The Value of Information from Sell-side Analysts*

Linying Lv

July 2024

Abstract

I investigate the value of information from sell-side analysts through textual analysis of a large corpus of their written reports. To quantify the dollar value of analyst information to a strategic investor, I leverage an imperfect competition insider trading model from Back et al. (2000). The aggregate annualized expected profit from receiving tips regarding the contents of an average S&P 100 constituent stock's forthcoming analyst reports is approximately $6.89 million. Using embeddings from state-of-the-art large language models, I demonstrate that textual information in analyst reports explains 10.19% of contemporaneous stock returns out-of-sample, a value that is both statistically and economically more significant than quantifiable analyst forecast revisions and traditional NLP approaches. A Shapley value decomposition is then performed to determine how different topics contribute to moving the market. The results show that income statement analyses from analysts account for more than half of the value of their reports.

JEL Classification: G11, G14, G24

Keywords: Sell-side Analysts, Value of Information, Large Language Model, Explainable AI

*Linying Lv is at Zhejiang University (linyinglv@zju.edu.cn). I thank Asaf Manela, Songrun He, Xiumin Martin, Zachary Kaplan, and Guofu Zhou at Washington University in St. Louis for helpful comments.

1 Introduction

In the United States, top-tier investment banks allocate more than $100 million each year to equity research (Groysberg et al., 2011). Meanwhile, brokerage firms provide early access to analysts' reports to selected clients in exchange for commission payments (Irvine et al., 2007). This raises crucial questions: Do the analysts generate value for clients? Do their insights enhance the informational efficiency of financial markets, or are some merely peddling expensive noise? As highlighted in recent review articles, e.g., Kothari et al. (2016) and Bradshaw et al. (2017), these questions form the core focus of this paper.

This study focuses on quantifying the dollar value of analyst outputs for insider traders and scrutinizes the heterogeneity in analyst information content. It builds on the theoretical framework developed by Back et al. (2000), which extends Kyle (1985)'s continuous-time model by incorporating imperfect competition of strategic investors. This multi-client model provides a realistic foundation for analyzing the complex dynamics of information dissemination in analyst reports. Leveraging this framework, I propose an estimation method for the ex-ante dollar expected profit of a client receiving insider information from analyst reports, calculated as the ratio of explained return volatility to price impact (Kyle's lambda). Empirically, this estimation can be interpreted as the total value strategic investors are willing to pay for analyst tips on a particular stock.

I quantify the value of information in analyst reports in three general steps: analyst outputs representation, econometric modeling, and value decomposition. To measure analyst output, the earlier research mainly focused on the numerical measures from I/B/E/S, which contains three major quantifiable outputs from analysts: stock recommendations, earnings forecasts, and target prices.[1] However, according to the annual survey of 'Institutional Investor' magazine, investors consistently rank 'Written Reports' as more valuable than summary quantifiable measures. Due to the unstructured nature of language, research on the value of written reports to the stock market

[1] Barber et al. (2001) construct portfolios based on the consensus buy/sell recommendations of security analysts and find that trading on analyst recommendations generates abnormal returns. Brav and Lehavy (2003) provide evidence that target prices contain valuable short-term and long-term information.

remains limited. Asquith et al. (2005) and Huang et al. (2014) provide early evidence on the value of analyst report content in explaining contemporaneous stock prices through sentiment analysis and 'bag-of-words' representation of texts.

The advent of large language models (LLMs) now enables far more accurate quantification of textual meaning, capturing more granular and rich information. This advancement facilitates a systematic analysis of a comprehensive set of outputs from sell-side analysts and their value to the market. Leveraging this technology, I extract contextualized representations of analyst report text using state-of-the-art LLMs such as Meta's LLaMA and OpenAI's text embedding models. These advanced linguistic tools are capable of capturing both contextual information and reasoning logic from text (Chen et al., 2022; Li et al., 2023). Each report text is thus mapped into a structured embedding space using LLMs' transformer architecture.

The theoretical framework requires understanding the implications of analyst information for stock returns. Specifically, I need to project stock returns onto analyst output representations. In the second step, following Gu et al. (2020), I employ a wide array of machine learning (ML) algorithms—Ridge regression, Partial Least Squares regression, XGBoost, and Neural Networks—to project contemporaneous market returns on the analyst outputs. These ML algorithms are particularly effective in extracting both linear and non-linear relationships between a large number of inputs and the target variable, which, in this study, is the contemporaneous abnormal stock returns.

I compile a comprehensive corpus of analyst reports from the Mergent Investext database, merging these written reports with detailed analyst forecast summaries from I/B/E/S, daily intraday data from NYSE TAQ, stock returns from CRSP, and financial characteristics from Compustat. The final sample comprises 122,252 analyst reports on S&P 100 constituent firms spanning from 2000 to 2023. Analysis of this sample reveals significant dollar value in analyst information. On a three-day window basis, the equilibrium ex-ante dollar expected profits for strategic investors receiving tips about the contents of soon-to-be-released analyst reports are \$0.34 million for numerical information, \$0.38 million for textual information, and \$0.47 million for a combination

of both. Considering an average of 15 reports per stock, the annualized aggregate information value for being pre-informed on analyst reports for an average S&P 100 stock is $6.89 million. This estimate represents a conservative lower bound, as per the theoretical framework of Kadan and Manela (2020), which demonstrates that the total value of information is at least 92% of this calculated figure. Notably, these values increase for large stocks, bold forecasts, and reports released promptly following corporate earnings announcements.

By definition, the dollar value of information in analyst reports is determined by two aspects: first, how much variance in stock return can be explained by the analyst information, and second, the stock-specific price impact. Of primary interest is the first aspect, which closely relates to the notion of "R-squared". I therefore conduct analyses on an out-of-sample R^2 measure to deepen the understanding of information content in analyst reports. In the third step, leveraging tools from explainable AI, I quantify the importance of different topics in analysts' reports. Specifically, I design a Shapley value decomposition approach to fully attribute the explanatory power of report text on stock returns to 17 major topics discussed in analyst reports.

Previous literature employs two popular metrics to measure the information content of analyst reports to the stock market. The first measure is the regression coefficient of stock returns on analysts' quantifiable outputs in a narrow window surrounding the publication of analyst reports (Asquith et al., 2005; Chen et al., 2010; Twedt and Rees, 2012; Huang et al., 2014). For example, Brav and Lehavy (2003) find that a one standard deviation increase in target price revision increases the event-day abnormal return by 2.9 percentage points, while Huang et al. (2014) document a 41 basis point increase in two-day cumulative abnormal return associated with a one standard deviation increase in text opinion. Another frequently used measure, closely related to value relevance, is the R^2 from the same regression (Brown et al., 1999). The R^2 provides an intuitive notion of how much variation in stock return can be explained by certain information. It is used for model comparison of information content in Lo and Lys (2000) and Asquith et al. (2005).

One significant challenge in my setting when using either of these two measures is in-sample overfitting due to the high-dimensional nature of text embeddings. For example, the LLaMA-

2-13B model uses a 5120-dimensional vector to represent texts. This complexity can lead to overfitting when using OLS. Models that perform well on training data may fail to generalize to unseen data, potentially inflating performance metrics and providing misleading conclusions. To address this concern, I assess the explanatory power of quantitative and qualitative measures for abnormal stock returns using out-of-sample R^2. The model's performance is evaluated on a testing subsample, whose data are never included in the expanding training samples. Using out-of-sample R^2, my analysis offers a viable and credible assessment of the true informational content embedded in analyst reports.

In the baseline case, the textual information in analyst reports explains 10.19% of the out-of-sample R^2 for contemporaneous stock returns, a value that is both statistically and economically greater than that of analyst forecast revision summaries. Additionally, I demonstrate that analyst report text contains distinct information from quantitative summary measures. When both report text and numerical measures are combined, the explainable variation of *CAR* increases to 15.6%. The value of out-of-sample R^2 is significantly larger than when using each type of information individually under the Diebold and Mariano (2002) test. Economically, a one standard deviation increase in earnings forecast revision, target price revision, and report text favorableness increases the three-day cumulative abnormal return (*CAR*) by 33.95, 91.12, and 122.58 basis points, respectively. These findings are robust across different LLM text representations, ML algorithms, and various *CAR* windows.

I then explore which information content in the reports generates the largest market reactions. Analysts play both information discovery and interpretation roles, gathering information not readily available to investors or clarifying publicly available information with their opinions. These roles may not be equally valued by the market, as argued in Ivković and Jegadeesh (2004) and Huang et al. (2018). To investigate this, I use a systematic chain-of-thought (CoT) prompting with GPT-4 from OpenAI to build a meaningful taxonomy of 17 topics discussed in analysts' reports. I then categorized each sentence in the corpus of analysts' reports into these 17 topics.

To provide a value for each topic, I perform a Shapley value decomposition of out-of-sample R^2

from the CAR regression following Shapley (1953). This approach enables an additive breakdown of the contribution of each topic to the overall out-of-sample R^2 of the report. The results show that over half of the explanatory power of the report, in terms of its impact on stock returns, is attributed to Income Statement Analyses in analyst reports, particularly the interpretation of realized income numbers. This evidence highlights the importance of detailed financial analysis and the interpretation role of analysts in providing value to market investors.

Since a notable amount of research emphasizes the role of analysts in information production around corporate earnings announcements, e.g., Livnat and Zhang (2012), Keskek et al. (2014), Kim and Song (2015), Lobo et al. (2017), and Barron et al. (2017), I investigate the information content of analyst reports within and beyond the earnings announcement window. Focusing on earnings announcements and studying the interaction between earnings conference calls and analyst reports, I show that the information content of analyst reports is most pronounced in the first week following earnings announcements. This finding echoes Huang et al. (2018), who highlight that timely reactions to corporate disclosures provide information advantages to clients. Using separately trained models on samples within and beyond earnings announcement windows, I find that the magnitude of the out-of-sample R^2 almost doubles around earnings announcement dates compared to other periods, further emphasizing the incremental value of analysts in interpreting earnings announcements. By controlling for the text embeddings of corresponding earnings conference call transcripts, I mitigate the concern that the explanatory power of analyst reports is driven merely by analysts 'piggybacking' the content of earnings conference calls.

The information content varies in analyst forecast revisions and reiterations. I document that both qualitative and quantitative information are more valuable to the financial market in revision cases. Specifically, the combination of numerical and textual input generates an impressive out-of-sample R^2 of 22.63%. Conditional on earnings forecast reiterations, the information content of report text disappears. These findings support the view that analyst reports are perceived as more informative in forecast revisions than in reiterations (Jegadeesh et al., 2004; Huang et al., 2014).

The evidence above stems from the increased understanding of language by state-of-the-art

LLMs. I explicitly compare the incremental information content of analyst report text captured by LLMs with that documented in Huang et al. (2014) using a bag-of-words (BoW) representation and Naive Bayes classification. The superiority of LLMs in extracting text information is evident in two main aspects. Firstly, these models can better capture the overall tone or sentiment. Moreover, they capture richer textual content beyond the tone. To quantify these gains, I train a BERT-based sentiment classifier and show that it increases the *CAR* regression R^2 of tone from approximately 1% in the Naive Bayes model to above 3%. Additionally, transitioning from tone alone to the full representation of text using embeddings from the language model further increases the R^2 measure to above 10%. Both pieces of evidence showcase the limitations of the BoW-based approaches employed in previous studies, which ignore contextual information and primarily rely on term usage frequency.

This paper relates to three strands of literature. First, this paper deepens the understanding of the economics of tipping - a well-documented phenomenon between analysts and institutional investors (Irvine et al., 2007; Christophe et al., 2010). Tipping, which involves brokerage firms rewarding high-commission clients with early research access, allows these firms to recover research costs through commissions from clients who benefit from this privileged information. Green (2006) examined the short-term profitability associated with early access to stock analyst recommendation revisions and found average 2-day returns of 1.02% by buying stocks after upgrades and 1.50% by short-selling after downgrades. This work extends previous studies that established the existence and incentives of tipping by quantifying its economic dollar value. It proposes an estimate based on general assumptions, eliminating the need for proprietary data.

Second, it contributes to the literature on the nature of information content conveyed by analysts. Most early studies focused on examining three quantitative forecast summaries and documented the informativeness of forecast revisions (e.g., Barber and Loeffler, 1993; Womack, 1996; Michaely and Womack, 1999; Brav and Lehavy, 2003; Sorescu and Subrahmanyam, 2006). Interest then shifted to assessing the usefulness and informativeness of narratives in analyst reports to the market. Researchers began with hand-collected small samples of full-text analyst reports and started to acknowledge the importance of written reports in assessing analysts' contributions to the

market (Previts et al., 1994; Asquith et al., 2005). With the development of textual analysis tools, Kothari et al. (2009) and Huang et al. (2014) examined the effect of sentiment in analyst reports on capital markets. This paper extends this line of research by employing state-of-the-art LLMs to systematically analyze the textual content of analyst reports, demonstrating the substantial value these texts add to the market. Additionally, it appears to be the first to systematically examine the value of different topics discussed by analysts.

Third, this work interacts with the literature on the application of machine learning and natural language processing in finance. Recent advancements in LLMs, such as BERT, LLaMA, and ChatGPT, have revolutionized the ability to extract nuanced information from text. Chen et al. (2022), Li et al. (2023), Jha et al. (2024), and Beckmann et al. (2024) have showcased the potential of these models in financial contexts. This research leverages these advanced tools to provide a comprehensive and detailed analysis of analyst reports, highlighting their superior ability to improve our understanding of the value of information produced by analysts. The outperformance of LLMs with large parameter spaces and deep ML models in explaining stock returns underscores the virtue of complexity in a linguistic context, as argued in Kelly et al. (2022).

The rest of the paper is organized as follows. Section 2 introduces methodology, measures of information value, and interpretation. Section 3 summarizes data construction and empirical results. Section 4 concludes.

2 Methodology

In this section, I introduce the large language models used for topic classification and text embedding, the strategic information value estimation of analyst reports, the information content estimation of out-of-sample R^2, and the decomposition metric used to evaluate the information content of specific topics.[2]

[2]The machine learning models used to explain stock returns are discussed in the Online Appendix C.

2.1 Large Language Models

Large Language Models such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and Large Language Model Meta AI (LLaMA) are trending significantly in literature for understanding textual data. These models represent a paradigm shift in natural language processing, offering capabilities that go beyond traditional methods like word-to-vec (W2V) or Latent Dirichlet Allocation (LDA) topic models. In this study, I harness the power of these generative AI models to analyze analyst reports. By employing fine-tuned BERT for classification, ChatGPT for topic extraction, and LLaMA for text embedding, I aim to extract valuable insights from the extensive textual data contained within analyst reports. BERT's classification capabilities allow for the categorization of sentences based on predefined criteria. ChatGPT is capable of understanding a broad context of text and extracting the major topics discussed by analysts. Text embedding functionality enables the representation of reports as structured high-dimensional vectors.

2.1.1 ChatGPT Prompting

To effectively categorize the topics discussed in analyst reports, I employ ChatGPT prompting to extract exclusive and consistent topics and avoid ad-hoc classification. By leveraging AI, my goal is to identify the primary topics covered in these reports and assign each sentence to a single, specific topic category. This approach ensures a systematic and coherent classification of the content within analyst reports.

In the first step, to obtain manageable and meaningful topics in analyst report content, I feed ChatGPT a random sample of 100 analyst reports and ask about the information contained, aiming to identify high-level categories.

Prompt 1: Please read the provided text file of sell-side analyst reports carefully. What are high-level mutually exclusive topics covered in these reports? Make sure that each sentence from the text file can be assigned to one of the topics. Here is the report content: {text}.

It ends up with 13 categories defined as follows.

- **Executive Summary:** Provides a high-level overview of the report's key findings and conclusions; includes a brief description of the company, its industry, and the purpose of the report; highlights the most important points from the analysis, such as the company's financial performance, competitive position, and growth prospects.
- **Company Overview:** Offers a comprehensive description of the company, including its history, business model, and key products or services; discusses the company's organizational structure, management team, and corporate governance; analyzes the company's mission, vision, and strategic objectives.
- **Industry Analysis:** Provides an in-depth analysis of the industry in which the company operates; includes information on market size, growth trends, and key drivers; discusses the regulatory environment, technological advancements, and other external factors affecting the industry; analyzes the industry's competitive dynamics and the company's position within the industry.
- **Competitive Landscape:** Identifies the company's main competitors and their market share; compares the company's products, services, and pricing strategies with those of its competitors; analyzes the strengths and weaknesses of the company and its competitors; discusses potential new entrants and substitutes that could disrupt the competitive landscape.
- **Financial Analysis:** Analyzes the company's revenue, expenses, profitability, assets, liabilities, shareholders' equity, liquidity, solvency, capital structure, cash flows, and key financial ratios; compares the company's financial performance with that of its competitors and industry benchmarks; discusses trends in the company's financial performance over time.
- **Business Segments:** Provides a detailed analysis of the company's various business segments or divisions; discusses the financial performance, growth prospects, and challenges of each segment; analyzes the contribution of each segment to the company's overall revenue and profitability.

- **Growth Strategies:** Discusses the company's strategies for driving future growth, such as organic growth initiatives, product innovations, and geographic expansions; analyzes the company's mergers and acquisitions (M&A) strategy and potential targets; examines the company's investments in research and development (R&D) and marketing.

- **Risk Factors:** Identifies and analyzes the key risks facing the company, such as market risks, operational risks, financial risks, and legal/regulatory risks; discusses the potential impact of these risks on the company's financial performance and growth prospects; examines the company's risk management strategies and mitigation measures.

- **Management and Governance:** Provides an overview of the company's management team, including their experience, expertise, and track record; analyzes the company's corporate governance practices, such as board composition, executive compensation, and shareholder rights; discusses the company's succession planning and key person risks.

- **Environmental, Social, and Governance (ESG) Factors:** Analyzes the company's performance and initiatives related to environmental sustainability, social responsibility, and corporate governance; discusses the potential impact of ESG factors on the company's reputation, risk profile, and financial performance; examines the company's compliance with ESG regulations and industry standards.

- **Valuation:** Estimates the intrinsic value of the company's shares using various valuation methodologies, such as discounted cash flow (DCF) analysis, relative valuation multiples, and sum-of-the-parts analysis; compares the company's valuation with that of its peers and historical benchmarks; discusses the key assumptions and sensitivities underlying the valuation analyses.

- **Investment Thesis:** Summarizes the key reasons for investing (or not investing) in the company's shares; discusses the potential catalysts and risks that could impact the company's valuation and stock price performance; provides a target price or price range for the company's shares based on the valuation analyses and investment thesis.

- **Appendices and Disclosures:** Includes additional supporting materials, such as financial statements, ratio calculations, and detailed segment data; provides important disclosures, such as the analyst's rating system, potential conflicts of interest, and disclaimers; discusses the limitations and uncertainties of the analysis and the need for further due diligence by investors.

The Financial Analyses category is further divided into four subcategories: Income Statement Analyses, Balance Sheet Analyses, Cash Flow Analyses, and Financial Ratios. Additionally, a category labeled "None of the Above" is included to accommodate sentences that do not fit into any of the predefined topics. In total, this process generates 17 distinct topics for classification. In the Appendix, Figure A3 displays word clouds for each topic while Table A7 provides illustrative sentences from the analyst reports, exemplifying the content discussed within each category.

In the second step, a systematic approach is required to assign topic classifications to a large corpus of 6,975,114 sentences. To begin with, a random sample of 17,028 sentences from 100 analyst reports is selected from analyst reports. I use ChatGPT API to call the GPT-4-Turbo model and categorize each sentence into a single, relevant topic.

Prompt 2: Please read the following sentence from the sell-side analyst report of the company {firm} ({ticker}) carefully. Determine which category of information it belongs to in the following 17 categories: {categories}. Pay attention to the sentence in the context of the report. Output your response in JSON format.
Here is the sentence from the analyst report: {sentence}.

This training sample is used to fine-tune a BERT model for categorizing all sentence segments.[3] Figure 3 displays the stacked area plot of 17 topics across years. Income Statement Analyses and Financial Ratios are the most frequently discussed topics in analyst reports, representing 17.23% and 15.65% of all sentences, respectively. The next most popular topics are Risk Factors, Valuation, and Investment Thesis, each comprising 8.32%, 8.10%, and 7.87% of the content. The least discussed topics are Appendices and Disclosures (0.50%), Executive Summary (0.48%), and

[3]The fine-tuned BERT model exhibited an accuracy score of 89% on a testing sample.

ESG (0.23%). The low frequency of Appendices and Disclosures is due to pre-boilerplate cleaning procedures. Executive Summary is difficult to identify at the sentence level, and ESG is mainly discussed after 2020.

2.1.2 Text Embedding

Large language models represent the state-of-the-art methodology in textual analysis. Unlike word-based techniques (such as bag-of-words or word-to-vector), the LLM framework begins with a contextual representation of text tokenizations. The tokenization procedure maps words into a token space. Before tokenization, text is often normalized, which can include converting to lowercase, removing punctuation, and handling special characters. Depending on the strategy, tokenization involves breaking down text into smaller units, which can be words, subwords, or even characters. Within LLMs, tokens are mapped to unique identifiers (IDs) based on a predefined vocabulary. This vocabulary is learned during the model's training phase and contains all the tokens the model can understand.[4]

Here is an example of tokenized analyst reports for NVIDIA: "We expect to see more GPU/SmartNIC integration as next-gen workloads grow and CPUs become a bigger bottleneck in the data center." The LLaMA tokenizer breaks this sentence into a sequence of sub-words: "'We', 'expect', 'to', 'see', 'more', 'GPU', '/', 'Sm', 'art', 'N', 'IC', 'integration', 'as', 'next', '-', 'gen', 'work', 'loads', 'grow', 'and', 'CPU', 's', 'become', 'a', 'bigger', 'bott', 'l', 'ene', 'ck', 'in', 'the', 'data', 'center', '.'". The word "next-gen workloads" is broken into five tokens in this example: "next", "-", "gen", "work" and "loads".

At the heart of LLMs lies the transformer architecture. Introduced by Vaswani et al. (2017), the transformer architecture revolutionized the field of natural language processing (NLP) with its novel approach to sequence-to-sequence tasks. Unlike traditional recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), transformers rely entirely on a mechanism called self-attention to process input sequences in parallel rather than sequentially.

[4] BERT uses 30,000 tokens, OpenAI GPT models use around 50,257 tokens, and LLaMA models use approximately 32,000 tokens.

This self-attention mechanism allows the model to weigh the significance of different words in a sentence relative to each other, effectively capturing context and dependencies regardless of distance. The transformer consists of an encoder-decoder structure where the encoder maps an input sequence into a set of continuous representations, and the decoder uses these representations to generate an output sequence. Each encoder and decoder layer comprises multi-head self-attention mechanisms and feed-forward neural networks, along with residual connections and layer normalization to enhance training stability. This architecture not only enables efficient training on large datasets but also achieves state-of-the-art performance on a wide range of NLP tasks, allowing for model scalability.

Text embedding is a crucial process in NLP that transforms text data into dense, continuous vectors, capturing the semantic meaning and syntactic structure of the text. In the context of transformer architectures like BERT, OpenAI GPT, and LLaMA, these embeddings are generated through the self-attention mechanisms within the model, effectively encoding contextual information about each token in relation to the entire sequence. BERT generates embeddings with a dimension of 768 for BERT Base and 1,024 for BERT Large, allowing the model to capture intricate patterns and relationships in the text. OpenAI's GPT models include an embedding dimension of 768 for GPT-2 Small, enabling the understanding and generation of human-like text. LLaMA models, designed to be efficient yet powerful, use an embedding dimension of 5,120 for their largest configurations, balancing computational efficiency with rich, contextual embeddings. These embeddings serve as the foundational representations for various downstream NLP tasks, enabling models to perform exceptionally well in applications like text classification, translation, and sentiment analysis.

To obtain structured representations of analyst reports, I used the pre-trained LLaMA-2-13B model and its tokenizer from MetaAI. The tokenizer processes the text data, converting it into tokens that the model can understand. By processing the analyst reports through this model, I obtained rich, contextual embeddings that capture the nuanced information within the reports. These embeddings are suitable for various downstream NLP tasks, such as sentiment analysis and text classification. The reason I employ LLaMA2 as a baseline Language Model is that it enables

an input limit of 4,096 tokens, which is sufficient for most analyst reports, given their median length of 1,393 tokens and mean length of 2,055 tokens.

2.2 Strategic Value of Information

When multiple reports are released on the same day following corporate events, it is unrealistic to assume that the information from these reports is uncorrelated. Thus, I apply a general measure that accounts for multiple informed traders, each possessing an imperfect signal based on the theoretical framework of Back, Cao, and Willard (2000), henceforth BCW, which corresponds to correlated information from distinct analyst reports about an asset.

The framework estimates the total value of information from analyst reports for all informed traders. Simply, the value of this information approximates the ratio of the variation in the asset's fundamental value explained by the trader's information to the asset's illiquidity (Kyle's lambda).

To understand this estimation, it's crucial to revisit the key elements of the BCW model. In their framework, there are $n \geq 1$ risk-neutral informed traders. Each informed trader i receives a signal $\tilde{s}^i$ about the fundamental value of an asset. Signals $\tilde{s}^i$ are symmetrically joint-normally distributed with a correlation coefficient $0 \leq \rho < 1$ between $\tilde{s}^i$ and $\tilde{s}^j$ for $i \neq j$.

The expected value $\tilde{v}$ of the asset at the end of the trading period is given by the sum of the signals from all informed traders:

$$\tilde{v} = \sum_{i=1}^{n} \tilde{s}^i. \tag{1}$$

Each informed trader i trades according to a strategy linear in the price $P(t)$ and their signal $\tilde{s}^i$:

$$\theta^i(t) = \alpha(t)P(t) + \beta(t)\tilde{s}^i. \tag{2}$$

Informed traders aim to maximize their expected profit from trading. The profit for trader i is given by:

$$E\left[\int_0^1 (\tilde{v} - P(t))\left(\alpha(t)P(t) + \beta(t)\tilde{s}^i\right) dt\right]. \tag{3}$$

Risk-neutral market makers set the asset price $P(t)$ based on the total order flow from informed and noise traders. The cumulative noise trading process $Z(t)$ follows a Wiener process:

$$dZ(t) = \sigma_z dW(t), \tag{4}$$

where W_t is a Wiener process, and σ_z is the volatility of the noise trading. The price $P(t)$ evolves according to the combined order flow and follows the stochastic differential equation:

$$dP(t) = \lambda(t)\left[dZ(t) + \sum_{i=1}^{N}\left(\alpha(t)P(t) + \beta(t)\tilde{s}^i\right)dt\right]. \tag{5}$$

That is, the price set by market makers adjusts to the information revealed through trading.

An equilibrium in the model is defined by two primary conditions: first, the price $P(t)$ at any time t must equal the conditional expectation of the asset's liquidation value $\tilde{v}$ given the information generated by the aggregate order process up to that time, ensuring the market makers' price reflects all available information; second, each trader's strategy θ^i must be optimal given the strategies of all other traders θ^{-i} and the function λ, meaning no trader can increase their expected profit by unilaterally changing their strategy. These conditions ensure that the market price accurately incorporates the collective information of all trades and that traders' strategies are mutually consistent and profit-maximizing.

Given by Theorem 1 in BCW, the linear equilibrium satisfies:

$$\begin{aligned}
\beta(t) &= \sigma_z(\kappa/\Sigma(0))^{1/2}\left(\frac{\Sigma(t)}{\Sigma(0)}\right)^{(n-2)/n}\exp\left[\frac{\Sigma(0)-\Sigma(t)}{\Sigma(t)}\right], \\
\alpha(t) &= -\beta(t)/n, \\
\lambda(t) &= \frac{\beta(t)\Sigma(t)}{\sigma_z^2},
\end{aligned} \tag{6}$$

where κ is a constant defined in the model, specifically involving an integral related to the equilibrium conditions, and $\Sigma(t)$ is the conditional variance of $\tilde{v}$ given the market maker's time

t information. Mathematically, κ is given by

$$\kappa = \int_1^{\infty} \frac{2(n-2)}{n} x^{\frac{n-2}{n}} e^{-2x\zeta} dx. \tag{7}$$

Here, $\zeta(n,\rho)$ and $\phi(n,\rho)$ are denoted as

$$\zeta(n,\rho) = \frac{1-\phi(n,\rho)}{n\phi(n,\rho)},$$

and

$$\phi(n,\rho) = \frac{\text{var}(\tilde{v})}{\text{var}(n\tilde{s}^i)} = \frac{1}{n} + \frac{n-1}{n}\rho.$$

This integral essentially captures the impact of the correlation among the signals received by the informed traders and the number of informed traders on the trading intensity and market depth.

Finally, the value of information to all informed traders combined, denoted by Ω, is given by the Corollary 3 in BCW as

$$\Omega = \sigma_z \left(\frac{\text{var}(\tilde{v})}{\kappa}\right)^{1/2} \int_1^{\infty} x^{-2/n} \exp[-\zeta x] dx. \tag{8}$$

Kadan and Manela (2020) build upon the work of BCW and express the total value of information in a more practical form. Considering versions of Kyle's lambda in which order flow affects returns, they derive the equilibrium ex-ante expected dollar value of information about a specific asset as

$$\Omega = c(n,\rho)\frac{\text{var}(\tilde{v})}{\lambda(0)} P(0), \tag{9}$$

where $\text{var}(\tilde{v})$ is the variance of returns, $\lambda(0)$ is the marginal impact of share order flow on stock returns, and $P(0)$ is the initial price of the asset.[5] The coefficient c depends on the number of informed traders, n, and the correlation ρ, encapsulating how their presence and signal

[5]The theoretical framework by Kyle (1985) offers insights into the value of insider information relative to market liquidity, but practical application to real stock prices is challenged by the assumption of normally distributed asset values, leading to potential negative prices, and by difficulties in estimating price volatility due to non-stationary price behaviors (Kadan and Manela, 2020). Due to these issues, a measure derived from a log-normal framework is applied.

correlation influence the overall value of information in the market. Specifically, the coefficient of proportionality $c(n,\rho) > 0$ does not depend on σ_z and is given by

$$c(n,\rho) = e^{\zeta} \int_1^{\infty} x^{-\frac{2}{n}} e^{-x\zeta} dx. \tag{10}$$

According to the discussion of the tightness of the bound in terms of $c(n,\rho)$ in Kadan and Manela (2020), the total value of information cannot be less than 92% of the ratio of the information variance to the initial price impact, regardless of competition among informed traders and the correlation between their signals. Thus, $\frac{\text{var}(\tilde{v})}{\lambda(0)} P(0)$ can serve as an approximate lower bound for the total value of information.

My goal is to estimate the value of information from analysts using Equation 9. To approximate information from analysts, I use the explained three-day window return variance from all reports issued on day t. Thus, for each stock i on date t, I estimate a dollar value of information as follows

$$\widehat{\Omega}_{it} = \frac{r_{it}^2 - \left(r_{it} - \frac{\sum_{j=1}^{N} \widehat{r}_{ijt}}{N} \right)^2}{\lambda_{it}} \cdot p_{it_-}, \tag{11}$$

where $\widehat{\Omega}_{it}$ is the ratio of explainable realized return variance estimated using Ridge regression to price impact estimated from 1-minute log returns and order flow, multiplied by the closing price of the day prior to the [t-1, t+1] window.[6]

Specifically, $\widehat{r}_{ijt}$ is the estimated return of analyst report j issued on day t, and r_{it} is the realized cumulative abnormal return $CAR_{[-1,+1]}$. N denotes the number of analyst reports about stock i on day t. To estimate the price impact λ_{jt}, I perform a regression of 1-minute log returns (r_{itk}) on contemporaneous share order flow (y_{itk}). Here, y_{itk} is calculated as the change in cumulative signed order flows $Y_{it\tau_k}$ over intraday intervals. The signed order flow is classified as buys (+1) or

[6]It should be noted that, due to risk neutrality, the equilibrium of the continuous-time Kyle model holds if $\tilde{v}$ is an unbiased signal of the asset price. The intuition is illustrated with an extended single-auction Kyle model incorporating incomplete information in the Online Appendix D.

sells (-1) based on algorithms from Lee and Ready (1991).[7] The regression model used is

$$r_{itk} = \widehat{\lambda}_{it} y_{itk} + \varepsilon_{it}, \tag{12}$$

where $\widehat{\lambda}_{it}$ represents the estimated price impact. In this context, $r_{itk} = p_{it\tau_k} - p_{it\tau_{k-1}}$ and $y_{itk} = Y_{it\tau_k} - Y_{it\tau_{k-1}}$, with $p_{it\tau_k}$ denoting the log price of stock i within the window [t-1, t+1] at time $\tau \in [0, T]$. The interval T is divided into one-minute sections $\tau_0, \tau_1, \ldots, \tau_K$. Thus, each regression has $1170 (= 390 * 3)$ observations.

Given that the measure is a ratio of two random variables measured with error, interpretation for information value estimation is inherently nuanced. Directly computing the sample mean of the value of information Ω, which is the ratio of mean return variance (μ_{v_s}) to mean price impact per dollar (μ_{λ_s}), can lead to biased and unreliable estimates. This approach fails because the mean of a ratio is not equal to the ratio of the means, leading to skewed and incorrect results, particularly when μ_{λ_s} is small or close to zero, which can cause extreme and highly variable values (Kadan and Manela, 2020). I use a first-order estimator for the mean and variance of information value over sample s[8]

$$E\widehat{\Omega}_s = \frac{\mu_{vs}}{\mu_{\lambda s}}, \tag{13}$$

$$\operatorname{Var}\widehat{\Omega}_s = \frac{1}{\mu_{\lambda s}^2}\left(\Sigma_{\lambda s}\frac{\mu_{vs}^2}{\mu_{\lambda s}^2} + \Sigma_{vs} - 2\Sigma_{v\lambda s}\frac{\mu_{vs}}{\mu_{\lambda s}}\right), \tag{14}$$

where

$$\mu_{\lambda s} = \frac{1}{|s|}\sum_{it\in s}\widehat{\lambda}_{it}/P_{it_-}, \tag{15}$$

$$\mu_{vs} = \frac{1}{|s|}\sum_{it\in s}\left(r_{it}^2 - \left(r_{it} - \frac{\sum_{j=1}^{N}\widehat{r}_{ijt}}{N}\right)^2\right), \tag{16}$$

and

$$\Sigma_s = \begin{bmatrix} \Sigma_{vs} & \Sigma_{v\lambda s} \\ \Sigma_{v\lambda s} & \Sigma_{\lambda s} \end{bmatrix} = \operatorname{Cov}\left(\left[\left(r_{it} - \frac{\sum_{j=1}^{N}\widehat{r}_{ijt}}{N}\right)^2, \frac{\widehat{\lambda}_{it}}{P_{it_-}}\right]\right). \tag{17}$$

[7] The algorithms from Ellis et al. (2000) and Chakrabarty et al. (2007) are also used in robustness tests.

[8] The delta method used to derive the first-order estimator is discussed in the Online Appendix E.

2.3 Out-of-sample R-sqaured

I use out-of-sample R^2 as a measure for the information content of analyst reports. Following Gu et al. (2020), I calculate the out-of-sample R^2 as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathscr{T}} \left(CAR_{i,t} - \widehat{CAR}_{i,t}\right)^2}{\sum CAR^2_{i,t}}, \tag{18}$$

where $\mathscr{T}$ is the test set of (i,t) which has not been used in training and validation of the model. R^2_{oos} aggregates estimation across analyst reports and captures the portion of CAR that can be explained by textual information.

The out-of-sample R^2 metric has two key properties. Firstly, it mitigates in-sample overfitting, which is particularly important given the high dimensionality of LLM embeddings. Secondly, it encapsulates the notion of information value to the market, a focus of existing literature Brown et al. (1999).

To evaluate the statistical significance of the out-of-sample R^2, I employ a modified version of the Diebold and Mariano (2002) test for comparing predictive accuracy between two models. Given the strong cross-sectional dependence in the stock-level analysis, I follow Gu et al. (2020) and adapt the Diebold-Mariano (DM) test by comparing the cross-sectional average of prediction errors from each model rather than individual returns. Specifically, I assess the forecast performance of method (1) versus method (2) using the test statistic $DM = \frac{\bar{d}_{12}}{\widehat{\sigma}_{d_{12}}}$. The numerator can be written as

$$d_{12,t} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(\left(\widehat{e}^{(1)}_{i,t}\right)^2 - \left(\widehat{e}^{(2)}_{i,t}\right)^2\right), \tag{19}$$

where $\widehat{e}^{(1)}_{i,t}$ and $\widehat{e}^{(2)}_{i,t}$ represent the model (1) and (2)'s residual for CAR at time t for each method, respectively. The number of stocks in the testing sample (year t) is denoted by M_t. The terms $\bar{d}_{12}$ and $\widehat{\sigma}_{d_{12}}$ denote the mean and the Newey-West standard error of $d_{12,t}$ over the testing sample.

This modified test statistic, constructed from a single time series $d_{12,t}$ of error differences, is robust to strong cross-sectional dependencies. It ensures compliance with the necessary regularity

conditions for asymptotic normality, thereby providing reliable t-statistics for the evaluation of R^2_{oos} evaluation and model comparisons.

2.4 Shapley Value Decomposation

To assess the impact of topics within analyst reports, I employ the Shapley Additive exPlanations (SHAP) methodology, as detailed by Lundberg and Lee (2017) and utilized by Chen et al. (2022). SHAP is a solution concept in cooperative game theory that distributes the total payoff generated by a coalition of players among the players themselves. Considering the favorable nature of additivity, I design a decomposition method based on word embeddings.

First, LLMs pool report-level embeddings by averaging embeddings across all tokens and over all layers. The textual embedding can be represented as:

$$y^{emb} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_i, \tag{20}$$

where e_i is the embedding of token i across all the layers, N is the total number of words in the report. If the tokens can be classified into different topics, the embedding can be fully decomposed into different topic-specific embeddings:

$$y^{emb} = \sum_{p=1}^{P} y_p^{emb} = \frac{1}{N}\sum_{p=1}^{P}\sum_{i_p=1}^{N_p} e_{i_p}, \tag{21}$$

where P is the number of topics, y_p^{emb} is the embedding for topic p, N_p is the number of tokens belonging to topic p, and i_p is the index of words within topic p.

One concern in the architecture of a transformer model is that each token in the input sequence can pay attention to other words through a mechanism called "self-attention" or "scaled dot-product attention." This mechanism allows the model to weigh the importance of each word relative to every other word in the sequence, thereby capturing contextual information. Consequently, the embedding of a word contains information from other words, which may pertain to different topics.

The embedding of the word might predominantly reflect the context information from other topics. To isolate the self-attention across topics, I process the embedding for each sentence in analyst reports and take the token-weighted mean as the report embedding.

$$y^{emb} = \sum_{i=1}^{n} \frac{Token_i}{\sum_{i=1}^{n} Token_i} y_i^{emb}, \tag{22}$$

where $Token_i$ represents the number of tokens in i-th sentence and y_i^{emb} represents the embedding of the i-th sentence.

The loss in R^2_{oos} from full-context embedding to sentence-segmented embedding reflects the information content of context in analyst reports.

Second, I calculate the Shapley value of the sentence-segmented embeddings distributed to each topic. Empirically, the Shapley value can be written as:

$$\varphi_p\left(R^2_{\text{oos}}\right) = \sum_{S \subseteq P \setminus \{p\}} \frac{|S|!\,(P-|S|-1)!}{P!} \left[R^2_{\text{oos}}\left(y_s^{\text{emb}} + y_p^{\text{emb}}\right) - R^2_{\text{oos}}\left(y_s^{\text{emb}}\right)\right], \tag{23}$$

where the sum extends over all subsets S of P not containing topic p, including the empty set. Note that $\binom{n}{k}$ is the binomial coefficient. y_p^{emb} is the token-weighted sentence embedding of topic p, and y_s^{emb} is the token-weighted sentence embedding of all topics in subset S. For each combination of topic embeddings, I calculate the R^2_{oos} accordingly.[9]

3 Data and Empirical Results

In this section, I introduce the data and report the empirical results of the analysts' information content. I then decompose the value across topics. Next, I quantify the strategic dollar value of information from analyst reports. Finally, I conduct various robustness checks.

[9] Shapley value requires an input for X when it is missing. In line with Gu et al. (2020), I assign a value of zero, treating the embedding as having all dimensions set to zero for model estimation.

3.1 Data

Sell-side analyst reports are obtained from the Mergent Investext Database. I download 223,091 analyst reports of S&P 100 firms covering the period from 2000 to 2023. The initial reports are in PDF format, and I use Adobe Acrobat to extract structured text from each report. The segmented text is then chunked into sentences. To get clean content of analyst discussions, I train a BERT model to remove the boilerplate segments from each report.[10]

The Mergent Investext Database provides information on analyst names and the release date for each report. To connect Investext with the I/B/E/S database, I first match the lead analyst name in reports to Analyst ID (AMASKCD) following the procedure of Li et al. (2023). I then match specific reports to the corresponding EPS forecasts, price targets, and recommendation files in the I/B/E/S database. As noted in Huang et al. (2014), the timing of a forecast is identified by its announcement date and revision date. Multiple reports are released between the announcement date and the revision date. Following Huang et al. (2014), the matching window for each report is defined as 2 days before the announcement date to 2 days after the revision date. The specific report is then linked to I/B/E/S by matching analyst names, firm ticker identifiers, and the matching window with specific forecasts. The numerical variables include earnings forecasts, stock recommendations, and price target information from analyst reports. I also compile firm characteristics via WRDS, particularly from the CRSP and Compustat databases.[11] My final sample comprises 122,252 reports of S&P 100 firms from 1,305 analysts across 140 unique brokerage firms.

Table 1 presents the summary statistics of the analyst report data, including the number of reports, the number of unique brokerage firms, and the number of unique analysts for each year, categorized by the Fama-French 12 industries. Additionally, it reports the average number of pages and tokens per report. Overall, the time trend in the number of analyst reports aligns with

[10] Boilerplate segments disclose information about the brokerage firm and analyst research team and do not reflect analyst opinions. Following Li et al. (2023), I randomly sampled one report from each of the top 20 brokerage firms producing the largest number of analyst reports and manually labeled each sentence in those reports. The BERT model trained for boilerplate classification reached an accuracy score of 97.31% on a testing sample.

[11] See Table A1 for detailed definitions.

the findings of Bonini et al. (2023), and the average number of pages per report is consistent with that reported by Huang et al. (2014).[12]

The daily stock-level price impact (Kyle's lambda) is estimated using intraday data from the NYSE TAQ database. The analysis covers an out-of-sample period from January 2, 2015, to December 28, 2023. To ensure comparability over time, I adjust all prices for inflation using the Consumer Price Index (CPI), with December 2020 as the base period.

In Section 3.3.3, I explore textual information from earnings conference calls and subsequent analyst reports. The earnings call transcripts are sourced from Seeking Alpha, which has been collecting the transcripts since 2005[13]. Seeking Alpha provides a comprehensive account of what was discussed during the call, including presentations by company executives and the subsequent question-and-answer session with analysts. The conference calls include a question-and-answer session where analysts and investors can ask questions, offering valuable insights into live interactions between managers and analysts. I sampled earnings call transcripts of S&P 100 constituent firms spanning 2005 to 2023 and merged them with ticker and Report Date for Quarterly Earnings (RDQ) data from the Compustat database. The matching window between Seeking Alpha's publishing date and RDQ is set to two weeks. This process resulted in a final sample of 7,909 firm-quarter-level conference calls.

Given my goal to estimate the information value of analyst outputs, I need to define the information set available to insider investors. The numerical information and variables as inputs to machine learning models are described as follows.

Report-Level Measures: REC_REV denotes recommendation revision, calculated as the current report's recommendation minus the last recommendation in I/B/E/S issued by the same analyst for the same firm. EF_REV represents earnings forecast revision, calculated as the current report's EPS forecast minus the prior EPS forecast in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release. TP_REV indicates target

[12]The data reveals a peak in the number of analyst reports in 2013, followed by a subtle decline. This trend can be attributed to the regulatory impacts of the Dodd-Frank Wall Street Reform in the United States and the introduction of MiFID II in Europe (Bradshaw et al., 2017).

[13]Available at: https://seekingalpha.com/earnings/earnings-call-transcripts

price revision, calculated as the current report's target price minus the last target price in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release[14]. *Boldness* is an indicator variable that takes the value of 1 if revisions are either above both the consensus and previous forecasts or below both. *SR* represents stock recommendation, following the I/B/E/S rating system, where a rating of 5 denotes a strong sell and a rating of 1 indicates a strong buy. *ERet* represents the 12-month return forecast, calculated by scaling the 12-month price target by the stock price 50 days before the report release. *Prior_CAR* represents the cumulative abnormal return over a ten-day period ending two days before the report release date.

Firm-Level Measures: *Size* is measured by market equity. The Book-to-Market Ratio ($BtoM$) compares the firm's book value to its market value. Following Huang et al. (2018), the Earnings Suprise (SUE) is calculated as the actual earnings per share (EPS) minus the last consensus EPS forecast prior to the earnings announcement date (EAD), with $AbsSUE$ representing its absolute value. The *Miss* variable is a binary indicator set to one if SUE is positive. $TradingVolume$ is the stock's trading volume on the latest earnings announcement days standardized by the number of shares outstanding. The Distance to Default (*DD*) variable is sourced from the NUS Credit Research Initiative (CRI). $Fluidity$ measures the firm's product market fluidity.

Industry-Level Measures: *IndustryRecession* is an indicator variable that equals one if the firm's industry, as defined by the Fama-French 48 industry classification, experiences a negative monthly return in the bottom quintile.

Macroeconomic Measures: The $TimeTrend$ variable captures the temporal dimension by recording the number of years elapsed since the beginning of the sample period.

3.2 Dollar Value of Information in Analyst Reports

In this Section, I quantify the information value of analyst reports using the measure introduced in Section 2.2. I first summarize the value of information from analyst reports in Table 2, covering

[14]Reports are labeled as revision reports or reiteration reports according to the criterion of Huang et al. (2014).

the period from 2015 to 2023. Both the mean and standard deviation are derived using the delta method introduced in Section 2.2. For S&P 100 constituent stocks, the average three-day window information value from sell-side analysts is \$0.47 million, with a standard error of \$0.05 million. The mean information value of text alone is \$0.38 million, and the mean value for revisions is \$0.34 million. Considering a stock typically has an average of 15 days with report releases each year, the lower bound an investor would be willing to pay for insights from analyst reports on an S&P 100 firm over a year is \$6.89 million. The narrow standard errors and confidence intervals across different information values indicate consistent estimates. The average price impact ($\widehat{\lambda}_{it}/P_{it\_}$) is 0.34, indicating that \$1 million dollar order pushes the stock price up by 0.034%. These statistics illustrate that information from analysts is highly valuable, with both textual information and revisions contributing significantly to return volatility.

Figure 1 describes the evolution of information value from analyst reports over time. The solid line represents the strategic value of both numerical and textual information from analyst reports, showing fluctuations over the period from 2015Q1 to 2023Q4. Despite these fluctuations, there is a general upward trend in information value, particularly noticeable from around 2018 onwards. The dashed line, which indicates the strategic value of textual information alone, follows a similar pattern but is consistently slightly lower. This suggests that numerical information complements the textual information in the reports.

The shaded area represents the 95% confidence interval. After 2020, there appears to be greater variability in the information value, as indicated by the wider confidence intervals and more pronounced peaks and troughs. This period corresponds with the global COVID-19 pandemic, which likely introduced additional uncertainty and volatility in financial markets, impacting the value of information from analyst reports.

These patterns suggest that while the overall information value of analyst reports has increased over time, there are significant periods of variability influenced by external events and market conditions. The strategic value derived from combining numerical and text information generally surpasses that from text alone, highlighting the importance of a comprehensive analysis of analyst

reports.

In the Online Appendix, Table A3 provides the summary statistics of alternative estimates for information value. Using high-frequency measures of realized return volatility and different algorithms for signing trades as buys or sells, the estimates for information value from analyst reports range from $0.42 million to $0.64 million. Figure A1 illustrates the time trends of these three alternative estimates, revealing consistent patterns in the evolution of information value over time.

3.2.1 The Value of Information in the Cross-section of Stocks

It is likely that the information value of sell-side analysts will vary in the cross-section of stocks. Specifically, firms of different sizes may have different capabilities and costs for producing and disseminating information (Zeghal, 1984). For example, Kadan and Manela (2020) argue that the value of information is consistently higher for large stocks due to higher liquidity. On the other hand, the opinions from analysts for smaller stocks may be associated with larger market reactions since small stocks typically have less information publicly available overall (Stickel, 1995). I plot the scatter of information value and market capitalization by stock in Figure A2. There is a general upward trend, indicating that stocks with higher market equity tend to have higher information values. However, the relationship is not strictly linear, as there are notable exceptions. I formally test the relationship by regressing the value of information on stock size, firm fixed effects, and year fixed effects. I also control for the book-to-market ratio in the regressions.

Table 3 shows the results. Stocks with larger market equity have significantly higher information value from analysts. Economically, column (4) means a 1% increase in market equity is associated with a 0.864% increase in analyst information value. Columns (5)-(8) show that the information value of report text is slightly more sensitive to size.

The positive size elasticity of information value could be attributed to either higher explained return volatility or lower price impact for large stocks. This decomposition is expressed

logarithmically as:

$$\log \widehat{\Omega}_{it} = \underbrace{\log \left[r_{it}^2 - \left(r_{it} - \frac{\sum_{j=1}^{N} \widehat{r}_{ijt}}{N} \right)^2 \right]}_{\text{log return variance}} - \underbrace{\frac{\log \widehat{\lambda}_{it}}{P_{it\_}}}_{\text{log price impact}} . \tag{24}$$

Panel B of Table 3 demonstrates that the main reason behind the positive size elasticity of information value is the higher liquidity of large stocks. This finding aligns with the results reported by Kadan and Manela (2020). Overall, analyst information does not appear to explain the higher contemporaneous return volatility of large stocks, yet it remains more valuable due to the higher liquidity of these stocks.

3.2.2 The Value of Information in the Cross-section of Analysts

I next examine how the information value of analysts differs with herding behavior. Prior research has stated that analysts making bold forecasts incorporate their private information more completely and provide more relevant information to investors than those making herding forecasts (Clement and Tse, 2005).

To investigate the cross-section of analyst characteristics, I regress the stock-analyst-day information value on an indicator of forecast boldness, controlling for firm fixed effects, analyst fixed effects, and year fixed effects. I also control for broker size, as Jacob et al. (1999) suggests that larger brokerage firms provide better research. The analyst-specific information value measure is defined as

$$\widehat{\Omega}_{ijt} = \frac{r_{it}^2 - \left(r_{it} - \widehat{r}_{ijt} \right)^2}{\lambda_{it}} \cdot p_{it\_}. \tag{25}$$

Table 4 presents the results. In columns (1)-(4) of Panel A, the coefficient estimates on *Bold* are significantly positive, suggesting that analysts' information is more strategically valuable when they provide bold forecasts. On average, information from bold reports is 29% more valuable than that from herding reports. The magnitude remains largely unchanged when testing with

information value derived from report text. However, the insignificant coefficient estimates for *Brokersize* suggest that the value of analyst information is not necessarily tied to the size of the brokerage firm.

The subsequent research question is whether the association between boldness and information value is driven by higher information content or lower price impact. In Panel B, I regress the two decompositions of information value on the boldness indicator. As shown in columns (1)-(4), information from analysts who issue bold forecasts explains higher return volatility, supporting the argument of Clement and Tse (2005) that analysts provide bold forecasts when they possess relevant private information. In contrast, the coefficient estimates in columns (5)-(8) are insignificant, indicating that bold forecasts do not significantly affect price impact.

3.2.3 The Value of Information around Earnings Announcements

Section 3.3.3 reveals that information content is higher for analyst reports issued promptly following corporate earnings announcements. I, therefore, further study the information value of sell-side analysts around this most significant corporate event.

Figure 2 illustrates the information value of analyst reports between earnings announcements on a weekly basis, from the 1st week after the earnings announcement up to the 13th week. The information value of analyst reports is highest immediately following earnings announcements, particularly in the first week, peaking at a sample mean of $0.84 million. This pattern aligns with the previously observed trend in information content, suggesting that analyst reports offer the most valuable insights immediately after earnings releases. However, the information value sharply declines by the second week and remains relatively stable, with minor fluctuations, for the remainder of the 13-week period. Overall, the data suggests that the market places the most importance on analyst reports shortly after earnings announcements, with diminishing incremental value in the following weeks.

To formally validate the observed pattern, I conduct a regression analysis of log information value on an indicator variable *Week* and its interaction with earnings announcement day trading

volume ($TradingVolume$). This model is motivated by prior literature suggesting that analysts' interpretative role becomes more prominent with the complexity of financial accounting information (Chen et al., 2010). The trading volume serves as a proxy for investor disagreement, following Banerjee (2011) and Beckmann et al. (2024). The interaction term ($Week \times TradingVolume$) captures how the relationship between information value and post-earnings timing varies with the level of investor disagreement. To account for time-invariant stock characteristics and market-wide yearly trends, I include both stock and year-fixed effects in the model.

Table 5 presents regression results showing the influence of earnings announcements and trading volume on information value. Panel A indicates that the information value is significantly higher for reports released within one week of earnings announcements ($Week$), with positive and significant coefficients across all models. The interaction between Week and Trading Volume also shows a significant positive effect, suggesting that the information value increases with higher trading volumes around earnings announcements. The rationale is that investors may rely more on analyst opinions when there are large disagreements in the interpretation of earnings announcement information.

Panel B reveals significant increases in both log return variance and log price impact following earnings announcements, as evidenced by positive coefficients for the $Week$ variable. The interaction term ($Week \times TradingVolume$) is significantly positive in columns (1)-(4), indicating that explained return volatility rises with higher trading volumes during post-earnings periods. Notably, the increase in the information value of analyst reports following earnings announcements is primarily attributed to higher explained return volatility, rather than changes in price impact. These findings suggest that analyst reports provide more valuable insights in the immediate post-earnings period, particularly when there is higher trading activity. This increased value likely stems from analysts' ability to interpret complex earnings information and its implications, thereby explaining a larger portion of return volatility during these periods.

3.3 Information Content of Analyst Report Text

In this section, I provide yearly out-of-sample R^2 of analyst reports for explaining contemporaneous cumulative abnormal returns to examine the information content of quantitative and qualitative measures. I also compare the in-sample R^2 performance of analyst revisions and analyst report text. Furthermore, I attribute the information content of analyst report text across topics using a Shapley value approach.

3.3.1 Information Content of Quantitative and Qualitative Information

Table 6 presents a comparison of the information content of numerical and text input in terms of their out-of-sample R^2. The estimation model used in this table is ridge regression, where the estimation object is $CAR_{[-1,+1]}$ centered at the analyst report release date. For reports released on non-trading days, I adjust t=0 to the next trading day as recorded in the CRSP database. In Panel A, I compare the performances of models with four inputs: revision only, numerical only, text only, and revision plus text measures. The revision variables denote three quantitative measures of analyst revision regarding recommendations, earnings forecasts, and target prices. The numerical input includes all report-level, firm-level, industry-level, and macroeconomic measures described in Section 3.1. The text input represents the 5,120-dimensional full-context embedding of the LLaMA-2-13B model for each report. For each year, the training data consists of all reports starting from 2000 to the prior year. I estimate Ridge regression with the four sets of inputs and use the model to estimate $CAR_{[-1,+1]}$ yearly with expanding training sets.

Column (1) reports the R^2_{oos} for the revision input, which reaches an overall level of 9.01%. As well-documented in the literature (Brav and Lehavy, 2003; Asquith et al., 2005; Bradshaw et al., 2013), analyst forecast revisions contain news and impact contemporaneous stock prices. After including other numerical variables, the R^2_{oos} in column (2) exhibits a negligible increase, with the Ridge regression using 18 numerical features producing an R^2_{oos} of 9.06%. I compare the estimation with a zero benchmark and report the Diebold-Mariano test t-statistics in columns (2) and (4). The Diebold-Mariano test follows a standard normal distribution under the null hypothesis of equal

predictive accuracy between two competing forecasting models, allowing similar interpretations of t-statistics as in regression analysis (Gu et al., 2020). Thus, models with revision input and numerical input significantly outperform a zero-estimation benchmark model.

Columns (5) and (6) present the R^2_{oos} and DM t-statistic for the text embedding input. The text input exhibits a significant overall R^2_{oos} of 10.19% and a t-statistic of 8.20, underscoring the information content of report text in explaining contemporaneous stock returns. The magnitude of R^2_{oos} remains relatively stable across the period 2015-2023, with a noticeable drop in 2020 due to the pandemic recession.[15] The Diebold-Mariano test for comparing the revision-only and text-only inputs yields a t-statistic of 1.66, indicating that the information in the report text is substantial.

In columns (7) and (8), I further investigate the performance of combining both revision and text input in the Ridge regression. The combination achieves an R^2_{oos} of 12.28%, which represents a 3.27% improvement over the R^2_{oos} of the revision-only model and a 2.09% improvement over the R^2_{oos} of the text-only model. I report the DM t-statistic of the comparison between combined input with revision input and text input in columns (7)-(1) and (7)-(5). The overall t-statistics of 3.95 and 3.77 indicate a significant superiority in the combination of text data and revision measures in explaining contemporaneous *CAR*. The DM tests confirm that report text and forecast revision contain distinct information sets that are valued by market investors.[16]

LLMs are known to exhibit improved language skills as the size of their parameters increases. I provide a quantitative comparison of LLMs with varying numbers of parameters, including the BERT model, OpenAI GPT-3 embedding model, and LLaMA-3 model. The OpenAI and LLaMA models have a substantially larger number of parameters compared to the BERT model. Panel B reports the R^2_{oos} and DM t-stats of pairwise comparisons between each model and the LLaMA

[15] Sarkar and Vafa (2024) raise the look-ahead bias concern associated with pre-trained LLMs. Language model outputs may incorporate future information that should not be available. Since the training sample of models I employ ends before 2023, the R^2_{oos} in 2023 provides a true out-of-sample performance. The main findings are robust in this period.

[16] To address the concern that the explanatory power might be derived from quantitative information in analyst reports, I conduct robustness checks using un-numbered report content. In Table A4, I remove numbers from analyst reports before generating LLaMA2 embeddings. The R^2_{oos} of the number-free text embedding input model is 10.95%, remaining significantly higher than quantitative revision measures (with a t-stat of 2.40). This finding suggests that the superior performance of the text-based model is not driven by numerical information in the reports.

2 model. The input in Panel B is report text, comparable to column (3) in Panel A. A positive t-statistic indicates that the model outperforms the benchmark, which is the LLaMA 2 model in the panel. Overall, the BERT model and OpenAI underperform the LLaMA2 model at the 1% level for each year, while LLaMA3 exhibits a slightly better yet statistically insignificant performance with an R^2_{oos} of 9.66%. The results in Table 6 suggest that the out-of-sample information content of analyst report text is fairly robust with respect to different LLMs. The pairwise comparisons between LLMs also shed light on how the scale of the model influences its performance capabilities.[17]

In Table A6, I present the information content of analyst report text across the Fama-French 12 industries. The R^2_{oos} exceeds 10% in five industries: Shops, Other, Manufacturing, Chemicals, and Durables. Notably, the magnitude and t-statistics are significantly lower in the Non-Durables, Telecommunications, Energy, and Utilities sectors. The industries with higher R^2_{oos} typically involve complex products and services that benefit from detailed analyst insights, whereas those with lower R^2_{oos} may operate in more stable or regulated environments where analyst reports have less incremental value. In Figure A6, I plot the information value of analysts by industry, revealing a pattern that aligns with the information content results.

To get a sense of economic magnitude, I run the following empirical tests with OLS regression:

$$CAR_{[-1,+1],it} = \alpha + \beta_1 \widehat{CAR}_{txt,it} + \beta_2 REC\_REV_{it} + \beta_3 EF\_REV_{it} + \beta_4 TP\_REV_{it} + \varepsilon_{it}. \tag{26}$$

In Equation 26, I include the out-of-sample rolling estimation of $CAR_{[-1,+1]}$ using report text embeddings as input. For example, to fit observations in the year 2015, the Ridge model is trained with samples from 2000 to 2014. The fitted value of $\widehat{CAR}_{txt}$ condenses the 5120-dimension text information into a single dimension. For comparison, I estimate $\widehat{CAR}_{rev}$ using the three forecast revision measures, following the same procedure.

[17] To test the robustness of the information content in these contemporaneous measures, I also used *CAR* measures in a 2-day window and a 5-day window, as applied in Huang et al. (2014) and Asquith et al. (2005), respectively. Additionally, I examined the abnormal stock return on the day of the report release. In Table A5, I find a similar magnitude in R^2_{oos} for explaining $AR_{[0]}$, $CAR_{[0,+1]}$, and $CAR_{[-2,+2]}$, ranging from 6.92% to 7.57% for the out-of-sample period from 2015 to 2023.

Table 7 summarizes the summary statistics of the sample used in the regressions and the estimated results. The out-of-sample period starts in 2015. I constrain the samples to include only those with valid previous and present stock recommendations, earnings forecasts, and target prices. The final sample for estimating Equation 26 contains 28,837 observations.

The regression results reported in column (1) of Panel B indicate that quantitative measures of recommendation revisions, earnings forecast revisions, and target price revisions all account for some variation in contemporaneous abnormal stock returns, which aligns with the findings in Brav and Lehavy (2003) and Huang et al. (2014). Column (2) reports the results of $\widehat{CAR}_{txt}$, which alone yields an R^2 of 10.5%, close to the R^2_{oos} value in Table 6. In column (3), I include both analyst revision measures and $\widehat{CAR}_{txt}$ as regressors. The coefficient estimate of REC_REV is no longer significant. This implies that recommendation revisions are prone to be captured in the narrative of analyst reports. The information from the report text raises R^2 from 9.1% to 15.6%, establishing the incremental value of text information.

The positive and significant coefficient on $\widehat{CAR}_{txt}$ indicates that the information in the report text is not subsumed by quantitative analyst revisions. Economically, a one standard deviation increase in earnings forecast revision, target price revision, and report text favorableness increases the three-day abnormal return by 33.95 basis points, 91.12 basis points, and 122.58 basis points, respectively. In terms of economic magnitude, this demonstrates the significant impact of text information on stock returns.

In columns (4) and (5) of Panel B, I report the coefficient estimates for $\widehat{CAR}_{rev}$, which combines the information from all revision metrics. The $\widehat{CAR}_{rev}$ model alone achieves an R^2 of 8.9%, while the combined input model improves this to 15.4%. Economically, a one standard deviation increase in $\widehat{CAR}_{rev}$ and $\widehat{CAR}_{txt}$ corresponds to increases of 101.78 basis points and 122.24 basis points in CAR, respectively. Overall, the information from the report text is more significant than the information in forecast revisions, both statistically and economically.

3.3.2 What Content in Analyst Reports is Valued by Market Investors?

Given the broad and comprehensive discussion in equity reports, what specific content in analyst reports is valuable to the market? I employ the Shapley value metric in Section 2.4 to attribute the R^2_{oos} of the full report to 17 topics extracted by ChatGPT. In this section, the report embeddings are taken as the token-weighted average of each sentence embedding. The sentence-segmented embedding eliminates the concern of contextual information from topic B being captured by embeddings of tokens in topic A. The R^2_{oos} of text input decreases by 2.25% when I alternate from full-context embeddings to sentence-segmented embeddings, roughly quantifying the magnitude of the information content of context in analyst reports.

Intuitively, Shapley values are calculated as the reduction in R^2_{oos} from deriving the embedding of a specific topic within all possible topic combinations and averaging these into a single importance measure. By design, the summation of Shapley values of 17 topics exactly equals the R^2_{oos} of sentence-segmented report embeddings. Figure 4 ranks the Shapley values of 17 topics. I order the topics by their Shapley values so that the most important topics are on the left and the least important ones are displayed on the right. The color gradient (from lightest to darkest) within each bar indicates the magnitude of topic importance in terms of Shapley values.

As demonstrated in Figure 4, the most valuable topics in analyst reports are Income Statement Analyses and Financial Ratio Analyses, each comprising 67% and 45% of text-based R^2_{oos}. The second most important categories are Investment Thesis and Valuation, with Shapley values of 1.76% and 1.51%, respectively. The remaining topics exhibit a minor or even negative Shapley value, which means they can barely explain the market's contemporaneous reaction to analyst reports. Figure 4 provides a distinct picture of topic importance, demonstrating the value of analyst income statement analyses and financial ratio analyses to the market.[18]

Figure A4 in the Online Appendix shows the text-based R^2_{oos} measure each year. The relative magnitude of R^2_{oos} closely matches that of full-context embedding reported in Table 6. The sharp

[18] In Figure A7, I scale the Shapley values by the number of sentences and length of tokens, the relative ranking of Income Statement Analyses remains unchanged.

drop in the R^2_{oos} between 2020 and 2021 can be attributed to the COVID-19 pandemic and its impact on financial markets and the economy. The red line of R^2_{oos} scaled Shapley value of the Income Statement category is relatively flat across years, indicating stable information content of analyst discussion on corporate income statements. In Figure A5, I find a similar pattern of topic importance across industries.

I further proceed to study the interpretation role and discovery role of analysts in analyzing income statements. Specifically, analysts play an information discovery role to the extent that they 'collect and generate information that is otherwise not readily available to investors' (Huang et al., 2018). Analysts' interpretation role refers to their ability to integrate available information and process it into a more interpretable and clear signal (Blankespoor et al., 2020). Inspired by the literature, I further categorize sentences in the Income Statement Analyses topic into two groups of sub-topics: (1) Income Acquisition versus Income Interpretation; (2) Income Realization versus Income Expectation using the following prompts:

Prompt 3: Please read the following sentence from a sell-side analyst report for the company {firm} ({ticker}) and classify it based on two criteria:

Information Type:

Information Acquisition (0): Sentences that directly report quantitative financial data, such as earnings, revenue, expenses, or other metrics. Example: 'Reported EPS of $1.40 beat consensus of $0.94 and our estimate of $1.17.'

Information Integration (1): Sentences that provide analysis, context, or interpretation of the financial data, such as discussing trends, comparing performance to forecasts, assessing market impacts, or considering strategic implications. Example: 'Despite compounded double-digit rate increases in commercial P&C pricing, reserve shortfalls from poor 1997-2000 underwriting and 1970s asbestos and environmental losses kept insurance stock prices in check.'

Time Reference:

Actual (0): Sentences referring to concrete, historical results from completed periods. Example: 'The company's revenue for the fiscal year 2022 was $500 million.'

Prompt 3: Please read the following sentence from a sell-side analyst report for the company {firm} ({ticker}) and classify it based on two criteria:
Forecast (1): Sentences containing subjective predictions, or expectations for future periods. Example: 'The company expects to achieve revenue of $550 million in fiscal year 2023 based on current market conditions.'
Output your classification as two comma-separated numbers: the first for Information Type (0 for Acquisition, 1 for Integration) and the second for Time Reference (0 for Actual, 1 for Forecast).
Output format: InfoType, TimeRef
Sentence to classify: {report_sent}.

In Figure 5, the R^2_{oos} of sentence-segmented text embeddings is attributed to 18 topics, with the Income Statement category being fully decomposed into two sub-topics. By the additive nature of Shapley values, the sum of the Shapley values of the two sub-topics equals the Shapley value of the Income Statement topic. The decomposition of topic importance demonstrates the value of analyst interpretation of realized income. Overall, I find that the information interpretation role is more valuable in analysts' analyses of corporate income statements.

3.3.3 Comparison of Earnings Announcement and Non-earnings Announcement Periods

A number of studies have emphasized the role of analysts surrounding corporate earnings announcements (Frankel et al., 2006; Chen et al., 2010; Huang et al., 2018), in terms of how these two information events reinforce each other in sequence. As the previous evidence underlines the value of income statement discussion by analysts, in the next stage, I evaluate the information content of analyst reports issued promptly following earnings announcements and beyond that.

First, I follow Chen et al. (2010) and group analyst reports into 13 weekly bins surrounding earnings announcement dates (EADs). I then aggregate the information content of analyst reports by week. Panel A in Table 8 presents the R^2_{oos} and DM t-stats for each week. A finding

consistent with Huang et al. (2018) is that an overwhelming number of analyst reports are released immediately following the earnings announcements. Another notable feature is that the information content is only significant in the first week, approximating a R^2_{oos} of 10% and a notable t-stat of 10.17. The performance disappears in the second week, turns negative in the third and fourth weeks, and never rebounds.

The distinct performance observed in weekly bins can be attributed to Ridge models effectively capturing the dominant patterns in analyst reports during earnings announcement weeks. Consequently, I train Ridge models using separate samples for periods surrounding and beyond earnings announcements, enabling the models to extract distinct information content based on the timing of the reports. Panel B in Table 8 compares the model performance during earnings announcement periods and non-earnings announcement periods, using 1-day, 2-day, 3-day, and 7-day windows. The R^2_{oos} of models fitted during earnings announcement periods is consistently higher than those fitted outside these periods. The R^2_{oos} and t-statistics for models trained within these periods diminish with longer windows, indicating that analysts react promptly to earnings announcements, as the information content declines from 11.84% to 7.29% over the course of the earnings announcement week.

If analysts merely restate the information from earnings announcements, the information content measure of their reports could reflect this text information in earnings conference calls. To examine this, I used earnings conference call transcripts downloaded from Seeking Alpha to differentiate the information content between the two text sources. The sample was narrowed to reports released promptly within a one-day window following earnings announcements, as this is when piggybacking on earnings announcements is most likely to occur.

In column (1) of Panel C, I present the R^2_{oos} for $CAR_{[-1,+1]}$ using embeddings from earnings conference call transcripts as input.[19] The R^2_{oos} is 4.20% for the S&P 100 firms, with a DM t-stat of 5.16. with a Diebold-Mariano (DM) t-statistic of 5.16. Since the conference call transcript

[19]The pre-trained LLM used for generating text embeddings in Panel C is LLaMA-3-8B, considering the long text lengths of earnings conference call transcripts (with a median length of 13,029 tokens). Specifically, the maximum input lengths of the LLaMA-2-13B model and the LLaMA-3-8B model are 4,096 tokens and 8,192 tokens, respectively. I truncate the text at the length limits to maintain maximum information.

data began in 2005, I retrained the Ridge regression model with analyst report text input to align the sample periods of both texts. The input is the mean of all report text embeddings on the day following the earnings announcement.

The firm-time level Ridge regression delivers an R^2_{oos} of 9.72%, as shown in column (2). When incorporating both analyst report embeddings and earnings conference call embeddings in the Ridge regression, there is a significant improvement in R^2_{oos} to 11.96%. The comparison t-statistic between the transcript-only model and the combined model is 5.12, as indicated in column (4), demonstrating a significant incremental value.

Overall, the results in Table 8 highlight the information content of analyst reports released promptly following earnings announcements, as evidenced by the better performance in R^2_{oos} within a narrow window post-earnings-announcement. Moreover, reports released beyond earnings announcement periods also convey valuable information to the market, as indicated by a significant R^2_{oos} that is comparable in magnitude to that of earnings announcement transcripts.

3.3.4 Comparison of Revision Reports and Reiteration Reports

Researchers have documented a robust relationship between market prices and analysts' forecast revisions (e.g., Gleason and Lee, 2003; Asquith et al., 2005; Bradshaw et al., 2021). Analysts' forecast revisions provide a significant and timely source of information to the financial market and may constitute a more valuable signal than reports that merely reiterate previous forecasts.

Therefore, I compare the information content of analyst report text associated with forecast revisions and forecast reiterations. Following the criteria of Huang et al. (2014), I consider a report as a revision report if its release date is within two days of the I/B/E/S forecast announcement date[20]. For each forecast target in recommendations, target prices, and earnings forecasts, I categorize the report independently. In this procedure, a report may be classified as both an earnings forecast revision report and a recommendation reiteration report.

[20]In I/B/E/S, each analyst forecast for a specific firm has two time stamps. The announcement date (ANNDATS) is the date the analyst announces the new forecast and revises the old one. The revision date (REVDATS) is the date the forecast becomes invalid or is revised.

Table 9 summarizes the out-of-sample performance of numerical and textual input in the reiteration sample and the revision sample. Recommendation revisions are rare events compared to target price revisions and earnings forecast revisions. In such events, the R^2_{oos} of revision numerical measures alone reaches 14.99%. Report text in recommendation revision analyst reports explains 16.80% of CAR variation. The combination of both quantitative and qualitative information sets improves upon this further, generating an R^2_{oos} of 22.63%. This finding is consistent with the literature that attributes large stock price changes to analyst recommendation revisions (Loh and Stulz, 2011; Bradley et al., 2014).

Target price revisions are considered incrementally informative due to discreteness and potential biases in analysts' recommendations (Bradshaw, 2002; Brav and Lehavy, 2003). In Table 9, the information content of analyst reports in target price reiteration conditions drops sharply. The combination of numerical and text information explains 3.58% of three-day CAR. However, conditional on target price revisions, both numerical measures and text information are significantly valuable. The combined information value is evident in the economically large magnitude of R^2_{oos} (20.88%, t-stats = 11.18).

The analysis of earnings forecast target sub-samples reveals distinct patterns. Earnings forecast revisions are the most frequent, making up 56.57% of the reports, while recommendation revision reports and target price revision reports constitute only 2.29% and 31.10%, respectively. Analysts' earnings forecast revisions convey both public and private information about cash flow news in financial markets (Kothari et al., 2016) and are the most studied type of analyst output. I find that the R^2_{oos} of report text in earnings forecast reiteration reports is negative, suggesting no explanatory power for contemporaneous stock returns. However, when considering analyst earnings forecast revisions, the R^2_{oos} for both numerical and textual inputs is significant. The combination of both inputs delivers an R^2_{oos} of 16.37% (t-stat = 11.42).

Taken together, Table 9 reveals that report content is more informative when it accompanies forecast revisions rather than reiterations. The information content of numerical measures and text embeddings is complementary. Additionally, analyst reports are most valuable when associated

with recommendation revisions.

3.3.5 Comparison with Tone in Analyst Report Text

Huang et al. (2014) emphasize the importance of analyst report text by extracting the tone of sentences using a Naive Bayes machine learning approach. They manually classified 10,000 randomly selected sentences into positive, negative, or neutral categories to create a training dataset. The Naive Bayes algorithm then learns the probabilistic relationships between words and opinion categories from this training data, based on the frequency of words appearing in each category. Although more advanced in its learning algorithm than BoW and W2V, the Naive Bayes approach remains a word-count-based method, which can be susceptible to context-related misunderstandings. In this section, I compare the relative performance of report text tone and report text embedding using the R^2_{oos} measure.

First, I construct report-level tone measures using both the Naive Bayes approach and BERT classification. Following Huang et al. (2014), I manually label 10,000 sentences as a training sample to train both the Naive Bayes model and the BERT model. For each sentence, I categorize its tone into positive, neutral, or negative using the pre-trained models. I then average the tone at the report level to derive the *Tone_NB* and *Tone_BERT* measures, respectively. Additionally, I constructed tone measures separately within Income Statement topics and non-Income Statement topics.

Table 10 reports the comparison of performance between text tone and text embedding. Panel A summarizes the OLS regression of three-day *CAR* on the tone measures. The coefficient estimate for *Tone_NB* in column (1) is 0.019, comparable to the 0.0208 reported by Huang et al. (2014). A one standard deviation increase in the favorableness of the text tone results in an additional *CAR* of 49 basis points, which is comparable with the 41 basis points increase in two-day abnormal return reported by Huang et al. (2014). Column (2) summarizes the results of regressing *CAR* on tone measures extracted within Income Statement topics and non-Income Statement topics. Both coefficients of tone are significantly positive. In column (3), I report

the regression results with firm-fixed effects and year-fixed effects, which are consistent with the findings in column (2). Columns (4)-(6) present the regression results with BERT-based tone measures as explanatory variables. The magnitude of coefficient estimates increases dramatically, along with the level of significance. In terms of economic magnitude, a one standard deviation increase in the favorableness of the BERT-based text tone measure results in an additional three-day *CAR* of 86 basis points. When both measures of text tone are included in an OLS regression, the coefficient estimates for Naive Bayes-based tone measures are no longer significant, suggesting that the BERT-based tone measures capture more nuanced information.

Confirming the value of text tone conveyed in Huang et al. (2014), I proceed to compare the relative information content of extracted tone measures and embedding representations of report text. Panel B of Table 10 presents the performance of different text measures by running Ridge regression with input containing tone measures, both tone-only and tone-plus. The R^2_{oos} of tone inputs ranges from 0.05% to 3.68% in the first four rows, with similar magnitudes of in-sample adjusted R^2. The information content increases from 9.49% to 12.27% after incorporating analyst revision and text embeddings. This further confirms the superior performance of state-of-the-art LLMs, which dominate benchmarks across various NLP tasks (Chen et al., 2022). The bottom two rows of Panel B report comparisons of R^2_{oos} and DM tests between models with "Tone + numerical + text" versus "Numerical + text." The difference in R^2_{oos} between the two models is not statistically distinguishable from zero for both Naive Bayes-based and BERT-based tone measures.

In general, the empirical results in Table 10 support two key findings regarding the superiority of LLMs compared to conventional machine learning approaches. First, the overall report text tone measure extracted using small-size LLMs is more explanatory for contemporaneous stock returns than that extracted using the Naive Bayes algorithm. Second, the tone measures (extracted with both Naive Bayes and BERT) add no incremental value to text embeddings, suggesting a more comprehensive contextualized representations provided by LLMs.

3.3.6 Information Content with Machine Learning Models

In this Section, I discuss the main results based on the four machine learning algorithms described in Section 4. For comparison, I follow all processes applied in Ridge regressions. For each machine learning model, the input consists of full-context report embeddings, with the target variable being the three-day *CAR*. To ensure out-of-sample evaluation, I train the models on a yearly basis and obtain estimations for the subsequent year's sample. The R^2_{oos} is averaged using the estimated and realized values of *CAR*. Additionally, I conduct pairwise Diebold-Mariano tests using a zero benchmark to assess the predictive accuracy of the models.

I begin with PLS regression results summarized in Table 11. The primary advancement in the domain of PLS is its ability to extract the most relevant common components from sentiment proxies that align with the objective of fitting the target variable (Huang et al., 2015). If there is a strong common component in the text embeddings that is informative of stock returns, the PLS regression architecture should effectively capture this relevant information. Examining the model performance, I observe that PLS regressions produce an overall R^2_{oos} of 8.26%. This measure of information content ranges from 2.14% in the 2021 sample to 12.23% in the 2015 sample, showing a similar pattern across years as seen in Ridge models. The Diebold-Mariano test statistics indicate that the R^2_{oos} is significant at the 1% level across all periods. Thus, the two linear regression methods provide highly consistent evidence of information content in the analyst report text.

Next, I replicate the analyses with the XGBoost algorithm. As introduced in Section 2, Gradient boosting combines weak estimations to form a strong estimation by constructing sequential trees. Each tree is responsible for training a small sample to fit the residuals of the previous trees and updating the model by learning new information. Therefore, if the high-dimensional feature spaces of text embeddings contain segmented information with discrete importance, XGBoost can potentially identify the most informative aspects of the analyst report embeddings for explaining contemporaneous stock returns. Compared to Ridge and PLS models, XGBoost is capable of capturing complex nonlinear relationships between analyst report embeddings and the target variable. As summarized in Table 11, the XGBoost algorithm

underperforms with an overall R^2_{oos} of 5.97% (DM t-stat = 6.76). The relatively poor performance could be attributed to the inherent complexity of language, where information is distributed across the embedding space rather than concentrated in specific dimensions.

Neural Networks, forming the core of large language models (LLMs) based on the transformer architecture, have shown significant advancements in various natural language processing (NLP) tasks (Cao et al., 2024; Li et al., 2023). In a Neural Network, nodes are interconnected, mimicking the neuron structure of the human brain. Since embeddings are encoded as nodes within the Neural Network layers, I expect this architecture to capture long-range relationships between words and the complex contextual information embedded in the text. I construct the Neural Network estimations as an ensemble of five models, each containing 1-5 layers. As shown in Table 11, the R^2_{oos} of Neural Network models ranges from 10.42% at the lowest end to 12.12% at the highest end, providing the best performance among the four categories of machine learning models. In line with the virtue of complexity demonstrated by Kelly et al. (2022), I observe that model performance generally increases with the number of layers.

The experiments with machine learning models provide noteworthy results. First, the information content of report text, as measured by R^2_{oos}, is approximately 10% across various models. This is significant given the low signal-to-noise ratio in stock returns. Second, the performance and characteristics of these machine learning models highlight that contextual deep learning and interdependencies between words are crucial for accurately representing and evaluating textual information.

4 Conclusion

In this paper, I quantify the dollar value of strategic informed traders receiving tipping from analysts. Using a measure of explained return volatility relative to price impact, I find the annualized expected profit for strategic investors to be informed about the analyst reports of an average S&P 100 constituent firm is $6.89 million dollars. Cross-sectional analysis reveals that the

information value from analysts is higher for large stocks, bold forecasts, and reports released in the week following earnings announcements.

I examine both quantitative and qualitative content of analyst reports using out-of-sample R^2. The results show that report text contains more market-valued information content than quantifiable forecast revisions, both statistically and economically. Combining both sources of information yields an R^2_{oos} of 15.6%. The robustness of report text R^2_{oos}, around 10%, is consistent across various models and *CAR* windows. I also find that 'deep' learning outperforms 'shallow' learning in extracting text information. Neural Network models increase the R^2_{oos} of text input from 10.42% in NN1 to 12.12% in NN5. I also observe the information from analyst report text exhibits a strong relationship with stock returns both within and beyond earnings announcement periods, while the magnitude of R^2_{oos} is about twice as large during earnings announcement periods compared to non-earnings periods.

Leveraging a Shapley value decomposition framework and CoT prompting with LLMs, I find that Income Statement Analyses and Financial Ratio Analyses are the most valuable topics among a taxonomy of 17 information content categories, contradicting the findings in Huang et al. (2014) that non-financial topics are emphasized by investors. Moreover, embedding representations provide a more systematic and comprehensive analysis of report text than sentiment analysis based on limited word-count algorithms used in Asquith et al. (2005) and Huang et al. (2014).

Overall, the trading profits derived from analyst information appear to be economically meaningful. It's important to note that this dollar value is estimated using a sample of S&P 100 stocks, so readers should exercise caution when applying these findings to broader markets. The primary objectives of this article are threefold: to demonstrate the value of information from sell-side analysts, to propose a straightforward framework for quantifying the dollar value of a specific information set, and to emphasize the importance of written reports as a prominent output from analysts.

References

Asquith, Paul, Michael B Mikhail, and Andrea S Au, 2005, Information content of equity analyst reports, *Journal of financial economics* 75, 245–282.

Back, Kerry, C Henry Cao, and Gregory A Willard, 2000, Imperfect competition among informed traders, *The journal of finance* 55, 2117–2155.

Banerjee, Snehal, 2011, Learning from prices and the dispersion in beliefs, *The Review of Financial Studies* 24, 3025–3068.

Barber, Brad, Reuven Lehavy, Maureen McNichols, and Brett Trueman, 2001, Can investors profit from the prophets? security analyst recommendations and stock returns, *The Journal of finance* 56, 531–563.

Barber, Brad M, and Douglas Loeffler, 1993, The "dartboard" column: Second-hand information and price pressure, *Journal of Financial and Quantitative Analysis* 28, 273–284.

Barron, Orie E, Donal Byard, and Yong Yu, 2017, Earnings announcement disclosures and changes in analysts' information, *Contemporary Accounting Research* 34, 343–373.

Beckmann, Lars, Heiner Beckmeyer, Ilias Filippou, Stefan Menze, and Guofu Zhou, 2024, Unusual financial communication-evidence from chatgpt, earnings calls, and the stock market, *Earnings Calls, and the Stock Market (January 15, 2024)* .

Blankespoor, Elizabeth, Ed deHaan, and Ivan Marinovic, 2020, Disclosure processing costs, investors' information choice, and equity market outcomes: A review, *Journal of Accounting and Economics* 70, 101344.

Bonini, Stefano, Thomas Shohfi, Majeed Simaan, and Guofu Zhou, 2023, The value of data: Analyst vs. machine, *Machine (December 11, 2023)* .

Bradley, Daniel, Jonathan Clarke, Suzanne Lee, and Chayawat Ornthanalai, 2014, Are analysts' recommendations informative? intraday evidence on the impact of time stamp delays, *The Journal of Finance* 69, 645–673.

Bradshaw, Mark, Yonca Ertimur, Patricia O'Brien, et al., 2017, Financial analysts and their contribution to well-functioning capital markets, *Foundations and Trends® in Accounting* 11, 119–191.

Bradshaw, Mark T, 2002, The use of target prices to justify sell-side analysts' stock recommendations, *Accounting Horizons* 16, 27–41.

Bradshaw, Mark T, Lawrence D Brown, and Kelly Huang, 2013, Do sell-side analysts exhibit differential target price forecasting ability?, *Review of Accounting Studies* 18, 930–955.

Bradshaw, Mark T, Brandon Lock, Xue Wang, and Dexin Zhou, 2021, Soft information in the financial press and analyst revisions, *The accounting review* 96, 107–132.

Brav, Alon, and Reuven Lehavy, 2003, An empirical analysis of analysts' target prices: Short-term informativeness and long-term dynamics, *The Journal of Finance* 58, 1933–1967.

Brown, Stephen, Kin Lo, and Thomas Lys, 1999, Use of r2 in accounting research: measuring changes in value relevance over the last four decades, *Journal of Accounting and Economics* 28, 83–115.

Cao, Sean, Wei Jiang, Junbo Wang, and Baozhong Yang, 2024, From man vs. machine to man+ machine: The art and ai of stock analyses, *Journal of Financial Economics* 160, 103910.

Chakrabarty, Bidisha, Bingguang Li, Vanthuan Nguyen, and Robert A Van Ness, 2007, Trade classification algorithms for electronic communications network trades, *Journal of Banking & Finance* 31, 3806–3821.

Chen, Xia, Qiang Cheng, and Kin Lo, 2010, On the relationship between analyst reports and corporate disclosures: Exploring the roles of information discovery and interpretation, *Journal of Accounting and Economics* 49, 206–226.

Chen, Yifei, Bryan T Kelly, and Dacheng Xiu, 2022, Expected returns and large language models, *Available at SSRN 4416687* .

Christophe, Stephen E, Michael G Ferri, and Jim Hsieh, 2010, Informed trading before analyst downgrades: Evidence from short sellers, *Journal of Financial Economics* 95, 85–106.

Clement, Michael B, and Senyo Y Tse, 2005, Financial analyst characteristics and herding behavior in forecasting, *The Journal of finance* 60, 307–341.

Diebold, Francis X, and Robert S Mariano, 2002, Comparing predictive accuracy, *Journal of Business & economic statistics* 20, 134–144.

Ellis, Katrina, Roni Michaely, and Maureen O'hara, 2000, When the underwriter is the market maker: An examination of trading in the ipo aftermarket, *The Journal of Finance* 55, 1039–1074.

Frankel, Richard, SP Kothari, and Joseph Weber, 2006, Determinants of the informativeness of analyst research, *Journal of Accounting and Economics* 41, 29–54.

Gleason, Cristi A, and Charles MC Lee, 2003, Analyst forecast revisions and market price discovery, *The Accounting Review* 78, 193–225.

Green, T Clifton, 2006, The value of client access to analyst recommendations, *Journal of Financial and Quantitative Analysis* 41, 1–24.

Groysberg, Boris, Paul M Healy, and David A Maber, 2011, What drives sell-side analyst compensation at high-status investment banks?, *Journal of Accounting Research* 49, 969–1000.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala, 2014, Product market threats, payouts, and financial flexibility, *The Journal of Finance* 69, 293–324.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong Zheng, 2018, Analyst information discovery and interpretation roles: A topic modeling approach, *Management science* 64, 2833–2855.

Huang, Allen H, Amy Y Zang, and Rong Zheng, 2014, Evidence on the information content of text in analyst reports, *The Accounting Review* 89, 2151–2180.

Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou, 2015, Investor sentiment aligned: A powerful predictor of stock returns, *The Review of Financial Studies* 28, 791–837.

Irvine, Paul, Marc Lipson, and Andy Puckett, 2007, Tipping, *The Review of Financial Studies* 20, 741–768.

Ivković, Zoran, and Narasimhan Jegadeesh, 2004, The timing and value of forecast and recommendation revisions, *Journal of Financial Economics* 73, 433–463.

Jacob, John, Thomas Z Lys, and Margaret A Neale, 1999, Expertise in forecasting performance of security analysts, *Journal of Accounting and Economics* 28, 51–82.

Jegadeesh, Narasimhan, Joonghyuk Kim, Susan D Krische, and Charles MC Lee, 2004, Analyzing the analysts: When do recommendations add value?, *The journal of finance* 59, 1083–1124.

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang, 2024, Chatgpt and corporate policies, Technical report, National Bureau of Economic Research.

Kadan, Ohad, and Asaf Manela, 2020, Liquidity and the strategic value of information, *Available at SSRN 3645137* .

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou, 2022, The virtue of complexity everywhere, *Available at SSRN 4166368* .

Keskek, Sami, Senyo Tse, and Jennifer Wu Tucker, 2014, Analyst information production and the timing of annual earnings forecasts, *Review of Accounting Studies* 19, 1504–1531.

Kim, Yongtae, and Minsup Song, 2015, Management earnings forecasts and value of analyst forecast revisions, *Management Science* 61, 1663–1683.

Kothari, Sabino P, Xu Li, and James E Short, 2009, The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis, *The Accounting Review* 84, 1639–1670.

Kothari, Sagar P, Eric So, and Rodrigo Verdi, 2016, Analysts' forecasts and asset pricing: A survey, *Annual Review of Financial Economics* 8, 197–219.

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica: Journal of the Econometric Society* 1315–1335.

Lee, Charles MC, and Mark J Ready, 1991, Inferring trade direction from intraday data, *The Journal of Finance* 46, 733–746.

Li, Kai, Feng Mai, Rui Shen, Chelsea Yang, and Tengfei Zhang, 2023, Dissecting corporate culture using generative ai–insights from analyst reports, *Available at SSRN 4558295* .

Livnat, Joshua, and Yuan Zhang, 2012, Information interpretation or information discovery: Which role of analysts do investors value more?, *Review of Accounting Studies* 17, 612–641.

Lo, Kin, and Thomas Z Lys, 2000, Bridging the gap between value relevance and information content, *Sauder School of Business Working Paper* .

Lobo, Gerald J, Minsup Song, and Mary Harris Stanford, 2017, The effect of analyst forecasts during earnings announcements on investor responses to reported earnings, *The Accounting Review* 92, 239–263.

Loh, Roger K, and René M Stulz, 2011, When are analyst recommendation changes influential?, *The review of financial studies* 24, 593–627.

Lundberg, Scott M, and Su-In Lee, 2017, A unified approach to interpreting model predictions, *Advances in neural information processing systems* 30.

Merton, Robert C, 1974, On the pricing of corporate debt: The risk structure of interest rates, *The Journal of finance* 29, 449–470.

Michaely, Roni, and Kent L Womack, 1999, Conflict of interest and the credibility of underwriter analyst recommendations, *The review of financial studies* 12, 653–686.

Previts, Gary John, Robert J Bricker, Thomas R Robinson, and Stephen J Young, 1994, A content analysis of sell-side financial analyst company reports, *Accounting Horizons* 8, 55.

Sarkar, Suproteem K, and Keyon Vafa, 2024, Lookahead bias in pretrained language models, *Available at SSRN* .

Shapley, Lloyd S, 1953, A value for n-person games .

Sorescu, Sorin, and Avanidhar Subrahmanyam, 2006, The cross section of analyst recommendations, *Journal of Financial and Quantitative Analysis* 41, 139–168.

Stickel, Scott E, 1995, The anatomy of the performance of buy and sell recommendations, *Financial Analysts Journal* 51, 25–39.

Twedt, Brady, and Lynn Rees, 2012, Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports, *Journal of Accounting and Public Policy* 31, 1–21.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need, *Advances in neural information processing systems* 30.

Womack, Kent L, 1996, Do brokerage analysts' recommendations have investment value?, *The journal of finance* 51, 137–167.

Zeghal, Daniel, 1984, Firm size and the informational content of financial statements, *Journal of Financial and Quantitative Analysis* 19, 299–310.

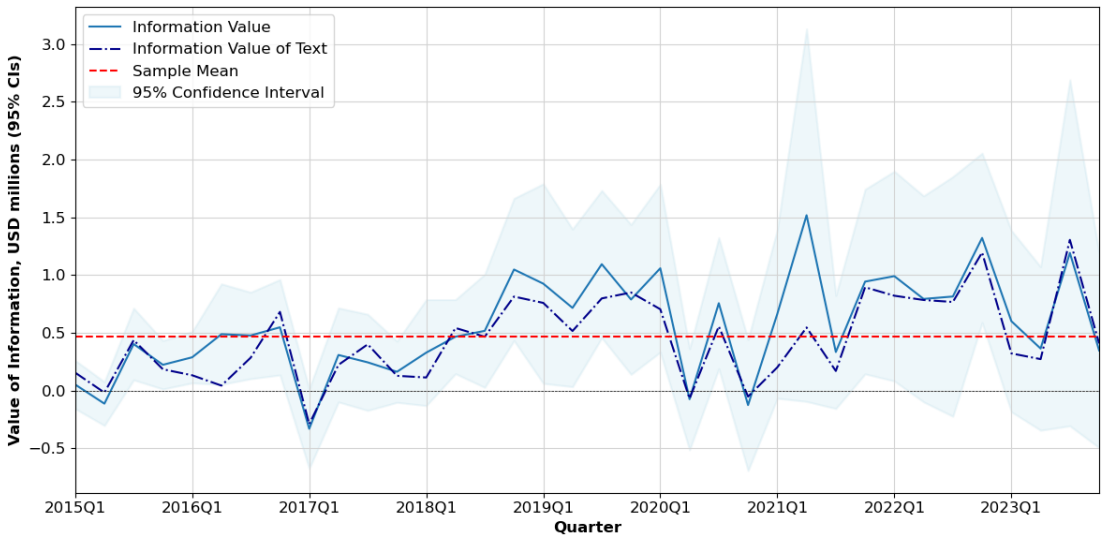


Figure 1 Dollar Value of Analyst Reports Over Time

This figure presents the estimated value of information, reported in millions of dollars and adjusted to 2020 dollars, derived from analyst reports between the first quarter of 2015 (2015Q1) and the last quarter of 2023 (2023Q4). Each quarter-mean approximation is calculated using the delta method. The solid line represents the strategic value of both numerical and text information from analyst reports, while the dashed line indicates the strategic value of report text information. the horizontal line indicates the sample mean over the 2015 to 2023 period. The light blue shaded area denotes the 95% confidence interval.

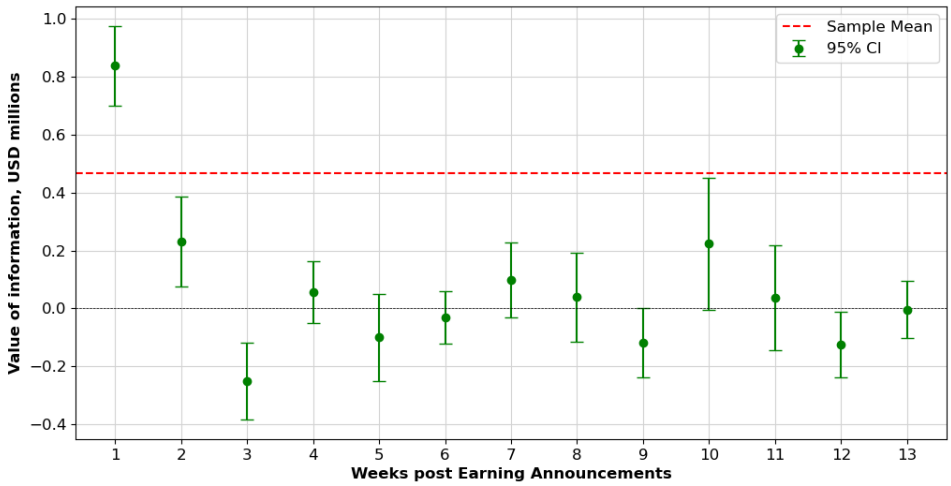


Figure 2 Dollar Value of Analyst Reports after Earnings Announcements

This figure presents the estimated information value, reported in millions of dollars and adjusted to 2020 dollars, derived from analyst reports released 1-13 weeks following the most recent earnings announcement. The sample spans from 2015 to 2023. The vertical axis indicates the value of information, while the horizontal axis represents weeks post-earnings announcements. The solid green line denotes the sample mean value of information, with the accompanying green bars representing the 95% confidence intervals. Each point reflects the average value for the respective week, calculated using the delta method. The nodes represent the strategic value derived from both numerical and textual information in analyst reports. The red dashed horizontal line represents the overall sample mean for the period.

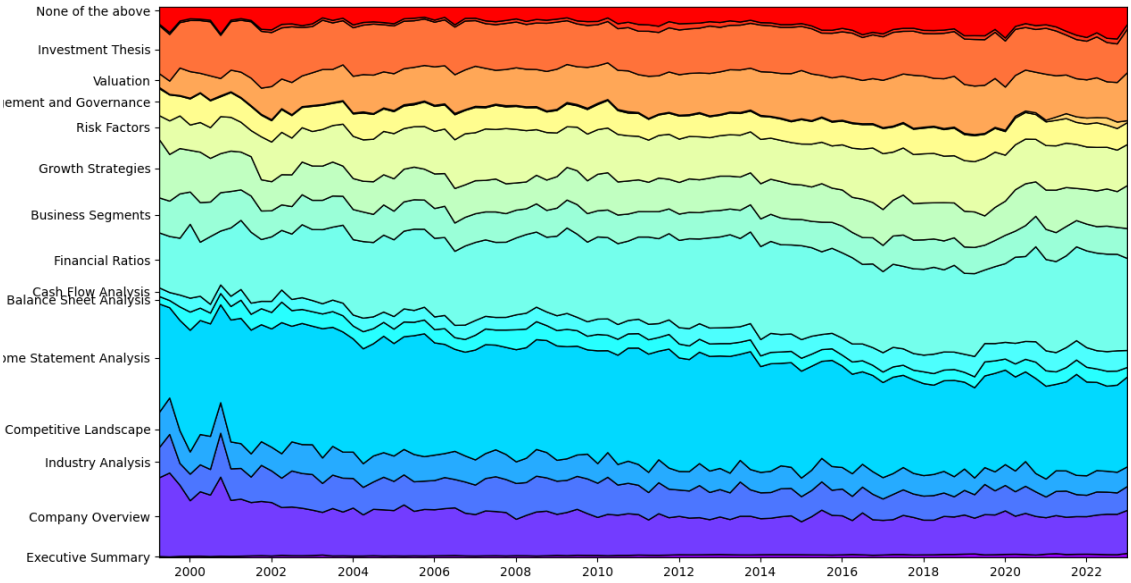


Figure 3 Analyst Discussion Across Topics

This figure shows the distribution of report sentences across 17 topics from 2000 to 2023. The stacked plot depicts the frequency of discussion for each topic based on sentence proportion. The topic categories are Executive Summary, Company Overview, Industry Analysis, Competitive Landscape, Income Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Financial Ratios, Business Segments, Growth Strategies, Risk Factors, Management and Governance, ESG Factors, Valuation, Investment Thesis, Appendices and Disclosures, and None of Above.

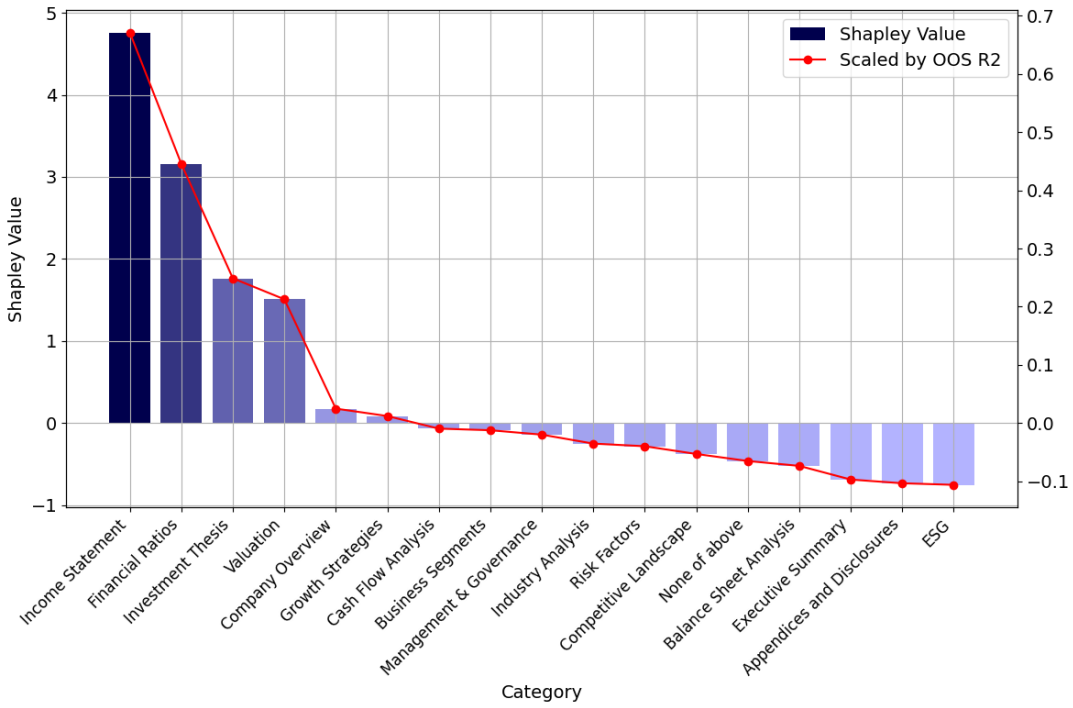


Figure 4 Shapley Values for Topic Importance

This figure shows the importance of topics discussed in analyst reports, calculated using the SHAP framework (Lundberg and Lee, 2017). The sum of SHAP values for the 17 topics equals the R^2_{OOS} of the sentence-segmented report embeddings. The red line represents the SHAP value scaled by the summed R^2_{OOS} for each topic.

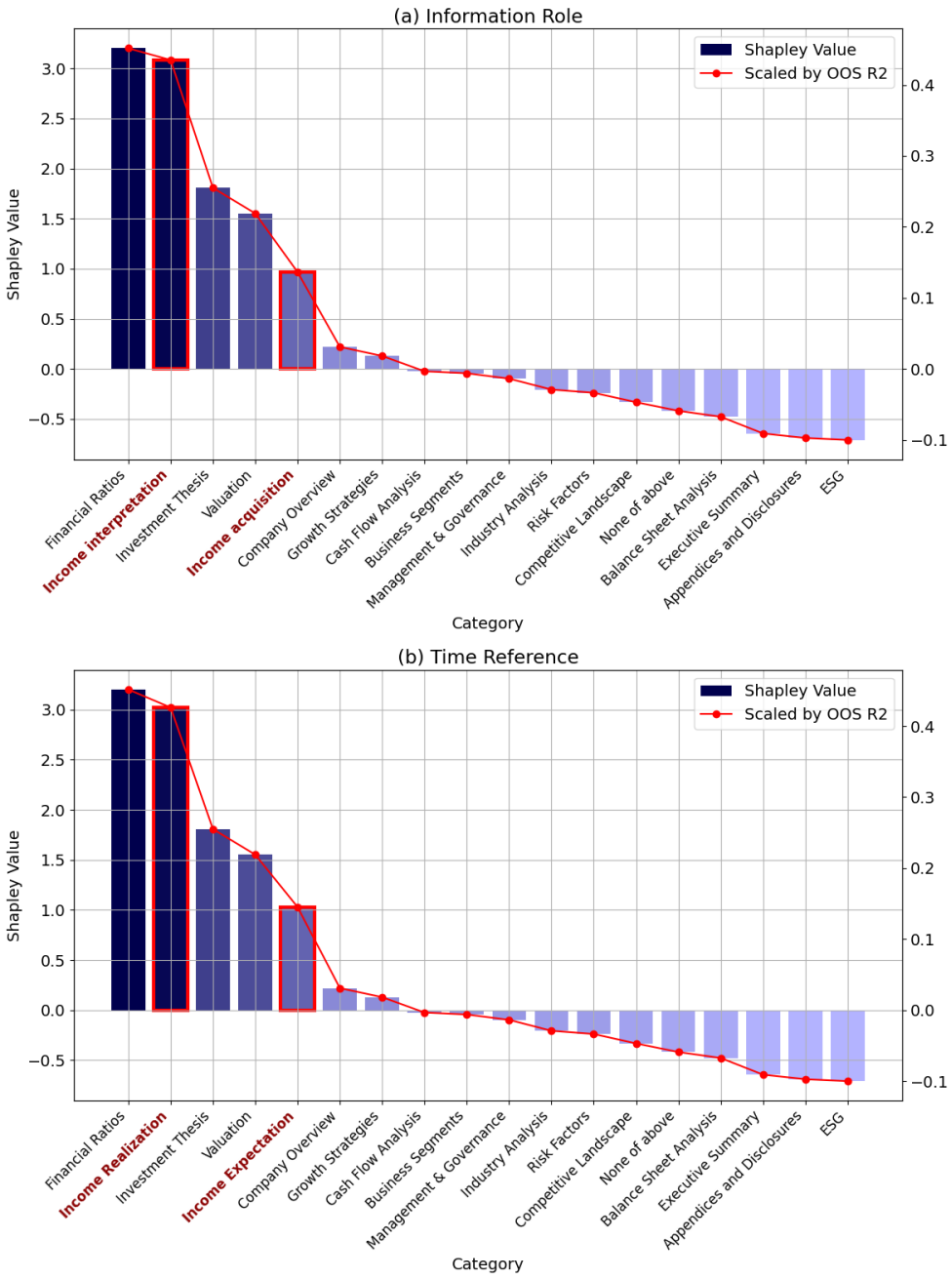


Figure 5 Shapley Values for Sub-Topic Importance

This figure illustrates the importance of sub-topics within income statement analyses discussed in analyst reports, calculated using the SHAP framework (Lundberg and Lee, 2017). The sum of SHAP values for the 18 topics equals the R^2_{OOS} of the sentence-segmented report embeddings. The red line represents the SHAP value scaled by the summed R^2_{OOS} for each topic. Panel (a) categorizes income statement analyses into the income acquisition and income interpretation categories, while Panel (b) categorizes them into the income realization and income expectation categories.

Table 1 Summary Statistics of Analyst Reports

This table presents the summary statistics of the distribution of the analyst report sample of S&P 100 firms over the years 2000-2023 and across industries. Panel A shows the number of reports, brokerage firms, sell-side analysts, average number of pages, and tokens per report for each year. Panel B shows the statistics across Fama-French 12 (FF12) industries. The definition of FF12 industries is from: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library.html

Panel A: Sell-side analyst reports between 2000-2023

Year	Reports	Brokerage firms	Analysts	Pages	Tokens
2000	582	28	78	5	1992
2001	1080	34	116	5	2043
2002	1973	42	167	5	1806
2003	2735	45	192	6	2006
2004	3622	49	262	7	1910
2005	4434	50	269	7	2154
2006	4716	45	234	7	2065
2007	4872	44	243	7	2259
2008	5834	51	279	8	2468
2009	5745	62	326	8	2322
2010	4957	67	313	7	2171
2011	7327	56	381	8	2164
2012	7534	54	379	8	1943
2013	7936	52	403	8	1941
2014	7350	50	407	8	1824
2015	7534	50	380	8	1866
2016	7009	50	374	8	1987
2017	6628	48	344	9	2039
2018	5481	37	282	8	2034
2019	5958	38	305	8	1992
2020	6049	40	299	8	2017
2021	3967	27	196	9	2047
2022	4536	25	210	9	2146
2023	4393	34	212	9	2094

Panel B: Sell-side analyst reports in Fama-French 12 industries

Industry	Reports	Brokerage firms	Analysts	Pages	Tokens
BusEq	25258	101	411	8	2133
Hlth	20257	66	191	8	2086
Money	19689	63	230	7	1955
Shops	11602	75	186	7	1968
Manuf	10531	57	170	8	1954
Other	9763	79	259	9	2303
Telcm	6022	60	89	9	2149
Utils	5068	30	54	6	1827
Enrgy	5049	41	67	8	1927
NoDur	3862	36	65	8	2209
Durbl	2915	29	40	8	1867
Chems	2236	33	53	8	2141

Table 2 Sumamry Statistics for Information Values

This table presents the summary statistics for the information values of analyst reports targeting common constituent firms of the S&P 100 index, covering the period from 2015 to 2023. The value of information is measured as the explained return volatility divided by the price impact, with the results reported in millions of dollars. All dollar values are adjusted for inflation to reflect 2020 values using the Consumer Price Index (CPI). Panel A reports the dollar value of analysts' information, information value of text, and information value of revisions. The mean and standard deviation are estimated using the delta method. Panel B reports the price impact per billion dollars and the stock price. The price impact is estimated by regressing one-minute log returns on contemporaneous share order flow, divided by the closing stock price two trading days prior to the report release (t-2), and is reported in billions of dollars. The stock price refers to the closing stock price two trading days prior to the report release (t-2).

Panel A: Dollar value of analysts information.

	Mean	SE	95%CI	99%CI	N
Information value, $M	0.47	0.05	[0.38, 0.56]	[0.35, 0.58]	17672
Information value of text, $M	0.38	0.04	[0.30, 0.46]	[0.28, 0.48]	17672
Information value of revisions, $M	0.34	0.04	[0.26, 0.43]	[0.23, 0.46]	17672

Panel B: Price impact and stock price.

	Mean	Std	p25	P50	P75	N
Price impact per $B	0.34	1.29	0.05	0.13	0.31	17672
Stock price	118.49	130.34	51.15	82.0	143.58	17672

Table 3 Value of Information and Stock Characteristics

This table shows the results of regressing the log information value and its component on stock characteristics. Log information value (text) measures the value of information estimated using analyst report text. The log information value equals log explained return volatility less log price impact, which are used as dependent variables in Panel B. Observations with negative price impact and negative explained return variance are omitted. Variable definitions are presented in Table A1. The t-statistic clustered by stock and year are presented in parentheses. $*p < .1$, $**p < .05$, $***p < .01$.

Panel A: Information value from analysts is higher for large firms.

	log Information value				log Information value (text)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Size	0.574***	0.853***	0.573***	0.864***	0.667***	0.922***	0.669***	0.934***
	(8.01)	(9.25)	(8.33)	(9.62)	(9.20)	(8.45)	(9.41)	(8.38)
Book-to-market	-0.222	0.416	-0.139	0.457	-0.218	0.326	-0.161	0.335
	(-1.38)	(1.27)	(-0.94)	(1.36)	(-1.34)	(0.93)	(-1.05)	(0.92)
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
N	8030	8030	8030	8030	8509	8509	8509	8509
Adjusted R^2	0.072	0.155	0.088	0.168	0.093	0.167	0.103	0.176

Panel B: Price impact is smaller for large firms.

	log Return variance				log Price impact			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Size	-0.229***	0.028	-0.228***	0.044	-0.801***	-0.831***	-0.798***	-0.826***
	(-5.58)	(0.53)	(-6.07)	(0.78)	(-13.15)	(-9.24)	(-13.36)	(-9.46)
Book-to-market	-0.273**	0.153	-0.192*	0.180	-0.059	-0.260	-0.060	-0.274
	(-2.31)	(0.62)	(-1.82)	(0.70)	(-0.51)	(-1.26)	(-0.56)	(-1.33)
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
N	8030	8030	8030	8030	8030	8030	8030	8030
Adjusted R^2	0.015	0.076	0.036	0.092	0.318	0.449	0.318	0.450

Table 4 Value of Information and Boldeness

This table shows the results of regressing the log information value and its component on analyst features. Log information value (text) measures the value of information estimated using analyst report text. The log information value equals log explained return volatility less log price impact, which are used as dependent variables in Panel B. Observations with negative price impact and negative explained return variance are omitted. Variable definitions are presented in Table A1. The t-statistic clustered by stock and year are presented in parentheses. $*p < .1$, $**p < .05$, $***p < .01$.

Panel A: Information value from analysts is higher for bold reports.

	log Information value				log Information value (text)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bold	0.300**	0.305***	0.275***	0.290***	0.240**	0.249**	0.230**	0.248**
	(3.000)	(4.150)	(3.420)	(4.090)	(2.360)	(3.120)	(2.810)	(3.100)
Brokersize	-0.000	0.000	0.001	0.001	-0.000	0.000	0.002	0.002
	(-0.340)	(0.760)	(0.890)	(0.880)	(-0.570)	(1.180)	(1.010)	(1.140)
Year FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
Analyst FE	No	No	Yes	Yes	No	No	Yes	Yes
N	13652	13650	13607	13606	13520	13518	13477	13476
Adjusted R^2	0.003	0.182	0.141	0.201	0.002	0.181	0.141	0.204

Panel B: Explained return volatility is higher for bold reports.

	log Return variance				log Price impact			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bold	0.290***	0.241***	0.251***	0.231***	-0.017	-0.065	-0.025	-0.059
	(5.770)	(5.960)	(5.910)	(5.350)	(-0.280)	(-1.350)	(-0.420)	(-1.340)
Brokersize	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	-0.002	-0.002
	(-0.700)	(-0.050)	(-0.570)	(-0.220)	(0.180)	(-1.130)	(-1.730)	(-1.680)
Year FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
Analyst FE	No	No	Yes	Yes	No	No	Yes	Yes
N	13642	13640	13588	13587	12867	12865	12813	12812
Adjusted R^2	0.004	0.095	0.091	0.122	-0.000	0.416	0.228	0.442

Table 5 Value of Information and Earnings Announcements

This table shows the results of regressing the log information value and its component on an earnings announcement indicator and the trading volume measure. Log information value (text) measures the value of information estimated using analyst report text. The log information value equals log explained return volatility less log price impact, which are used as dependent variables in Panel B. Observations with negative price impact and negative explained return variance are omitted. *Week* is an indicator variable that takes a value of 1 when the reports are released within 1 week following earnings announcement day, and 0 otherwise. *TradingVolume* is the trading volume of the stock on the earnings announcement day scaled by its shares outstanding. The t-statistic clustered by stock and year are presented in parentheses. $*p < .1$, $**p < .05$, $***p < .01$.

Panel A: Information value from analysts is higher following earnings announcements, especially when trading volumes are high.

	log Information value				log Information value (text)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Week	0.246*	0.392***	0.225*	0.380***	0.199	0.354***	0.173	0.335***
	(1.910)	(3.560)	(1.880)	(3.540)	(1.550)	(3.200)	(1.450)	(3.190)
Trading Volume	0.003	-0.005	0.003	-0.004	0.000	-0.005	-0.000	-0.004
	(0.770)	(-1.050)	(0.600)	(-0.960)	(0.110)	(-0.940)	(-0.020)	(-0.890)
Week × Trading Volume	0.018***	0.019***	0.015***	0.017***	0.013***	0.015**	0.010*	0.013**
	(3.390)	(3.540)	(2.740)	(2.950)	(2.720)	(2.510)	(1.970)	(2.130)
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
N	8030	8030	8030	8030	8509	8509	8509	8509
Adjusted R^2	0.013	0.149	0.029	0.158	0.006	0.152	0.017	0.158

Panel B: Explained return volatility is higher following earnings announcements, especially when trading volumes are high.

	log Return variance				log Price impact			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Week	0.501***	0.491***	0.477***	0.473***	0.243***	0.088**	0.241***	0.083**
	(5.090)	(5.280)	(5.120)	(5.380)	(3.430)	(2.110)	(3.640)	(2.000)
Trading Volume	0.020***	0.004	0.019***	0.005	0.016***	0.009**	0.016***	0.009**
	(5.910)	(1.170)	(5.740)	(1.410)	(3.080)	(2.020)	(3.090)	(2.030)
Week × Trading Volume	0.014***	0.013***	0.012**	0.011**	-0.003**	-0.005**	-0.002	-0.005**
	(3.120)	(3.160)	(2.530)	(2.410)	(-2.050)	(-2.290)	(-1.350)	(-2.370)
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Stock FE	No	Yes	No	Yes	No	Yes	No	Yes
N	8030	8030	8030	8030	8030	8030	8030	8030
Adjusted R^2	0.059	0.104	0.073	0.115	0.025	0.393	0.033	0.396

Table 6 Information Content of Analyst Reports

This table presents the out-of-sample performance of a ridge regression model in predicting market reactions to quantitative information in numerical measures and qualitative information in analyst reports. *Revision only* means the input contains three analyst forecast revision measures. *Numerical only* supplements *revision only* with numerical measures described in Section 3.1. *Text only* implies the input is comprised of analyst report text embeddings. *Rev + text* combines three analyst forecast revision measures and text embeddings. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020). In Panel A, the benchmark prediction is set to zero. In Panel B, the reported t-statistic compares the predictions of alternative large language models with those of LLaMA-2-13B. *BERT* denotes the `bert-base-uncased` model. *OpenAI* denotes the `text-embedding-3-small` model. *LLaMA 3* denotes the `LLaMA-3-8B` model.

Panel A: Information content of numerical and textual information

Year	Revision only	t-stat	Numerical only	t-stat	Text only	t-stat	Rev + text	t-stat	t-stat	t-stat	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(5)-(1)	(7)-(1)	(7)-(5)
2015	10.31%	4.27	7.79%	2.63	12.63%	6.39	10.63%	2.98	3.63	0.16	-1.25
2016	14.61%	11.26	15.30%	11.44	11.98%	3.93	17.08%	9.67	-1.15	3.82	2.93
2017	8.99%	6.23	9.95%	6.44	11.11%	6.40	11.98%	7.09	4.99	5.96	2.38
2018	10.05%	3.28	10.55%	3.59	10.87%	5.85	13.64%	5.95	0.87	5.77	4.47
2019	9.94%	20.14	9.93%	14.82	12.16%	17.68	14.44%	26.17	2.68	19.77	4.83
2020	5.52%	3.88	6.11%	4.53	3.82%	5.18	6.34%	6.16	-1.75	1.57	7.11
2021	5.43%	5.83	5.99%	6.75	8.50%	6.21	11.94%	27.16	1.48	7.48	3.28
2022	9.78%	6.03	9.12%	4.95	14.88%	10.34	16.95%	10.09	9.21	8.01	5.30
2023	6.68%	5.48	7.15%	4.01	9.30%	4.17	10.76%	6.09	2.13	4.87	3.08
Overall	9.01%	9.45	9.06%	9.46	10.19%	8.20	12.28%	8.87	1.66	3.95	3.77

Panel B: Information content of textual information using alternative LLMs

Year	BERT	t-stat	OpenAI	t-stat	LLaMA 3	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)
2015	7.24%	-13.45	6.76%	-9.97	11.28%	-8.04
2016	6.50%	-2.22	5.94%	-4.05	10.48%	2.23
2017	5.48%	-5.95	5.88%	-7.72	10.25%	0.11
2018	6.94%	-14.05	6.48%	-8.15	10.21%	1.29
2019	6.16%	-13.80	5.41%	-16.94	10.48%	-6.19
2020	3.92%	-0.33	4.10%	0.09	6.07%	6.27
2021	2.63%	-18.53	3.07%	-4.11	6.98%	-1.26
2022	7.75%	-9.93	7.89%	-7.09	13.02%	-2.96
2023	4.18%	-5.65	4.03%	-8.04	8.68%	12.05
Overall	5.72%	-5.54	5.57%	-5.66	9.66%	0.19

Table 7 Infomation Content of Analyst Forecast Revisions

This table presents the results of OLS regressions. Panel A shows the summary statistics of the variables in the regression sample. Panel B provides the coefficient estimates and t-statistics from the OLS regression. The dependent variable is $CAR_{[-1,+1]}$, which represents the cumulative three-day abnormal returns centered around the release day. REC_REV denotes recommendation revision, calculated as the current report's recommendation minus the last recommendation in I/B/E/S issued by the same analyst for the same firm. EF_REV represents earnings forecast revision, calculated as the current report's EPS forecast minus the last EPS forecast in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release. TP_REV indicates target price revision, calculated as the current report's target price minus the last target price in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release. $\widehat{CAR}_{txt}$ is the out-of-sample fitted value of Ridge regression using full-context report embeddings. $\widehat{CAR}_{rev}$ is the out-of-sample fitted value of Ridge regression using the three report revision measures. t-Statistics based on standard errors clustered at the firm and year levels are reported in parentheses. $*p < .1$, $**p < .05$, $***p < .01$.

Panel A: Summary statistics						
	Mean	Std	p25	P50	P75	N
$CAR_{[-1,+1]}$	0.002	0.048	-0.019	0.001	0.021	28837
REC_REV	0.002	0.151	0.000	0.000	0.000	28837
EF_REV	0.000	0.005	0.000	0.000	0.001	28837
TP_REV	0.010	0.067	0.000	0.000	0.000	28837
$\widehat{CAR\_rev}$	0.002	0.014	-0.001	-0.000	0.003	28837
$\widehat{CAR\_txt}$	0.002	0.018	-0.009	0.002	0.013	28837

Panel B: Market reaction to analyst revision and report text					
	(1)	(2)	(3)	(4)	(5)
REC_REV	0.010***		0.004		
	(3.26)		(1.29)		
EF_REV	1.078***		0.679***		
	(7.72)		(5.12)		
TP_REV	0.177***		0.136***		
	(14.95)		(11.45)		
$\widehat{CAR_{txt}}$		0.835***	0.681***		0.680***
		(24.23)	(26.25)		(26.03)
$\widehat{CAR_{rev}}$				1.003***	0.727***
				(18.29)	(13.69)
Intercept	-0.001	0.000	-0.001***	-0.000	-0.001*
	(-1.56)	(0.05)	(-2.60)	(-0.70)	(-1.93)
N	28837	28837	25698	28837	25698
Adjusted R^2	0.091	0.105	0.156	0.089	0.154

Table 8 Information Content of Earnings Announcement Transcript and Analyst Report Text

This table reports the information content of the text in earnings announcement transcripts and sell-side analyst reports. Panel A shows the out-of-sample R-squared (R^2_{OOS}) and t-statistic of analyst report text in 13 weekly bins following earnings announcements. Panel B presents sub-sample analyses of earnings announcement periods and non-earnings announcement periods using 1-day, 2-day, 3-day, and 7-day windows. Panel C examines the information content of both earnings conference call transcripts and analyst reports released promptly (1-day window) following earnings announcements. The out-of-sample period is 2015-2023. *Transcript* means the input contains corporate earnings conference call embeddings. *Reports* implies the input is comprised of analyst report text embeddings. *Rev + text* combines transcript embeddings and text embeddings. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020). The benchmark prediction is set to zero. In Panel C column Diff, the reported t-statistic compares the predictions of models with conference call transcript as input and with both report text and transcript text.

Panel A: Weekly bins

Weeks	R^2_{OOS}	t-stat	N
1	9.80%	10.17	26490
2	0.60%	-0.08	4586
3	-8.71%	-2.95	4144
4	-3.61%	-1.36	4320
5	0.23%	0.53	3796
6	-4.07%	-1.53	4230
7	-4.26%	-1.62	4404
8	2.88%	1.57	4286
9	-4.89%	-3.50	3722
10	2.06%	0.83	3528
11	2.74%	0.83	4116
12	-0.02%	-0.42	4764
13	-0.11%	-0.76	5642

Panel B: Sub-sample analyses of earnings announcement periods

Window	Earnings announcement	t-stat	Non earnings announcement	t-stat
1 day	11.84%	3.50	4.59%	2.55
2 day	11.97%	3.38	4.38%	2.47
3 day	11.69%	3.38	4.13%	2.38
7 day	7.29%	2.57	5.08%	2.59

Panel C: Information content of earnings announcement transcripts

	Transcripts (1)	Reports (2)	Reports + Transcripts (3)	Diff (3) - (1)
R^2_{OOS}	4.20%	9.72%	11.96%	7.76%
t-stat	5.16	3.24	6.42	5.12

Table 9 Information Content of Revision Reports and Reiteration Reports

This table reports the information content of the text in reiteration analyst reports and revision analyst reports. The revision and reiteration reports are labeled following the criteria of Huang et al. (2014). *Revision only* means the input contains three analyst forecast revision measures. *Text only* implies the input is comprised of analyst report text embeddings. *Rev + text* combines three analyst forecast revision measures and text embeddings. Panel A presents the out-of-sample R-squared (R^2_{OOS}) and corresponding t-statistic for analyst reports that maintain the same price target as the previous report (i.e., reiteration). Panel B presents sub-sample analyses of analyst reports that change the price target from the previous report (i.e., revision). The out-of-sample period is 2015-2023. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020).

Panel A: Sub-sample analyses of reiteration reports

Target	Revision only	t-stat	Text only	t-stat	Rev + text	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)
Recommendation	8.71%	9.86	9.85%	8.21	11.76%	9.12
Target price	1.20%	1.63	2.57%	2.26	3.58%	3.40
Earnings forecast	2.89%	2.96	-2.71%	-2.05	-0.66%	0.06

Panel B: Sub-sample analyses of revision reports

Target	Revision only	t-stat	Text only	t-stat	Rev + text	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)
Recommendation	14.99%	2.92	16.80%	4.8	22.63%	4.83
Target price	16.72%	11.14	17.72%	11.1	20.88%	11.18
Earnings forecast	10.94%	7.48	14.27%	10.9	16.37%	11.42

Table 10 Infomation Content of Text Embedding and Text Tones

This table reports the market reaction to the tone-based information and embedding-based information of analyst reports. *Tone_NB* is the tone measure for the whole report, constructed using the Naive Bayes approach. *Tone_Income_NB* is the tone measure for the income statement analyses topic, also constructed using the Naive Bayes approach. *Tone_NonIncome_NB* is the tone measure for non-income statement topics, constructed using the same Naive Bayes approach. *Tone_BERT*, *Tone_Income_BERT* and *Tone_NonIncome_BERT* are tone measures constructed using the BERT model. Panel A reports the coefficient estimates and t-statistics from OLS regression. t-Statistics based on standard errors clustered at the firm and year levels are reported in parentheses. Panel B presents the out-of-sample performance of the ridge regression models in predicting market reactions to various information content in analyst reports. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020). $*p < .1$, $**p < .05$, $***p < .01$.

Panel A: Market reaction to report tones							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Tone_NB	0.019***						
	(19.43)						
Tone_Income_NB		0.006***	0.006***				0.000
		(15.32)	(15.56)				(1.16)
Tone_Nonincome_NB		0.014***	0.015***				-0.001
		(15.30)	(18.87)				(-1.56)
Tone_BERT				0.031***			
				(31.16)			
Tone_Income_BERT					0.011***	0.011***	0.010***
					(23.73)	(24.03)	(22.89)
Tone_Nonincome_BERT					0.021***	0.021***	0.022***
					(23.34)	(25.57)	(23.33)
Intercept	-0.007***	-0.007***	-0.008***	-0.007***	-0.007***	-0.007***	-0.007***
	(-11.94)	(-11.70)	(-14.07)	(-15.69)	(-14.35)	(-15.68)	(-13.82)
Firm FE	No	No	Yes	No	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes	Yes
N	122248	106488	99705	122248	106488	99705	99705
Adjusted R^2	0.010	0.011	0.018	0.031	0.035	0.041	0.041

Panel B: Comparison Results for Report Tones and Report Embedding		
Model Input	R^2_{OOS}	t-stat
Tone (Naive Bayes)	0.05%	-0.30
Tone (BERT)	3.78%	10.55
Tone (Naive Bayes) + numerical	9.49%	10.19
Tone (BERT) + numerical	10.58%	11.11
Tone (Naive Bayes) + numerical + text	12.28%	8.87
Tone (BERT) + numerical + text	12.27%	8.91
”Tone (Naive Bayes) + numerical + text” vs ”numerical + text”	0.00%	0.67
”Tone (BERT) + numerical + text” vs ”numerical + text”	-0.01%	-0.79

Table 11 Information Content of Analyst Reports using Machine Learning Models

This table reports the out-of-sample R-square (R^2_{OOS}) for various machine learning models estimating the information content of analyst report text. The R^2_{OOS} is calculated annually using a training sample from 2000 to the preceding year. PLS represents the Partial Least Squares regression model, with the number of components tested being 5, 10, 15, 20, and 25. XGBoost (Extreme Gradient Boosting) implements the concept of gradient-boosted decision trees. NN1 to NN5 specify the number of layers in the neural network model. The Overall row reports the R^2_{OOS} and t-statistics for the sample period of 2015-2023. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020). The benchmark for cumulative abnormal return (CAR) estimates is set to 0 for the t-statistic calculation.

	PLS		XGBoost		NN1		NN2		NN3		NN4		NN5	
year	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat
2015	12.23%	10.95	8.87%	11.25	15.81%	9.72	13.84%	14.53	15.30%	14.10	11.58%	12.32	16.61%	16.95
2016	10.27%	15.59	6.66%	5.63	12.59%	24.64	11.78%	13.02	10.99%	9.10	12.03%	14.75	12.46%	17.41
2017	6.60%	6.94	4.92%	5.26	9.08%	4.64	10.01%	4.64	9.58%	4.76	8.57%	4.51	10.03%	5.49
2018	8.08%	13.37	5.15%	8.07	10.73%	26.65	10.54%	16.42	10.41%	11.43	9.86%	17.44	11.98%	13.97
2019	11.54%	38.70	6.32%	16.40	13.43%	17.37	15.18%	14.73	15.59%	18.06	15.89%	10.48	15.37%	16.94
2020	3.21%	2.16	3.02%	3.64	4.45%	2.92	4.73%	2.78	3.85%	2.27	4.25%	3.08	5.51%	4.06
2021	2.14%	1.76	4.30%	3.34	4.07%	6.44	5.88%	7.13	9.28%	13.59	8.47%	9.44	8.83%	7.67
2022	11.42%	7.40	10.04%	7.80	13.78%	8.31	15.67%	6.38	16.44%	6.64	17.63%	8.55	17.56%	7.12
2023	8.14%	9.26	4.23%	3.82	9.54%	8.50	12.16%	10.76	11.06%	8.59	12.18%	11.59	11.70%	6.78
Overall	8.26%	6.66	5.97%	6.76	10.42%	8.11	11.04%	7.54	11.18%	7.29	11.05%	6.62	12.12%	7.93

Internet Appendix for
"The Value of Information from Sell-side Analysts"

Not for Publication

A. Additional Figures

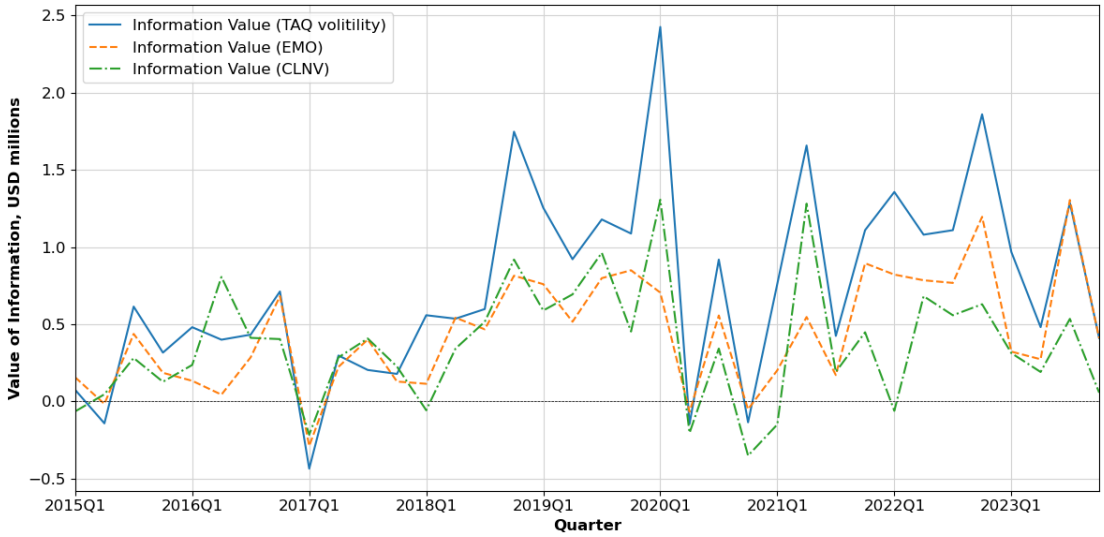


Figure A1 Alternative Dollar Values of Analyst Reports Over Time

This figure presents the alternative estimations of the value of information, reported in millions of dollars and adjusted to 2020 dollars, derived from analyst reports between the first quarter of 2015 (2015Q1) and the last quarter of 2023 (2023Q4). Each quarter-mean approximation is calculated using the delta method. TAQ volatility refers to the explained volatility calculated as $\frac{r^2-(r-\widehat{r})^2}{r^2} \cdot \sigma_v^2$, where σ_v^2 is the realized volatility of the sum of squared one-minute log returns. EMO and CLNV show the mean value of information across days and stocks for three different algorithms for signing trades as buys or sells, specifically those developed by Ellis et al. (2000) and Lee and Ready (1991).

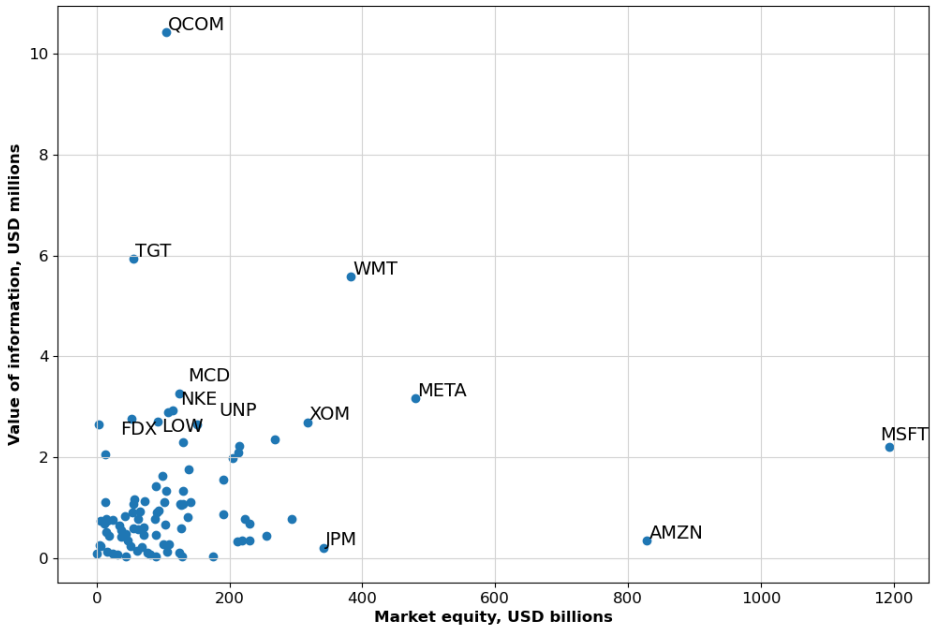


Figure A2 Dollar Value of Analyst Reports by Stock

This figure demonstrates the estimated information value of analyst reports for individual stocks, quantified in millions of dollars. The data spans from the first quarter of 2015 (2015Q1) to the fourth quarter of 2023 (2023Q4). Information value estimates are derived using the delta method. Market equity averaged daily at the time of each analyst report release, is reported in billions of dollars. All monetary values are adjusted to 2020 levels using the Consumer Price Index (CPI) for inflation adjustment.

(1) Executive Summary

(2) Company Overview

(3) Industry Analysis

(4) Competitive Landscape

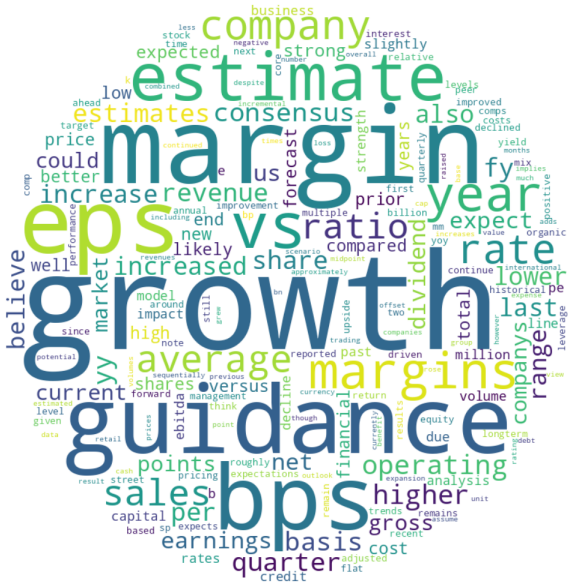
Figure A3 Word Clouds of Topics

This figure presents word clouds for 16 key topics commonly found in analyst reports, excluding the "None of the Above" category. The topics include Executive Summary, Company Overview, Industry Analysis, Competitive Landscape, Income Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Financial Ratios, Business Segments, Growth Strategies, Risk Factors, Management and Governance, ESG Factors, Valuation, Investment Thesis, and Appendices and Disclosures.

(5) Income Statement Analysis

(6) Balance Sheet Analysis

(7) Cash Flow Analysis



(8) Financial Ratios

Figure A3 Word Clouds of Topics

This figure presents word clouds for 16 key topics commonly found in analyst reports, excluding the "None of the Above" category. The topics include Executive Summary, Company Overview, Industry Analysis, Competitive Landscape, Income Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Financial Ratios, Business Segments, Growth Strategies, Risk Factors, Management and Governance, ESG Factors, Valuation, Investment Thesis, and Appendices and Disclosures.

(9) Business Segments

(10) Growth Strategies

(11) Risk Factors

(12) Management and Governance

Figure A3 Word Clouds of Topics

This figure presents word clouds for 16 key topics commonly found in analyst reports, excluding the "None of the Above" category. The topics include Executive Summary, Company Overview, Industry Analysis, Competitive Landscape, Income Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Financial Ratios, Business Segments, Growth Strategies, Risk Factors, Management and Governance, ESG Factors, Valuation, Investment Thesis, and Appendices and Disclosures.

(13) ESG Factors

(14) Valuation

(15) Investment Thesis

(16) Appendices and Disclosures

Figure A3 Word Clouds of Topics

This figure presents word clouds for 16 key topics commonly found in analyst reports, excluding the "None of the Above" category. The topics include Executive Summary, Company Overview, Industry Analysis, Competitive Landscape, Income Statement Analysis, Balance Sheet Analysis, Cash Flow Analysis, Financial Ratios, Business Segments, Growth Strategies, Risk Factors, Management and Governance, ESG Factors, Valuation, Investment Thesis, and Appendices and Disclosures.

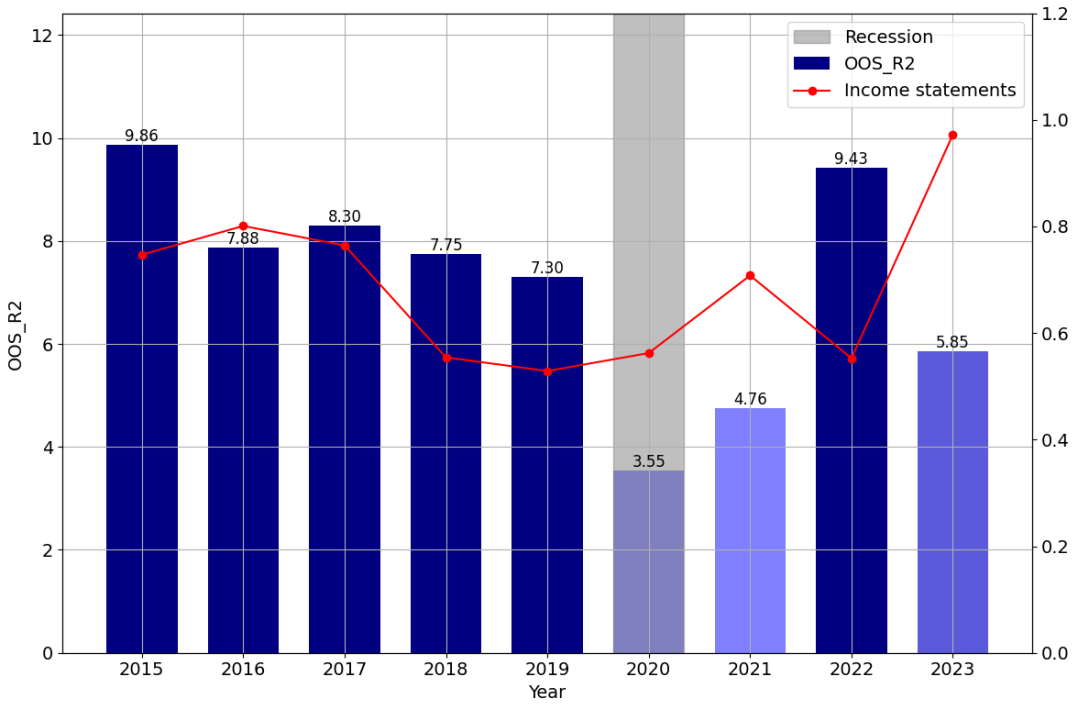


Figure A4 R^2_{oos} of Sentence-Segmented Embeddings by Year

This figure displays the R^2_{oos} of sentence-segmented report embeddings from 2015 to 2023. The red line represents the SHAP value of the Income Statements Analyses topic across the years. The shaded area indicates the pandemic recession in 2020. The color gradient shows the magnitude of R^2_{oos}.

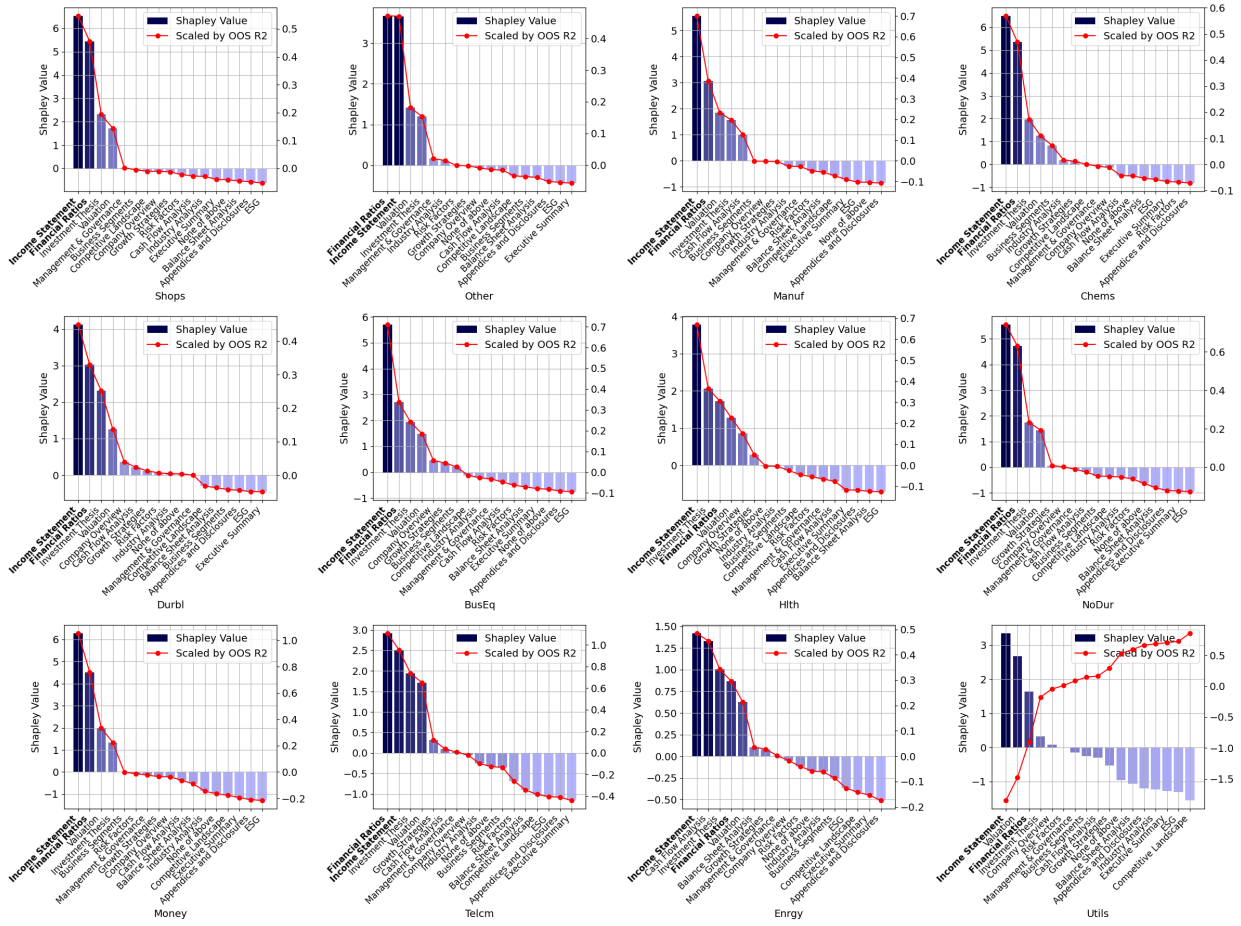


Figure A5 Shapley Values for Topic Importance across Industries

This figure illustrates the importance of topics discussed in analyst reports across Fama-French 12 industries, calculated using the SHAP framework (Lundberg and Lee, 2017). The sum of SHAP values for the 17 topics equals the R^2_{OOS} of the sentence-segmented report embedding. The red line represents the SHAP value scaled by the summed R^2_{OOS} for each topic.

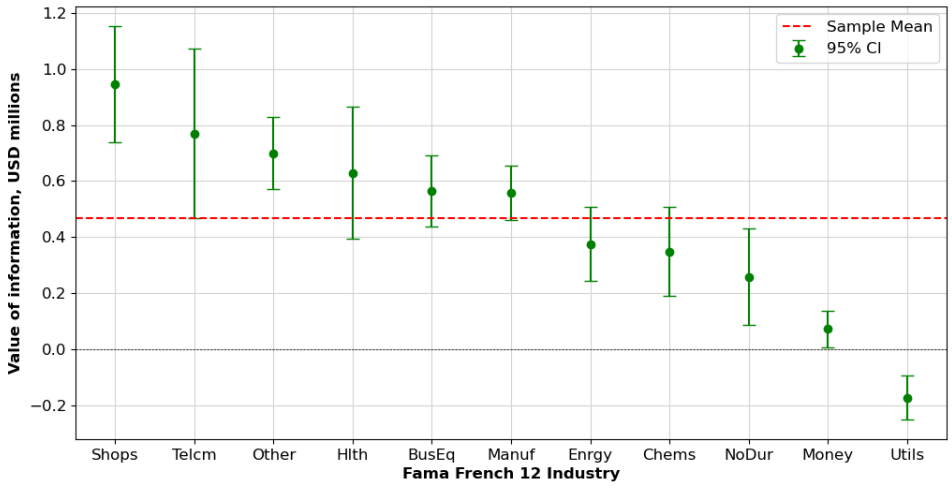


Figure A6 Dollar Value of Analyst Reports by Industry

This figure demonstrates the estimated information value of analyst reports for Fama-French 12 industries, quantified in millions of dollars. The data spans from the first quarter of 2015 (2015Q1) to the fourth quarter of 2023 (2023Q4). Information value estimates are derived using the delta method and adjusted to 2020 levels using the Consumer Price Index (CPI) for inflation adjustment.

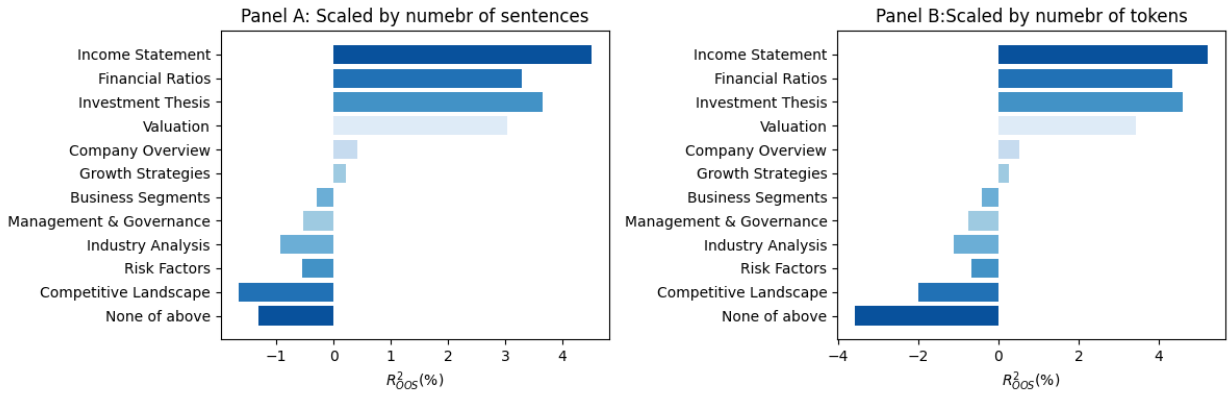


Figure A7 Shapley Values Scaled by Length

This figure presents the Shapley values of topics scaled by the total number of sentences (Panel A) and the total number of tokens (Panel B) for topics with more than 100,000 sentences. The scaling process involves two steps: First, normalize the Shapley value of each topic by dividing it by the number of sentences in that topic. This yields the topic's contribution per sentence. Second, adjust the normalized values to maintain the total sum of Shapley values. Multiply each normalized value by the ratio of the total sum of original Shapley values to the sum of normalized values. This ensures that the scaled values reflect the relative importance of topics while preserving the overall magnitude of contributions.

B. Additional Tables

Table A1 Numerical Variables Description

This table shows the variable definitions of numerical measures.

Numerical Measures	Definition and/or sources
Panel A: Firm Level Measures	
Size	The market value equity of the firm ($CSHOQ * PRCCQ$) at the end of the quarter prior to report release.
BtoM	The book value of equity ($SEQ + TXDB + ITCB - PREF$) scaled by the market value of equity ($CSHOQ * PRCCQ$) t the end of quarter prior the report release.
Prior_CAR	The cumulated 10-day abnormal return ending 2 days before release. The abnormal return is calculated as the raw return minus the buy-and-hold market value-weighted return.
SUE	Earnings surprise, calculated as the actual EPS minus the last consensus EPS forecast before the earnings announcement. Consensus EPS is the median value of 1-year EPS forecast within 90-days window of all analysts following the firm. The unexpected earnings is scaled by price per share at the fiscal quarter end.
AbsSUE	Absolue value of SUE, representing the distance between realized EPS and EPS expectation.
Miss	Dummy variable that equals one if the actual EPS is less than the last consensus forecast, and 0 otherwise.
Trading volume	Trading volume at earnings announcement day (or the first trading day post earnings announcement), calculated as $VOL/SHROUT$.
Distance to default	The distance to default calculated following Merton (1974). The proxy is compiled from the National University of Singapore's Credit Research Initiative (NUS CRI).
Fluidity	A measure of firms' product market competition introduced by Hoberg et al. (2014). The data is compiled from the Hoberg and Phillips database.
Panel B: Industry Level Measures	
Industry recession	An indicator variable that equals one if the FF-48 industry return is negative and in the bottom quintile of FF-48 industry returns and zero otherwise.
Panel C: Macroeconomic Measures	
Time trend (ttr)	The number of years elapsed from the beginning of the sample.
Panel D: Report Level Measures	
REC_REV	Recommendation revision, calculated as the current report's recommendation minus the last recommendation in I/B/E/S issued by the same analyst for the same firm.
EF_REV	Earnings forecast revision, calculated as the current report's EPS forecast minus the last EPS forecast in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release.
TP_REV	Target price revision, calculated as the current report's target price minus the last target price in I/B/E/S issued by the same analyst for the same firm, scaled by the stock price 50 days before the report release.
Boldness	An indicator variable that takes the value of 1 if the EPS forecast revision is above both the analyst's own prior forecast and the consensus forecast, or else below both, and zero otherwise.
SR	Stock recommendation from I/B/E/S rating, with 1 being the most bullish (Strong Buy) and 5 being the most bearish (Sell), based on the ratings provided by the Institutional Brokers' Estimate System (I/B/E/S).
ERet	12-month return forecast by scaling the 12-month price target by the stock price 1-day before release.

Table A2 Sumamry Statistics for Numerical Measures

This table reports summary statistics of numerical measures.

	Mean	Std	p25	P50	P75	N
Miss	0.35	0.48	0.00	0.00	1.00	102776
Trading volume	18.19	21.98	7.24	11.46	20.29	99908
Prior_CAR	0.00	0.05	-0.02	0.00	0.02	122251
Size	2.39	0.11	2.33	2.39	2.46	117534
BtoM	0.46	0.43	0.18	0.32	0.61	117534
Distance to default	0.00	0.01	0.00	0.00	0.00	86693
Fluidity	7.17	3.66	4.40	6.57	9.44	114942
Time trend	13.02	5.71	9.00	13.00	17.00	122252
Industry recession	0.14	0.35	0.00	0.00	0.00	122252
BrokerSize	71.92	56.71	26.00	53.00	113.00	119233
Firm Experience	7.68	6.72	3.00	6.00	11.00	111192
Number of firms	20.08	17.08	14.00	18.00	23.00	119233
REC_REV	0.00	0.16	0.00	0.00	0.00	120673
EF_REV	0.00	0.00	0.00	0.00	0.00	90625
TP_REV	0.01	0.07	0.00	0.00	0.00	84108
ERet	0.19	0.23	0.07	0.18	0.31	84108
Boldness	0.75	0.43	1.00	1.00	1.00	81182
SR	2.77	0.84	2.00	3.00	3.00	122252

Table A3 Sumamry Statistics for Alternative Information Value Measures

The value of information is measured as the explained return volatility divided by the price impact, with the results reported in millions of dollars. All dollar values are adjusted for inflation to reflect 2020 values using the Consumer Price Index (CPI). The sample comprises analyst reports targeting common constituent firms of the S&P 100 index, covering the period from 2015 to 2023. The mean and standard deviation are estimated using the delta method. TAQ volatility refers to the explained volatility calculated as $\frac{r^2-(r-\widehat{r})^2}{r^2} \cdot \sigma_v^2$, where σ_v^2 is the realized volatility of the sum of squared one-minute log returns. Intraday denotes that σ_v^2 does not include overnight return volatility. EMO and CLNV show the mean value of information across days and stocks for three different algorithms for signing trades as buys or sells, specifically those developed by Ellis et al. (2000) and Lee and Ready (1991).

	Mean	SE	95%CI	99%CI	N
Information value (TAQ volitility), $M	0.47	0.05	[0.38, 0.56]	[0.35, 0.58]	17669
Information value of text (TAQ volitility), $M	0.47	0.05	[0.38, 0.56]	[0.35, 0.58]	17672
Information value of revisions (TAQ volitility), $M	0.42	0.04	[0.34, 0.50]	[0.31, 0.53]	17672
Information value (Intraday), $M	0.42	0.04	[0.34, 0.51]	[0.32, 0.53]	17672
Information value of text (Intraday), $M	0.38	0.04	[0.30, 0.46]	[0.28, 0.48]	17669
Information value of revisions (Intraday), $M	0.38	0.04	[0.30, 0.46]	[0.28, 0.48]	17672
Information value (EMO), $M	0.34	0.04	[0.27, 0.41]	[0.25, 0.43]	17672
Information value of text (EMO), $M	0.35	0.04	[0.27, 0.42]	[0.25, 0.44]	17672
Information value of revisions (EMO), $M	0.34	0.04	[0.26, 0.43]	[0.23, 0.46]	17669
Information value (CLNV), $M	0.34	0.04	[0.26, 0.43]	[0.23, 0.46]	17672
Information value of text (CLNV), $M	0.31	0.04	[0.23, 0.39]	[0.21, 0.41]	17672
Information value of revisions (CLNV), $M	0.31	0.04	[0.23, 0.39]	[0.21, 0.42]	17672

Table A4 Information Content of Analyst Reports with Numbers Removed

This table presents the out-of-sample R^2 using analyst reports with numbers removed. The input sets include revision numerical measures, text embeddings generated by the LLaMA 2 model, and a combination of both. The estimation targets are three-day cumulative abnormal returns centered around the analyst report release date. The machine learning model employed is Ridge regression, with training samples expanding on a yearly rolling basis. The t-statistic for the R^2_{OOS} is calculated following the procedure described by Gu et al. (2020). In columns (2), (4), and (6), the benchmark estimation is set to zero. Columns (3)-(1) and (5)-(3) present pairwise tests comparing the performance of revision-only input versus text-only input, and text-only input versus combined text and revision input, respectively.

Year	Revision only	t-stat	Text only	t-stat	Revision plus text	t-stat	t-stat	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)	(3)- (1)	(5)- (3)
2015	10.30%	3.59	13.19%	5.87	11.10%	3.24	3.97	-1.45
2016	15.47%	11.80	13.73%	5.00	17.91%	9.02	-0.81	3.13
2017	9.16%	5.28	11.66%	5.98	12.33%	6.12	10.58	1.56
2018	9.79%	2.98	11.46%	9.12	13.62%	7.97	1.22	4.09
2019	9.76%	15.69	12.61%	14.73	13.90%	15.17	4.26	2.79
2020	5.20%	3.43	4.42%	4.38	6.98%	5.81	-0.70	8.37
2021	5.31%	4.80	8.77%	5.06	10.66%	29.67	1.54	1.56
2022	10.24%	6.16	16.71%	10.47	17.90%	8.94	9.93	2.38
2023	5.09%	3.86	9.27%	3.60	10.39%	5.09	2.60	1.81
Overall	8.93%	8.51	10.95%	7.87	12.51%	8.83	2.40	3.00

Table A5 Cumulative Return with Alternative Windows

This table reports the out-of-sample R^2 (R^2_{OOS}) of using analyst report text embeddings to estimate contemporaneous stock returns with Ridge regressions. The R^2_{OOS} is calculated annually using a training sample from 2000 to the preceding year. The T_0 in all *CAR* measures refers to the release day of analyst reports. Panel A summarizes the results of models trained with the entire sample, while Panel B and Panel C report the results of models trained within and beyond earnings announcement periods (1-day window). The Overall row reports the R^2_{OOS} and t-statistics for the sample period of 2015-2023. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020). The benchmark for pairwise model comparison is set to 0 for the t-statistic calculation.

	AR[0]		CAR[0,+1]		CAR[-2,+2]	
year	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat	R^2_{OOS}	t-stat
Panel A: Full sample period						
2015	8.03%	8.53	8.24%	8.36	10.87%	9.40
2016	7.88%	8.27	6.09%	4.68	6.60%	3.11
2017	8.75%	14.35	8.56%	9.29	9.00%	8.78
2018	9.09%	14.95	9.39%	11.26	9.75%	10.54
2019	9.89%	99.70	7.68%	9.45	8.66%	18.93
2020	2.33%	2.33	1.76%	3.15	4.13%	2.14
2021	5.75%	7.32	5.80%	6.62	4.05%	3.57
2022	10.54%	12.32	11.23%	18.20	10.91%	13.65
2023	5.38%	15.44	3.39%	7.00	0.53%	1.48
Overall	7.55%	8.07	6.92%	6.77	7.57%	6.79
Panel B: Earnings announcement period						
Overall	9.78%	5.58	7.87%	5.10	13.14%	8.02
Panel C: Non-earnings announcement period						
Overall	5.77%	5.78	5.62%	5.64	4.96%	4.91

Table A6 Information Content of Analyst Reports across Industries

This table reports the out-of-sample R^2 of Fama-French 12 industries, using the definition from Kenneth French's website. The t-statistic for the R^2_{OOS} is calculated using the procedure outlined by Gu et al. (2020).

year		Shops			Other			Manuf			Chems	
	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N
2015	13.27%	6.03	657	17.69%	5.03	585	10.84%	3.19	808	15.05%	1.74	147
2016	10.43%	2.46	556	3.34%	0.48	593	10.37%	5.38	811	23.91%	3.25	111
2017	16.65%	3.49	506	20.45%	2.67	544	25.03%	13.64	722	11.04%	0.59	82
2018	10.86%	4.32	406	21.65%	3.40	405	3.32%	1.55	553	25.80%	1.87	77
2019	21.04%	9.24	394	7.49%	2.37	610	16.25%	12.07	568	5.00%	3.22	106
2020	4.39%	2.55	499	-0.30%	-0.53	499	4.06%	1.41	542	2.65%	1.60	115
2021	10.27%	2.13	312	9.70%	3.66	362	8.66%	4.32	367	1.46%	0.30	102
2022	22.75%	3.83	320	15.98%	2.66	437	14.26%	11.46	420	7.00%	2.10	110
2023	17.78%	1.76	353	3.49%	2.00	339	7.15%	2.71	360	19.31%	6.54	76
Overall	13.88%	6.03	4003	11.27%	3.69	4374	11.39%	5.57	5151	11.35%	3.55	926
year		Durbl			BusEq			Hlth			NoDur	
	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N
2015	16.83%	5.77	216	15.66%	6.41	1378	5.75%	2.72	1242	0.68%	0.31	241
2016	4.95%	0.79	209	10.24%	6.09	1303	11.07%	2.90	1185	17.14%	5.18	242
2017	-2.17%	-0.30	282	10.36%	3.12	1154	3.37%	2.39	1060	-2.14%	-0.68	278
2018	18.29%	5.57	208	0.32%	0.31	1002	12.47%	6.58	918	1.29%	0.13	324
2019	15.31%	2.26	222	10.17%	6.13	1036	10.02%	7.68	884	18.51%	2.30	277
2020	5.73%	0.79	168	8.11%	4.14	1100	2.86%	1.84	950	-0.37%	-0.28	265
2021	10.01%	5.31	150	11.43%	1.49	731	7.66%	4.58	592	-6.38%	-0.60	195
2022	7.99%	2.58	162	12.08%	4.77	822	12.05%	2.71	850	5.24%	1.11	187
2023	12.05%	2.37	140	10.45%	11.80	962	3.52%	1.32	844	-5.69%	-0.71	158
Overall	11.02%	3.24	1757	9.68%	6.77	9488	7.53%	5.71	8525	7.50%	1.38	2167
year		Money			Telcm			Enrgy			Utils	
	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N	R^2_{OOS}	t-stat	N
2015	18.13%	4.41	1153	1.68%	0.00	403	-4.21%	-1.45	303	18.47%	3.69	401
2016	7.64%	3.53	1108	-10.18%	-1.22	329	15.80%	2.83	252	-1.84%	-0.23	310
2017	1.77%	0.46	1071	8.60%	5.16	307	-18.18%	-8.75	307	-39.46%	-6.99	315
2018	13.55%	8.48	879	10.65%	2.46	219	8.38%	1.29	275	-23.76%	-5.53	215
2019	2.10%	2.25	977	1.53%	0.52	351	-10.28%	-9.07	340	-2.36%	-0.27	193
2020	1.42%	0.76	952	-8.84%	-2.83	360	5.72%	2.62	434	-7.29%	-2.36	165
2021	2.62%	0.47	614	6.15%	3.27	188	-15.04%	-6.83	251	-4.01%	-1.16	103
2022	4.34%	2.13	667	12.84%	3.50	143	16.14%	4.44	286	8.31%	2.71	132
2023	-0.63%	-0.18	569	8.99%	6.89	182	-27.33%	-5.03	290	-6.43%	-1.12	120
Overall	5.18%	3.31	7990	3.32%	1.39	2482	1.90%	0.31	2738	-1.97%	-0.88	1954

Table A7 Examples of Sentences within Topic Categories

Topics	Examples
Executive Summary	• Our key takeaways from CEO Jeff Immelt's presentation at EPG were: Outlook for substantial EPS growth over 2010/12 driven by abatement of credit losses ($8-9bn in 2010E tapering to $4bn run-rate) and CRE impair. • Looking ahead, guidance was tightened, essentially framing the Street, and commentary suggests a strong outlook for the balance of 2007. • Key topics: 1) Update on ABTÕs portfolio of COVID-19 tests; 2) Libre trends, Libre 2 launch and expectations for Libre 3; 3) Elective surgery trends exiting Q1, expectations for 2021 and an update on recent and upcoming new product approvals; and 4) Global trends and impact on EPD and Nutrition.
Company Overview	• Oracle Corporation, founded in 1977 and headquartered in Redwood Shores, California, is one of the largest and most prominent companies in the software space – and a technology bellwether. • As it has grown, Microsoft has expanded into enterprise software with Windows Server, SQL Server, Dynamics CRM, SharePoint, Azure and Lync; hardware with the Xbox gaming/media platform and the Surface tablet; and online services through MSN and Bing. • Altria Group, Inc., is the world's largest producer and marketer of consumer products, and had revenues of $80 billion in 2002.
Industry Analysis	• According to our global Immunology market model, US Psoriasis (PsO) represented a $7.7B market in 2016 and is expected to grow at a low-teens CAGR to $11.8B in 2019E and $13B by 2021E driven by more highly effective therapies. • Despite record prices, oil demand continues to grow, while supply growth lags and spare production and refining capacity is almost nonexistent. • Add to that that some new advertising expectations from PricewaterhouseCoopers of a decline in ad spending of 12% worldwide and 15% in the U.S. for 2009 and continue to decline in 2010.
Competitive Landscape	• According to Mercury Research, NVIDIA is now the 3rd largest chipset supplier (consisting of desktop and mobile chipsets, and integrated and non-integrated chipsets), shipping 5.4 million units in calendar Q3 for an 8.2% market share, versus Intel's shipments of 51 million units (62% share), VIA's shipments of 9.6 million units (14.4% share), SiS's shipments of 5.3 million units (8% share), and ATI's shipments of 4.4 million units (6.6% share). • Further, competition in the CDK-4/6 space is rising with Verzenio (abemaciclib) & Kisqali launches placing downward pressure on Ibrance trajectory. • We believe that Accenture has recognized that web services could compete directly with client/server as the new systems architecture.
Financial Statement Analysis: Income Statement Analysis	

Table A7 – continued from previous page

Topics	Examples
	• IPG OM% was 18.2% vs. 17.0% and 19.8% in prior and year-ago periods.The company saw a continued recovery in O.ce print amid o.ce reopening, with a q/q increase in both share and backlog. • At the respective midpoints, sales of $8.5 billion would be down 22% annually and 8% sequentially; and non-GAAP EPS would be down 39%. • Adjusted operating expenses rose 5% in 2022, and management projects 4% growth in 2023 (4.5% on a constant-currency basis).
Financial Statement Analysis: Balance Sheet Analysis	• Long-term borrowings of $6.59 billion at September 30, 2022 were modest compared to shareholders' equity of $37.3 billion. • The company finished 2Q16 with $21.4 billion in cash and short-term investments, up from $15.6 Growth & Valuation Analysis GROWTH ANALYSIS RISK ANALYSIS billion at the end of 4Q15. • Inventory turns improved to 4.5x in 4Q from 4.4x in the same period last year.
Financial Statement Analysis: Cash Flow Analysis	• UPS generated $2.3 billion in operating cash for the quarter. • Assuming that the company completes a large portion of its current $1 billion stock buyback plan in 2008, we estimate that cash per share will be about $6.20 by the end of the year. • We view eBays FCF generation as relatively defensible even in the case of a revenue shortfall.
Financial Statement Analysis: Financial Ratios	• The company achieves average scores on our three main measures of financial strength: leverage based on debt-to-cap, profitability and interest coverage. • P/S is at 0.6x and EV/EBITDA is 7.7x. • The index members currently trade at an average of 16.3-times trailing earnings, which is below the five-year average of 19-times.
Business Segments	• The company is organized into three businesses, software, representing the majority of the company's total revenues, hardware systems and services. • Results in 2008 also benefited from the absence of the significant level of mark-to-market losses in the company's Gas Marketing segment in 2007. • The company operates five distinct segments: Americas (71% of FY15 profit); Europe, Middle East, and Africa (4% of FY15 profit); China and Asia Pacific (12% of FY15 profit); Channel Development (13% of FY15 profit).
Growth Strategies	

Table A7 – continued from previous page

Topics	Examples
	• Visa has been especially active on the acquisition front over the last several months. • Combined with the Horizon assets and an emerging pipeline, there is enough in AMGN's portfolio to offset potential headwinds and allow the company to grow through its patent cycle. • Specific areas where more investment may be needed: Over the past decade, ROK management often claim that Process automation represents the growth opportunity for the company, but its sales in this market have barely grown in recent years, despite its much smaller sales base compared with established incumbents.
Risk Factors	• Risks to our BUY thesis have to do with global competition, changing user behavior, global macro uncertainty, and anything else that can affect FB's relationship with members, its advertisers or its publishing partners. • Risks to achieving our price target include: 1) Apple crushing PayPal; 2) increasing competition in the payments space; 3) heavy investment spending on marketing, point of sale, or technology; and 4) legislative action. • In addition to the expenses incurred by patent challenges, product liability and other legal suits could occur and lead to additional liabilities and revenue loss, which could substantially change our financial assumptions.
Management and Governance	• Top management changes can be unsettling, and the resulting uncertainty has caused 3M shares to decline. • Chairman, President, and CEO Charles Ergen beneficially owns about 53.6% of DISH's equity securities and has 90.5% voting power. • Bill Johnson, the present CEO of Progress Energy, will become president and chief executive officer of the new company.
Environmental, Social, and Governance (ESG) Factors	• The assessment of ESG (Environmental, Social & Governance) risks by Baptista Research includes a wide range of considerations that pertain to the long-term sustainability of a company. • Sustainalytics assesses the degree to which a company's enterprise business is affected by ESG issues. • Failure to adequately address social risks like labor disputes and community relations could jeopardize the company's social license to operate in certain regions.
Valuation	• Our DCF derives an intrinsic value of ¿$100 for ABBV by discounting cash flows through 2024E, and assuming a -5% terminal growth & 7.6% WACC. • We value MET based on a Sum-Of-The-Parts (SOTP) analysis based on our 2021 EPS estimate and using peer comps across each business segment. • Our target price is based on a five-year discounted cash flow (DCF) valuation that employs a 5% discount rate and 20x terminal-year FCF multiple.
Investment Thesis	

Table A7 – continued from previous page

Topics	Examples
	• The results for the back half of the year will still be complex and confusing given the purchase accounting impacts and the full quarter of HSBC, but we do believe that there is a pay-off at the end of the road. • As good as it gets: With its record multi-year backlog, Boeing's revenue profile over the rest of decade is generally considered secure, and expectations for execution and cash already appear high. • Clearly, our Ford Investment Thesis, which was based in large part on our belief that Ford would be able to offset headwinds (slowing cyclical tailwinds in North America, weakness in South America, weak growth in Europe, slowing growth in China, and regulatory cost headwinds), has been thrown into question.
Appendices and Disclosures	• Although the information contained in the subject report has been obtained from sources, we believe to be reliable, its accuracy and completeness cannot be guaranteed. • For a complete discussion of the risk factors that could affect the market price of a company's shares, refer to the most recent Form 10-Q or 10-K that a company has filed with the Securities and Exchange Commission. • The Benchmark Company, LLC makes every effort to use reliable, comprehensive information, but we make no representation that it is accurate or complete.

C. Machine Learning Models

Ridge Regression

Ridge regression addresses multicollinearity by adding a regularization term to the least squares objective function. The ridge regression estimator is given by:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \|\beta\|_2^2 \right\}, \tag{27}$$

where α is the regularization parameter that controls the trade-off between fitting the data and shrinking the coefficients.

To find the optimal value of α, cross-validation is used over a grid of values ranging from 10^{-10} to 10^{10}. The cross-validation process ensures that the chosen model generalizes well to unseen data, preventing overfitting while capturing the predictive power of the text embeddings.

Partial Least Square Regression

To mitigate the risk of overfitting inherent in high-dimensional text embeddings, I employ Partial Least Squares (PLS) for dimensionality reduction.

The optimization problem can be expressed as follows:

$$\theta = \operatorname{argmin}((\Omega' X_i)'\theta - y_i), \tag{28}$$

where Ω is a $K \times P$ transformation matrix that reduces the K predictors in X_i to P lower-dimensional components.

The extraction of the j-th PLS component is guided by the following objective function:

$$\omega_j = \underset{\omega}{\operatorname{argmax}} \operatorname{Cov}(Y, X\omega), \quad \text{s.t. } \omega'\omega = 1,\ \operatorname{Cov}(X\omega, X\omega_i) = 0\ \forall i < j. \tag{29}$$

In essence, this approach sequentially extracts components that maximize the covariance with the outcome variable, while ensuring orthogonality to previously extracted components.

Extreme Gradient Boosting

Tree-based approaches are commonly applied in stock return forecasting literature (see, e.g., Gu et al., 2020; Cao et al., 2024; Bonini et al., 2023). XGBoost is an advanced implementation of tree-based machine learning models that builds an ensemble of decision trees. In XGBoost, each tree is added sequentially to correct the errors of previous trees. The main idea is to combine the outputs of multiple weak learners (decision trees) to create a strong learner. XGBoost incorporates regularization techniques, such as L1 (Lasso) and L2 (Ridge), to prevent overfitting and enhance model generalization.

In comparison to random forests, which build a multitude of independent trees and aggregate their predictions, XGBoost constructs trees sequentially, with each tree designed to correct the errors of the preceding ones. While random forests rely on bagging, a method that combines the predictions of various trees to reduce variance, XGBoost uses boosting, an approach that aims to reduce both bias and variance by focusing on difficult-to-predict instances in subsequent iterations. This boosting approach allows XGBoost to effectively capture and leverage the nuanced information embedded in textual data, leading to a more accurate estimation of stock returns.

Neural Networks

To extend beyond the linear modeling approach, I explore the use of Neural Networks to estimate *CAR* using text embeddings derived from analyst reports. A Neural Network can capture complex non-linear relationships between the text embeddings and the *CAR*, potentially improving the model's fitting capabilities. Consider a three-layer Neural Network as the prediction model. The prediction problem can be formulated as follows:

$$f(X_i;\theta) = W_3\sigma\left(W_2\sigma\left(W_1X_i + b_1\right) + b_2\right) + b_3, \tag{30}$$

where $\sigma(\cdot)$ is the ReLU activation function, W_i and b_i represent the weights and biases for layer i, respectively. The Neural Network architecture consists of an input layer representing the text embeddings, followed by multiple hidden layers. The specific architecture employed includes 32 neurons in the first hidden layer, followed by 16, 8, 4, and 2 neurons in subsequent layers. This structure is flexible and can be adjusted by adding or removing layers as necessary to optimize performance.

The training process involves optimizing the weights and biases to minimize the loss function, which, in this case, is the mean-square loss. To regularize the model and prevent overfitting, early stopping is implemented, halting training once the validation loss ceases to decrease. Additionally,

following Gu et al. (2020), the models are retrained five times, and the final estimation is obtained by averaging the outputs of these five models, forming an ensemble estimation.

D. Theory

I provide an intuition for the measure of strategic value by discussing a simple extension of Kyle (1985) model. In the Kyle model, there is one risky asset with a payoff $\tilde{v} \sim N(p_0, \Sigma_0)$. Three types of traders exist: a strategic trader with insider information, a market maker who sets prices in a perfectly competitive market, and an uninformed trader who trades $\tilde{u} \sim N\left(0, \sigma_u^2\right)$, where $\tilde{u}$ is independent of $\tilde{v}$. Illiquidity is measured by Kyle's lambda (λ). Kyle's lambda depends on private information and liquidity trading. I extend the model by considering a case where ϕ percentage of variance in $\tilde{v}$ is explained by the informed trader's information.

A Single Auction Model

There are two periods, t_0 and t_1. The asset is traded with asymmetric information at period t_0, and the value $\tilde{v}$ is realized at period t_1. I assume without loss of generality that

$$\tilde{v} = P_0 + \tilde{s} + \tilde{\varepsilon},$$

where $\tilde{s}$ is a mean-zero signal observed by the informed trader at t_0, and $\tilde{\varepsilon}$ is the combination of residual information and noise. I assume that the signal $\tilde{s}$ and residual $\tilde{\varepsilon}$ are uncorrelated, that is, $\sigma_v^2 = \sigma_s^2 + \sigma_\varepsilon^2$.

Let

$$\phi = \frac{\text{var}(\tilde{s})}{\text{var}(\tilde{v})} = \frac{\sigma_s^2}{\sigma_v^2}.$$

This measure ϕ is the "R-square" of projecting $\tilde{v}$ on $\tilde{s}$, or the percentage of explainable variance in $\tilde{v}$ using signal $\tilde{s}$.

After observing $\tilde{s}$, the informed trader submits a market order $\tilde{x} = X(\tilde{s})$, and the uninformed trader trades a zero-mean random variable $\tilde{z}$ that is normally distributed and independent of $\tilde{v}$. The profit of the informed trader is given by: $\tilde{\pi} = (\tilde{v} - \tilde{p})\tilde{v}$. The insider has a rational guess of $P(\tilde{x} + \tilde{u})$ and understands that his order $\tilde{x}$ will move the price against him.

The market maker observes the order flow $\tilde{y} \stackrel{\text{def}}{=} \tilde{x} + \tilde{z}$ and determines the equilibrium price $\tilde{p} = P(\tilde{x} + \tilde{u})$ to break even. The assumptions for the market maker are risk neutrality and perfect

competition, which drives the profits for market makers to zero.

Equilibrium

An equilibrium is a set of X and P satisfying

$$E[\tilde{\pi}(X,P) \mid \tilde{s} = s],$$

$$\tilde{p}(X,P) = E[\tilde{v} \mid \tilde{x} + \tilde{u}].$$

The first condition is profit maximization, stating that given the market maker's pricing rule, the insider chooses a strategy X maximizing her conditional expected profit, taking into account the pricing rule. The second condition is market efficiency. Given the insiders' trading strategy, the market maker sets the price to be the expected value of the security.

Conjecture:

$$P(\tilde{y}) = \mu + \lambda \tilde{y},$$
$$X(\tilde{s}) = \alpha + \beta \tilde{s}.$$

The profit of the insider can be written as:

$$\begin{aligned} E[\tilde{\pi}(X,P) \mid \tilde{s} = s] &= E[(\tilde{v} - \mu - \lambda(\tilde{u} + x))x \mid \tilde{s} = s] \\ &= (P_0 + s + -\mu - \lambda x)x. \end{aligned}$$

Traders take into account that her order flow will move the price against her, which serves to restrain her position size.

Solving for optimal profit, I get:

$$x^* = \frac{P_0 + s - \mu}{2\lambda} = \alpha + \beta v.$$

Hence, I can express α and β :

$$\beta = \frac{1}{2\lambda},$$
$$\alpha = \frac{P_0 - \mu}{2\lambda} = (P_0 - \mu)\beta.$$

When the market maker puts a higher weight on the order flow in setting the price, the trader

puts a lower weight on her information.

I now look at the price-setting rule:

$$\mu + \lambda y = E\{\tilde{v} \mid \alpha + \beta\tilde{s} + \tilde{u} = y\}.$$

Essentially, the market maker observes a normally distributed signal about

$$\begin{aligned}\mu + \lambda y &= E[\tilde{v} \mid y] \\ &= \bar{v} + \frac{\text{cov}(\tilde{v}, \alpha + \beta\tilde{s} + \tilde{u})}{\text{var}(\alpha + \beta\tilde{s} + \tilde{u})}(\beta\tilde{s} - \beta\bar{s} + \tilde{u}) \\ &= p_0 + \frac{\beta\sigma_s^2}{\beta^2\sigma_s^2 + \sigma_u^2}(y - \alpha).\end{aligned}$$

Hence, I can express λ and μ :

$$\begin{aligned}\lambda &= \frac{\beta\sigma_s^2}{\beta^2\sigma_s^2 + \sigma_u^2}, \\ \mu &= p_0 - \alpha\lambda.\end{aligned}$$

There is a unique linear equilibrium given by

$$\begin{aligned}\mu &= P_0, \\ \lambda &= \frac{\sigma_s}{2\sigma_u}, \\ \alpha &= 0, \\ \beta &= \frac{\sigma_u}{\sigma_s}.\end{aligned}$$

Discussion

Kyle's Lambda The parameter λ is universally known as Kyle's lambda. Formally, it is the impact on the equilibrium price of a unit order. Its reciprocal $(1/\lambda)$ measures the liquidity (or depth) of the market. If $1/\lambda = \frac{2\sigma_u}{\sigma_s}$ is larger, then the market is more liquid, either because there is less private information in σ_s or there is more liquid trade in the sense of σ_u.

Information Relevance The variance of $\tilde{s}$ measures the informational advantage of the informed trader, with a larger σ_s indicating a significant advantage due to the trader's estimate $\tilde{s}$ differing considerably from the market makers' perceived value $\bar{s}$. The information revealed to market

makers by the order flow $\tilde{y}$ in linear equilibrium indicates that the variance of $\tilde{s}$ conditional on $\tilde{y}$ is half of its unconditional variance. This is because the equilibrium price, being affine in $\tilde{y}$, conveys the same information, allowing the market to learn half of the trader's private information. To compute the conditional variance of $\tilde{s}$ given $\tilde{y}$, I use the formula $\sigma_{s|y}^2 = \sigma_s^2 \left[1 - \text{corr}(\tilde{s},\tilde{y})^2\right]$. The correlation between $\tilde{s}$ and $\tilde{y}$ is derived from $\text{corr}(\tilde{s},\tilde{y}) = \frac{\sigma_s^2/(2\lambda)}{\sigma_s \sigma_y}$, and with $\sigma_y = \sqrt{\sigma_s^2/(4\lambda^2) + \sigma_z^2}$, the correlation simplifies to $\frac{1}{\sqrt{2}}$. Substituting this back, the conditional variance becomes $\sigma_{s|y}^2 = \sigma_s^2 \left[1 - \left(\frac{1}{\sqrt{2}}\right)^2\right] = \frac{\sigma_s^2}{2}$. Thus, the conditional variance of $\tilde{s}$ given $\tilde{y}$ is half of its unconditional variance, showing that the market learns half of the informed trader's private information.

Value of Private Information Notice that the equilibrium strategy of the informed trader is $\tilde{x} = \beta s$. The unconditional expected gain of the informed trader is

$$\begin{aligned}
\text{E}[\tilde{x}[\tilde{v} - p(\tilde{x} + \tilde{z})]] &= \beta \text{E}[s(\tilde{v} - \mu - \lambda\beta s - \lambda\tilde{u})] \\
&= \beta(1 - \lambda\beta)\sigma_s^2 \\
&= \frac{\sigma_s \sigma_u}{2} \\
&= \frac{\phi \sigma_v^2}{4\lambda}.
\end{aligned}$$

The expected gain for the informed trader is maximized when she has more private information or when there is more liquidity trading. On the other hand, liquidity traders incur losses equivalent to the informed trader's gains, but they accept these losses due to other motives for trading. The equilibrium price ensures that market makers do not profit or lose in expectation.

E. Delta Method Approximation

I use the delta method to approximate the expected value and variance of a ratio of two random variables. This method relies on a first-order Taylor expansion. Specifically, consider two random variables X and Y with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2, respectively. We are interested in the ratio $Z = \frac{X}{Y}$ and want to approximate the mean $E[Z]$ and the variance $\text{Var}(Z)$.

Delta Method Approximation

The key idea is to approximate the function $g(X,Y) = \frac{X}{Y}$ using a Taylor series expansion around the means μ_X and μ_Y. For the function $g(X,Y)$, I use a first-order Taylor expansion around (μ_X, μ_Y):

$$g(X,Y) \approx g(\mu_X,\mu_Y) + \left.\frac{\partial g}{\partial X}\right|_{(\mu_X,\mu_Y)} (X-\mu_X) + \left.\frac{\partial g}{\partial Y}\right|_{(\mu_X,\mu_Y)} (Y-\mu_Y).$$

The partial derivatives of $g(X,Y) = \frac{X}{Y}$ are:

$$\frac{\partial g}{\partial X} = \frac{1}{Y}, \quad \frac{\partial g}{\partial Y} = -\frac{X}{Y^2}.$$

Evaluating these at (μ_X, μ_Y), I get:

$$\left.\frac{\partial g}{\partial X}\right|_{(\mu_X,\mu_Y)} = \frac{1}{\mu_Y}, \quad \left.\frac{\partial g}{\partial Y}\right|_{(\mu_X,\mu_Y)} = -\frac{\mu_X}{\mu_Y^2}.$$

Substituting the partial derivatives into the Taylor expansion, I obtain:

$$\frac{X}{Y} \approx \frac{\mu_X}{\mu_Y} + \frac{1}{\mu_Y}(X-\mu_X) - \frac{\mu_X}{\mu_Y^2}(Y-\mu_Y).$$

Taking the expectation on both sides:

$$E\left[\frac{X}{Y}\right] \approx \frac{\mu_X}{\mu_Y} + \frac{1}{\mu_Y}E[X-\mu_X] - \frac{\mu_X}{\mu_Y^2}E[Y-\mu_Y].$$

Since $E[X-\mu_X] = 0$ and $E[Y-\mu_Y] = 0$, the approximation simplifies to:

$$E\left[\frac{X}{Y}\right] \approx \frac{\mu_X}{\mu_Y}.$$

Using the delta method, the variance of $Z = \frac{X}{Y}$ is approximated by:

$$\text{Var}(Z) \approx \left(\frac{\partial g}{\partial X}\right)^2 \text{Var}(X) + \left(\frac{\partial g}{\partial Y}\right)^2 \text{Var}(Y) + 2\left(\frac{\partial g}{\partial X}\right)\left(\overline{\partial Y}\right)\text{Cov}(X,Y).$$

Substituting the partial derivatives:

$$\text{Var}\left(\frac{X}{Y}\right) \approx \frac{\text{Var}(X)}{\mu_Y^2} + \frac{\mu_X^2 \,\text{Var}(Y)}{\mu_Y^4} - 2\frac{\mu_X \,\text{Cov}(X,Y)}{\mu_Y^3}.$$

Application to Value of Information

In the context of the provided study, I estimate the mean value of information for a subsample s using the delta method.

Let $\widehat{\Omega}_{it} = \frac{\widehat{\sigma}_{it}^2}{\widehat{\lambda}_{it}/P_{it\_}} = \frac{r_{if}^2 - \left(r_{it} - \frac{\Sigma_{j=1}^N \widehat{r}_{ijt}}{N}\right)^2}{\widehat{\lambda}_{it}/P_{it\_}}$ be the value of information for stock i on date t. Define the mean and variance over subsample s as:

$$\mu_{vs} = \frac{1}{|s|}\sum_{it\in s}\left(r_{if}^2 - \left(r_{it} - \frac{\sum_{j=1}^N \widehat{r}_{ijt}}{N}\right)^2\right),$$

and the mean price impact per dollar over subsample s as:

$$\mu_{\lambda s} = \frac{1}{|s|}\sum_{it\in s}\frac{\widehat{\lambda}_{it}}{P_{it\_}}.$$

Using the delta method approximation for the mean of the ratio, the mean value of information over subsample s is given by:

$$E\widehat{\Omega}_s \approx \frac{\mu_{vs}}{\mu_{\lambda s}}.$$

The variance of the ratio $\widehat{\Omega}_{it} = \frac{\widehat{\sigma}_{it}^2}{\widehat{\lambda}_{it}/P_{it\_}}$ over the subsample s is estimated as:

$$\text{Var}\left(\widehat{\Omega}_s\right) \approx \frac{1}{\mu_{\lambda s}^2}\left(\Sigma_{vs} + \frac{\mu_{vs}^2}{\mu_{\lambda s}^2}\Sigma_{\lambda s} - 2\frac{\mu_{vs}}{\mu_{\lambda s}}\Sigma_{v\lambda s}\right),$$

where Σ_{vs}, $\Sigma_{\lambda s}$, and $\Sigma_{v\lambda s}$ are defined as:

$$\Sigma_{vs} = \frac{1}{|s|} \sum_{it \in s} \left(\widehat{\sigma}_{it}^2 - \mu_{vs} \right)^2,$$

$$\Sigma_{\lambda s} = \frac{1}{|s|} \sum_{it \in s} \left(\frac{\widehat{\lambda}_{it}}{P_{it_-}} - \mu_{\lambda s} \right)^2,$$

$$\Sigma_{v\lambda s} = \mathrm{Cov}\left(\widehat{\sigma}_{it}^2, \frac{\widehat{\lambda}_{it}}{P_{it_-}} \right).$$

This provides a first-order approximation for the mean and variance of the value of information over the subsample s using the delta method.