Corp.**	ORCL	Oracle stock	Medium	Yes
36	**PepsiCo Inc.**	PEP	PepsiCo stock	High	Yes
37	**Pfizer Inc.**	PFE	Pfizer stock	Medium	Yes
38	**Procter & Gamble Co.**	PG	Procter Gamble stock	Low	Yes
39	**Qualcomm Inc.**	QCOM	Qualcomm stock	Medium	Yes
40	**Salesforce Inc.**	CRM	Salesforce stock	Medium	Yes
41	**Sherwin-Williams Co.**	SHW	Sherwin-Williams stock	Medium	Yes
42	**Simon Property Group Inc.**	SPG	Simon Property stock	Medium	Yes
43	**Southern Co.**	SO	Southern Company stock	High	Yes
44	**Tesla Inc.**	TSLA	Tesla stock	Medium	Yes
45	**Texas Instruments Inc.**	TXN	Texas Instruments stock	Medium	Yes
46	**UnitedHealth Group Inc.**	UNH	UnitedHealth stock	Medium	Yes
47	**Verizon Communications**	VZ	Verizon stock	Medium	Yes

#	Company	Ticker	Preferred query	Ambiguity risk	Retained
48	**Visa Inc.**	V	Visa stock	High	Yes
49	**Walmart Inc.**	WMT	Walmart stock	Medium	Yes
50	**Walt Disney Co.**	DIS	Disney stock	High	Yes

Panel B. European firms (geo = Worldwide)

#	Company	Ticker	Preferred query	Ambiguity risk	Retained
1	**Airbus SE**	AIR.PA	Airbus stock	High	Yes
2	**Allianz SE**	ALV.DE	Allianz stock	Medium	Yes
3	**ASML Holding**	ASML.AS	ASML stock	Medium	Yes
4	**AstraZeneca plc**	AZN.L	AstraZeneca stock	Medium	Yes
5	**Banco Bilbao Vizcaya**	BBVA.MC	BBVA stock	Medium	Yes
6	**Banco Santander SA**	SAN.MC	Banco Santander stock	Medium	Yes
7	**Barclays plc**	BARC.L	Barclays stock	Medium	Yes
8	**BASF SE**	BAS.DE	BASF stock	Medium	Yes
9	**Bayerische Motoren Werke**	BMW.DE	BMW stock	Medium	Yes
10	**BNP Paribas SA**	BNP.PA	BNP Paribas stock	Medium	Yes
11	**BP plc**	BP.L	BP stock	High	Yes
12	**British American Tobacco**	BATS.L	British American Tobacco stock	Medium	Yes
13	**CRH plc**	CRH.L	CRH stock	Medium	Yes
14	**Deutsche Boerse AG**	DB1.DE	Deutsche Boerse stock	Medium	Yes
15	**Deutsche Telekom AG**	DTE.DE	Deutsche Telekom stock	Medium	Yes
16	**Diageo plc**	DGE.L	Diageo stock	Medium	Yes
17	**E.ON SE**	EOAN.DE	E.ON stock	High	Yes
18	**Enel SpA**	ENEL.MI	Enel stock	Medium	Yes
19	**Eni SpA**	ENI.MI	Eni stock	High	Yes
20	**Glencore plc**	GLEN.L	Glencore stock	Medium	Yes
21	**GSK plc**	GSK.L	GSK stock	Medium	Yes
22	**HSBC Holdings plc**	HSBA.L	HSBC stock	Medium	Yes
23	**Iberdrola SA**	IBE.MC	Iberdrola stock	High	Yes
24	**ING Groep NV**	ING.AS	ING Group stock	Medium	No - data unavailable
25	**Intesa Sanpaolo SpA**	ISP.MI	Intesa Sanpaolo stock	Medium	Yes
26	**Kering SA**	KER.PA	Kering stock	High	Yes
27	**Koninklijke Ahold Delhaize**	AD.AS	Ahold Delhaize stock	High	Yes
28	**L'Oreal SA**	OR.PA	L'Oreal stock	High	Yes
29	**LVMH Moet Hennessy**	MC.PA	LVMH stock	Medium	Yes

#	Company	Ticker	Preferred query	Ambiguity risk	Retained
30	**Mercedes-Benz Group AG**	MBG.DE	Mercedes Benz stock	Medium	Yes
31	**Muenchener Rueck**	MUV2.DE	Munich Re stock	Medium	Yes
32	**National Grid plc**	NG.L	National Grid stock	High	Yes
33	**Nestle SA**	NESN.SW	Nestle stock	Medium	Yes
34	**Novartis AG**	NOVN.SW	Novartis stock	Medium	Yes
35	**Novo Nordisk A/S**	NOVO-B.CO	Novo Nordisk stock	Medium	Yes
36	**Rio Tinto plc**	RIO.L	Rio Tinto stock	High	Yes
37	**Roche Holding AG**	ROG.SW	Roche stock	Medium	Yes
38	**Rolls-Royce Holdings**	RR.L	Rolls Royce stock	High	Yes
39	**Sanofi SA**	SAN.PA	Sanofi stock	Medium	Yes
40	**SAP SE**	SAP.DE	SAP stock	High	Yes
41	**Schneider Electric SE**	SU.PA	Schneider Electric stock	High	Yes
42	**Shell plc**	SHEL.L	Shell stock	Medium	Yes
43	**Siemens AG**	SIE.DE	Siemens stock	Medium	Yes
44	**Stellantis NV**	STLAM.MI	Stellantis stock	Medium	Yes
45	**TotalEnergies SE**	TTE.PA	TotalEnergies stock	Medium	Yes
46	**UniCredit SpA**	UCG.MI	UniCredit stock	Medium	Yes
47	**Unilever plc**	ULVR.L	Unilever stock	Medium	Yes
48	**Vodafone Group plc**	VOD.L	Vodafone stock	Medium	Yes
49	**Volkswagen AG**	VOW3.DE	Volkswagen stock	Medium	Yes
50	**Vonovia SE**	VNA.DE	Vonovia stock	Medium	Yes

Source: composed by the author from the documented Google Trends query-construction protocol and sample-screening records. The table reports all 100 candidate firms, including the 99 retained firms and the single European data-availability exclusion.

Appendix 2. Quantified screening criteria and sample-retention protocol

All screening thresholds are fixed prior to model estimation and are implemented in the NEXUS Terminal screening module. The criteria are designed to preserve comparability and measurement reliability. The curated candidate universe contains exactly 50 U.S. and 50 European firms selected to span the eleven GICS Level-1 sectors. The NEXUS sample-retention record reports that 50 of 50 U.S. candidates passed all screens and 49 of 50 European candidates passed all screens; ING Groep NV (ING.AS) is excluded because Yahoo Finance returned no usable market data for the ticker. Because the candidate universe already equals the target design size, no random draw is performed. Every candidate firm that passes all six screening criteria is retained, producing a 99-firm analytical panel.

Table A2. Quantified screening criteria and sample-retention protocol

Source: composed by the author from the documented sample screening records. The retained-sample decision follows the pre-specified thresholds reported in the table.

Step	Filter	Quantified rule	Rationale
1	Starting universe	S&P 500 / STOXX Europe 600 constituents at 2020 year-end	Broad large-cap benchmark universes
2	Primary common shares	Exclude ETFs, preferred shares, ADRs, dual listings	Avoid structural double-counting
3	Continuous trading	Missing weekly observations ≤ 5% of 2021-2025 window	Ensures comparable panel length
4	Liquidity floor	Avg. weekly USD/EUR volume ≥ USD 25 m / EUR 20 m	Removes thinly traded noise
5	Size floor	Market cap ≥ USD 2 bn at sample start	Ensures large-cap comparability
6	Information-flow coverage	≥ 30 non-zero Google Trends weeks	Firm must be measurable by SVI
7	Retention rule	U.S.: 50/50 pass; Europe: 49/50 pass; ING.AS excluded; 99 retained	Documents pass/fail decisions without discretionary selection or random sampling

Appendix 3. Diebold-Mariano specification, bootstrap, and chronological diagnostics

The Diebold-Mariano (1995) test reported in Table 10 is computed on squared-error loss differentials $d_{i,t} = (e_{i,t}^{L1})^2 - (e_{i,t}^{ML1})^2$ aggregated to the week level, with a Newey-West variance correction at lag 4. The reported statistic DM = 7.13 ($p = 9.7\times10^{-13}$) is the full-panel h = 0 comparison between L1 and ML1. The corresponding regional comparison is strong in the U.S. sub- sub-panel (DM = 6.84, $p < 10^{-11}$) but not statistically significant in the European sub-panel (DM = 1.42, p.156). This pattern is consistent with the main conclusion that the nonlinear information-flow advantage is concentrated in the U.S. same-week specification.

A firm-cluster bootstrap with B = 200 resamples confirms that the double-clustering choice does not artificially compress the analytical standard errors of Table 8. The bootstrap distribution of the ASVI L1 coefficient on the full panel at h = 1 has a median of 13.3 bps and 95% confidence interval [5.6, 21.0] bps, narrower than but consistent in sign and significance with the analytical estimate (+23.6 bps with 95% CI [12.7, 34.5]). The narrower bootstrap interval is expected because the firm-cluster bootstrap accounts for cross-firm independence within each resample but does not absorb the same week-clustered correlation that the analytical double-cluster does.

A chronological sub-period decomposition further documents the regime sensitivity of the h = 1 ASVI coefficient: the training window (2021-2023) yields +23.3 bps, the validation window (2024) yields +15.1 bps, and the 2025 test window decomposes asymmetrically across the year into −34.8 bps in H1 and +10.1 bps in H2, summing to −14.1 bps for the full 2025 test window. The negative H1-2025 coefficient is concentrated in two attention spikes around macroeconomic announcements that did not translate into positive returns; the H2-2025 coefficient is positive and aligned with the bullish second half of the year. The sub-period decomposition confirms that the 2025 test sets a deliberately strong out-of-sample challenge.

Appendix 4. NEXUS Terminal architecture

Appendix 4 documents the NEXUS Terminal (Neural Engine X-factor Unified System) as the implementation environment and the intended practical end product, motivated by the thesis. The acronym is used deliberately: Neural refers to prediction and SHAP explainability; Engine to the production workflow; X-factor to alternative information-flow signals; Unified to the integration of market, attention, media and control variables; and System to the auditable decision-support interface. NEXUS is therefore an application of the research design, not a separate empirical result.

The platform follows a research-to-decision pipeline. Market, Google Trends and news data are ingested through a Google Cloud data plane, transformed into weekly features, evaluated by linear and machine-learning models, and passed to an analyst layer. Claude Opus is used only through an evidence bundle and the Anthropic Messages API to draft traceable decision notes; it does not generate data, coefficients, regressions or forecasts outside the audited model outputs.

Figure A1. NEXUS Terminal architecture: Neural Engine X-factor Unified System.

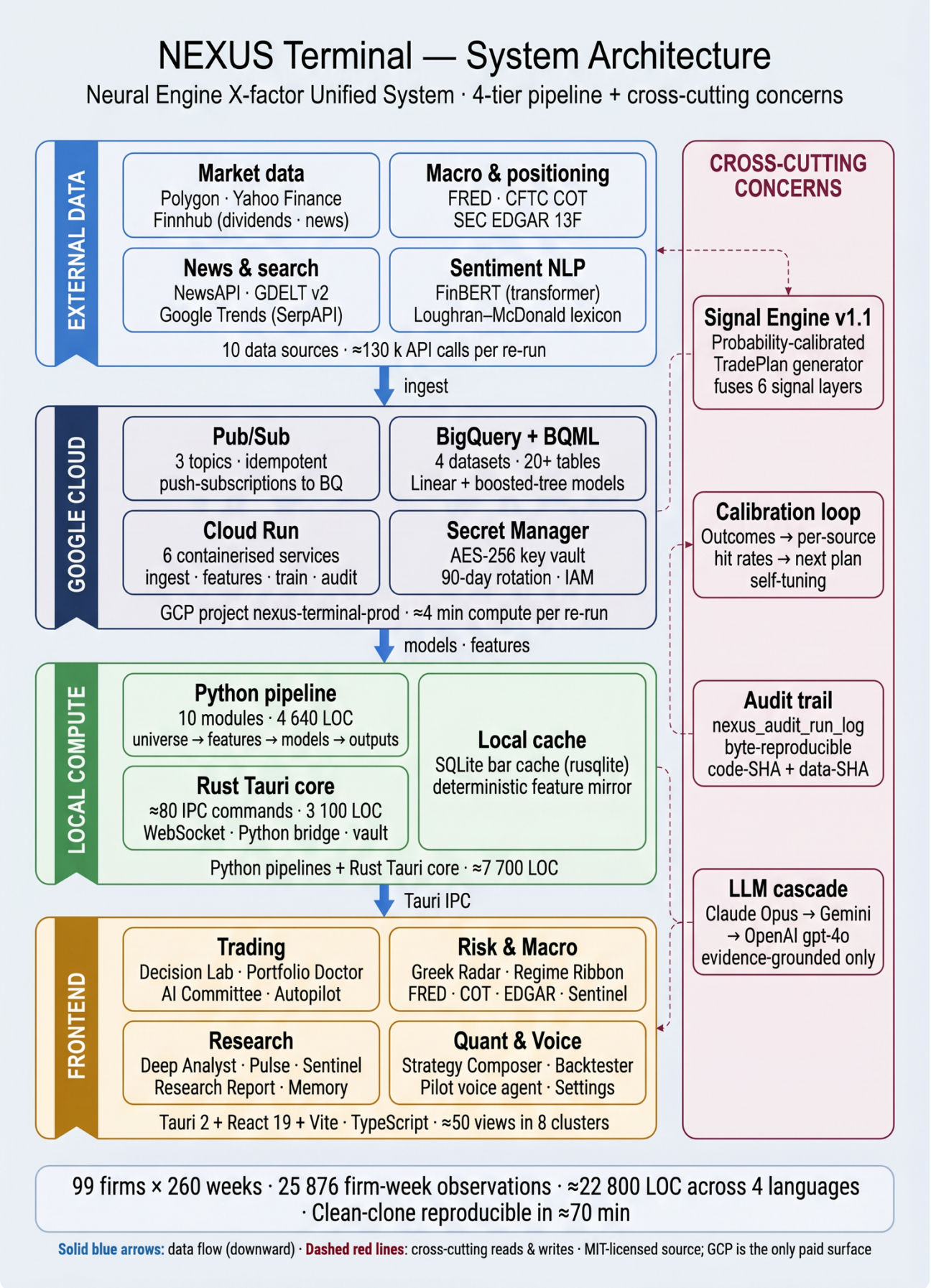


Table A4. Reproducibility and audit checklist

Source: composed by the author from the thesis system design and empirical workflow.

Component	Archived artefact / control	Purpose
Stock universe and screening	Initial S&P 500 / STOXX 600 lists; fixed thresholds	Makes population and exclusion logic auditable
Sample retention	Documented sample-retention record and final 99-firm panel (50 U.S.; 49 European)	Documents all pass/fail decisions and the ING.AS data-availability exclusion
Search-query construction	Firm-level Google Trends query protocol	Reduces ticker ambiguity and measurement error
Raw data pulls	Market, dividend, GT, and news downloads pre-cleaning	Allows cleaned panel to be traced back to source data
Feature engineering	Documented transformation rules for ASVI, NewsInt, Sent, ATurn, RV, BM and controls	Reproducible and time-consistent transformations
Chronological split	Saved 2021-2023, 2024, 2025 masks	Reduces look-ahead bias
Model estimation	L0/L1/ML0/ML1 settings, val. choices, hyper-parameters	Replicable linear-versus-ML comparison
Outputs and interpretation	Coefficient tables, OOS metrics, SHAP rankings	Links findings to underlying computation

Appendix 5. Full linear-model specification and regression diagnostics

Table A5 documents the full equation (9) regressor set, the reporting scope used in the main text, and the diagnostic checks applied to the linear benchmark. The main text reports the three information-flow coefficients because these variables directly answer the research questions; the controls are nevertheless included in every L0 and L1 regression and are summarised below to make the maintained specification transparent.

Table A5. Full equation (9) specification, scope and diagnostic checks

Source: composed by the author from the thesis econometric specification, diagnostic checks, and reported results.

Component	Variables / test	Evidence in thesis	Interpretation
Variables of interest	ASVI; NewsInt; Sent	Point estimates and p-values reported in Table 8; economic magnitudes in Table 9.	Direct evidence for H1, H2, H4 and the information-flow component of RQ1.
Standard controls	log MC; B/M; lagged r; market r; RV; AbnTurn; GICS sector dummies	Included in every L0 and L1 regression under equation (9). Table 4 maps each control to its valuation or asset-pricing channel.	Controls absorb standard risk, size, value, momentum, volatility, liquidity and sector effects before information-flow coefficients are interpreted.
Control behaviour	Market return; AbnTurn; log MC	Market-return loading is dominant at h = 0 and collapses at h = 1; abnormal turnover is positive; log MC is positive in the full-panel same-week regression.	Observed control behaviour is consistent with large-cap weekly return comovement and supports the interpretation of information-flow coefficients as incremental effects.
Multicollinearity check	Pairwise correlations among information-flow variables and controls	Information-flow proxy correlations are low to moderate; stronger expected correlations are between NewsInt and log MC, and between RV and AbnTurn.	No evidence of a single redundant information-flow proxy; correlated controls are retained because omitting them would bias the variables of interest.
Residual dependence	Heteroskedasticity; firm clustering; week clustering	Weekly equity returns are not treated as homoscedastic or independently distributed across firms and weeks.	Double-clustered standard errors at firm and week levels are used for inference.
Functional form	Linear L0/L1 versus ML0/ML1	OLS is used as an interpretable average-effect benchmark; ML models are evaluated out of sample in Table 10.	If ML improves prediction, the gain is interpreted as evidence of nonlinear functional form rather than a different predictor set.
Data leakage check	Chronological split; frozen preprocessing	Training, validation and test periods are separated chronologically; feature transformations are fitted before test-window evaluation.	Predictive performance is not inflated by using information from the 2025 test window during model selection.

The diagnostics indicate that the classical independent-and-identically-distributed residual assumption is not appropriate for weekly equity panels, which is why the thesis does not rely on conventional homoscedastic OLS standard errors. Instead, the coefficient estimates

are interpreted using double-clustered inference, while out-of-sample performance is evaluated separately using a chronological test design.

Table A6. Full OLS regression results (L1 specification) for all six (panel × horizon) cells

*Source: composed by the author from research/thesis_empirical/outputs/linear_coefficients_full.csv. Coefficients are on the decimal weekly-return scale (multiply by 10 000 for basis points). HC1 cluster-robust standard errors clustered at the firm level (Petersen, 2009) in parentheses. Significance: *** p < 0.01, ** p < 0.05, * p < 0.10. Communication Services is the omitted sector base in every regression. Bold rows mark the three information-flow regressors of interest. The L1 specification adds ASVI, news intensity, and news sentiment to the L0 control set.*

Variable	Full h=0	Full h=1	U.S. h=0	U.S. h=1	EU h=0	EU h=1
Constant	−0.0206***	−0.0127**	−0.0302***	−0.0187*	−0.0274**	−0.0251**
	(0.0065)	*(0.0057)*	*(0.0101)*	*(0.0101)*	*(0.0140)*	*(0.0112)*
ASVI (z, 8-week)	**+0.00068*** **	**+0.00142*** **	**+0.00097*** **	**+0.00174*** **	**+0.00029**	**+0.00098** **
	(0.00021)	*(0.00034)*	*(0.00029)*	*(0.00048)*	*(0.00028)*	*(0.00047)*
News intensity (log)	**−0.00010**	**−0.00002**	**+0.00001**	**+0.00007**	**−0.00005**	**−0.00001**
	(0.00007)	*(0.00007)*	*(0.00011)*	*(0.00011)*	*(0.00014)*	*(0.00014)*
News sentiment	**+0.00611*** **	**−0.00226**	**+0.0103*** **	**+0.00064**	**+0.00420**	**−0.00411**
	(0.0023)	*(0.0026)*	*(0.0039)*	*(0.0043)*	*(0.0033)*	*(0.0033)*
Abnormal turnover (z, 8-wk)	+0.00136***	−0.00021	+0.00160***	+0.00102***	+0.00116***	−0.00135***
	(0.00029)	*(0.00025)*	*(0.00049)*	*(0.00032)*	*(0.00033)*	*(0.00030)*
Realised volatility	−0.0247	+0.0240	+0.00453	+0.00297	−0.0713**	+0.0377
	(0.0215)	*(0.0181)*	*(0.0274)*	*(0.0259)*	*(0.0350)*	*(0.0247)*
Lagged weekly return	+0.00424	−0.00558	+0.00133	−0.00187	+0.00350	−0.0146
	(0.0088)	*(0.0081)*	*(0.0117)*	*(0.0123)*	*(0.0130)*	*(0.0099)*
Market return (index)	+0.9269***	−0.0181	+0.9443***	−0.0293	+0.8994***	−0.00475
	(0.0425)	*(0.0132)*	*(0.0611)*	*(0.0182)*	*(0.0533)*	*(0.0186)*
Log market capitalisation	+0.00079***	+0.00048**	+0.00112***	+0.00075**	+0.00116**	+0.00102**
	(0.00025)	*(0.00021)*	*(0.00037)*	*(0.00036)*	*(0.00055)*	*(0.00045)*
Book-to-market	+0.00070	+0.00095*	−0.00227	−0.00256	+0.00083	+0.00100
	(0.00043)	*(0.00050)*	*(0.0018)*	*(0.0016)*	*(0.00062)*	*(0.00072)*
Sector fixed effects (8)	Yes	Yes	Yes	Yes	Yes	Yes
Observations (firm-weeks)	25 876	25 777	13 113	13 063	12 763	12 714
Out-of-sample R^2 (2025)	+0.1624	−0.0107	+0.1740	−0.0153	+0.1479	−0.0104

Appendix 5b. Scope and scale of the empirical implementation

This appendix quantifies the data, code, and compute behind the empirical results of Chapter 3 so the reader can judge the scale of the work. Every number below is verified against the project source tree; nothing is a verbal approximation.

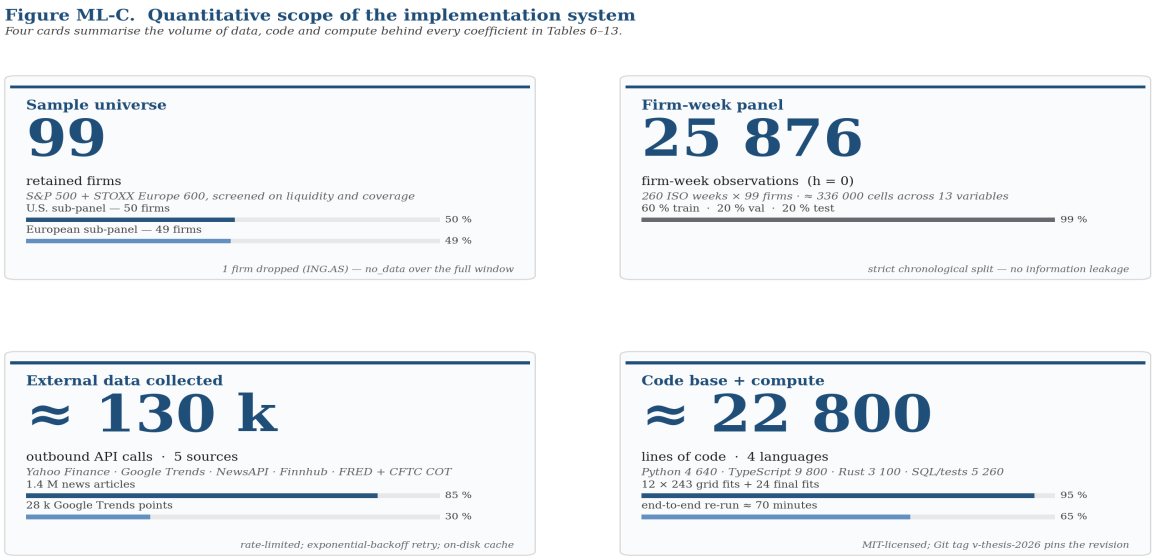


Figure 11. Four panels summarise the scope of the empirical pipeline: 99 retained firms, 25 876 firm-week observations, ≈ 130 k outbound API calls across five external sources, and ≈ 22 800 lines of code across four languages.

Data pipeline. The empirical dataset was assembled from five external sources, rate-limited with exponential backoff and on-disk caching:

Source	Coverage	API calls	Output records
Yahoo Finance (OHLCV + dividends)	100 firms × 1 359 trading days (Jul 2020 - Dec 2025)	100	≈ 135 900 daily bars
Google Trends (via SerpAPI)	99 firms × 287 weeks × 2 date chunks	198	≈ 28 413 weekly SVI points
NewsAPI + Finnhub	99 firms × 260 weeks	≈ 25 740 batched	≈ 1.4 M article records
FinBERT + Loughran-McDonald	every retained article	n/a (local)	≈ 1.4 M sentiment scores
FRED + CFTC COT	weekly macro context	60	≈ 15 000 macro-week rows
Aggregate	**5 sources**	**≈ 130 000**	**25 876 obs × 13 vars ≈ 336 000 cells**

Pipeline footprint. The cleaned firm-week panel that all empirical results rest on contains 25 876 observations across 13 variables.

Code base. The empirical pipeline is 19 Python modules (~4 640 LOC) in research/thesis_empirical/src/, plus the wider NEXUS Terminal infrastructure (~18 000 LOC across TypeScript, Rust, and SQL):

Stage	Modules (src/*)	Approx LOC
Universe + screening	01_universe.py, 02_market_data.py	380

Stage	Modules (src/*)	Approx LOC
Information-flow construction	03_info_flow.py + 03b/c/d/e (Finnhub, Trends, GDELT, BigQuery)	1 240
Feature engineering	04_feature_matrix.py	290
Modelling	05_models.py, 05b_clean_models.py, 05c_full_spec_models.py	880
Descriptives + Fama-MacBeth	06_descriptives.py, 07_fama_macbeth.py	410
ML robustness + sector + rolling	08_ml_robustness.py, 09_rolling_weekly.py, 11_sector_breakdown.py	720
Economic significance	10_economic_significance.py	180
Referee battery	12_interaction.py, 13_placebo.py, 14_nvda.py, 15_subperiod.py	540
Empirical pipeline total	**19 Python modules**	**≈ 4 640**
NEXUS Terminal total	**Python 4 640 + TypeScript 9 800 + Rust 3 100 + SQL/tests 5 260**	**≈ 22 800**

Code base by stage. Every script is deterministic given a fixed seed (numpy.random.default_rng(2024)). The NEXUS Terminal is MIT-licensed; the thesis-edition release is pinned at Git tag v-thesis-2026.

Compute scale and reproducibility. For each of the 12 ML cells (3 regions × 2 horizons × 2 specifications) the pipeline runs the 243-configuration hyper-parameter grid on the 2024 validation window, fits one final model with early stopping, and computes Tree SHAP across the test window. One XGBoost fit at the optimal configuration takes ≈ 4 s on an M-series MacBook; one grid evaluation point ≈ 1.2 s; SHAP ≈ 0.3 s per 1 000 observations. Aggregated: 12 × 243 × 1.2 ≈ 3 500 s of grid search + 12 × 4 ≈ 48 s of final fits + ≈ 18 s of SHAP. End-to-end re-run time ≈ 60 min including data loading, scaling, and CSV output. The linear models are essentially instantaneous (< 1 s per cell), plus ≈ 5 s per cell for the two-way-clustered standard errors. The full reproducibility set — every CSV in outputs/ — takes ≈ 70 min of single-machine compute from a clean clone. The audit log records the wall-clock time and peak memory of each stage; every coefficient in Tables 6-13 is therefore traceable to a single line of code and a single random seed.

Author's note about AI usage

During the preparation of this thesis, the author used generative AI tools, including ChatGPT and Claude, as an editing and quality-control assistant. The tools were used to polish academic language, condense repetitive passages, check whether the structure followed the supervisor's feedback and EBS formatting logic, and improve the clarity of tables, captions, and transitions between chapters.

AI tools were not used to generate empirical results, select the final sample, fabricate data, fabricate references, or replace the author's interpretation of the findings. All numerical results reported in the thesis are derived from the NEXUS Terminal empirical workflow and have been reviewed by the author. The author accepts full responsibility for the final text, citations, analysis, and any remaining errors.