



Sentiment trading with large language models

Kemal Kirtac ^{a,*}, Guido Germano ^{a,b}

^a Department of Computer Science, University College London, 66–72 Gower Street, London WC1E 6EA, United Kingdom

^b Systemic Risk Centre, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom



ARTICLE INFO

JEL classification:

C53
G10
G11
G12
G14
G17

Keywords:

Natural language processing (NLP)
Large language models
Generative pre-trained transformer (GPT)
Machine learning in stock return prediction
Artificial intelligence investment strategies

ABSTRACT

We analyse the performance of the large language models (LLMs) OPT, BERT, and FinBERT, alongside the traditional Loughran-McDonald dictionary, in the sentiment analysis of 965,375 U.S. financial news articles from 2010 to 2023. Our findings reveal that the GPT-3-based OPT model significantly outperforms the others, predicting stock market returns with an accuracy of 74.4%. A long-short strategy based on OPT, accounting for 10 basis points (bps) in transaction costs, yields an exceptional Sharpe ratio of 3.05. From August 2021 to July 2023, this strategy produces an impressive 355% gain, outperforming other strategies and traditional market portfolios. This underscores the transformative potential of LLMs in financial market prediction and portfolio management and the necessity of employing sophisticated language models to develop effective investment strategies based on news sentiment.

1. Introduction

Many texts aim to understand and predict economic and financial events. In recent decades, the financial literature has turned to various sources of text data, such as financial news articles, regulatory filings, and social media posts, to extract valuable insights. However, the integration of text mining into financial models is still in its early stages. So far, most research has only explored a small portion of text data related to financial markets, often focusing on a single data source. Typically, these studies use straightforward methods, such as determining sentiment scores from dictionaries (Loughran and McDonald, 2011; Malo et al., 2014). A common finance-specific dictionary used in sentiment analysis is the Loughran-McDonald master dictionary (Loughran and McDonald, 2022).

Utilizing economic and financial information in text data poses difficulties. Unlike numerical data, text data lacks consistent structure, making it more complex to handle and interpret. Language itself is inherently intricate. Therefore, sophisticated models are necessary to uncover concealed insights from text. However, these advanced models can pose challenges, particularly for researchers who lack the technical expertise, computational resources, or sufficient funding. Given this context, it becomes apparent that current efforts in textual analysis within finance and economics are only beginning to explore the possibilities. There is a significant untapped potential in textual data that can be harnessed to gain a deeper understanding of asset markets. As challenges in text analysis emerge, they signal an exciting avenue for further research. It is expected that, in the future, economists will expand their text datasets and refine their techniques to extract valuable insights (Acemoglu et al., 2022).

The primary goal of this paper is to get deeper into this exploration. We aim to create more refined representations of news text using large language models (LLMs) and leverage these representations to develop models for predicting stock returns. To appreciate the value of LLMs, it is essential to understand the current landscape of financial text mining. Most methods in current use

* Corresponding author.

E-mail address: kemal.kirtac.21@ucl.ac.uk (K. Kirtac).

rely on supervised machine learning and are tailored for specific objectives, such as predicting returns, volatility, or macroeconomic variables (Jegadeesh and Wu, 2013; Baker et al., 2016; Manela and Moreira, 2017). These methods typically involve two main steps: converting text into a numerical representation and then modelling it. In many cases, researchers opt for the dictionary approach, which transforms each document into a vector that accounts for term frequency (Loughran and McDonald, 2011; Malo et al., 2014). Sometimes, this representation is further refined to create summarized scores that represent the text according to a domain-specific word dictionary (Tetlock, 2007). The outcome of the initial step is a matrix, which is then input into the econometric model in the subsequent step to describe various economic or financial phenomena. Studies such as Baker and Wurgler (2006), Lemmon and Ni (2014) and Shapiro et al. (2022) have highlighted the impact of investor sentiment on asset prices leveraging macro- or market-level sentiment indicators. However, these contributions do not deal with sentiment dynamics at the individual stock level, which our research aims to do by employing LLMs for a granular sentiment analysis. This approach not only enhances our understanding of sentiment's role in financial markets, but also introduces a novel dimension to sentiment analysis by focusing on the predictive power of LLM-derived sentiment scores for individual stock performances.

Current methods for representing financial text data have inherent limitations. Starting with the dictionary method, it oversimplifies text by solely considering term frequency and overlooking important aspects like word order and contextual relationships between terms (Malo et al., 2014). Additionally, the high dimensionality of a dictionary can lead to statistical inefficiencies, requiring many parameters in the subsequent modelling phase, even when many terms might be irrelevant. Dimensionality reduction techniques like latent Dirichlet allocation (LDA) can help, but are still based on a dictionary and do not address the fundamental issue of information loss (Bybee et al., 2019). Moreover, these reduced representations are specific to a text source in a particular domain, even though incorporating multiple text sources might provide better insights (Devlin et al., 2019).

A LLM can play a crucial role in text analysis. A LLM is trained on a wide variety of texts that cover different topics (Devlin et al., 2019). Creating an LLM involves a specialized team that builds a versatile language model. This model is trained on a vast amount of text, including entire books, Wikipedia articles, and more. Once developed, LLMs are made accessible for broader research purposes (Hugging Face, 2023).

In our research, we utilize two distinct LLMs for our analysis: Bidirectional Encoder Representations from Transformers (BERT) developed by Google (Devlin et al., 2019) and Open Pre-trained Transformers (OPT) developed by Meta (Zhang et al., 2022). State-of-the-art LLMs have consistently outperformed in various natural language processing (NLP) tasks, primarily due to their extensive scale. They are often available pre-trained on platforms like Hugging Face (Hugging Face, 2023). Hugging Face is a leading open-source repository that offers a plethora of advanced NLP and AI models, making state-of-the-art machine learning techniques readily available and deployable for academic and research purposes. We constructed our analytical models by leveraging the Hugging Face transformers library (Wolf et al., 2020) and utilizing Python.

We chose BERT for its innovation in language understanding, which is critical for analysing complex financial texts. BERT's ability to analyse the context of text from all sides improves how we interpret financial documents (Devlin et al., 2019). We would have liked to use also GPT-4 as a LLM in our research. However, we could not use the most advanced GPT models because they are not publicly available. Instead, we used OPT as a substitute. OPT is similar to the GPT models, but is open to the public (Radford et al., 2018). We used smaller versions of OPT due to our limited computer resources. For our analysis we chose BERT and OPT models with 345 million and 2.7 billion parameters (Hugging Face, 2023). Our choice of OPT and BERT provides us with powerful LLM tools to better understand financial markets and predict their movements. Furthermore, our analysis includes FinBERT, a specialized variant of the BERT LLM, specifically pre-trained for financial contexts. FinBERT is an example of how open-source BERT models can be adapted for distinct tasks in finance. Huang et al. (2023) demonstrated this adaptability by fine-tuning a BERT model for classification tasks using the Financial PhraseBank dataset, initially compiled by Malo et al. (2014). Last, we employed also the Loughran and McDonald dictionary (Loughran and McDonald, 2011; Loughran and McDonald, 2022) to facilitate a comparative assessment across different modelling approaches, i.e. FinBERT and our fine-tuned versions of BERT and OPT.

Our analysis is a two-step process: first, we turn text into numbers, and then we use those numbers to model economic patterns. In the first part, we choose how to represent text data numerically for the model in the second part. A popular method is the dictionary approach, which represents each text as a list of all the words and how often they appear. Sometimes, this basic list is all we use (Loughran and McDonald, 2011; Jegadeesh and Wu, 2013). Other times, we use the advanced LLM capabilities to analyse the news. The result of this first step is a table of numbers where each row represents a news text and each column represents a sentiment score. In the second step, we use this table to help us understand financial outcomes like stock returns or market trends. Using LLMs in the initial phase improves text representations compared to existing dictionary methods. This improvement comes from the extensive parameter space of LLMs and their training on a diverse range of language samples (Devlin et al., 2019). By enabling the use of LLMs, a wealth of knowledge becomes accessible for financial research (Huang et al., 2023). Our research primarily focuses on demonstrating the usefulness of LLM representations in modelling stock returns.

To achieve this, we conduct two separate econometric studies that leverage text mining to gain insights into the financial market. Firstly, we assess the relationship between financial news sentiment and stock returns using sentiment analysis techniques. This involves categorizing news sentiment based on the aggregated 3-day excess returns of stocks, providing a more dynamic and precise reflection of market response to news. Secondly, we conduct a regression analysis to determine how effectively LLM-derived sentiment scores predict stock returns on the following day. We utilize linear regression models with firm-specific and date-specific fixed effects, enabling a detailed evaluation of the predictive accuracy of LLMs, including OPT, BERT, and FinBERT, against traditional dictionary models. Finally, we explore practical applications through the development of distinct trading strategies based on LLM-derived sentiment scores. These strategies encompass creating long, short, and long-short portfolios, guided by sentiment scores from OPT, BERT, FinBERT and the Loughran-McDonald dictionary models. Our methodological design accounts for real-world

trading conditions by incorporating transaction costs and aligning trade execution with news release timings. This comprehensive approach not only highlights the nuanced sentiment analysis capabilities of LLMs, but also demonstrates their practical value in formulating effective trading strategies in the financial market.

This investigation both complements and extends the scope of existing scholarly works that combine text processing and machine learning approaches to tackle a range of questions in financial research (Jegadeesh and Wu, 2013; Garcia, 2013; Hoberg and Phillips, 2016; Manela and Moreira, 2017; Hansen et al., 2018; Ke et al., 2020). Additionally, it contributes to the ongoing research focused on the correlation between news sentiment and collective stock market performance (Tetlock, 2007; Campbell et al., 2014; Baker et al., 2016; Calomiris and Mamaysky, 2019; Frankel et al., 2022). A unique aspect of our study lies in exploring the capabilities of LLMs. We hypothesize that these sophisticated LLMs have the potential to unearth more profound, and perhaps previously unrecognized, insights from textual data, leading to enhanced accuracy in predicting stock market reactions.

The paper is organized as follows: Section 2 describes the data and methods we used for our study. Section 3 reports the study's results and discusses them. Section 4 contains the conclusions.

2. Data and methods

2.1. Data

In our research, we primarily use two datasets: one from the Center for Research in Security Prices (CRSP) that includes daily stock returns, and another from Refinitiv with global news. The news data from Refinitiv comprises detailed articles and quick alerts, focusing on companies based in the U.S. The CRSP data provides daily return information for companies that trade on major U.S. stock exchanges. It includes details like stock prices, trading volumes and market capitalisation. We use this data to analyse the link between stock market returns and sentiment scores derived from LLMs.

Our analysis includes companies from the American Stock Exchange (AMEX), National Association of Securities Dealers Automated Quotations (NASDAQ) and New York Stock Exchange (NYSE) that appear in at least one news article. We apply filters to ensure the quality of our data. We only consider news articles related to individual stocks with available three-day returns. Moreover, we avoid redundancy by using a novelty score based on the similarity between articles: if a new article is too similar to an older article published within the past 20 days according to a cosine similarity score of 0.8 or more, we exclude it. This approach helps us to focus on unique information significant for our analysis.

Our study covers the period from January 1, 2010, to June 30, 2023. We matched 2,732,845 news about 6214 unique companies. After applying our filters, we were left with 965,375 articles. Our sample dataset is summarized in Table 1.

Table 1

Summary statistics of our U.S. news articles dataset after we applied filters. "All news" is the total count of news items from Refinitiv. "News for single stock" is the post-filtering count for articles exclusively associated with individual stocks. "Unique news" is the remaining count after excluding articles with a high degree of similarity (cosine similarity score higher than 0.8) to any other content published within the prior five business days, ensuring a dataset without redundant information.

All news	News for single stock	Unique news
2,732,845	1,865,372	965,375

Table 2 presents descriptive statistics of our dataset. We find that the mean daily return is 0.37% with a standard deviation of 0.18%. The sentiment scores derived from the OPT, BERT, and FinBERT models show a normal distribution around the median of 0.5, with slight variations in mean and standard deviation. In contrast, the Loughran-McDonald dictionary score exhibits a more positively skewed distribution with a mean of 0.68 and a higher standard deviation of 0.32, indicating a tendency towards more positive sentiment scores in our dataset.

Table 2

Descriptive statistics for daily stock returns expressed in percentage points and sentiment scores derived from the OPT, BERT, and FinBERT models, as well as the Loughran-McDonald (LM) dictionary. The table provides the mean, standard deviation (StdDev), minimum, median, maximum, and the total count of observations N for each variable related to stock returns and sentiment scores from the mentioned models and dictionary.

	Mean	StdDev	Minimum	Median	Maximum	N
Daily return (%)	0.37	0.18	-64.97	-0.02	237.11	965,375
BERT score	0.48	0.25	0	0.5	1	965,375
OPT score	0.53	0.24	0	0.5	1	965,375
FinBERT score	0.51	0.24	0	0.5	1	965,375
LM dictionary score	0.68	0.32	0	0.5	1	965,375

2.2. Methods

This study commences with the fine-tuning of pre-trained language models, specifically BERT and OPT, sourced from Hugging Face, to tailor their capabilities for specialized financial analysis (Hugging Face, 2023). LLMs, originally designed for broad linguistic comprehension, require significant adaptation to perform niche tasks such as producing a sentiment index through textual analysis

of financial news with the ultimate aim of forecasting stock returns. This necessity enforces the adaptation phase, where the models are recalibrated after their original training on extensive data to prepare them for specific analytical functions (Radford et al., 2018).

In addition to the OPT and BERT LLMs, our analysis includes FinBERT, a variant of BERT pre-trained specifically for financial texts, and the Loughran and McDonald dictionary. Notably, FinBERT and the Loughran and McDonald dictionary do not necessitate the fine-tuning process, as they are already tailored for financial text analysis. FinBERT leverages BERT's architecture but is fine-tuned on financial texts, providing nuanced understanding in this domain (Huang et al., 2023). The Loughran and McDonald dictionary, a specialized lexicon for financial texts, aids in traditional textual analysis without the complexity of machine learning models (Loughran and McDonald, 2022).

Guided by the methodologies introduced by Alain and Bengio (2016), our approach adopts a probing technique, which is a form of feature extraction. This method builds on the models' pre-existing parameters, harnessing them to create features pertinent to text data, thereby facilitating the downstream task of sentiment analysis. To enhance the precision of our LLMs, we adapted and modified the methodology proposed by Ke et al. (2020). In our methodology, the process of fine-tuning the pre-trained OPT and BERT language models involves a specific focus on the aggregated 3-day excess return associated with each stock. This excess return is calculated from the day a news article is first published and extends over the two subsequent days. To elaborate, excess return is defined as the difference between the return of a particular stock and the overall market return on the same day. This calculation is not limited to the day the news is published; instead, it aggregates the returns for the following two days as well, providing a comprehensive three-day outlook.

Sentiment labels are assigned to each news article based on the sign of this aggregated three-day excess return. A positive aggregated excess return leads to a sentiment label of '1', indicating a positive sentiment. Conversely, a non-positive aggregated excess return results in a sentiment label of '0', suggesting a negative sentiment. Our approach of using a 3-day aggregated excess return for sentiment labelling plays a crucial role in refining our analysis. It follows the common practice in economics and finance of studying events that span multiple days (MacKinlay, 1997). This approach entails evaluating returns spanning from the day of the article's publication through the two following days. This technique is particularly beneficial in understanding the relationship between the sentiment in financial news and the corresponding movements in stock prices. We allocated 20% of the data randomly for testing and, from the remaining data pool, allocated another 20% randomly for validation purposes, resulting in a training set of 193,070 articles.

After completing the language model fine-tuning, our analysis continues with an empirical evaluation of these models in the context of U.S. financial news sentiment. A subset of 20% of these articles was set aside as a test sample, allowing for an unbiased evaluation of the models' predictive accuracy. Our analysis focused on the ability of OPT, BERT, FinBERT, and the Loughran-McDonald dictionary to accurately forecast the direction of stock returns based on news sentiment, particularly over a three-day period post-publication. To assess the models' performance, we calculated these statistical measures: accuracy, precision, recall, specificity, and the F1 score. Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Precision (positive predictive value) is the proportion of positive identifications that were actually correct. Recall (sensitivity) is the proportion of actual positives that were identified correctly. Specificity is the proportion of actual negatives that were correctly identified. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both quantities.

We subsequently conducted a regression analysis with the objective of investigating the influence of language model scores on the subsequent day's stock returns. The regression is modelled as

$$r_{i,n+1} = a_i + b_n + \gamma \cdot \mathbf{x}_{i,n} + \epsilon_{i,n}, \quad (1)$$

where $r_{i,n+1}$ is the return of stock i on the subsequent trading day $n+1$, $\mathbf{x}_{i,n}$ is a vector of scores from language models, and a_i and b_n are the fixed effects for firm and date. We employ double clustering for standard errors by firm and date, addressing potential concerns related to heteroscedasticity and autocorrelation. This regression framework facilitates an in-depth comparison of the predictive efficacy of different LLMs, including OPT, BERT, FinBERT and Loughran-McDonald dictionary variants with respect to stock returns.

Our choice of the linear regression model corresponds to a standard panel regression approach where article features $\mathbf{x}_{i,n}$ are directly translated into the expected return $E(r_{i,n+1})$ of the corresponding stock on the next day. The simplicity of linear regression is chosen to emphasize the importance of text-based representations in financial analysis. By using linear models, we can focus on the impact of these representations without the added complexity of nonlinear modelling. This approach highlights the direct influence of textual data on financial predictions, ensuring a clear understanding of the role and effectiveness of text-based features in financial sentiment analysis.

Following our predictive analysis, our study extends to assess practical outcomes through the implementation of distinct trading strategies utilizing sentiment scores derived from OPT, BERT, FinBERT and the Loughran-McDonald dictionary models. To comprehensively evaluate these strategies, we construct various portfolios with a specific focus on market value-weighted approaches. For each language model, we create three types of portfolios: long, short, and long-short. The composition of these portfolios is contingent on the sentiment scores assigned to individual stocks every day. Specifically, the long portfolios comprise stocks with the highest 20% sentiment scores, while the short portfolios consist of stocks with the lowest 20% sentiment scores. Moreover, the long-short portfolios are self-financing strategies that simultaneously involve taking long positions in stocks with the highest 20% sentiment scores and short positions in stocks with the lowest 20% sentiment scores. We observe cumulative returns of these trading strategies considering transaction costs. We dynamically update these market value-weighted sentiment portfolios on a daily basis in response to changes in sentiment scores. This means that each day, we reevaluate and adjust the portfolios

by considering the latest sentiment data. By doing so, we aim to capture the most current market conditions and enhance the effectiveness of our trading strategies.

This method allows us to test the real-world application of sentiment analysis findings without the influence of overall market movements. We base our stock choices on their market value, giving preference to larger, more stable companies, as these often represent safer, more reliable investments, and help reduce trading costs. We synchronize our trading decisions with the timing of news releases. For news reported before 6 am, we initiate trades at the market opening on that day, exploiting immediate reaction opportunities and close the position at the same date. For news appearing between 6 am and 4 pm, we initiate a trade with closing prices of the same day and exit the trade the next trading day. Any news coming in after 4 pm was used for trades at the start of the next trading day, adapting to market operating hours. To make our simulation more aligned with actual trading conditions, we included a transaction cost of 10 basis points for each trade, accounting for the typical costs traders would encounter in the market.

3. Results

3.1. Sentiment analysis accuracy in U.S. financial news

In this study, we used LLMs to analyse sentiment in U.S. financial news. We processed a dataset of 965,375 articles from Refinitiv, spanning from January 1, 2010, to June 30, 2023. We used 20% of these articles as a test set. We measured the accuracy of each model in predicting the direction of stock returns based on news sentiment. This accuracy indicates how well the model links the sentiment in financial news with stock returns over a three-day period. We evaluated four models: OPT, BERT, FinBERT and the Loughran-McDonald dictionary. Their performance in sentiment analysis is shown in [Table 3](#).

Table 3

Language model performance metrics: accuracy, precision, recall, specificity, and the F1 score for each model. The OPT model is the most accurate, followed closely by BERT and FinBERT.

Metric	OPT	BERT	FinBERT	Loughran-McDonald
Accuracy	0.744	0.725	0.722	0.501
Precision	0.732	0.711	0.708	0.505
Recall	0.781	0.761	0.755	0.513
Specificity	0.711	0.693	0.685	0.522
F1 score	0.754	0.734	0.731	0.508

The results show that the OPT model is the most accurate, followed closely by BERT and FinBERT. The Loughran-McDonald dictionary, a traditional finance text analysis tool, has significantly lower accuracy. This indicates that language models like OPT, BERT, and FinBERT are better at understanding and analysing complex financial news. The precision and recall values further support the superiority of the OPT model; its F1 score, which combines precision and recall, also confirms its effectiveness in sentiment analysis. These findings confirm that language models, particularly OPT, are valuable tools for analysing financial news and predicting stock market trends.

3.2. Predicting returns with LLM scores

This section assesses the ability of various LLMs to predict stock returns for the next day using regression models. Our regression with Eq. (1) uses LLM-generated scores from news headlines as the main predictors. To account for unobserved variations, these regressions include fixed effects for both firms and time, and we cluster standard errors by date and firm for added robustness. [Table 4](#) provides our regression findings, focusing on how stock returns correlate with predictive scores from advanced LLMs, specifically OPT, BERT, FinBERT and the Loughran-McDonald dictionary models.

Our findings reveal the predictive capabilities of the advanced LLMs. The OPT model, in particular, demonstrates a strong correlation with next-day stock returns, as indicated by significant coefficients in different model specifications. The FinBERT model follows closely, showcasing its own robust predictive power. BERT scores, while more modest in their predictive strength, still show a statistically significant relationship with stock returns. We also observe that the predictive strength increases when both LLMs are used as independent variables in the same regression. In contrast, the Loughran-McDonald dictionary model exhibits the least predictive power among the models examined.

Our analysis suggests that several factors contribute to explain the different performance among OPT, BERT and FinBERT, notably model design, parameter scale, and the specificity of training data. OPT's expanded parameter space, exceeding that of BERT and FinBERT, alongside its advanced training methodologies, is likely to cause its superior forecasting accuracy in stock returns and portfolio management. Furthermore, the nuanced performance of FinBERT, despite its financial domain specialization, raises intriguing considerations. Our exploration detailed in Section 3.3 posits that the broader pre-training data diversity of BERT and the potential for overfitting in highly specialized models such as FinBERT might explain this unexpected outcome. These insights collectively emphasize the intricate balance between model specificity, scale, and training regimen in optimizing predictive performance within financial sentiment analysis.

The robustness of our regression models is further underscored by the inclusion of a substantial number of observations, ensuring a comprehensive and representative analysis. Additionally, the adjusted *R*-squared values, while moderate, indicate a reasonable level of explanatory power within the models. The reported AIC and BIC values aid in assessing model fit and complexity, further enriching our comparative analysis across different LLMs.

Table 4

Regression results of stock returns on LLM sentiment scores done with Eq. (1), which includes firm and time fixed effects (FE) represented by a_i and b_n . The independent variable $x_{i,n}$ includes prediction scores from the language models. This analysis compares scores from OPT, BERT, FinBERT and Loughran-McDonald dictionary models, providing insights into their predictive abilities for stock market movements based on news sentiment. This analysis encompasses all U.S. common stocks with at least one news headline about the firm. T -statistics are presented in parentheses. Regressions 1 and 2 include two scores, regressions 3–6 only one. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Regression	1	2	3	4	5	6
OPT score	0.274*** (5.367)		0.254*** (4.871)			
BERT score	0.142** (2.632)	0.091* (1.971)		0.129* (2.334)		
FinBERT score		0.257*** (5.121)			0.181*** (4.674)	
LM dictionary score						0.083 (1.871)
Observations	965,375	965,375	965,375	965,375	965,375	965,375
R2	0.221	0.217	0.195	0.145	0.174	0.087
R2 adjusted	0.183	0.184	0.195	0.145	0.174	0.087
R2 within	0.021	0.022	0.017	0.009	0.016	0.002
R2 within adjusted	0.020	0.020	0.017	0.009	0.016	0.002
AIC	64,378	77,884	62,345	97,473	67,345	135,783
BIC	117,231	132,212	115,655	114,746	109,272	123,382
RMSE	5.32	11.12	4.21	14.12	9.75	23.54
FE: date	X	X	X	X	X	X
FE: firm	X	X	X	X	X	X

3.3. Performance of sentiment-based portfolios

Next, we assess the effectiveness of sentiment analysis in portfolio management by constructing various sentiment-based portfolios, including market value-weighted portfolios. These portfolios are developed using sentiment scores derived from different language models: OPT, BERT, FinBERT, and the Loughran-McDonald dictionary model. The investment strategies employed in our analysis can be described as follows: each LLM is utilised to create three distinct portfolios, one composed of stocks with top 20 percentile positive sentiment scores (long), another comprising stocks with top 20 percentile negative sentiment scores (short), and a self-financing long-short portfolio (L-S) based on both top 20 percentile negative and positive scores. Additionally, we include benchmark comparisons with value-weighted and equal-weighted market portfolios without considering sentiment scores. Value-weighted portfolios distribute investments based on the market capitalisation of each stock, while equal-weighted portfolios allocate investments equally to all stocks, regardless of market capitalisation. The selection of value-weighted and equal-weighted market portfolios was made to align with passive trading strategies, a widely acknowledged method in financial research (Fama and French, 1993; Carhart, 1997). We evaluate these strategies using key financial metrics, including the Sharpe ratio, mean daily returns, standard deviation of daily returns, and maximum drawdown.

The long-short OPT strategy demonstrates the most robust risk-adjusted performance, as evidenced by its superior Sharpe ratio indicated in Table 5. The Loughran-McDonald dictionary model-based strategy (L-S LM dictionary) lags behind, particularly when compared to the value-weighted market portfolio. This highlights the varying effectiveness of different sentiment analysis models in guiding investment decisions and underscores the significance of model selection in sentiment-based trading.

Table 5

Descriptive statistics of trading strategies. The table presents the Sharpe ratio, mean daily return (MDR), daily standard deviation (StdDev) and the maximum daily drawdown (MDD) for the trading strategies based on the sentiment analysis models OPT, BERT, FinBERT, and Loughran-McDonald dictionary (LM dictionary), each comprising long (L), short (S), and long-short (L-S) portfolios. The portfolios are value-weighted for comparison to a value-weighted (VW) market portfolio, which is provided for benchmarking, as well as an equal-weighted (EW) portfolio. LM dictionary refers to a sentiment analysis approach that uses a dictionary of finance-specific terms developed by Loughran and McDonald.

	OPT			BERT			FinBERT		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Sharpe ratio	1.81	1.42	3.05	1.59	1.28	2.11	1.51	1.19	2.07
MDR (%)	0.32	0.25	0.55	0.25	0.21	0.45	0.22	0.18	0.39
StdDev (%)	2.18	2.91	2.49	2.49	3.19	2.68	2.59	3.31	2.81
MDD (%)	-14.76	-24.69	-18.57	-17.89	-27.95	-21.95	-19.71	-29.94	-23.82
LM dictionary			EW			VW			
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Sharpe ratio	0.87	0.66	1.23	1.25	1.05	1.40	1.28	1.08	1.45
MDR (%)	0.12	0.13	0.22	0.18	0.15	0.33	0.19	0.16	0.35
StdDev (%)	3.54	4.13	3.74	2.90	3.70	3.20	2.95	3.75	3.25
MDD (%)	-35.47	-45.39	-38.29	-31.13	-42.21	-32.87	-28.76	-38.95	-31.87

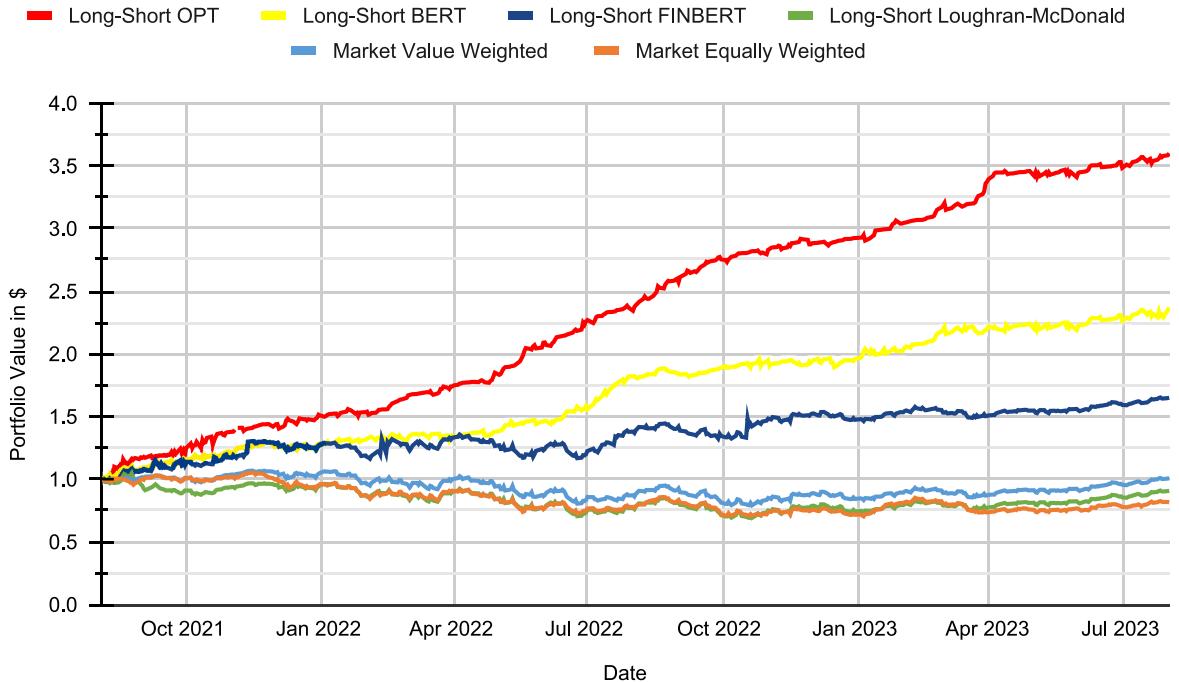


Fig. 1. Cumulative returns from investing \$1 with value-weighted, zero-cost long-short portfolios based on OPT (red), BERT (yellow), FinBERT (dark blue) and the Loughran-McDonald dictionary (green), rebalanced daily with a 10 bps transaction cost. For comparison, we also show a value-weighted market portfolio (light blue) and an equal-weighted market portfolio (orange), both without transaction costs.

Finally, we examine the outcomes of trading strategies based on news sentiment including a 10 bps trading cost from August 2021 to July 2023. Fig. 1 illustrates the performance of various strategies, notably highlighting the long-short OPT strategy with an impressive 355% gain. This underscores the powerful predictive capability of advanced language models in forecasting market movements. Other strategies, such as long-short BERT and long-short FinBERT, also register significant gains of 235% and 165%, in stark contrast to traditional market portfolios, which barely exceed 1%. Conversely, the Loughran-McDonald dictionary model, extensively employed in finance research, managed only a 0.91% return. This pronounced disparity suggests that dictionary-based models do not effectively interpret the nuanced sentiments present in contemporary financial news as efficiently as more advanced language models. This analysis substantiates the importance of employing sophisticated language models in developing investment strategies based on news sentiment.

4. Conclusion

Our study has far-reaching implications for the financial industry, offering insights that could reshape market prediction and investment decision-making methodologies. By demonstrating the application of OPT and BERT models, we enhance the understanding of LLM applications in financial economics. This encourages further research into integrating artificial intelligence and LLMs in financial markets.

Notably, the advanced capabilities of LLMs surpass traditional sentiment analysis methods in predicting and explaining stock returns. We compared the performance of OPT, BERT and FinBERT scores to sentiment scores derived from conventional methods such as the sentiment score provided by the Loughran-McDonald dictionary model. Our analysis reveals that the latter basic model exhibits limited stock forecasting capabilities, with little to no significant positive correlation between their sentiment scores and subsequent stock returns. In contrast, complex models like OPT demonstrate the highest prediction power. For instance, a self-financing strategy based on OPT scores, buying stocks with positive scores and selling stocks with negative scores after news announcements, achieves a remarkable Sharpe ratio of 3.05 over our sample period, compared to a Sharpe ratio of 1.23 for the strategy based on the dictionary model.

The implications of our research reach beyond the financial industry to inform regulators and policymakers. Our research enhances our knowledge of the advantages and risks linked to the increasing use of LLMs in financial economics. As LLM usage expands, it becomes crucial to focus on their impact on market behaviour, information dissemination, and price formation. Our results add insights to the dialogue surrounding regulatory policies that oversee the use of AI in finance, thereby aiding in the establishment of optimal practices for incorporating LLMs into the operations of financial markets.

Our research offers tangible benefits to asset managers and institutional investors, presenting empirical data that demonstrates the strengths of LLMs in forecasting stock market trends. In our analysis covering August 2021 to July 2023, we observed that strategies

utilizing news sentiment with advanced language models, notably the long-short OPT strategy, achieved remarkable gains, with an impressive 355% return. This starkly contrasts with the modest performance of traditional market portfolios and the Loughran-McDonald dictionary model, which only managed a 0.91% return. These findings highlight the significant advantage of employing sophisticated language models in developing effective investment strategies, marking a pivotal shift away from traditional sentiment analysis methods. Such evidence enables these professionals to make more informed choices regarding the integration of LLMs into their investment strategies. This could not only improve their performance but also decrease their dependence on traditional methods of analysis.

Our study contributes to the debate about the role of AI in finance, particularly through our investigation into how well LLMs predict stock market returns. By investigating both the possibilities and the boundaries of LLMs in the domain of financial economics, we open the way for further research aimed at creating more advanced LLMs specifically designed for the distinctive needs of the finance sector. Our goal in highlighting the potential roles of LLMs in financial economics is to foster ongoing research and innovation in the field of finance that is driven by artificial intelligence.

CRediT authorship contribution statement

Kemal Kirtac: Writing – original draft, Conceptualization, Methodology, Data Curation, Software, Investigation, Formal analysis.
Guido Germano: Writing — review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

- Acemoglu, D., Autor, D., Hazell, J., Restrepo, P., 2022. Artificial intelligence and jobs: Evidence from online vacancies. *J. Labor Econ.* 40 (S1), S293–S340. <http://dx.doi.org/10.1086/718327>.
- Alain, G., Bengio, Y., 2016. Understanding intermediate layers using linear classifier probes. [arXiv:1610.01644](https://arxiv.org/abs/1610.01644). <http://dx.doi.org/10.48550/arXiv.1610.01644>.
- Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. *Q. J. Econ.* 131 (4), 1593–1636. <http://dx.doi.org/10.1093/qje/qjw024>.
- Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *J. Finance* 61 (4), 1645–1680. <http://dx.doi.org/10.1111/j.1540-6261.2006.00885.x>.
- Bybee, L., Kelly, B.T., Manela, A., Xiu, D., 2019. The structure of economic news. *SSRN* 3446225. <http://dx.doi.org/10.2139/ssrn.3446225>.
- Calomiris, C.W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. *J. Financ. Econ.* 133 (2), 299–336. <http://dx.doi.org/10.1016/j.jfineco.2018.11.009>.
- Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H.-m., Steele, L.B., 2014. The information content of mandatory risk factor disclosures in corporate filings. *Rev. Account. Stud.* 19 (1), 396–455. <http://dx.doi.org/10.1007/s11142-013-9258-3>.
- Carhart, M.M., 1997. On persistence in mutual fund performance. *J. Finance* 52 (1), 57–82. <http://dx.doi.org/10.1111/j.1540-6261.1997.tb03808.x>.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. <http://dx.doi.org/10.18653/v1/N19-1423>.
- Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* 33 (1), 3–56. [http://dx.doi.org/10.1016/0304-405X\(93\)90023-5](http://dx.doi.org/10.1016/0304-405X(93)90023-5).
- Frankel, R., Jennings, J., Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. *Manage. Sci.* 68 (7), 5514–5532. <http://dx.doi.org/10.1287/mnsc.2021.4156>.
- Garcia, D., 2013. Sentiment during recessions. *J. Finance* 68 (3), 1267–1300. <http://dx.doi.org/10.1111/jofi.12027>.
- Hansen, S., McMahon, M., Prat, A., 2018. Transparency and deliberation within the FOMC: A computational linguistics approach. *Q. J. Econ.* 133 (2), 801–870. <http://dx.doi.org/10.1093/qje/qjx045>.
- Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. *J. Polit. Econ.* 124 (5), 1423–1465. <http://dx.doi.org/10.1086/688176>.
- Huang, A.H., Wang, H., Yang, Y., 2023. FinBERT: A large language model for extracting information from financial text. *Contemp. Account. Res.* 40 (2), 806–841. <http://dx.doi.org/10.1111/1911-3846.12832>.
- Hugging Face, 2023. Hugging face's transformer models. URL: <https://huggingface.co>.
- Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. *J. Financ. Econ.* 110 (3), 712–729. <http://dx.doi.org/10.1016/j.jfineco.2013.08.018>.
- Ke, Z., Kelly, B.T., Xiu, D., 2020. Predicting returns with text data. *SSRN* 3389884. <http://dx.doi.org/10.2139/ssrn.3389884>.
- Lemmon, M., Ni, S.X., 2014. The impact of investor sentiment on the market's reaction to stock splits. *Rev. Financ. Stud.* 27 (5), 1367–1401. <http://dx.doi.org/10.1093/rfs/huu006>.
- Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66 (1), 35–65. <http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x>.
- Loughran, T., McDonald, B., 2022. Master Loughran-McDonald word dictionary. URL: <https://sraf.nd.edu/loughranmcdonald-master-dictionary>.
- MacKinlay, A.C., 1997. Event studies in economics and finance. *J. Econ. Lit.* 35 (1), 13–39, URL: <http://www.jstor.org/stable/2729691>.
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P., 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.* 65 (4), 782–796. <http://dx.doi.org/10.1002/asi.23062>.

- Manela, A., Moreira, A., 2017. News implied volatility and disaster concerns. *J. Financ. Econ.* 123 (1), 137–162. <http://dx.doi.org/10.1016/j.jfineco.2016.01.032>.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training. OpenAI Blog, URL: <https://openai.com/research/language-unsupervised>.
- Shapiro, A.H., Sudhof, M., Wilson, D.J., 2022. Measuring news sentiment. *J. Econometrics* 228 (2), 221–243. <http://dx.doi.org/10.1016/j.jeconom.2020.07.053>.
- Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. *J. Finance* 62 (3), 1139–1168. <http://dx.doi.org/10.1111/J.1540-6261.2007.01232.X>.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 38–45. <http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6>.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. OPT: Open pre-trained transformer language models. [arXiv:2205.01068](https://arxiv.org/abs/2205.01068). <http://dx.doi.org/10.48550/arXiv.2205.01068>.



Predicting stock price trends using language models to extract the sentiment from analyst reports: Evidence from IBEX 35-listed companies

Alejandro Moreno^{a,b} *, Joaquín Ordieres-Meré^c

^a Universidad Pontificia de Comillas, Alberto Aguilera 23, Madrid, 28015, Spain

^b Fogram Analytics, Agustín Durán 24, Madrid, 28028, Spain

^c PMQ Research Group. ETSII UPM, José Gutiérrez Abascal 2, Madrid, 28006, Spain

ARTICLE INFO

Keywords:

Natural language processing
Large language models
Stock market prediction
Analyst recommendations

ABSTRACT

This study investigates the utility of large language models to extract sentiment from sell-side equity analysts' reports and their potential ability to predict stock price trends, using the IBEX 35 index as a case study. The RoBERTa, FinBERT, and GPT natural language processing models are employed to analyze a corpus of analysts' equity research reports over 2016–2022. The results indicate that the extracted sentiment can serve as a valuable tool for forecasting stock price movements, avoiding the potential bias in analyst reports when assigning a target price. This highlights the transformative potential of language models in the financial industry and their role in assisting investors in making informed investment decisions.

1. Introduction

Research on stock market prediction has received significant attention because of the potential benefits of successful strategies. However, the inherent unpredictability of the market, combined with the vast amount of information, numerous influencing variables and unanticipated noise, makes stock market forecasting highly challenging (Henrique et al., 2019). These complexities hinder investors' ability to make well-informed decisions (Bernales et al., 2021).

Stock price trends are essentially driven by the convergence of buyers' and sellers' economic decisions driven by their expectations. The efficient market hypothesis (EMH) (Fama, 1970) states that available information is already reflected in the stock price and that investors are rational. Specifically, the semi-strong form of the EMH suggests that all publicly available information is assumed to be incorporated into stock prices. The research of sell-side analysts does not reject the semi-strong form of the EMH. Instead, it explains the mechanism through which public information is incorporated into stock prices, supporting the principle of the EMH.

Still, the EMH has been challenged since it was proposed. Several works by behavioral economists and econometricians (Brown, 1999; Hsu et al., 2016) posit reasons to question this hypothesis, as demonstrated by the development of consistently profitable factors based on stock market anomalies (Azevedo and Hoegner, 2022). Investor expectations may differ from rational forecasts, built on personal beliefs

subject to social influence. Traditional models for predicting market behavior are based on either fundamental (company trends) or technical analysis (price trends) (McMillan, 2016).

Analyst forecasts frequently challenge the premise of the EMH that all relevant information has been incorporated into stock prices. Analysts may have access to private information not yet publicly available or fully comprehended by the market. They also provide expert interpretations. Given these two factors, analyst forecasts can offer a broader market understanding and even cause price movements that the EMH may not completely account for, indicating the potential inefficiency. Essentially, analyst forecasts may affect the EMH through the speed at which information is incorporated into stock prices. We can distinguish the following three effects both before and after the date on which a company releases its results:

- Earnings management. Analyst forecast can significantly shape firms' behavior, particularly their earnings management. Companies often face pressure to meet or exceed analyst forecasts, inducing them to adopt strategies to manage reported earnings. Givoly et al. (2011) highlight that such managerial responses can distort stock prices in the short-term, which may lead them to deviate from the EMH. These practices demonstrate that analyst forecasts not only reflect public information but also actively influence managerial decisions, potentially distorting the mechanism through which stock prices adjust to fundamentals.

* Corresponding author at: Universidad Pontificia de Comillas, Alberto Aguilera 23, Madrid, 28015, Spain.
E-mail address: amalonso@comillas.edu (A. Moreno).

- Information dissemination speed. The speed at which information is incorporated into stock prices often hinges on analysts' expertise and access to information. Analysts with domain-specific knowledge (Bradley et al., 2017) and privileged relationships with companies (Brown et al., 2015) can interpret complex data more rapidly than the average market participant. This accelerated information dissemination reduces asymmetry and facilitates quicker stock price adjustments, aligning with the EMH. However, temporary inefficiencies may arise when analysts' interpretations (Graaf, 2023) dominate market perceptions before independent validation by other participants.
- Impact of analysts forecast revisions. Analysts forecast revisions frequently cause significant market movements. As Beaver et al. (2008) discusses, revisions based on evolving interpretations of data can cause price movements that the market does not immediately predict. This challenges the EMH by demonstrating that even publicly available information can generate unexpected price movements because of analysts' interpretations. Moreover, the market's reliance on these revisions underscores a vulnerability to biases and inaccuracies in analyst forecasts, further complicating the principle of the EMH.

To examine a company's fundamental prospects, sell-side analysts produce equity reports for clients based on their familiarity with industry dynamics and companies in the sector (Abarbanell and Bushee, 1997). When a substantial number of analysts cover these stocks, this reduces market uncertainty and enhances investor rationality (Hou and Hu, 2023). Greater analyst coverage is critical in emerging markets in which the disclosure of company information is generally of low quality (Gao et al., 2020).

Equity analysts typically analyze a company's financial statements and management teams, industry trends, and other factors to determine the target price for the company's stock. They then issue recommendations such as "overweight", "neutral", and "underweight" (or any similar terminologies) based on their analysis and prediction of stock price trends. Analysts' influence on stock prices is highlighted by the Bloomberg rankings of the analysts covering a certain stock. The effectiveness of equity analysts' recommendations is significant for portfolio management (Markowitz, 1991). Moreover, the distribution of stock recommendations is skewed toward the positive side (Morgan and Stocken, 2003) because of the following factors:

- An investment bank issuing a negative recommendation on a company may see some influence on its other business with that company.
- An analyst issuing a non-positive recommendation on a company may not benefit from access to its senior management and other crucial information in the future.
- Positive recommendations may attract investment interest and increase brokerage fees for the analyst's bank.

Analysts are not robots and have biases (Pursiainen, 2018, Li, 2022, Karmaziene, 2023), or at least limited time and resources, as stated by Le and Trinh (2022) and Kim et al. (2022). According to Thas Thaker et al. (2018), analyst reports explain 66% of stock price trends. Similarly, Bandyopadhyay et al. (1995) show that while profits determine stock price trends in the long-term, which explains 60% of price variation, short-term stock price trends are determined by the sentiment of non-earnings variables (Nyakurukwa and Seetharam, 2023).

Artificial intelligence (AI) and machine learning (ML) are starting to be used to analyze market trends and help investors make better decisions. One such application of ML is natural language processing (NLP). NLP has the potential to enable human-like language interpretation in various applications, including analyzing stock-related news and earnings reports. Analyzing language essentially is analyzing sentiment. Sentiment analysis is an NLP technique that identifies the polarity of

a given text (i.e. positive, negative, or neutral). Although finance has involved AI since its early stages (Bickley et al., 2022), the full potential of conversational AI technologies is yet to be uncovered (Yue et al., 2023).

NLP research shows that *transformer* models have achieved remarkable performance in language modeling, surpassing previous dictionary-based algorithms. The release of large language models (LLMs) such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and generative pre-trained transformers (GPTs) (Radford et al., 2019), represented a substantial advance in NLP. Sentiment extraction by LLMs is far more insightful than previous sentiment analysis techniques, such as bag-of-words. The seminal study by Vaswani et al. (2017) on the attention mechanism was the first on transfer learning models based on transformer architectures. Such models include GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), Bloom (Laurençon et al., 2022), GPT-NeoX-20B (Black et al., 2022), and the META OPT-66B language model (Zhang et al., 2022).

The majority of financial studies on sentiment analysis employ lexicon-based approaches that concentrate on extensive word corpora. Wisniewski and Yekini (2015) review 1262 annual reports of 209 UK-listed firms over 2006–2012 to extract sentiment using a lexicon-based approach. The authors count the word frequency under three categories (praise, concreteness, and activity) to forecast future stock performance and find weak correlation coefficients with stock price trends (0.0708–0.0803). Loughran et al. (2011) study 50,155 annual reports of US companies over 1994–2008 and find that classifying lexicon-based sentiment may not extract sentiment properly in the financial domain, thereby reducing its meaningfulness. Li (2006) investigates 10,000 US annual reports using word frequency counting and reveals that reports with negative sentiment are followed by share price reductions for the companies concerned within 12-months.

Sul et al. (2017) suggests that sentiment analysis has a greater impact on small companies than on larger ones. Similarly, Bukovina (2015) finds that sentiment can influence stock prices and trade volumes. Duz Tan and Tas (2021) find that firm-specific *Twitter* sentiment contains information for predicting stock returns. The positive tone of *Twitter* sentiment is more pronounced in small and emerging market firms, consistent with the literature stating that small firms are hard to value and that emerging market firms contain high information asymmetry.

Olof (2019) processes equity analysts' reports using bag-of-words analysis, term frequency-inverse document frequency (TF-IDF), and Word2Vec to extract sentiment features. The author finds that analysts' upgrade and downgrade recommendations are the most informative labels, whereas neutral labels convey little information. Subsequently, the author applies logistic regressions jointly with convolutional neural networks to classify the sentiment. Schlaubitz (2021) trains a Distil-BERT model using data from the Financial PhraseBank to conduct sentiment analysis on Swiss earnings reports and news articles. The model predicts sentiment with an accuracy of 90%. Further, the sentiment analysis reveals that a relatively low percentage of earnings reports express negative sentiment, even when companies report lower earnings. Conversely, news articles are more balanced between classes. The author argues that text sentiment's ability to predict stock returns is weak, as only 4 of the 15 companies under study show a significant connection in linear regression analysis.

Meanwhile, the predictive ability of StonkBERT – an LLM in the financial domain (Pasch and Ehnes, 2022) – depends on the informational value of the underlying text data with the news sample outperforming both the blog and the annual report samples. Firms predicted to be "good", "average", and "bad" by StonkBERT showed average performances of 16.83%, 4.72%, and -3.17% in the 12-month post-prediction period, respectively. Experiments with PIXIU (Xie et al., 2023) show that the quality of the instructions, rather than the model size, is critical for LLM performance. Models that are not fine-tuned using financial datasets have a limited ability to predict stock price movement. Authors van Binsbergen et al. (2021) find that FinBERT (Araci,

2019; Yang et al., 2020) performs poorly when applied to hedge fund reports. This is because 95% of the reports on short-selling included words as “fraud” and “misconduct”, while FinBERT is trained on companies’ annual reports and conference calls using a language not mentioning those words. Fatouros et al. (2023) find that ChatGPT performs approximately 35% better in sentiment classification than FinBERT and shows a 36% higher correlation with market returns in the short term.

ChatGPT (Yue et al., 2023) is an LLM developed by OpenAI. We work with ChatGPT’s version based on GPT-4, which is trained on a vast corpus of text data and can generate coherent and meaningful responses to diverse questions and prompts. GPT-4 exhibits distinct behavior compared with the older GPT-3 model. For instance, GPT-3 followed a text-in and text-out approach, wherein it received a prompt string and provided a completion to be appended to the prompt. However, the GPT-4 model operates on a conversation-in and message-out basis.

This study seeks to advance research showing that the sentiment embedded in analyst reports has the power to predict stock price trends (Suzuki et al., 2022). Among related studies, Rybinski (2020b) reveals that the predictive power of NLP sentiment improves when it is combined with traditional macroeconomic data. Corbet et al. (2015) study the Greek and German markets to investigate the impact of analyst recommendations in three market states: rising, falling, and crisis. They find that “sell” recommendations increase volatility, whereas the influence of “buy” recommendations depends on the market state.

Owing to the popularity of applying LLMs in different research fields and their relatively limited use in financial studies, this study aims to answer the following research questions:

- RQ1: Is the sentiment of sell-side analysts consistent with their recommendations?
- RQ2: To what extent does the sentiment extracted from financial reports by LLMs improve their ability to predict stock price trends?

To assess analysts’ potential biases and determine their true intent, we evaluate whether the explicit sentiment in sell-side analysts’ reports is consistent with the implicit sentiment extracted by language models.

We then examine the predictive ability of both analysts and language models for different ranges of price movements and over different forecasting horizons. Specifically, in addition to analysts’ explicit recommendations, we compare the sentiment extracted by three LLMs: (a) the fine-tuned RoBERTa language model, (b) the financial domain-specific FinBERT, and (c) GPT-4 .

2. Data

Our corpus comprises sell-side analysts’ research reports in English for the 10 IBEX companies listed in Table 2 extracted from the Bloomberg database for 2016–2022.

We select sell-side rather than buy-side analysts’ research reports for the following reasons:

- Public availability: Sell-side reports are accessible, while buy-side reports are typically private.
- Market influence: Sell-side reports can significantly impact market prices and investor behavior because of their wide dissemination. Buy-side reports are private.
- Research focus: Sell-side analysts provide detailed financial modeling, industry analysis, and investment recommendations, which are useful for understanding market trends. Buy-side reports are released for internal portfolio management.
- Regulatory standards: Sell-side reports are subject to strict regulations, ensuring transparency and reliability. Buy-side reports follow internal compliance guidelines.
- Consensus estimates: Sell-side analysts’ estimates are aggregated to form widely tracked consensus benchmarks. The buy-side target price is not openly shared.

Table 1

Sizes of the European indexes. October 2024 (Bloomberg)

Index	Market cap. (EUR Bn)	Country
FTSE 100	2.580	UK
CAC 40	2.410	France
DAX 30	1.840	Germany
IBEX 35	611	Spain

Table 2

Selected IBEX 35 companies: IBEX share, volatility (Beta) and industry breakdown.

Ticker	% Market cap.	Beta	Industry
IBE	13.96	0.62	Power generation
ITX	13.38	1.08	Apparel
SAN	11.63	1.32	Banking
AMS	5.84	1.28	Data processing
TEF	4.41	0.68	Telecoms
FER	4.33	0.94	Infrastructure
AENA	3.56	1.23	Transport
IAG	1.92	2.30	Airlines
ELE	1.63	0.63	Power generation
ACX	0.48	1.21	Steelmaking

Our 10 selected IBEX 35-listed companies (Tables 1 and 2) are suitable for the analysis because they: (a) are small. The literature shows that small firm size is a factor in the degree of impact of analyst recommendations’ because of their lower liquidity and coverage (Lo, 2017), (b) have sufficient sell-side coverage, and (c) show industry diversification. After the implementation of Markets in Financial Instruments Directive II (MiFID II) (Keohane, 2018), which demands that the direct payments made to sell-side analyst be explicitly stated, the number of reports available fell drastically, by 15% according to Reuters (Reid et al., 2019), 47%–53% according to the CFA Institute (CFA Institute, 2019) and 30%–40% according to the European Commission (Risk Control Limited, 2020). They were free of charge until 2018.

Our sample is representative of the IBEX 35 index, as the selected companies account for more than 60% of the market capitalization and cover different sectors despite a concentration on the banking industry. This makes the analysis robust and not dependent on the data.

Our training data are unusual for training LLMs because they include data from reliable sources rather than web-scraped data, which is the usual case. The text on the first page of the sell-side equity research report, which contains a detailed summary, is added to the model. We then perform preprocessing to eliminate upper cases and remove any mention of the label assigned by the analyst to avoid forward-looking bias. These labels include sell/underweight/underperform (UP), maintain/neutral (N), and buy/overweight/outperform (OP).

After extracting the available research reports issued by Barclays, Deutsche Bank, JP Morgan, and Credit Suisse provided by Bloomberg, and discarding those with errors, we examine 379 reports (Tables 3 and 4).

The corpus is divided into the training dataset (303 reports, 80%) and testing dataset (76 reports, 20%).

We conduct a bag-of-words analysis of the equity reports of the selected companies to examine the most frequently used terms (Table 5). The top four most frequently used words typically account for 18%–26% of the company corpus.

As shown in Table 6, of the 17 most frequently mentioned words in Table 5, only 10 are from the financial domain (Loughran and McDonald, 2020). This shows the low discriminatory power of bag-of-words analysis (Amin et al., 2023).

3. Methodology

Sentiment is assessed for each equity research report and classified into positive, neutral, and negative sentiments. Studies analyzing the

Table 3

Sell-side equity reports for the selected IBEX companies issued by Barclays (BAR), Deutsche Bank (DB), JP Morgan (JPM), and Credit Suisse (CS).

Ticker	No. of documents	BAR	DB	JPM	CS
IBE	34	18	0	4	12
ITX	43	13	2	8	20
SAN	31	4	4	10	13
AMS	42	6	3	16	17
TEF	36	12	3	10	11
FER	37	2	0	35	0
AENA	43	11	2	12	18
IAG	32	9	2	6	15
ELE	39	20	0	7	12
ACX	28	1	8	6	13
Others	14	8	0	0	6
TOTAL	379	104	24	114	137

sentiment of analyst reports have not previously distinguished between current companies current situations and forward-looking perspectives, which are closely intertwined. LLMs' understanding of the current situation is critical for assessing sentiment (Yang et al., 2020, 2023a).

Using such a large corpus of reports from different sectors and over different time horizons is also considered to smooth the potential noise of one-off effects from such factors as macro data and political events.

As mentioned before, the LLMs classify the sentiment in each report as positive, neutral, or negative, indicating an upward trend (OP), no change (N), and a downward trend (UP), respectively. The stock price movement is then analyzed from the day before each report's release across various future time horizons: 2, 8, 16, 30, and 60 days.

We focus on the degree of correlation between the sentiment in these reports and stock price movements over the different time horizons, considering various ranges of price movements $X \in [1\%, 3\%, 5\%, 7\%]$.

$Price_t$ is the closing stock price on the last day of the interval considered, and $Price_0$ is the stock price on the day before the report is released:

$$\frac{Price_t - Price_0}{Price_0} > X\% \rightarrow OP \quad (1)$$

$$\frac{Price_t - Price_0}{Price_0} < -X\% \rightarrow UP \quad (2)$$

$$[-X\%, +X\%] \rightarrow N \quad (3)$$

We compare the stock price trends over the selected time horizon $t \in [2, 8, 16, 30, 60]$ days for each range of price movements $X \in [1\%, 3\%, 5\%, 7\%]$ to classify the movement as OP Eq. (1), UP Eq. (2), or N Eq. (3) if the movement is above X, below -X, or inside +/-X, respectively.

We use the weighted F1 score to evaluate model performance. We weight the F1 scores across the classes based on the number of instances of each class. We follow the four-step method below, where the first two steps represent the practical implementation and the last two provide the explanation and interpretation:

1. Calculate the F1 score for each class:

- For each class (OP, N, and UP), compute the precision Eq. (4) and recall Eq. (5) as follows:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Where TP: True Positive, FP: False Positive, and FN: False Negative.

- Then, calculate the F1 Eq. (6) score for each class as follows:

$$F1 = \frac{2 \cdot (Precision + Recall)}{Precision \cdot Recall} \quad (6)$$

2. Weight the F1 score by the number of instances of each class:

- In the weighted F1 score Eq. (7), each class's F1 score is weighted by the number of instances of that class. Thus, classes with more samples will contribute more to the overall score.

- The weighted F1 score is computed as follows:

$$F1_{weighted} = \frac{\sum_{c=1}^3 F1_c \cdot support_c}{\sum_{c=1}^3 support_c} \quad (7)$$

where $F1_c$ is the F1 score for each class c (OP, N, UP) and $support_c$ is the number of samples in that class.

3. Interpret the weighted F1 score:

- The weighted F1 score provides an overall measure of the model performance, considering both the performance of each class and the distribution of samples across classes.
- This is especially useful if the classes are imbalanced (i.e. more N samples than OP or UP), as it prevents the metric from being overly influenced by a class with fewer samples.

4. Compare the analyst's classification with the model's prediction:

- Comparing the analyst classification with the models' predictions using the weighted F1 score sheds light on how well the model performs across all the classes, accounting for any imbalances in the dataset.
- This metric highlights whether the model struggles with certain classes or aligns closely with the analyst's judgments across all the classes.

In summary, the weighted F1 scores provide a balanced view of model performance across all the categories (OP, N, and UP) while adjusting for class imbalances which allows for a comprehensive comparison of the models' and analysts' classifications.

Using the BERT model's transfer learning capabilities (Devlin et al., 2018), we fine-tune our models using a sample of sell-side reports as a training set to link sentiment with the analyst recommendations. Using these fine-tuned models, we incorporate text from the unseen analyst reports to test the models' ability to categorize sentiment. As price is not used to fine-tune the model, forward-looking bias is excluded (Sun et al., 2019). This methodology is similar to those of Yang et al. (2023b), Kirtac and Germano (2024).

3.1. RoBERTa

RoBERTa (Robustly optimized BERT Liu et al., 2019) has been pre-trained on a massive corpus of text and code. It improves Google's BERT model including dynamic masking at each epoch, longer sequences, larger batches, and byte pair encoding (BPE). This helps increase the text handling efficiency of RoBERTa compared to BERT.

RoBERTa outperforms BERT on various NLP tasks, including natural language inference, question answering, and sentiment analysis. It is thus one of the most popular and widely used LLMs worldwide.

We perform numerous tests using this model, obtaining the best results with a training of 10 epochs, batch size 16, and length limited to 512 tokens (sliding window).

3.2. FinBERT

FinBERT (Financial BERT Araci, 2019; Yang et al., 2020) is a pre-trained language model specializing in analyzing financial text. It has been specifically fine-tuned to understand financial languages and concepts. Given that it is trained on a large corpus of financial documents, it is adept at financial NLP tasks such as sentiment

Table 4

Number of equity reports for the 10 selected IBEX 35 companies (2016–2022).

Ticker	2016	2017	2018	2019	2020	2021	2022	No. of documents
IBE	3	2	2	2	9	7	9	34
ITX	0	0	0	6	12	15	10	43
SAN	0	0	0	0	1	15	15	31
AMS	0	0	4	6	20	8	4	42
TEF	0	0	0	4	9	10	13	36
FER	0	0	0	0	0	15	22	37
AENA	0	0	9	4	9	9	12	43
IAG	0	0	2	8	7	9	6	32
ELE	1	1	2	5	8	11	11	39
ACX	1	5	4	2	2	6	8	28
Other	2	2	1	1	1	3	4	14
TOTAL	7	10	24	38	78	108	114	379

Table 5

Top four most frequently used words in the analyst reports of the selected IBEX 35 companies, 2016–2022 (bag-of-words analysis).

Ticker	#1	#2	#3	#4	% Top 4
IBE	EBITDA	Env. Issues	Wind	CAPEX	23.39
ITX	Revenue	Margins	Pricing	Leverage	22.31
SAN	Revenue	Equity	Dividends	Regulation	26.42
AMS	Revenue	EBITDA	Cash Flow	Margins	22.67
TEF	Revenue	EBITDA	Cash Flow	Competition	22.14
FER	Net Debt	Regulation	Cash Flow	EBITDA	17.78
AENA	Revenue	Cash Flow	EBITDA	Free Cash Flow	21.68
IAG	Revenue	Cash Flow	Margins	Free Cash Flow	20.32
ELE	EBITDA	Margins	CAPEX	Env. Issue	22.83
ACX	EBITDA	Steel	Net Debt	Inventory	21.00

Table 6Frequency analysis of the terms extracted from [Table 5](#).

Word	Count
EBITDA	7
Revenues	6
Cash Flow	5
Margins	4
CAPEX	2
Env. Issues	2
Free Cash Flow	2
Net Debt	2
Regulation	2
Competition	1
Dividends	1
Equity	1
Inventory	1
Leverage	1
Pricing	1
Steel	1
Wind	1

analysis and entity recognition. FinBERT is particularly valuable for applications in the financial industry, including analyzing stock market sentiment and tracking the sentiment in financial news because it offers state-of-the-art sentiment scoring using the Financial PhraseBank dataset.

We test both RoBERTa and FinBERT using a sliding window technique because some texts may exceed their 512 token capabilities.

3.3. GPT-4

Since its release in November 2022, ChatGPT has revolutionized NLP. GPT-4, the most recent version when writing, is larger and stronger ([Liu et al., 2023](#)) than GPT-3.5.

In the context of stock portfolio management ([Ko and Lee, 2023](#)), ChatGPT can be used to analyze market trends, provide insights into stocks and sectors, and answer questions on investment strategies. ChatGPT can also provide information on a company's financials, recent news, and market trends.

Table 7

Stock price trend predictions: Analyst recommendations versus the three LLMs (%).

	Analyst	RoBERTa	FinBERT	GPT-4
UP	21	17	14	21
N	49	53	50	45
OP	30	30	36	34
TOTAL	100	100	100	100

The GPT-4 training set comprises publicly available online information. By contrast, analyst reports include client specific data unavailable online, preventing forward-looking bias.

We perform inferences on GPT-4 to categorize each report. We upload the analyst reports through the GPT-4 API and ask GPT-4 to assign the above mentioned labels: OP, N, and UP.

4. Results

The sentiment analysis revealed that the most equity analysts' reports carried a neutral sentiment, with fewer UP and a considerable number of OP reports, consistent with the incentives mentioned earlier ([Grant et al., 2015](#)). This suggests that analysts exhibit a degree of optimism in their reports. Indeed, this trend is evidenced by the proportion of predicted OP/Buy labels in the testing group for which both the LLM-extracted sentiment and analysts' targets assign very low UP recommendations ([Table 7](#)).

Once the sentiment is extracted from the report text, we perform the following two steps:

- We compare the consistency of the labels assigned by the analyst with those extracted from the report text ([Table 8](#)). The weighted F1 score is then used to check the alignment between the sentiment labels assigned by the analyst and those of the three LLMs. Market capitalization, industry, volatility, and analyst background are considered, with the latter factor the most influential in ensuring the consistency between the sentiment in the report text and the analyst's explicit recommendation.

Table 8

Weighted F1 scores for the selected companies analyzed by the three LLMs (Eq. (7)).

	An. vs RoBERTa	An. vs FinBERT	An. vs. GPT-4
IBE	1.00	1.00	1.00
ITX	0.89	0.94	1.00
SAN	1.00	1.00	1.00
AMS	1.00	0.95	0.85
TEF	0.92	0.64	0.92
FER	1.00	1.00	1.00
AENA	1.00	0.76	1.00
IAG	1.00	1.00	1.00
ELE	0.76	0.91	0.91
ACX	0.91	0.91	1.00

Table 9

Precision of the stock price prediction compared with the real stock price (Eq. (4)) 1% range of price movement. 1: best, 0: worst.

	2d	8d	16d	30d	60d
Analysts	0.28	0.32	0.26	0.33	0.26
FinBERT	0.28	0.30	0.25	0.30	0.24
RoBERTa	0.29	0.30	0.28	0.34	0.25
GPT-4	0.25	0.30	0.30	0.37	0.28

Table 10

Precision of the stock price prediction compared with the real stock price (Eq. (4)) 3% range of price movement. 1: best, 0: worst.

	2d	8d	16d	30d	60d
Analysts	0.43	0.37	0.30	0.36	0.32
FinBERT	0.45	0.39	0.32	0.36	0.30
RoBERTa	0.47	0.42	0.32	0.37	0.30
GPT-4	0.39	0.34	0.29	0.39	0.30

Table 11

Precision of the stock price prediction compared with the real stock price (Eq. (4)) 5% range of price movement. 1: best, 0: worst.

	2d	8d	16d	30d	60d
Analysts	0.45	0.45	0.45	0.36	0.33
FinBERT	0.46	0.47	0.45	0.37	0.34
RoBERTa	0.49	0.50	0.46	0.39	0.38
GPT-4	0.41	0.42	0.41	0.39	0.32

Table 12

Precision of the stock price prediction compared with the real stock price (Eq. (4)) 7% range of price movement. 1: best, 0: worst.

	2d	8d	16d	30d	60d
Analysts	0.49	0.47	0.39	0.37	0.39
FinBERT	0.50	0.50	0.39	0.38	0.42
RoBERTa	0.53	0.53	0.41	0.39	0.45
GPT-4	0.45	0.45	0.36	0.37	0.38

Table 13

Comparison of stock price trend predictions by time horizon (AN: Analysts, FI: FinBERT, RO: Roberta, GP: GPT-4).

	2d	8d	16d	30d	60d
1%	RO	AN	GP	GP	GP
3%	RO	RO	FI-RO	GP	AN
5%	RO	RO	RO	RO-GP	RO
7%	RO	RO	RO	RO	RO

- Considering the different ranges of price movements for the different time horizons, we compare the predictive ability of the analyst recommendation with the label extracted by the LLMs (Table 13).

We find that analysts are not particularly good at predicting price movements (Bradshaw et al., 2013) for any horizon, as shown in Tables 9, 10, 11, and 12.

In the short term (2d-8d), RoBERTa demonstrates higher precision than the other models for the 1%, 3%, 5%, and 7% ranges of price

movements (Table 13), as it seems to better anticipate the market reaction after analyst reports are issued.

Over time, the stock market experiences increasingly pronounced fluctuations (5%-7%). In these cases, the fine-tuned RoBERTa model performs better because its estimates tend to be able to detect upward and downward trends rather than adopting a neutral position (Tables 11 and 12). GPT-4 shows good performance in the medium term (30d-60d) when markets are flat (1%-3%), likely because of its generalist training and reliance on neutral predictions. Within the 3% range of price movements (Table 10), RoBERTa and FinBERT show identical precision, particularly over the 16d horizon, with their predictions outperforming analysts and GPT-4. The improvement in precision for RoBERTa and FinBERT in this range suggests that these models are better equipped to capture moderate price movements. Analysts exhibit comparable performance to GPT-4 over the 60d horizon under flat markets (1%-3%), highlighting the limitations of general models without fine-tuning in financial contexts. The selected time horizons reflect different trading and forecasting scenarios relevant to market participants. The short horizons (2d and 8d) capture immediate market reactions to new information, while the medium-term and longer horizons (16d, 30d, and 60d) assess the sustainability of predictive models over broader time periods. These horizons reflect the real-world timings of investment decisions and market dynamics, ensuring their relevance across various potential trading strategies.

5. Conclusions

Language models provide a new approach to examining behavioral finance by enabling advanced sentiment analysis. This study analyzes the analyst reports of IBEX 35 companies over 2016–2022 to assess the consistency of the recommendations in analyst reports with their sentiment. Companies with higher market capitalization and fewer reports show the greatest alignment between sentiment and analyst recommendations regardless of the industry and volatility level. The consistency between analysts' explicit recommendation and the sentiment found by LLMs is notable for UK-based analysts. By time frame, we find that:

- Short term: The predictive ability of the RoBERTa model is marginally higher than that of the other two LLMs.
- Medium term: GPT-4 tends to outperform the other two models if markets stay flat; contrast, the RoBERTa model is more precise during larger price movements.

This study demonstrates the superiority of smaller but finance-related pretrained LLMs such as RoBERTa over larger general models such as GPT-4, which are not fine-tuned. This is similar to the findings of Xie et al. (2023). The predictive power of LLMs with no additional data is low, consistent with the results of Rybinski (2020a). LLMs pretrained in the finance domain such as FinBERT, but with no fine-tuning do not perform as well of other LLMs. Larger general models such as GPT-4 do not discriminate, as they tend to err on caution by using neutral labels that match with flattish markets (i.e. 1%-3%).

We also find that GPT-4 is not specifically trained on analyst reports but has more extensive generalist training, consistent with Gururangan et al. (2020). This may explain why RoBERTa better identifies trends over longer forecasting horizons.

Overall, while language models can provide valuable insights (Lopez-lira and Tang, 2023; Pelster and Val, 2024) and GPT-4 is suitable for summarizing vast amounts information for retail investors (Kim et al., 2023b), they cannot replace the experience and knowledge of human investors. Nonetheless, by combining the insights generated by human experts' analysis with those from LLMs, investors can make more informed decisions, thereby reducing the potential for errors and biases (Cao et al., 2021). However, we find that the precision of NLP sentiment stock price trends hardly exceeds 50%, indicating its

low predictive capability. These results are similar to those obtained by Rybinski (2020b).

Our research highlights the importance of the continued exploration and development of language models in the financial industry (Li et al., 2023). Moreover, scholars should examine how such models affect market dynamics.

Our research contributes to the literature in the following ways:

- We show the consistency of analyst recommendations and potential bias, in line with Frijns and Garel (2021).
- We also show that the bigger the firm and the higher the analyst coverage, the better the GPT-4 forecast. These results are consistent with those of Li et al. (2023).
- Crucially, our work is novel in its use of analyst reports as its dataset (i.e. research reports from IBEX 35-listed companies over 2016–2022). Similar studies have only covered the Korean stock market (Kim et al., 2023a; Cho et al., 2021).
- We also use a novel methodology, under which both domain-specific and general LLMs are used (FinBERT, RoBERTa, and GPT-4). Previous work on analyst reports based on ML has focused on discovering the importance of forecasting features (Sidogi et al., 2022).

Still, some issues related to the development of LLMs need further attention:

- Secrecy and limited access to the training corpora of LLMs. Controversy continues about whether companies should open their proprietary LLMs to the public (i.e. BloombergGPT Wu et al., 2023), with OpenAI and META exhibiting opposite views.¹ This remains a keystone in developing LLMs in the financial domain jointly with datasets such as The Pile (Biderman et al., 2022), C4, and Wikipedia.
- Noise and instability. LLM strategies, especially when used by major institutional investors and hedge funds, may influence wider financial markets. Substantial capital allocation into specific securities and sectors because of LLM strategies can sway prices and market sentiment, potentially inciting herd behavior or unexpected repercussions that could engender systemic risks.

Overall, this study provides novel evidence on the predictive ability of state-of-the-art NLP models, such as FinBERT, RoBERTa, and GPT-4, using sell-side equity analysts' reports. Focusing on companies in the IBEX 35 index, this study demonstrates the potential of such NLP models to enhance the accuracy of stock price forecasts, even in markets with relatively high efficiency. Therefore, it is valuable for the literature on behavioral finance and ML applications in equity markets. This study also compares explicit recommendations by analysts with the implicit sentiment extracted by LLMs to examine the relationship between analyst sentiment and stock price dynamics. The results underline the limitations of human analysis by showing the complementary function of ML in financial decisions. Crucially, our work enriches the debate on the EMH by showing how analysts influence market efficiency through their sentiments and forecasts. Further, we provide a framework for further research on the intersection between NLP and financial markets. Practitioners such as investors and portfolio managers, can leverage NLP-derived sentiment analysis as a tool to refine decision-making processes, especially in anticipating price movements beyond analysts' explicit recommendations. Finally, our insights can help policymakers by showing the need to carefully consider sell-side analysts' role in influencing market dynamics, particularly in terms of ensuring transparency and mitigating biases in financial reports. Future research

could consider more general applications of this approach to other markets, particularly emerging markets in which information asymmetry and imperfections are higher. Meanwhile, extending sentiment analysis by incorporating new alternative data sources, such as social media and macroeconomic indicators, may improve the predictive power. Overall, the main strength of this study, which shows the efficiency of using NLP models in analyst reports, is that it closes the gap between traditional financial analysis and advanced AI technologies. These findings contribute not only to the literature but also to actionable knowledge relevant to market participants striving for higher forecasting accuracy and optimized investment strategies.

Data availability

Data will be made available on request.

References

- Abarbanell, J.S., Bushee, B.J., 1997. Fundamental analysis, future earnings, and stock prices. *J. Account. Res.* 35 (1), 1.
- Amin, M.M., Cambria, E., Schuller, B.W., 2023. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intell. Syst.* 38 (2), 15–23.
- Araci, D., 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv*.
- Azevedo, V., Hoegner, C., 2022. Enhancing Stock Market Anomalies with Machine Learning. Number 0123456789. Springer US.
- Bandyopadhyay, S.P., Brown, L.D., Richardson, G.D., 1995. Analysts' use of earnings forecasts in predicting stock returns: Forecast horizon effects. *Int. J. Forecast.* 11 (3), 429–445.
- Beaver, W., Cornell, B., Landsman, W.R., Stubben, S.R., 2008. The impact of analysts' forecast errors and forecast revisions on stock prices. *J. Bus. Financ. Account.* 35 (5–6), 709–740.
- Bernales, A., Valenzuela, M., Zer, I., 2021. Effects of information overload on financial market returns: How much is too much? *SSRN Electron. J.* 2500 (1372).
- Bickley, S.J., Chan, H.F., Torgler, B., 2022. Artificial intelligence in the field of economics. *Scientometrics* 127 (4), 2055–2084.
- Biderman, S., Bicheno, K., Gao, L., 2022. Datasheet for the pile. pages 1–22.
- van Binsbergen, J.H., Han, X., Lopez-Lira, A., 2021. Textual analysis of short-seller research reports, stock prices and real investment. *SSRN Electron. J.*
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U.S., Purohit, S., Reynolds, L., Tow, J., Wang, B., Weinbach, S., 2022. GPT-NeoX-20B: An open-source autoregressive language model. In: *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 95–136.
- Bradley, D., Gokkaya, S., Liu, X., Xie, F., 2017. Are all analysts created equal? Industry expertise and monitoring effectiveness of financial analysts. *J. Account. Econ.* 63 (2–3), 179–206.
- Bradshaw, M.T., Brown, L.D., Huang, K., 2013. Do sell-side analysts exhibit differential target price forecasting ability? *Rev. Account. Stud.* 18 (4), 930–955.
- Brown, E., 1999. A non-random walk down wall street. *J. Econ. Surv.* 13 (4), 477–478.
- Brown, L.D., Call, A.C., Clement, M.B., Sharp, N.Y., 2015. Inside the black box of sell-side financial analysts. *J. Account. Res.* 53 (1), 1–47.
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33(NeurIPS) 1877–1901.
- Bukovina, J., 2015. Sentiment and Blue-Chip Returns. Firm Level Evidence from a Dynamic Threshold Model. Technical report, Mendel University in Brno, Faculty of Business and Economics.
- Cao, S.S., Jiang, W., Wang, J.L., Yang, B., 2021. From man vs. Machine to man machine: The art and ai of stock analyses. *SSRN Electron. J.*
- CFA Institute, 2019. MIFID II: ONE YEAR on Assessing the Market for Investment Research. Technical report.
- Cho, P., Park, J.H., Song, J.W., 2021. Equity research report-driven investment strategy in Korea using binary classification on stock price direction. *IEEE Access* 9, 46364–46373.
- Corbet, S., Dowling, M., Cummins, M., 2015. Analyst recommendations and volatility in a rising, falling, and crisis equity market. *Financ. Res. Lett.* 15, 187–194.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. vol. 1, (Mlm), pp. 4171–4186.

¹ <https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html>

- Duz Tan, S., Tas, O., 2021. Social media sentiment in international stock returns and trading activity. *J. Behav. Financ.* 22 (2), 221–234.
- Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *J. Financ.* 25 (2), 383.
- Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., Kyriazis, D., 2023. Transforming sentiment analysis in the financial domain with ChatGPT. *Mach. Learn. Appl.* 14, 100508, (October).
- Frijns, B., Garel, A., 2021. The effect of cultural distance between an analyst and a CEO on analysts' earnings forecast performance. *Econom. Lett.* 205, 109957.
- Gao, K., Lin, W., Yang, L., Chan, K.C., 2020. The impact of analyst coverage and stock price synchronicity: Evidence from brokerage mergers and closures. *Financ. Res. Lett.* 33 (2019), 101190.
- Givoly, D., Hayn, C., Yoder, T.R., 2011. What do analysts really predict? Inferences from earnings restatements and managed earnings. *SSRN Electron. J.* (November).
- Graaf, J., 2023. Can sell-side analysts compete using public information? Analysts as frame-makers revisited. *Eur. Account. Rev.* 32 (1), 141–167.
- Grant, A., Jarnevic, E., Su, M., 2015. Asymmetric effects of sell-side analyst optimism and broker market share by clientele. *J. Financ. Mark.* 24, 49–65.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Pages. Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 8342–8360.
- Henrique, B.M., Sobreiro, V.A., Kimura, H., 2019. Literature review: Machine learning techniques applied to financial market prediction. *Expert Syst. Appl.* 124, 226–251.
- Hou, Y., Hu, C., 2023. Understanding the role of aggregate analyst attention in resolving stock market uncertainty. *Financ. Res. Lett.* 57, 104183, (May).
- Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., Johnson, J.E., 2016. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Syst. Appl.* 61, 215–234, (May).
- Karmaziene, E., 2023. The greater the volume, the greater the analyst. *Financ. Res. Lett.* 51 (2022), 103377.
- Keohane, D., 2018. French watchdog calls for rethink of research rules.
- Kim, S.S.S., Kim, S.S.S., Kim, Y., Park, J., Kim, S.S.S., Kim, M., Sung, C.H., Hong, J., Lee, Y., 2023a. LLMs analyzing the analysts: Do BERT and GPT extract more value from financial analyst reports? In: 4th ACM International Conference on AI in Finance. New York, NY, USA. ACM, pp. 383–391.
- Kim, S., Lee, W.-J., Park, S., Sunwoo, H.-Y., 2022. Busy analysts in uncertain times. *Financ. Res. Lett.* 46 (PB), 102488.
- Kim, A.G., Muhn, M., Nikolaev, V.V., Kim, A.G., 2023b. Bloated disclosures: Can ChatGPT help investors process information? *SSRN Electron. J.* 23.
- Kirtac, K., Germano, G., 2024. Sentiment trading with large language models. *Financ. Res. Lett.* 62 (PB), 105227.
- Ko, H., Lee, J., 2023. Can ChatGPT improve investment decision? From a portfolio manager perspective.
- Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Moral, V., Scao, T.L., Werra, L.V., Mou, C., Ponferrada, E.G., Nguyen, H., Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova, A., 2022. The BigScience ROOTS corpus: A 1. 6TB composite multilingual dataset. In: Advances in Neural Information Processing Systems, Number. vol. 35, pp. 31809–31826.
- Le, T.D., Trinh, T., 2022. Distracted analysts and earnings management. *Financ. Res. Lett.* 49, 103038, (June).
- Li, F., 2006. Do stock market investors understand the risk sentiment of corporate annual reports? *SSRN Electron. J.*
- Li, T., 2022. Analyst's stock views and revision actions. *Financ. Res. Lett.* 44 (2021), 102033.
- Li, E.X., Tu, Z., Zhou, D., 2023. The promise and peril of generative AI: Evidence from ChatGPT as sell-side analysts. *SSRN Electron. J.* 1–31, (June).
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., Ge, B., 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiol.* 1 (2), 100017.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, (1).
- Lo, H.-C., 2017. Do firm size influence financial analyst research reports and subsequent stock performance. *Account. Financ. Res.* 6 (4), 181.
- Lopez-lira, A., Tang, Y., 2023. Can ChatGPT forecast stock price movements ? Return predictability and large language models. *SSRN Electron. Journa*.
- Loughran, T., McDonald, B., Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* 66 (1), 35–65.
- Loughran, T., McDonald, B., 2020. Textual analysis in finance. *Annu. Rev. Financ. Econ.* 12 (1), 357–375.
- Markowitz, H., 1991. Foundations of portfolio theory. *J. Financ.* 46 (2), 469–477.
- McMillan, D.G., 2016. Which variables predict and forecast stock market returns? *SSRN Electron. J.* 44, 0–28, (June).
- Morgan, J., Stocken, P.C., 2003. An analysis of stock recommendations. *Rand J. Econ.* 34 (1), 183.
- Nyakurukwa, K., Seetharam, Y., 2023. Can textual sentiment partially explain differences in the prices of dual-listed stocks? *Financ. Res. Lett.* 58, 104529, (PC).
- Olof, L., 2019. Sentiment Analysis of Equity Analyst Research Reports using Convolutional Neural Networks. (Ph.D. thesis). Uppsala University.
- Pasch, S., Ehnes, D., 2022. StonkBERT: Can language models predict medium-run stock price movements?. pp. 1–16, arXiv.
- Pelster, M., Val, J., 2024. Can ChatGPT assist in picking stocks?. *Financ. Res. Lett.* 59 (2023), 104786.
- Pursiainen, V., 2018. Cultural biases in equity analysis. *SSRN Electron. J.*
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
- Rae, J.W., Borgeaud, S., Cai, T., Milligan, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, Hendricks, L.A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X.L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sotiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., D'Autume, C.d.M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D.d.L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., Irving, G., 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv.
- Reid, B.H., Jessop, S., Mason, J., 2019. Pressure on small brokers grows a year after new EU rules.
- Risk Control Limited, 2020. The Impact of MiFID II on SME and Fixed Income Investment Research Final Report. Technical Report April, European Commission.
- Rybinski, K., 2020a. Should asset managers pay for economic research? A machine learning evaluation. *J. Financ. Data Sci.* 6, 31–48.
- Rybinski, K., 2020b. The forecasting power of the multi-language narrative of sell-side research: A machine learning evaluation. *Financ. Res. Lett.* 34 (2019), 101261.
- Schlaubitz, A., 2021. Natural Language Processing in Finance: Analysis of Sentiment and Complexity of News and Earnings Reports of Swiss SMEs and their Relevance for Stock Returns. (Ph.D. thesis). Zurich University of Applied Sciences.
- Sidogi, T., Mongwe, W.T., Mbuvha, R., Marwala, T., 2022. Fusing sell-side analyst bidirectional forecasts using machine learning. *IEEE Access* 10, 76966–76974, (July).
- Sul, H.K., Dennis, A.R., Yuan, L.I., 2017. Trading on Twitter: Using social media sentiment to predict stock returns. *Decis. Sci.* 48 (3), 454–488.
- Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How To Fine-Tune BERT for Text Classification? in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 11856 LNAI. pp. 194–206.
- Suzuki, M., Sakaji, H., Izumi, K., Ishikawa, Y., 2022. Forecasting stock price trends by analyzing economic reports with analyst profiles. *Front. Artif. Intell.* 5, (June).
- Thas Thaker, H.M., Mohamad, A., Mustaffa Kamil, N.K., Duasa, J., 2018. How useful are the equity analysts' report? Evidence from Malaysia. *Rep. Econ. Financ.* 4 (4), 221–246.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *IEEE Ind. Appl. Mag.* 8 (1), 8–15.
- Wisniewski, T.P., Yekini, L.S., 2015. Stock market returns and the content of annual report narratives. *Account. Forum* 39 (4), 281–294.
- Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023. BloombergGPT: A large language model for finance.
- Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., Huang, J., 2023. PIXIU: A large language model, instruction data and evaluation benchmark for finance.
- Yang, G., He, P., Liu, X., 2023a. FundRecILM: Fund recommendation based on financial news and research analyst report. *Frontiers Artificial Intelligence Appl.* 378, 515–523.
- Yang, H., Liu, X.-Y., Dan Wang, C., 2023b. Fingpt: Open-source financial large language models. *SSRN Electron. J.*
- Yang, Y., UY, M.C.S., Huang, A., 2020. Finbert: A pretrained language model for financial communications.
- Yue, T., Au, D., Au, C.C., Iu, K.Y., 2023. Democratizing financial knowledge with ChatGPT by OpenAI: Unleashing the power of technology. *SSRN Electron. J.* 1–26.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L., 2022. OPT: Open pre-trained transformer language models. arXiv.

DeepLOB: Deep Convolutional Neural Networks for Limit Order Books

Zihao Zhang, Stefan Zohren, and Stephen Roberts

Abstract—We develop a large-scale deep learning model to predict price movements from limit order book (LOB) data of cash equities. The architecture utilises convolutional filters to capture the spatial structure of the limit order books as well as LSTM modules to capture longer time dependencies. The proposed network outperforms all existing state-of-the-art algorithms on the benchmark LOB dataset [1]. In a more realistic setting, we test our model by using one year market quotes from the London Stock Exchange and the model delivers a remarkably stable out-of-sample prediction accuracy for a variety of instruments. Importantly, our model translates well to instruments which were not part of the training set, indicating the model’s ability to extract universal features. In order to better understand these features and to go beyond a “black box” model, we perform a sensitivity analysis to understand the rationale behind the model predictions and reveal the components of LOBs that are most relevant. The ability to extract robust features which translate well to other instruments is an important property of our model which has many other applications.

I. INTRODUCTION

IN today’s competitive financial world more than half of the markets use electronic Limit Order Books (LOBs) [2] to record trades [3]. Unlike traditional quote-driven marketplaces, where traders can only buy or sell an asset at one of the prices made publicly by market makers, traders now can directly view all resting limit orders¹ in the limit order book of an exchange. Because limit orders are arranged into different levels based on their submitted prices, the evolution in time of a LOB represents a multi-dimensional problem with elements representing the numerous prices and order volumes/sizes at multiple levels of the LOB on both the buy and sell sides.

A LOB is a complex dynamic environment with high dimensionality, inducing modelling complications that make traditional methods difficult to cope with. Mathematical finance is often dominated by models of evolving price sequences. This leads to a range of Markov-like models with stochastic driving terms, such as the vector autoregressive model (VAR) [4] or the autoregressive integrated moving average model (ARIMA) [5]. These models, to avoid excessive parameter spaces, often rely on handcrafted features of the data. However, given the billions of electronic market quotes that are generated

The authors are with the Oxford-Man Institute of Quantitative Finance, Department of Engineering Science, University of Oxford (e-mail: zihao@robots.ox.ac.uk).

Github: <https://github.com/zcakhaa>

¹Limit orders are orders that do not match immediately upon submission and are also called passive orders. This is opposed to orders that match immediately, so-called aggressive orders, such as a market order. A LOB is simply a record of all resting/outstanding limit orders at a given point in time.

everyday, it is natural to employ more modern data-driven machine learning techniques to extract such features.

In addition, limit order data, like any other financial time-series data is notoriously non-stationary and dominated by stochastics. In particular, orders at deeper levels of the LOB are often placed and cancelled in anticipation of future price moves and are thus even more prone to noise. Other problems, such as auction and dark pools [6], also add additional difficulties, bringing ever more unobservability into the environment. The interested reader is referred to [7] in which a number of these issues are reviewed.

In this paper we design a novel deep neural network architecture that incorporates both convolutional layers as well as Long Short-Term Memory (LSTM) units to predict future stock price movements in large-scale high-frequency LOB data. One advantage of our model over previous research [8] is that it has the ability to adapt for many stocks by extracting representative features from highly noisy data.

In order to avoid the limitations of handcrafted features, we use a so-called Inception Module [9] to wrap convolutional and pooling layers together. The Inception Module helps to infer local interactions over different time horizons. The resulting feature maps are then passed into LSTM units which can capture dynamic temporal behaviour. We test our model on a publicly available LOB dataset, known as FI-2010 [1], and our method remarkably outperforms all existing state-of-the-art algorithms. However, the FI-2010 dataset is only made up of 10 consecutive days of down-sampled pre-normalised data from a less liquid market. While it is a valuable benchmark set, it is arguable not sufficient to fully verify the robustness of an algorithm. To ensure the generalisation ability of our model, we further test it by using one year order book data for 5 stocks from the London Stock Exchange (LSE). To minimise the problem of overfitting to backtest data, we carefully optimise any hyper-parameter on a separate validation set before moving to the out-of-sample test set. Our model delivers robust out-of-sample prediction accuracy across stocks over a test period of three months.

As well as presenting results on out-of-sample data (in a timing sense) from stocks used to form the training set, we also test our model on out-of-sample (in both timing and data stream sense) stocks that are not part of the training set. Interestingly, we still obtain good results over the whole testing period. We believe this observation shows not only that the proposed model is able to extract robust features from order books, but also indicates the existence of universal features in the order book that modulate stock demand and price. The ability to transfer the model to new instruments opens up a

number of possibilities that we consider for future work.

To show the practicability of our model we use it in a simple trading simulation. We focus on sufficiently liquid stocks so that slippage and market impact are small. Indeed, these stocks are generally harder to predict than less liquid ones. Since our trading simulation is mainly meant as a method of comparison between models we assume trading takes place at mid-price² and compare gross profits before fees. The former assumption is equivalent to assuming that one side of the trade may be entered into passively and the latter assumes that different models trade similar volumes and would thus be subject to similar fees. Our focus here is using a simulation as a measure of the relative value of the model predictions in a trading setting. Under these simplifications, our model delivers significantly positive returns with a relatively small risk.

Although our network achieves good performance, a complex “black box” system, such as a deep neural network, has limited use for financial applications without some understanding of the rationale behind the model predictions. Here we exploit the model-agnostic LIME method [10] to highlight highly relevant components in the order book to gain a better understanding between our predictions and model inputs. Reassuringly, these conform to sensible (though arguably unusual) patterns of activity in both price and volume within the order book.

Outline: The remainder of the paper is as follows. Section II introduces background and related work. Section III describes limit order data and the various stages of data preparation. We present our network architecture in Section IV and give justifications behind each component of the model. In Section V we compare our work with a large group of popular methods. Section VI summarises our findings and considers extensions and future work.

II. BACKGROUND AND RELATED WORK

Research on the predictability of stock markets has a long history in the financial literature e.g., [11] [12]. Although opinions differ regarding the efficiency of markets, many widely accepted studies show that financial markets are to some extent predictable [13] [14] [15] [16]. Two major classes of work which attempt to forecast financial time-series are, broadly speaking, statistical parametric models and data-driven machine learning approaches [17]. Traditional statistical methods generally assume that the time-series under study are generated from a parametric process [18]. There is, however, agreement that stock returns behave in more complex ways, typically highly nonlinearly [19, 20]. Machine learning techniques are able to capture such arbitrary nonlinear relationships with little, or no, prior knowledge regarding the input data [21].

Recently, there has been a surge of interest to predict limit order book data by using machine learning algorithms [1, 22, 23, 24, 25, 26, 27, 20, 28, 29]. Among many machine learning techniques, pre-processing or feature extraction is often performed as financial time-series data is highly stochastic. Generic feature extraction approaches have been implemented, such as the Principal Component Analysis (PCA) and the

Linear Discriminant Analysis (LDA) in the work of [24]. However these extraction methods are static pre-processing steps, which are not optimised to maximise the overall objective of the model that observes them. In the work of [25, 24], the Bag-of-Features model (BoF) is expressed as a neural layer and the model is trained end-to-end using the back-propagation algorithm, leading to notably better results on the FI-2010 dataset [1]. These works suggest the importance of a data driven approach to extract representative features from a large amount of data. In our work, we advocate the end-to-end training and show that the deep neural network by itself not only leads to even better results but also transfers well to new instruments (not part of the training set) - indicating the ability of networks to extract “universal” features from the raw data.

Arguably, one of the key contributions of modern deep learning is the addition of feature extraction and representation as part of the learned model. The Convolutional Neural Network (CNN) [30] is a prime example, in which information extraction, in the form of filter banks, is automatically tuned to the utility function that the entire network aims to optimise. CNNs have been successfully applied to various application domains, for example, object tracking [31], object-detection [32] and segmentation [33]. However, there have been but a few published works that adopt CNNs to analyse financial microstructure data [34, 35, 26] and the existing CNN architectures are rather unsophisticated and lack of thorough investigation. Just like when moving from “AlexNet” [36] to “VGGNet” [37], we show that a careful design of network architecture can lead to better results compared with all existing methods.

The Long Short-Term Memory (LSTM) [38] was originally proposed to solve the vanishing gradients problem [39] of recurrent neural networks, and has been largely used in applications such as language modelling [40] and sequence to sequence learning [41]. Unlike CNNs which are less widely applied in financial markets, the LSTM has been popular in recent years, [42, 28, 43, 44, 45, 46, 47, 20] all utilising LSTMs to analyse financial data. In particular, [20] uses limit order data from 1000 stocks to test a four layer LSTM model. Their results show a stable out-of-sample prediction accuracy across time, indicating the potential benefits of deep learning methods. To the best of our knowledge, there is no work that combines CNNs with LSTMs to predict stock price movements and this is the first extensive study to apply a nested CNN-LSTM model to raw market data. In particular, the usage of the Inception Model in this context is novel and is essential in inferring the optimal “decay rates” of the extracted features.

III. DATA, NORMALISATION AND LABELLING

A. Limit Order Books

We first introduce some basic definitions of limit order books (LOBs). For classical references on market microstructure the reader is referred to [48, 49] and for a short review on LOBs in particular we refer to [7]. Here we follow the conventions of [7]. A LOB has two types of orders: bid orders and ask orders. A bid (ask) order is an order to buy (sell) an

²The average of the best buy and best sell prices in the market at the time.

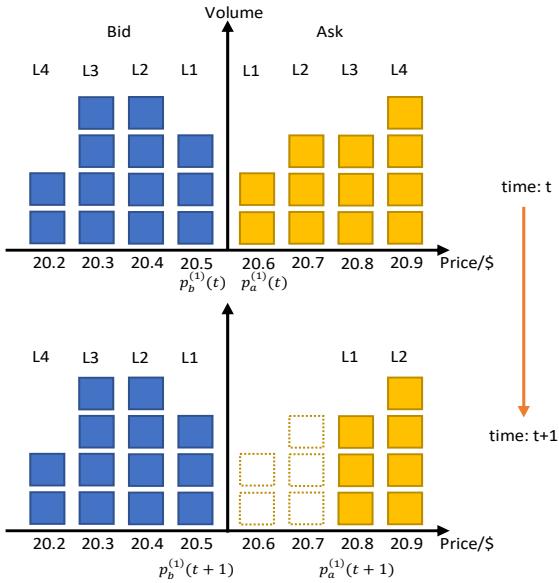


Figure 1. A slice of LOB at time t and $t + 1$. L1 represents the respective first level, L2 the second, etc. $p_a^{(1)}(t)$ is the lowest ask price (best ask) and $p_b^{(1)}(t)$ is the highest bid price (best bid) at time t .

asset at or below (above) a specified price. The bid orders have prices $\mathbf{P}_b(t)$ and sizes/volumes $\mathbf{V}_b(t)$, and the ask orders have prices $\mathbf{P}_a(t)$ and sizes/volumes $\mathbf{V}_a(t)$. Both $\mathbf{P}(t)$ and $\mathbf{V}(t)$ are vectors representing values at different price levels of an asset.

Figure 1 illustrates the above concepts. The upper plot shows a slice of a LOB at time t . Each square in the plot represents an order of nominal size 1. This is done for simplicity, in reality different orders can be of different sizes. The blue bars represent bid orders and the yellow bars represent ask orders. Orders are sorted into different levels based on their submitted prices, where L1 represents the first level and so on. Each level contains two values: price and volume. On the bid side, $\mathbf{P}_b(t)$ and $\mathbf{V}_b(t)$ are 4-vectors in this example. We use $p_b^{(1)}(t)$ to denote the highest available price for a buying order (first bid level). Similarly, $p_a^{(1)}(t)$ is the lowest available selling order (first ask level). The bottom plot shows the action of an incoming market order to buy 5 shares at time $t + 1$. As a result, the entire first and second ask-levels are executed against that order and $p_a^{(1)}(t + 1)$ moved to 20.8 from 20.6 at time t .

B. Input Data

We test our model on two datasets: the FI-2010 dataset [1] and one year length of limit order book data from the London Stock Exchange (LSE). The FI-2010 dataset [1] is the first publicly available benchmark dataset of high-frequency limit order data and extracted time series data for five stocks from the Nasdaq Nordic stock market for a time period of 10 consecutive days. Many earlier algorithms are tested on this dataset and we use it to establish a fair comparison to other algorithms. However, 10 days is an insufficient amount of data to fully test the robustness and generalisation ability of an algorithm as the problem of overfitting to backtest data

is severe and we often expect a signal to be consistent over a few months.

To address the above concerns, we train and test our model on limit order book data of one year length for Lloyds Bank, Barclays, Tesco, BT and Vodafone. These five instruments are among the most liquid stocks listed on the London Stock Exchange. It is generally more difficult to train models on more liquid stocks, but at the same time, those instruments are easier to trade without price impact so making the simple trading simulation used to assess performance more realistic. The data includes all LOB updates for the above names. It spans all trading days from 3rd January 2017 to 24th December 2017 and we restrict it to the interval between 08:30:00 and 16:00:00, so that only normal trading activities occur and no auction takes place. Each state of the LOB contains 10 levels on each side and each level contains information on both price and volume. Therefore, we have a total of 40 features at each timestamp. Note that the FI-2010 dataset is actually downsampled limit order book data because the authors followed [50] to create additional features by using every non-overlapping block of 10 events. We did not perform any processing on our data and only feed raw order book information to our algorithm.

Overall, our LSE dataset is made up of 12 months, and has more than 134 million samples. On average, there are 150,000 events per day per stock. The events are irregularly spaced in time. The time interval, $\Delta_{k,k+1}$, between two events can vary considerably from a fraction of a second to seconds, and $\Delta_{k,k+1}$ is on average 0.192 seconds in the dataset. We take the first 6 months as training data, the next 3 months as validation data and the last 3 months as test data. In the context of high-frequency data, 3 months test data corresponds to millions of observations and therefore provides sufficient scope for testing model performance and estimating model accuracy.

C. Data Normalisation and Labelling

The FI-2010 dataset [1] provides 3 different normalised datasets: z -score, min-max and decimal precision normalisation. We used data normalised by z -score without any emendation and found subtle difference when using the other two normalisation schemes. For the LSE dataset, we again use standardisation (z -score) to normalise our data, but use the mean and standard deviation of the previous 5 days' data to normalise the current day's data (with a separate normalisation for each instrument). We want to emphasize the importance of normalisation because the performance of machine learning algorithms often depends on it. As financial time-series usually experiences regime shifts, using a static normalisation scheme is not appropriate for a dataset of one year length. The above method is dynamic and the normalised data often falls into a reasonable range. We use the 100 most recent states of the LOB as an input to our model for both datasets. Specifically, a single input is defined as $X = [x_1, x_2, \dots, x_t, \dots, x_{100}]^T \in \mathbb{R}^{100 \times 40}$, where $x_t = [p_a^{(i)}(t), v_a^{(i)}(t), p_b^{(i)}(t), v_b^{(i)}(t)]_{i=1}^{n=10}$. $p^{(i)}$ and $v^{(i)}$ denote the price and volume size at i -th level of a limit order book.

After normalising the limit order data, we use the mid-price

$$p_t = \frac{p_a^{(1)}(t) + p_b^{(1)}(t)}{2}, \quad (1)$$

to create labels that represent the direction of price changes. Although no order can transact exactly at the mid-price, it expresses a general market value for an asset and it is frequently quoted when we want a single number to represent an asset price.

Because financial data is highly stochastic, if we simply compare p_t and p_{t+k} to decide the price movement, the resulting label set will be noisy. In the works of [1] and [26], two smoothing labelling methods are introduced. We briefly recall the two methods here. First, let m_- denote the mean of the previous k mid-prices and m_+ denote the mean of the next k mid-prices:

$$\begin{aligned} m_-(t) &= \frac{1}{k} \sum_{i=0}^k p_{t-i} \\ m_+(t) &= \frac{1}{k} \sum_{i=1}^k p_{t+i} \end{aligned} \quad (2)$$

where p_t is the mid-price defined in Equation 1 and k is the prediction horizon. Both methods use the percentage change (l_t) of the mid-price to decide directions. We can now define

$$l_t = \frac{m_+(t) - p_t}{p_t} \quad (3)$$

$$l_t = \frac{m_+(t) - m_-(t)}{m_-(t)} \quad (4)$$

Both are methods to define the direction of price movement at time t , where the former, Equation 3 was used in [1] and the latter, Equation 4, in [26].

The labels are then decided based on a threshold (α) for the percentage change (l_t). If $l_t > \alpha$ or $l_t < -\alpha$, we define it as up (+1) or down (-1). For anything else, we consider it as stationary (0). Figure 2 provides a graphical illustration of two labelling methods on the same threshold (α) and the same prediction horizon (k). All the labels classified as down (-1) are shown as red areas and up (+1) as green areas. The uncoloured (white) regions correspond to stationary (0) labels.

The FI-2010 dataset [1] adopts the method in Equation 3 and we directly used their labels for fair comparison to other methods. However, the produced labels are less consistent as shown on the top of Figure 2 because this method fits closer to real prices as smoothing is only applied to future prices. This is essentially detrimental for designing trading algorithms as signals are not consistent here leading to many redundant trading actions thus incurring larger transaction costs.

Further, the FI-2010 dataset was collected in 2010 and the instruments were less liquid compared to now. We experimented with this approach in [1] on our data from the London Stock Exchange and found the resulting labels are rather stochastic, therefore we adopt the method in Equation 4 for our LSE dataset to produce more consistent signals.

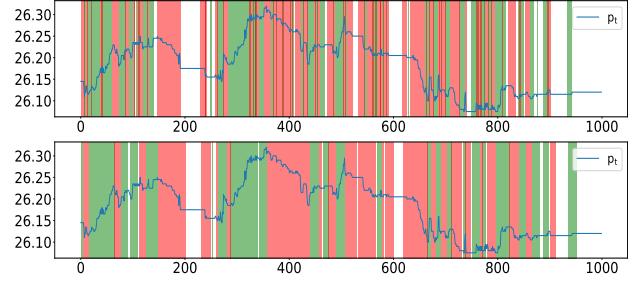


Figure 2. An example of two smoothed labelling methods based on a same threshold (α) and same prediction horizon (k). Green shading represents a +1 signal and red a -1. **Top:** [1]'s method and **Bottom:** [26]'s method.

IV. MODEL ARCHITECTURE

A. Overview

We here detail our network architecture, which comprises three main building blocks: standard convolutional layers, an Inception Module and a LSTM layer, as shown in Figure 3. The main idea of using CNNs and Inception Modules is to automate the process of feature extraction as it is often difficult in financial applications since financial data is notoriously noisy with a low signal-to-noise ratio. Technical indicators such as MACD and the Relative Strength Index are included as inputs and preprocessing mechanisms such as principal component analysis (PCA) [51] are often used to transform raw inputs. However, none of these processes is trivial, they make tacit assumptions and further, it is questionable if financial data can be well-described with parametric models with fixed parameters. In our work, we only require the history of LOB prices and sizes as inputs to our algorithm. Weights are learned during inference and features, learned from a large training set, are data-adaptive, removing the above constraints. A LSTM layer is then used to capture additional time dependencies among the resulting features. We note that very short time-dependencies are already captured in the convolutional layer which takes “space-time images” of the LOB as inputs.

B. Details of Each Component

a) Convolutional Layer: Recent development of electronic trading algorithms often submit and cancel vast numbers of limit orders over short periods of time as part of their trading strategies [52]. These actions often take place deep in a LOB and it is seen [7] that more than 90% of orders end in cancellation rather than matching, therefore practitioners consider levels further away from best bid and ask levels to be less useful in any LOB. In addition, the work of [53] suggests that the best ask and best bid (L1-Ask and L1-Bid) contribute most to the price discovery and the contribution of all other levels is considerably less, estimated at as little as 20%. As a result, it would be otiose to feed all level information to a neural network as levels deep in a LOB are less useful and can potentially even be misleading. Naturally, we can smooth these signals by summarising the information contained in deeper levels. We note that convolution filters used in any CNN architecture are discrete convolutions, or finite impulse response (FIR) filters, from the viewpoint of

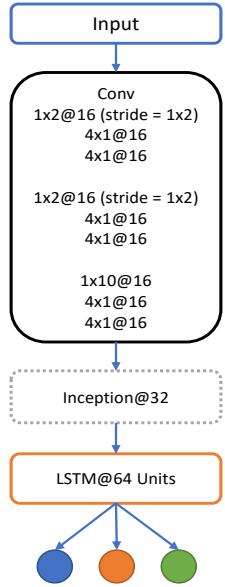


Figure 3. Model architecture schematic. Here $1 \times 2 @ 16$ represents a convolutional layer with 16 filters of size (1×2) . ‘1’ convolves through time indices and ‘2’ convolves different limit order book levels.

signal processing [54]. FIR filters are popular smoothing techniques for denoising target signals and they are simple to implement and work with. We can write any FIR filter in the following form:

$$y(n) = \sum_{k=0}^M b_k x(n-k) \quad (5)$$

where the output signal $y(n)$ at any time is a weighted sum of a finite number of past values of the input signal $x(n)$. The filter order is denoted as M and b_k is the filter coefficient. In a convolutional neural network, the coefficients of the filter kernel are not obtained via a statistical objective from traditional signal filtration theory, but are left as degrees of freedom which the network infers so as to extremise its value function at output.

The details of the first convolutional layer inevitably need some consideration. As convolutional layers operate a small kernel to “scan” through input data, the layout of limit order book information is vital. Recall that we take the most 100 recent updates of an order book to form a single input and there are 40 features per time stamp, so the size of a single input is (100×40) . We organise the 40 features as following:

$$\{p_a^{(i)}(t), v_a^{(i)}(t), p_b^{(i)}(t), v_b^{(i)}(t)\}_{i=1}^{n=10} \quad (6)$$

where i denotes the i -th level of a limit order book. The size of our first convolutional filter is (1×2) with stride of (1×2) . The first layer essentially summarises information between price and volume $\{p^{(i)}, v^{(i)}\}$ at each order book level. The usage of stride is necessary here as an important property of convolutional layers is parameter sharing. This property is attractive as less parameters are estimated, largely avoiding overfitting problems. However, without strides, we would apply same parameters to $\{p^{(i)}, v^{(i)}\}$ and $\{v^{(i)}, p^{(i+1)}\}$. In other words, $p^{(i)}$ and $v^{(i)}$ would share same parameters

because the kernel filter moves by one step, which is obviously wrong as price and volume form different dynamic behaviors.

Because the first layer only captures information at each order book level, we would expect representative features to be extracted when integrating information across multiple order book levels. We can do this by utilising another convolutional layer with filter size (1×2) and stride (1×2) . The resulting feature maps actually form the micro-price defined by [55]:

$$p^{\text{micro price}} = I p_a^{(1)} + (1 - I) p_b^{(1)} \quad (7)$$

$$I = \frac{v_b^{(1)}}{v_a^{(1)} + v_b^{(1)}}$$

The weight I is called the imbalance. The micro-price is an important indicator as it considers volumes on bid and ask side, and the imbalance between bid and ask size is a very strong indicator of the next price move. This feature of imbalances has been reported by a variety of researchers [56, 57, 58, 59, 60]. Unlike the micro-price where only the first order book level is considered, we utilise convolutions to form micro-prices for all levels of a LOB so the resulting feature maps are of size $(100, 10)$ after two layers with strides. Finally, we integrate all information by using a large filter of size (1×10) and the dimension of our feature maps before the Inception Module is $(100, 1)$.

We apply zero padding to every convolutional layer so the time dimension of our inputs does not change and Leaky Rectifying Linear Units (Leaky-ReLU) [61] are used as activation functions. The hyper-parameter (the small gradient when the unit is not active) of the Leaky-ReLU is set to 0.01, evaluated by grid search on the validation set.

Another important property of convolution is that of equivariance to translation [62]. Specifically, a function $f(x)$ is equivariant to a function g if $f(g(x)) = g(f(x))$. For example, suppose that there exists a main classification feature m located at (x_m, y_m) of an image $I(x, y)$. If we shift every pixel of I one unit to the right, we get a new image I' where $I'(x, y) = I(x-1, y)$. We can still obtain the main classification feature m' in I' and $m = m'$, while the location of m' will be at $(x_{m'}, y_{m'}) = (x_m - 1, y_m)$. This is important to time-series data, because convolution can find universal features that are decisive to final outputs. In our case, suppose a feature that studies imbalance is obtained at time t . If the same event happens later at time t' in the input, the exact feature can be extracted later at t' .

We do not use any pooling layer except in the Inception Modules. Although pooling layers help us find representations invariant to translations of the input, the smoothing nature of pooling can cause under-fitting. Common pooling layers are designed for image processing tasks, and they are most powerful when we only care if certain features exist in the inputs instead of where they exist [62]. Time-series data has different characteristics from images and the location of representative features is important. Our experiences show that pooling layers in the convolutional layer, at least, cause under-fitting problems to the LOB data. However, we think pooling is important and new pooling methods should be designed to process time-series data as it is a promising

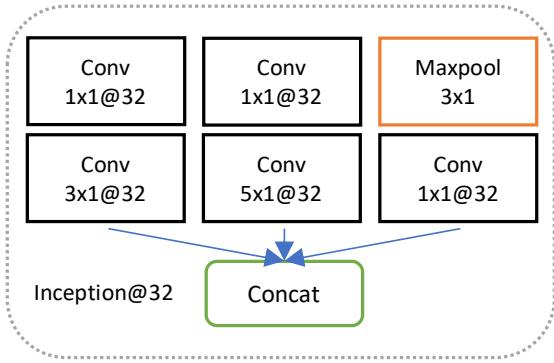


Figure 4. The Inception Module used in the model. For example, $3 \times 1@32$ represents a convolutional layer with 32 filters of size (3×1) .

solution to extract invariant features.

b) Inception Module: We note that all filters of a standard convolutional layer have fixed size. If, for example, we employ filters of size (4×1) , we capture local interactions amongst data over four time steps. However, we can capture dynamic behaviours over multiple timescales by using Inception Modules to wrap several convolutions together. We find that this offers a performance improvement to the resultant model.

The idea of the Inception Module can be also considered as using different moving averages in technical analysis. Practitioners often use moving averages with different decay weights to observe time-series momentum [63]. If a large decay weight is adopted, we get a smoother time-series that well represents the long-term trend, but we could miss small variations that are important in high-frequency data. In practice, it is a daunting task to set the right decay weights. Instead, we can use Inception Modules and the weights are then learned during back-propagation.

In our case, we split the input into a small set of lower-dimensional representations by using 1×1 convolutions, transform the representations by a set of filters, here 3×1 and 5×1 , and then merge the outputs. A max-pooling layer is used inside the Inception Module, with stride 1 and zero padding. “Inception@32” represents one module and indicates all convolutional layers have 32 filters in this module, and the approach is depicted schematically in Figure 4. The 1×1 convolutions form the Network-in-Network approach proposed in [64]. Instead of applying a simple convolution to our data, the Network-in-Network method uses a small neural network to capture non-linear properties of our data. We find this method to be effective and it gives us an improvement on prediction accuracy.

c) LSTM Module and Output: In general, a fully connected layer is used to classify the input data. However, all inputs to the fully connected layer are assumed independent of each other unless multiple fully connected layers are used. Due to the usage of Inception Module in our work, we have a large number of features at end. Just using one fully connected layer with 64 units would result in more than 630,000 parameters

to be estimated, not to mention multiple layers. In order to capture temporal relationship that exist in the extracted features, we replace the fully connected layers with LSTM units. The activation of a LSTM unit is fed back to itself and the memory of past activations is kept with a separate set of weights, so the temporal dynamics of our features can be modelled. We use 64 LSTM units in our work, resulting in about 60,000 parameters, leading to 10 times fewer parameters to be estimated. The last output layer uses a softmax activation function and hence the final output elements represent the probability of each price movement class at each time step.

V. EXPERIMENTAL RESULTS

A. Experiments Settings

We apply the same architecture to all our experiments in this section and the proposed model is denoted as DeepLOB. We learn the parameters by minimising the categorical cross-entropy loss. The Adaptive Moment Estimation algorithm, ADAM [65], is utilised and we set the parameter “epsilon” to 1 and the learning rate to 0.01. The learning is stopped when validation accuracy does not improve for 20 more epochs. This is about 100 epochs for the FI-2010 dataset and 40 epochs for the LSE dataset.

We train with mini-batches of size 32. We choose a small mini-batch size due to the findings in [66] in which they suggest that large-batch methods tend to converge to narrow deep minima of the training functions, but small-batch methods consistently converge to shallow broad minima. All models are built using Keras [67] based on the TensorFlow backend [68], and we train them using a single NVIDIA Tesla P100 GPU.

B. Experiments on the FI-2010 Dataset

There are two experimental setups using the FI-2010 dataset. Following the convention of [24], we denote them as Setup 1 and Setup 2. Setup 1 splits the dataset into 9 folds based on a day basis (a standard anchored forward split). In the i -th fold, we train our model on the first i days and test it on the $(i+1)$ -th day where $i = 1, \dots, 9$. The second setting, Setup 2, originates from the works [26, 28, 27, 25] in which deep network architectures were evaluated. As deep learning techniques often require a large amount of data to calibrate weights, the first 7 days are used as the train data and the last 3 days are used as the test data in this setup. We evaluate our model in both setups here.

Table I shows the results of our model compared to other methods in Setup 1. Performance is measured by calculating the mean accuracy, recall, precision, and F1 score over all folds. As the FI-2010 dataset is not well balanced, I suggests to focus on F1 score performance as fair comparisons. We have compared our model to all existing experimental results including Ridge Regression (RR) [1], Single-Layer-Feedforward Network (SLFN) [1], Linear Discriminant Analysis (LDA) [2], Multilinear Discriminant Analysis (MDA) [22], Multilinear Time-series Regression (MTR) [22], Weighted Multilinear Time-series Regression (WMTR) [22], Multilinear

Table I
SETUP 1: EXPERIMENT RESULTS FOR THE FI-2010 DATASET

Model	Accuracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
RR [1]	48.00	41.80	43.50	41.00
SLFN [1]	64.30	51.20	36.60	32.70
LDA [22]	63.83	37.93	45.80	36.28
MDA [22]	71.92	44.21	60.07	46.06
MCSDA [23]	83.66	46.11	48.00	46.72
MTR [22]	86.08	51.68	40.81	40.14
WMTR [22]	81.89	46.25	51.29	47.87
BoF [24]	57.59	39.26	51.44	36.28
N-BoF [24]	62.70	42.28	61.41	41.63
B(TABL) [25]	73.62	66.16	68.81	67.12
C(TABL) [25]	78.01	72.03	74.04	72.84
DeepLOB	78.91	78.47	78.91	77.66
Prediction Horizon k = 50				
RR [1]	43.90	43.60	43.30	42.70
SLFN [1]	47.30	46.80	46.40	45.90
BoF [24]	50.21	42.56	49.57	39.56
N-BoF [24]	56.52	47.20	58.17	46.15
B(TABL) [25]	69.54	69.12	68.84	68.84
C(TABL) [25]	74.81	74.58	74.27	74.32
DeepLOB	75.01	75.10	75.01	74.96
Prediction Horizon k = 100				
RR [1]	42.90	42.90	42.90	41.60
SLFN [1]	47.70	45.30	43.20	41.00
BoF [24]	50.97	42.48	47.84	40.84
N-BoF [24]	56.43	47.27	54.99	46.86
B(TABL) [25]	69.31	68.95	69.41	68.86
C(TABL) [25]	74.07	73.51	73.80	73.52
DeepLOB	76.66	76.77	76.66	76.58

Class-specific Discriminant Analysis (MCSDA) [23], Bag-of-Feature (BoF) [24], Neural Bag-of-Feature (N-BoF) [24], and Attention-augmented-Bilinear-Network with one hidden layer (B(TABL)) and two hidden layers (C(TABL)) [25]. More methods such as PCA and Autoencoder (AE) are actually tested in their works but, for simplicity, we only report their best results and our model achieves better performance.

However, the Setup 1 is not ideal for training deep learning models as we mentioned that deep network often requires a large amount of data to calibrate weights. This anchored forward setup leads to only one or two days' training data for the first few folds and we observe worse performance in the first few days. As training data grows, we observe remarkably better results as shown in Table II which shows the results of our network compared to other methods in Setup 2. In particular, the important difference between our model and CNN-I [26] and CNN-II [27] is due to network architecture and we can see huge improvements on performance here. In Table III, we compare the parameter sizes of DeepLOB with CNN-I [26]. Although our model has many more layers, there are far fewer parameters in our network due to the usage of LSTM layers instead of fully connected layers.

We also report the computation time (forward pass) in milliseconds (ms) for available algorithms in Table III. Due to the development of GPUs, training deep networks is now feasible and it is swift to make predictions, making it possible for high frequency trading. We will discuss this more in the next section.

Table II
SETUP 2: EXPERIMENT RESULTS FOR THE FI-2010 DATASET

Model	Accuracy %	Precision %	Recall %	F1 %
Prediction Horizon k = 10				
SVM [28]	-	39.62	44.92	35.88
MLP [28]	-	47.81	60.78	48.27
CNN-I [26]	-	50.98	65.54	55.21
LSTM [28]	-	60.77	75.92	66.33
CNN-II [27]	-	56.00	45.00	44.00
B(TABL) [25]	78.91	68.04	71.21	69.20
C(TABL) [25]	84.70	76.95	78.44	77.63
DeepLOB	84.47	84.00	84.47	83.40
Prediction Horizon k = 20				
SVM [28]	-	45.08	47.77	43.20
MLP [28]	-	51.33	65.20	51.12
CNN-I [26]	-	54.79	67.38	59.17
LSTM [28]	-	59.60	70.52	62.37
CNN-II [27]	-	-	-	-
B(TABL) [25]	70.80	63.14	62.25	62.22
C(TABL) [25]	73.74	67.18	66.94	66.93
DeepLOB	74.85	74.06	74.85	72.82
Prediction Horizon k = 50				
SVM [28]	-	46.05	60.30	49.42
MLP [28]	-	55.21	67.14	55.95
CNN-I [26]	-	55.58	67.12	59.44
LSTM [28]	-	60.03	68.58	61.43
CNN-II [27]	-	56.00	47.00	47.00
B(TABL) [25]	75.58	74.58	73.09	73.64
C(TABL) [25]	79.87	79.05	77.04	78.44
DeepLOB	80.51	80.38	80.51	80.35

Table III
AVERAGE COMPUTATION TIME OF STATE-OF-THE-ART MODELS

Models	Forward (ms)	Number of parameters
BoF [24]	0.972	86k
N-BoF [24]	0.524	12k
CNN-I [26]	0.025	768k
LSTM [28]	0.061	-
C(TABL) [25]	0.229	-
DeepLOB	0.253	60k

C. Experiments on the London Stock Exchange (LSE)

As we suggested, the FI-2010 dataset is not sufficient to verify a prediction model - it is far too short, downsampled and taken from a less liquid market. To perform a meaningful evaluation that can hold up to modern applications, we further test our method on stocks from the LSE of one year length with a testing period of three months. As mentioned in Section III, we train our model on five stocks: Lloyds Bank (LLOY), Barclays (BARC), Tesco (TSCO), BT and Vodafone (VOD). Recent work of [20] suggests that deep learning techniques can extract universal features for limit order data. To test this universality, we directly apply our model to five more stocks that were not part of the training data set (**transfer learning**). We select HSBC, Glencore (GLEN), Centrica (CNA), BP and ITV for transfer learning because they are also among the most liquid stocks in the LSE. The testing period is the same three months as before, and the classes are roughly balanced.

Table IV presents the results of our model for all stocks on different prediction horizons. To better investigate the results,

Table IV
EXPERIMENT RESULTS FOR THE LSE DATASET

Prediction Horizon	Accuracy %	Precision %	Recall %	F1 %
Results on LLOY, BARC, TSCO, BT and VOD				
k=20	70.17	70.17	70.17	70.15
k=50	63.93	63.43	63.93	63.49
k=100	61.52	60.73	61.52	60.65
Results on Transfer Learning (GLEN, HSBC, CNA, BP, ITV)				
k=20	68.62	68.64	68.63	68.48
k=50	63.44	62.81	63.45	62.84
k=100	61.46	60.68	61.46	60.77

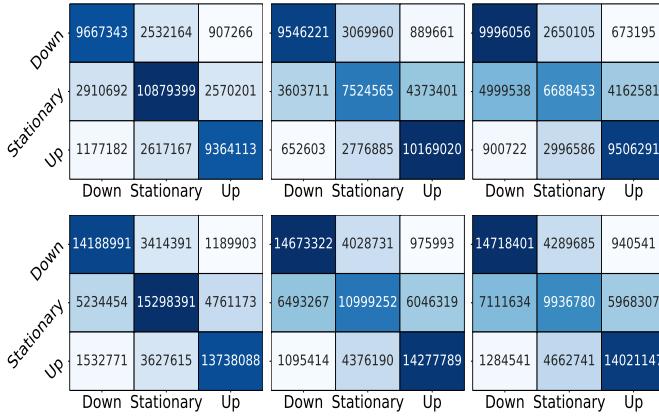


Figure 5. Confusion matrices. **Top:** results on LLOY, BARC, TSCO, BT and VOD. From the left to right, prediction horizon (k) equals 20, 50 and 100; **Bottom:** results on transfer learning (GLEN, HSBC, CNA, BP, ITV).

we display the confusion matrices in Figure 5 and calculate the accuracy for every day and for every stock across the testing period. We use the boxplots in Figure 6 to present this information and we can observe consistent and robust performance, with narrow interquartile range (IQR) and few outliers, for all stocks across the testing period. The ability of our model that generalises well to data not in the training set indicates that the CNN block in the algorithms, acting to extract features from the LOB, can capture universal patterns that relate to the price formation mechanism. We find this observation most interesting.

D. Performance of the Model in a Simple Trading Simulation

A simple trading simulation is designed to test the practicability of our results. We set the number of shares per trade, μ , to one both for simplicity and to minimise the market impact, ensuring orders to be executed at the best price. Although μ can be optimised to maximise the returns, for example, prediction probabilities are used to size the orders in [69], we would like to show that our algorithm can work even under this simple set-up.

To reduce the number of trades, we use following rules to take actions. At each time-step, our model generates a signal from the network outputs $(-1, 0, +1)$ to indicate the price movements in k steps. Signals $(-1, 0, +1)$ correspond to actions (sell, wait and buy). Suppose our model produces a prediction of $+1$ at time t , we then buy μ shares at time $t+5$

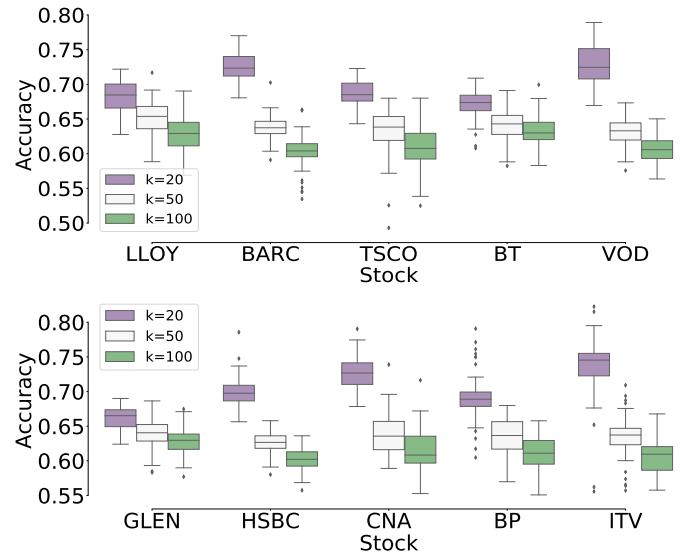


Figure 6. Boxplots of daily accuracy for the different prediction horizons. **Top:** results on LLOY, BARC, TSCO, BT and VOD; **Bottom:** results on transfer learning (GLEN, HSBC, CNA, BP, ITV).

(taking slippage into account), and hold until -1 appears to sell all μ shares (we do nothing if 0 appears). We apply the same rule to short selling and repeat the process during a day. All positions are closed by the end of the day, so we hold no stocks overnight. We make sure no trades take place at the time of auction, so no abnormal profits are generated.

As the focus of our work is on predictions and the above simple simulation is a way of showing that this prediction is in principle monetisable. In particular, our aim is not to present a fully developed, stand-alone trading strategy. Realistic high-frequency strategies often require a combination of various trading signals in particular to time the exact entry and exit points of the trade. For the purpose of the above simulation we use mid-prices without transaction costs. While in particular the second assumption is not a reasonable assumption for a standalone strategy, we argue that (i) it is enough for a relative comparison of the above models and (ii) it is a good indicator of the relative value of the above predictor to a more complex high-frequency trading model. Regarding the first assumption, a mid-mid simulation, we note that in high-frequency trading, many participants are involved in market making, as it is difficult to design profitable fully aggressive strategies with such short holding periods. If we assume that we are able to enter the trade passively, while we exit it aggressively, crossing the spread, then this is effectively equivalent to a mid-mid trade. Such a situation arises naturally for example in investment banks which are involved in client market making. Regarding the second assumption, careful timing of the entry points as well as more elaborate trading rules, such as including position upsizing, should be able to account for additional profits to cover the transaction costs. In any case, as merely a metric of testing predictability of our model, the above simple simulation suffices.

Figure 7 presents the boxplots for normalised daily profits (profits divided by number of trades in that day) for different

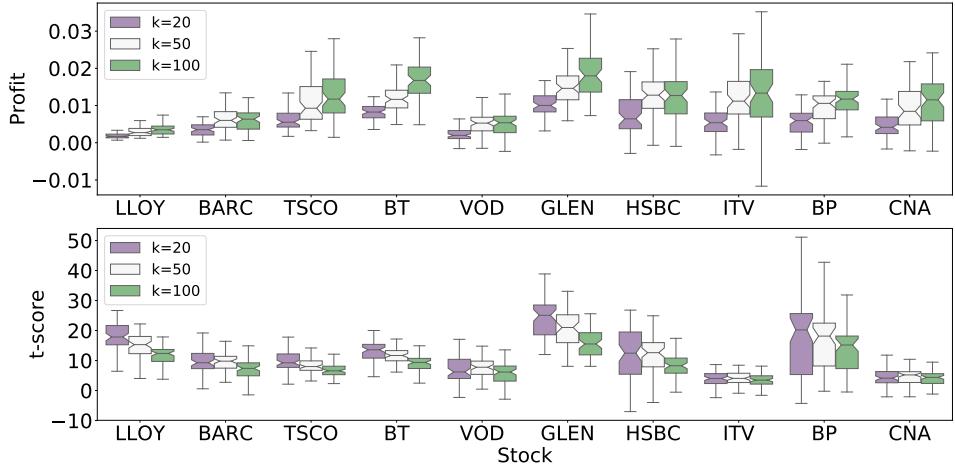


Figure 7. Boxplots for normalised daily profits and t -statistics for different stocks and prediction horizons (k). Profits are in GBX (= GBP/100).

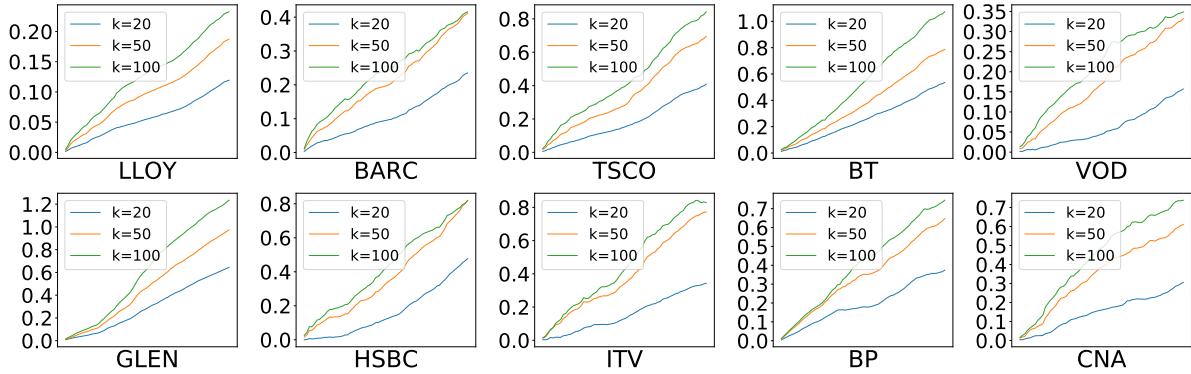


Figure 8. Normalised cumulative profits for test periods for different stocks and prediction horizons (k). Profits are in GBX(= GBP/100).

stocks and prediction horizons. We use a t -test to check if the profits are statistically greater than 0. The t -statistics is essentially the same as Sharpe ratios but a more consistent evaluation metric for high frequency trading. Figure 8 shows the cumulative profits across the testing period. We can observe consistent profits and significant t -values over the testing period for all stocks. Although we obtain worse accuracy for longer prediction horizons, the cumulative profits are actually higher as a more robust signal is generated.

E. Sensitivity Analysis

Trust and risk are fundamental in any financial application. If we take actions based on predictions, it is always important to understand the reasons behind those predictions. Neural networks are often considered as “black boxes” which lack interpretability. However, if we understand the relationship between the inputs’ components (e.g. words in text, patches in an image) and the model’s prediction, we can compare those relationships with our domain knowledge to decide if we can accept or reject a prediction.

The work of [10] proposes a method, which they call LIME, to obtain such explanations. In our case, we use LIME to reveal components of LOBs that are most important for predictions and to understand why the proposed model DeepLOB works

better than other network architectures such as CNN-I [26]. LIME uses an interpretable model to approximate the prediction of a complex model on a given input. It locally perturbs the input and observes variations in the model’s predictions, thus providing some measure of information regarding input importance and sensitivity.

Figure 9 presents an example that shows how DeepLOB and CNN-I [26] react to a given input. In the figure we show the top 10 areas of pros (in green) and cons (in red) for the predicted class (yellow being the boundary). Not coloured areas represent the components of inputs that are less influential on the predicted results or “unimportant”. We note that most components of the input are inactive for CNN-I [26]. We believe that this is due to two max-pooling layers used in that architecture. Because [26] used large-size filters in the first convolutional layer, any representation deep in the network actually represents information gleaned from a large portion of inputs. Our experiments applying LIME to many examples indicate this observation is a common feature.

VI. CONCLUSION

In this paper, we introduce the first hybrid deep neural network to predict stock price movements using high frequency limit order data. Unlike traditional hand-crafted models, where

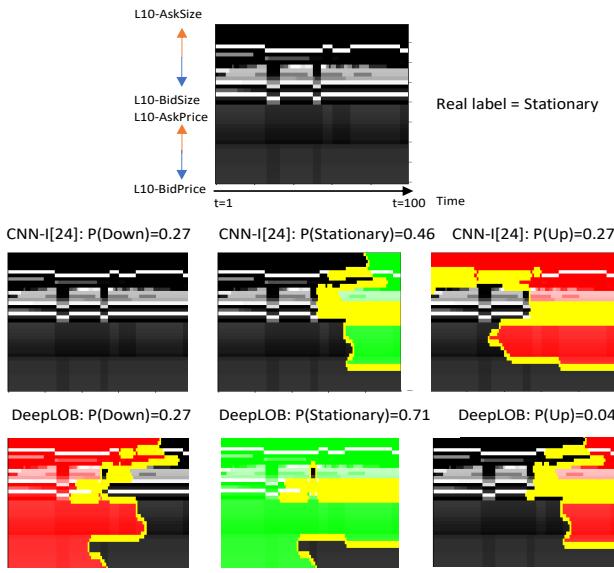


Figure 9. LIME plots. x-axis represents time stamps and y-axis represents levels of the LOB, as labelled in the top image. **Top:** Original image. **Middle:** Importance regions for CNN-I [26]. **Bottom:** Importance regions for DeepLOB model. Regions supportive for prediction are shown in green, and regions against in red. The boundary is shown in yellow.

features are carefully designed, we utilise a CNN and an Inception Module to automate feature extraction and use LSTM units to capture time dependencies.

The proposed method is evaluated against several baseline methods on the FI-2010 benchmark dataset and the results show that our model performs better than other techniques in predicting short term price movements. We further test the robustness of our model by using one year of limit order data from the LSE with a testing period of three months. An interesting observation from our work is that the proposed model generalises well to instruments that did not form part of the training data. This suggests the existence of universal features that are informative for price formation and our model appears to capture these features, learning from a large data set including several instruments. A simple trading simulation is used to further test our model and we obtain good profits that are statistically significant.

To go beyond the often-criticised “black box” nature of deep learning models, we use LIME, a method for sensitivity analysis, to indicate the components of inputs that contribute to predictions. A good understanding of the relationship between the input’s components and the model’s prediction can help us decide if we can accept a prediction. In particular, we see how the information of prices and sizes on different levels and horizons contribute to the prediction which is in accordance with our econometric understanding.

In a recent extension of this work we have modified the DeepLOB model to use Bayesian neural networks [69]. This allows to provide uncertainty measures on the network’s outputs which for example can be used to upsize positions as demonstrated in [69].

In subsequent continuations of this work we would like to investigate more detailed trading strategies, using Rein-

forcement Learning, which are based on the feature extraction performed by DeepLOB.

ACKNOWLEDGEMENTS

The authors would like to thank members of Machine Learning Research Group at the University of Oxford for their helpful comments on drafts of this paper. We are most grateful to the Oxford-Man Institute of Quantitative Finance, who provided limit order data and other support. Computation for our work was supported by Arcus Phase B and JADE HPC at the University of Oxford and Hartree national computing facilities, U.K. We also thank the Royal Academy of Engineering U.K. for their support.

REFERENCES

- [1] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, “Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods,” *Journal of Forecasting*, vol. 37, no. 8, pp. 852–866, 2018.
- [2] C. A. Parlour and D. J. Seppi, “Limit order markets: A survey,” *Handbook of financial intermediation and banking*, vol. 5, pp. 63–95, 2008.
- [3] I. Rosu *et al.*, “Liquidity and information in order driven markets,” Tech. Rep., 2010.
- [4] E. Zivot and J. Wang, “Vector autoregressive models for multivariate time series,” *Modeling Financial Time Series with S-PLUS®*, pp. 385–429, 2006.
- [5] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, “Stock price prediction using the ARIMA model,” in *Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on*. IEEE, 2014, pp. 106–112.
- [6] C. Carrie, “The new electronic trading regime of dark books, mashups and algorithmic trading,” *Trading*, vol. 2006, no. 1, pp. 14–20, 2006.
- [7] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison, “Limit order books,” *Quantitative Finance*, vol. 13, no. 11, pp. 1709–1742, 2013.
- [8] W.-C. Chiang, D. Enke, T. Wu, and R. Wang, “An adaptive stock index trading decision support system,” *Expert Systems with Applications*, vol. 59, pp. 195–207, 2016.
- [9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.
- [11] A. Ang and G. Bekaert, “Stock return predictability: Is it there?” *The Review of Financial Studies*, vol. 20, no. 3, pp. 651–707, 2006.

- [12] P. Bacchetta, E. Mertens, and E. Van Wincoop, "Predictability in financial markets: What do survey expectations tell us?" *Journal of International Money and Finance*, vol. 28, no. 3, pp. 406–426, 2009.
- [13] T. Bollerslev, J. Marrone, L. Xu, and H. Zhou, "Stock return predictability and variance risk premia: Statistical inference and international evidence," *Journal of Financial and Quantitative Analysis*, vol. 49, no. 3, pp. 633–661, 2014.
- [14] M. A. Ferreira and P. Santa-Clara, "Forecasting stock market returns: The sum of the parts is more than the whole," *Journal of Financial Economics*, vol. 100, no. 3, pp. 514–537, 2011.
- [15] B. Mandelbrot and R. L. Hudson, *The Misbehavior of Markets: A fractal view of financial turbulence*. Basic books, 2007.
- [16] B. B. Mandelbrot, "How Fractals Can Explain What's Wrong with Wall Street," *Scientific American*, vol. 15, no. 9, p. 2008, 2008.
- [17] J. Agrawal, V. Chourasia, and A. Mittra, "State-of-the-art in stock prediction techniques," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 4, pp. 1360–1366, 2013.
- [18] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira, "Computational intelligence and financial markets: A survey and future directions," *Expert Systems with Applications*, vol. 55, pp. 194–211, 2016.
- [19] Q. Cao, K. B. Leggio, and M. J. Schniederjans, "A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market," *Computers Operations Research*, vol. 32, no. 10, pp. 2499–2512, 2005.
- [20] J. Sirignano and R. Cont, "Universal features of price formation in financial markets: perspectives from deep learning," *arXiv preprint arXiv:1803.06917*, 2018.
- [21] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market forecasting techniques—Part II: Soft computing methods," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5932–5941, 2009.
- [22] D. T. Tran, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Tensor representation in high-frequency financial data for price change prediction," in *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*. IEEE, 2017, pp. 1–7.
- [23] D. T. Tran, M. Gabbouj, and A. Iosifidis, "Multilinear class-specific discriminant analysis," *Pattern Recognition Letters*, vol. 100, pp. 131–136, 2017.
- [24] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018.
- [25] D. T. Tran, A. Iosifidis, J. Kanniainen, and M. Gabbouj, "Temporal attention-augmented bilinear network for financial time-series data analysis," *IEEE transactions on neural networks and learning systems*, 2018.
- [26] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting stock prices from the limit order book using convolutional neural networks," in *Business Informatics (CBI), 2017 IEEE 19th Conference on*, vol. 1. IEEE, 2017, pp. 7–12.
- [27] —, "Using Deep Learning for price prediction by exploiting stationary limit order book features," *arXiv preprint arXiv:1810.09965*, 2018.
- [28] —, "Using deep learning to detect price change indications in financial markets," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 2511–2515.
- [29] M. Dixon, D. Klabjan, and J. H. Bang, "Classification-based financial markets prediction using deep neural networks," *Algorithmic Finance*, vol. 6, no. 3-4, pp. 67–77, 2017.
- [30] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.
- [31] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in neural information processing systems*, 2013, pp. 809–817.
- [32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [34] J.-F. Chen, W.-L. Chen, C.-P. Huang, S.-H. Huang, and A.-P. Chen, "Financial time-series data analysis using deep convolutional neural networks," in *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on*. IEEE, 2016, pp. 87–92.
- [35] J. Doering, M. Fairbank, and S. Markose, "Convolutional neural networks applied to high-frequency market microstructure forecasting," in *Computer Science and Electronic Engineering (CEEC), 2017*. IEEE, 2017, pp. 31–36.
- [36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [37] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.
- [38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [39] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

- [40] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [42] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PloS one*, vol. 12, no. 7, p. e0180944, 2017.
- [43] S. Selvin, R. Vinayakumar, E. Gopalakrishnan, V. K. Menon, and K. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in *Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on*. IEEE, 2017, pp. 1643–1647.
- [44] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.
- [45] L. Di Persio and O. Honchar, "Artificial neural networks architectures for stock price prediction: Comparisons and applications," *International Journal of Circuits, Systems and Signal Processing*, vol. 10, pp. 403–413, 2016.
- [46] M. Dixon, "Sequence classification of the limit order book using recurrent neural networks," *Journal of computational science*, vol. 24, pp. 277–286, 2018.
- [47] D. M. Nelson, A. C. Pereira, and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1419–1426.
- [48] L. Harris, *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press, USA, 2003.
- [49] M. O'Hara, *Market microstructure theory*. Blackwell Publishers Cambridge, MA, 1995, vol. 108.
- [50] A. N. Kercheval and Y. Zhang, "Modelling high-frequency limit order book dynamics with Support Vector Machines," *Quantitative Finance*, vol. 15, no. 8, pp. 1315–1329, 2015.
- [51] A. Abraham, B. Nath, and P. K. Mahanti, "Hybrid intelligent systems for stock market analysis," in *International Conference on Computational Science*. Springer, 2001, pp. 337–345.
- [52] T. Hendershott, C. M. Jones, and A. J. Menkveld, "Does algorithmic trading improve liquidity?" *The Journal of Finance*, vol. 66, no. 1, pp. 1–33, 2011.
- [53] C. Cao, O. Hansch, and X. Wang, "The information content of an open limit-order book," *Journal of futures markets*, vol. 29, no. 1, pp. 16–41, 2009.
- [54] S. J. Orfanidis, *Introduction to signal processing*. Prentice-Hall, Inc., 1995.
- [55] J. Gatheral and R. C. Oomen, "Zero-intelligence realized variance estimation," *Finance and Stochastics*, vol. 14, no. 2, pp. 249–283, 2010.
- [56] Y. Nevmyvaka, Y. Feng, and M. Kearns, "Reinforcement learning for optimized trade execution," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 673–680.
- [57] M. Avellaneda, J. Reed, and S. Stoikov, "Forecasting prices from Level-I quotes in the presence of hidden liquidity," *Algorithmic Finance*, vol. 1, no. 1, pp. 35–43, 2011.
- [58] Y. Burlakov, M. Kamal, and M. Salvadore, "Optimal limit order execution in a simple model for market microstructure dynamics," 2012.
- [59] L. Harris, "Maker-taker pricing effects on market quotations," *USC Marshall School of Business Working Paper*. Available at <http://bschool.huji.ac.il/upload/hujibusiness/Maker-taker.pdf>, 2013.
- [60] A. Lipton, U. Pesavento, and M. G. Sotiropoulos, "Trade arrival dynamics and quote imbalance in a limit order book," *arXiv preprint arXiv:1312.0514*, 2013.
- [61] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.
- [62] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [63] T. J. Moskowitz, Y. H. Ooi, and L. H. Pedersen, "Time series momentum," *Journal of financial economics*, vol. 104, no. 2, pp. 228–250, 2012.
- [64] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2014.
- [65] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the International Conference on Learning Representations 2015*, 2015.
- [66] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017.
- [67] F. Chollet *et al.*, "Keras," <https://keras.io>, 2015.
- [68] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <https://www.tensorflow.org/>
- [69] Z. Zhang, S. Zohren, and S. Roberts, "BDLOB: Bayesian Deep Convolutional Neural Networks for Limit Order Books," in *Third workshop on Bayesian Deep Learning (NeurIPS 2018)*, 2018.

Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?

Haohan Zhang^{1,2}, Fengrui Hua^{1,3}, Chengjin Xu¹, Hao Kong¹, Ruiting Zuo⁴, Jian Guo^{1*}

¹International Digital Economy Academy (IDEA)

²Artificial Intelligence Thrust, The Hong Kong University of Science and Technology (Guangzhou)

³Data Science and Analytics Thrust, The Hong Kong University of Science and Technology (Guangzhou)

⁴Financial Technology Thrust, The Hong Kong University of Science and Technology (Guangzhou)

{hzhang760, fhua430}@connect.hkust-gz.edu.cn

{xuchengjin,konghao}@idea.edu.cn

ruitingzuo@hkust-gz.edu.cn

guojian@idea.edu.cn

Abstract

The rapid advancement of Large Language Models (LLMs) has spurred discussions about their potential to enhance quantitative trading strategies. LLMs excel in analyzing sentiments about listed companies from financial news, providing critical insights for trading decisions. However, the performance of LLMs in this task varies substantially due to their inherent characteristics. This paper introduces a standardized experimental procedure for comprehensive evaluations. We detail the methodology using three distinct LLMs, each embodying a unique approach to performance enhancement, applied specifically to the task of sentiment factor extraction from large volumes of Chinese news summaries. Subsequently, we develop quantitative trading strategies using these sentiment factors and conduct back-tests in realistic scenarios. Our results will offer perspectives about the performances of Large Language Models applied to extracting sentiments from Chinese news texts.

1 Introduction

At present, an overwhelming volume of news articles and columns are being generated on a daily basis, especially pertaining to companies being traded. Given this landscape, considerable attention has been given to investigating the feasibility of employing Large Language Models (LLMs) for sentiment analysis and processing of these news texts. The aim is to derive a quantifiable technical indicator, or factor, that effectively reflects the desirability of investing in a particular company's stock at a given moment in time. It should be clear that the endeavor we described is a very well-defined down-stream task operating within a language environment (the Chinese language) that is drastically different from the English language that the main-stream LLMs have been predominantly trained on. Although several approaches

have been proposed to enhance the performance of LLMs on down-stream tasks in alternative language environment, questions remain as to which LLM or method of improvement is the most optimal in the specific context of extracting sentiment factor from Chinese financial news texts. For such a comparative analysis to be possible however, we must adopt a standardized experiment procedure that can be easily applied to all LLMs and yield quantitative results based on metrics selected with sufficient domain knowledge in the field of quantitative trading.

On the other hand, even though works such as [Lopez-Lira and Tang, 2023] have successfully demonstrated that LLMs like ChatGPT [Brown *et al.*, 2020] can be used to extract sentiment factors from English news texts that are highly correlated to the returns of US stocks, it is our contention that the smooth implementation of the same framework applied on the Chinese text still faces two major concerns. Firstly, it is noteworthy that the dominant and leading LLMs have predominantly been trained on English corpora. Consequently, the transferability of sentiment mining techniques from English texts to Chinese texts remains uncertain. Secondly, while the effectiveness of sentiment factor mining by LLMs has been established in prior research, discrepancies arise due to variations in parameter selection for constructing stock trading simulations back-tests and the utilization of diverse raw news data-sets, encompassing disparate sizes, scopes, and sources. As a consequence, objectively evaluating and comparing the efficacy of different LLMs when applied to the specific task of Chinese financial text sentiment factor building still poses significant challenges.

To address these concerns, we propose an innovative approach that combines sentiment extraction with realistic back-tests of quantitative strategies. This approach allows us to directly assess the effectiveness of LLMs' sentiment extraction capabilities using a comprehensive and standardized experimental procedure with quantifiable metrics such as excess return, risk adjusted return and win ratio. Before elaborating on the experimental procedure, we outline the scope and coverage of data, data pre-processing, and the parameters that the back-tests should adhere to. By integrating these el-

*Corresponding author

ements, we aim to provide a robust framework for evaluating and comparing the performance of various LLMs in sentiment extraction from Chinese financial news texts. As practical illustrations, we will subsequently conduct sentiment extraction from 394,426 items of Chinese news summaries about Chinese publicly traded companies. This process will be executed using three distinct LLMs, representing: 1) the baseline model, 2) the Chinese language-specific pre-trained LLM, and 3) the financial domain-specific pre-trained LLM. We will then construct investment portfolios and run stock trading simulation back-tests according to our defined settings and parameters in order to rigorously test the correlation between the sentiment factors and return of investment, which, from our perspective, is the best way to reflect the LLMs' effectiveness in correctly extracting sentiments from Chinese financial texts. Finally, we will discuss the results of the back-tests and the insights gained from such comparative analysis. To the best of our knowledge, we are the first group to conduct sentiment analysis on such an extensive source of Chinese news text using prevalent LLMs such as ChatGPT [Brown *et al.*, 2020] and back-test the acquired sentiment factors by deploying quantitative trading strategies on platforms that lead in the quantitative finance industry.

2 Related Works

Over the past decade, significant advancements have been made in the field of Natural Language Processing (NLP), leading to the development of powerful language models. One such groundbreaking model is the Transformer [Vaswani *et al.*, 2017], which introduced an attention-based encoder-decoder architecture that has consistently yielded superior performance compared to Recurrent Neural Network (RNN) based architectures, including Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] and Gated Recurrent Unit (GRU) [Cho *et al.*, 2014], in various language tasks. The Transformer model's key innovation lies in its ability to effectively capture long-range dependencies within sequences by employing self-attention mechanisms. This enables the model to assign importance to different parts of the input text and establish contextual relationships, resulting in improved language understanding and generation capabilities. Consequently, the Transformer architecture has paved the way for a new generation of state-of-the-art Large Language Models that have inherited its powerful framework exemplified by the Generative Pre-trained Transformer (GPT) [Radford *et al.*, 2018] series, developed by OpenAI, which focus on generative language modeling and utilize a variant of the Transformer architecture, where tokens attend to tokens that appear before in the sentence. This unidirectional attention mechanism, often referred to as left-to-right attention or auto-regressive decoding, allows the models to generate coherent and contextually relevant text referencing the attention assigned to preceding words. Another LLM that is based on the Transformer architecture is the BERT (Bidirectional Encoder Representations from Transformers [Devlin *et al.*, 2019]) model. BERT, introduced by Google AI, is designed to capture bidirectional contextual information from the input text. Unlike the unidirectional attention of GPT,

BERT employs a bidirectional attention mechanism. It enables tokens to attend not only to preceding words but also to words appearing after in latter parts of the sentence. By considering the entire context, BERT can effectively capture the dependencies and relationships between words, resulting in a better understanding of the text. Building on BERT, authors of RoBERTa [Liu *et al.*, 2019] seek to optimize the performance of parent model by conducting various pre-training improvements such as increasing the data size, training duration as well as adopting dynamic masking during training. Such methods are proven to have pushed to boundaries of the performance of BERT even further.

These LLMs are designed in such a way as to greatly facilitate down-stream task-specific fine-tuning and additional pre-training on supplementary corpora. From our perspective, user-defined improvement on parent LLMs includes two key aspects: language-specific and domain-specific. While certain commercialized prototypes, such as ChatGPT [Brown *et al.*, 2020], support multilingual reasoning and text generation, many language models predominantly rely on training with English texts. To achieve comparable performance across other language environments, including our specific focus Chinese, extensive training on alternative language environment becomes imperative. The Erlangshen-RoBERTa [Wang *et al.*, 2022] model, which inherits from the RoBERTa architecture and is further pre-trained on the 180 Giga-byte version of the Wudao Corpora [Yuan *et al.*, 2021], marks one of the most notable efforts in Chinese language-specific pre-training, achieving top performance on state-of-the-art NLP benchmarks.

The other aspect pertains to domain-specific improvements which has to do with inheriting a parent LLM architecture and either continuously training it on additional corpus data-sets related to the target technical domain or fine-tuning it with expertly constructed labels, this process effectively leverages the pre-existing linguistic knowledge of the LLMs while allowing them to acquire domain-specific nuances and intricacies, thus enabling heightened proficiency in the desired domain. One key example is the FinBERT [Huang *et al.*, 2023] model, which trains the BERT model further on U.S. Securities and Exchange Commission (SEC) filing data, resulting in a model well-versed in financial domain knowledge and well-suited for English financial text sentiment classifications. Similar efforts also include [Wu *et al.*, 2023], a LLM trained on extensive English financial copora.

Related works on the task of applying LLM for stock prediction or investment decision optimization include [Ding *et al.*, 2023], which proposes new approaches for integrating the information from LLMs into traditional quantitative factors. [Koa *et al.*, 2023] proposed a new paradigm for training LLMs to give explainable stock predictions, while [Yu *et al.*, 2023] also worked on using LLM to generate explainable stock forecasting but based on financial time series data. About LLM based sentiment analysis for stock prediction, recent works include [Steinert and Altmann, 2023] and [Jaggi *et al.*, 2021]

3 Data

The data we use take in the form of news summaries regarding Chinese publicly traded companies and is mainly acquired through web crawling. A total of 394,429 news summary items were acquired, spanning a time period spanning from October 1, 2021, to February 22, 2023 and covering 5,021 publicly traded companies listed on Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE). As an additional filtering criterion, we retain only those news summaries that are generated prior to the market opening at 9:30 am. Doing so helps us ensure that the technical indicators extracted from these news summaries are promptly available for utilization as soon as the market commences trading.

The October 2021 to February 2023 period that the data spans is stipulated to ensure that it falls chronologically after the data used for training GPT-3.5, which serves as our baseline model for sentiment analysis. If we had not chosen to do this, it could potentially result in the ChatGPT model referencing "future data" that might not be available at the time the news summaries are generated. This scenario could lead to inaccurate sentiment analysis results and create situations that do not align with reality. To procure professional financial information and analysis, we have identified a set of prime sources which are selected based on their credibility and reputation within the financial domain. Consequently, we mainly conduct data crawling from these prime sources. These sources are: Sina Finance [Sina, 2023], Hithink RoyalFlush [Hithink, 2023], Tencent [Tencent, 2023], Dazhong Securities [DZZQ, 2023], Hexun [Hexun, 2023], Netease Finance [Netease, 2023], Caifuhao [Caifuhao, 2023]. We offer Table 1 to illustrate the distribution of these prime sources within our data-set.

Source Name	Proportion (%)
Hithink RoyalFlush	59.57
Sina Finance	33.65
Tencent	4.88
Hexun	1.55
Caifuhao	0.23
Netease Finance	0.1
Dazhong Securities	0.02

Table 1: The Distribution of News Sources within Our Data-set

4 Methodology

4.1 Using ChatGPT to Extract Sentiment Factors

We use ChatGPT, and the GPT-3.5 behind it as a baseline language model without any additional pre-training or fine-tuning. Similar to [Lopez-Lira and Tang, 2023], we conduct the sentiment factor mining by fusing the news summary information as well as the instructions about the task that we expect ChatGPT to perform for us into prompts. A user's prompt is essential in guiding the model's response and determining the context of the conversation. When generating a response, the model takes into account the user's input, including the specific words used, the overall tone, and the desired

information or assistance being sought. The prompt serves as the primary input that helps shape the model's understanding of the user's intent and the direction the conversation is taking.

Upon receiving our prompt, ChatGPT decides whether the item of news summary in question contains good, bad, or neutral sentiment (this also includes cases where ChatGPT is unsure of the sentiment and unable to produce an informed rating) for the company involved. One sample prompt as well as the associated response is illustrated in Fig 1. For this particular example, we also invite ChatGPT to elaborate on the reason behind forming such a sentiment analysis. For the sentiment analysis procedure on our data-set, we used Chinese prompts, the English version of the prompt is given for demonstration purposes.

After the responses for all 394,429 items of news summary have been given. We translate the response into numeric values with GOOD NEWS being assigned as 1, NOT SURE being assigned as 0 and BAD NEWS being assigned as -1. For cases where a listed company is being mentioned by multiple news sources on the same day, we take the average of the ChatGPT's responses across the different news sources. These averaged ratings by ChatGPT are stored as the Chinese ChatGPT Factor to be referenced by our trading simulation back-tests.

4.2 Using Language-Specific Pre-Trained LLM to Extract Sentiment Factors

As previously stated, a fundamental objective of this research is to examine the degree to which the effectiveness of Language Models (LLMs) in generating sentiment factors from analysis of news that are conducive to the development of quantitative trading strategies with high returns can be extended to Chinese financial textual data. By such explorations, our research aims to contribute insights into the potential adaptability and performance of LLM-based approaches in the realm of Chinese language sentiment analysis for the exclusive purpose of informing quantitative trading strategies. Therefore, in addition to applying the baseline model, GPT-3.5, we also turn to the Erlangshen-RoBERTa-110M-Sentiment [Wang *et al.*, 2022] model which was pre-trained on the 180 GB version of the WuDao Chinese Corpora. The Erlangshen-RoBERTa series were pre-trained on Chinese texts in such a way as to take into account the unique characteristics of the Chinese language and the traits of the Chinese character contextual relationships. This aspect sets them apart from many other pre-trained LLMs that, although supporting Chinese textual input and response generation, may not fully leverage such language-specific traits.

Through a comparative analysis of the performance of the quantitative strategies from both the baseline GPT-3.5 and the language-specific pre-trained model on the same benchmark, we seek to answer the question of whether such language-specific pre-training approaches have any contribution to yielding a superior returns. The Erlangshen-RoBERTa-110M-Sentiment is open-source and can be easily downloaded and applied for the sentiment analysis of any section of Chinese texts within the maximum token limit. The return of the last soft-max activation layer is a tuple contain-

Prompt (Chinese): 忘掉之前的所有指令，你现在是一个金融专家去给投资建议。我给你一个新闻，请回答这条新闻对于该上市公司来说是GOOD NEWS还是BAD NEWS，请只在GOOD NEWS, BAD NEWS, NOT SURE 三个选项中选择一个回答，不要有更多的回答。新闻是：公司持续推动传统滤光片业务产品结构升级，在竞争较为激烈的业态下彰显自身韧性。展望未来，我们看好公司手机光学业务夯实基本盘，HUD 及 AR 等前瞻布局有望打开第二成长曲线。上市公司是：浙江水晶光电科技股份有限公司

Prompt (English): Forget all previous instructions. You are now a financial expert giving investment advice. I'll give you a news summary, and you need to answer whether this news is GOOD NEWS or BAD NEWS for the listed company. Please choose only one option from GOOD NEWS, BAD NEWS, NOT SURE, and do not provide any additional responses. The news content is: "The company continues to promote the upgrading of its traditional filter business product structure, demonstrating its resilience in a fiercely competitive market. Looking ahead, we are optimistic about the company's solid foundation in the mobile optics business and expect forward-looking layouts such as HUD and AR to open up a second growth curve." The listed company is Zhejiang Crystal Optoelectronics Technology Co., Ltd.

Response: GOOD NEWS

Elaboration (Chinese): 根据新闻，公司持续推动传统滤光片业务产品结构升级，在竞争激烈的行业中展现出自身的韧性。除此以外，展望未来，公司看好手机光学业务的发展，并有前瞻布局如HUD和AR等，有望打开第二成长曲线。基于这些信息，我认为对于该上市公司来说，这条新闻是GOOD NEWS（好消息）。

Elaboration (English): Based on the news, the company is continuously pushing for the upgrade of its traditional filter film business and demonstrating resilience in a highly competitive industry. Furthermore, looking ahead, the company is optimistic about solidifying its foundation in the mobile optical business and expects future growth through forward-looking initiatives such as HUD and AR. Based on this information, I believe this news is GOOD NEWS for the listed company.

Figure 1: Demonstration of Prompts Structured for Sentiment Analysis and the Response by ChatGPT

ing the probability of the input text's sentiment being negative or positive. When we applied Erlangshen-RoBERTa-110M-Sentiment model to the same news summary as shown in section 4.1, it rated the sentiment as being 98.33% positive. It can be seen that for this particular sample, the sentiment analysis of Erlangshen-RoBERTa-110M-Sentiment aligns with that of ChatGPT. We conduct sentiment classification on all the rest of the samples in the data-set and store the results as the Erlangshen Facto

4.3 Using Domain-Specific Pre-trained LLM to Extract Sentiment Factors

For this part, we start with the open-source variation of BERT and conduct further pre-training on Chinese financial corpora data. After the tokenizations and the embedding layer updates have been finalized, we connect a hidden layer and a soft-max activation layer of size 3 to the original attention layers of the BERT model to acquire our own BERT-based sentiment classifier. We then ask colleagues of financial expertise to manually label the news summary data into three classes in terms of the sentiment manifested by the text (Positive (+1), Neutral (0), Negative (-1)) and conduct supervised training to derive what we shall call a Chinese FinBERT classifier. It is guaranteed that the training data used for manual sentiment labeling is strictly separated from the news summary data employed in our experimental analysis. After fine-tuning has been completed, we use our acquired BERT-based sentiment classifier to predict the sentiment of the 394,429 items of news summary. The model will output a predicted class based on the probabilities output by the final soft-max layer. We document the results as Chinese FinBERT factor.

5 Experiment and Parameters of Trading Strategies

As stated above, we intend to provide a comprehensive and rigorous experimental procedure so that any LLM's efficacy

in extracting sentiment from Chinese financial text for the purpose of building quantitative trading strategies can be objectively measured no matter which model or method was used. In order to ensure an unbiased assessment in accordance with our benchmark, we establish a requirement that the quantitative trading strategies adhere to uniform settings and parameters. While a seasoned quantitative trading expert may identify significant potential for optimizing these parameters to achieve higher returns, we deliberately refrain from delving extensively into the intricacies of trading strategy technicalities within this study. Our primary objective lies in the evaluation of LLMs' effectiveness within the specific realm of Chinese text sentiment mining which can be achieved as long as the same set of parameters are employed across all tests.

We will now proceed to enumerate and explain the standardized settings all trading strategies in our experiments adhere to:

- Portfolios are adjusted daily and only at market open, 9:30 am Beijing Time (UTC+8:00)
- We only use news generated or acquired before market open. In this way, the sentiment factors extracted can be directly used at trading time.
- We adjust the investment portfolios by buying in the stocks who have the highest ranking sentiment factors and selling those in our portfolio that have the lowest ranking factors. The maximum number of stocks to be bought or sold in one day is 500.
- The maximum turnover ratio for our portfolio to be 1.0, which means we allow all of the previously held stocks to be sold to be replenished by entirely new stocks. Although this rarely happens during the back-test.
- In order to account for the slippage and delays commonly encountered in real-world trading scenarios, we have chosen to deviate from using the straight mar-

ket open price in our back-test. Instead, we have implemented a more realistic approach by utilizing the Volume-Weighted Average Price (VWAP) between 9:30 am and 9:35 am. This VWAP is calculated by summing the values of all trades that occurred during this specific five-minute period and dividing the sum by the total volume traded within that time-frame.

- We avoid being overly-optimistic about our simulated returns by imposing a transaction fee of 0.15% of transaction value. This includes a 0.05% commission charged by the stock brokerage firms and a 0.1% stamp duty fee paid to the stock exchanges. In reality at present, Chinese brokerage firms rarely charge above 0.03% and the stamp duty fee is only charged at selling transactions instead of all transactions. Therefore, it should be clear that a 0.15% of transaction value emulates a more stringent trading environment than what is observed in reality.
- We use the CSI 300 index as basis when calculating the excess returns.

6 Results and Discussions

After the back-tests have been run, we collect the results and the performance of all portfolios built around all three sentiment factors on a series of metrics which comprise our benchmark. We now proceed to give definitions of these metrics:

- Annual Excess Return: Annual excess return refers to the additional or extra return earned by an investment or portfolio over and above what was expected or compared to a stock index. As stated before, in our experiment we have selected the CSI 300 index to be the baseline, therefore our excess return is calculated over the performance of the CSI 300 index.
- Annual Net Asset Return: The annual percentage increase or decrease in the value of our portfolio in terms of net market capitalization of all stocks held.
- Win Rate: Percentage of trade days where portfolio encounters a positive return over all trade days.
- Sharpe Ratio: The Sharpe Ratio [Sharpe, 1966] was invented by William Sharpe in 1966. The formula is given as

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p} \quad (1)$$

The R_p is the return of the portfolio, R_f is the risk-free return and the σ_p is the standard deviation of the return. The Sharpe ratio gives a measurement of risk-adjusted returns.

- Average Stocks Held per Day: The average number of stocks present in the investment portfolio every day
- Turn-over Ratio: Average percentage of the investment portfolio adjusted daily in terms of market capitalization.

We group these metrics into to parts and show the results under Annual Excess Return, Annual Net Asset Return, Win Rate and Sharpe Ratio in Table 2. These would serve as the

main metrics and are primary indicators of performance. We put Average Stocks Held per Day and Turn-over Ratio into Table 3, these would serve as supplementary metrics that provide insights into the portfolio adjustment characteristics of the trading strategy. We also plot the excess returns of all three factors during the entire back-test period in Figure 2.

The net asset return of all three factors during the entire back-test period is given in Figure 3

It is clear to see that in terms of annual return, risk adjusted return and excess return, the Erlangshen sentiment factor outperforms the remaining factors. To further elucidate the correlation between the values of the sentiment factor derived from the Erlangshen-110M-Sentiment model and portfolio excess returns, we have partitioned our held stocks into three distinct groups based on their rankings according to the Erlangshen sentiment factor. Group 1, on average, exhibits the lowest Erlangshen factor value while Group 3 displays the highest. The excess returns for Group 1, Group 2, and Group 3 are then plotted in Figure 4. Notably, we observe that after an initial period of fluctuation, the three groups become distinctly separated. Furthermore, Group 3, characterized by the highest Erlangshen factor value, consistently demonstrates the highest returns, while Group 1, with the lowest Erlangshen factor value, consistently exhibits the lowest excess returns. This observation provides further substantiation for the notion that the Erlangshen sentiment factor extracted through the Erlangshen-110M-Sentiment model is closely associated and correlated with investment returns.

It is remarkable to witness how the comparatively smaller Erlangshen model, with a modest 110 million parameters, manages to exhibit slightly superior performance within our benchmark for the specific task at hand. This outcome serves as a testament to the fact that practitioners and researchers working on Chinese quantitative stock trading strategies may not always need to invest substantial resources into larger models. Instead, by employing strategic fine-tuning and extensive pre-training techniques tailored to the intricacies of the Chinese language, desired outcomes can be effectively achieved. This revelation underscores the significance of considering language-specific characteristics and employing targeted methodologies, illustrating that optimal results can be attained without solely relying on sheer model size for the particular task of Chinese financial sentiment extraction.

7 Conclusions

This study explores the potential of Large Language Models (LLMs) in constructing quantitative stock trading strategies by extracting sentiment factors from Chinese financial news texts. Our research addresses the need for successful implementation of LLMs in the Chinese stock market and provides a rigorous standardized back-testing framework to objectively evaluate the efficacy of various LLMs in this specialized domain.

By employing three distinct LLMs, including a baseline all-purpose LLM (ChatGPT) with extensive parameters, a Chinese language-specific pre-trained LLM (Erlangshen-RoBERTa), and a financial domain-specific pre-trained LLM (Chinese FinBERT), we extract sentiment factors from a large

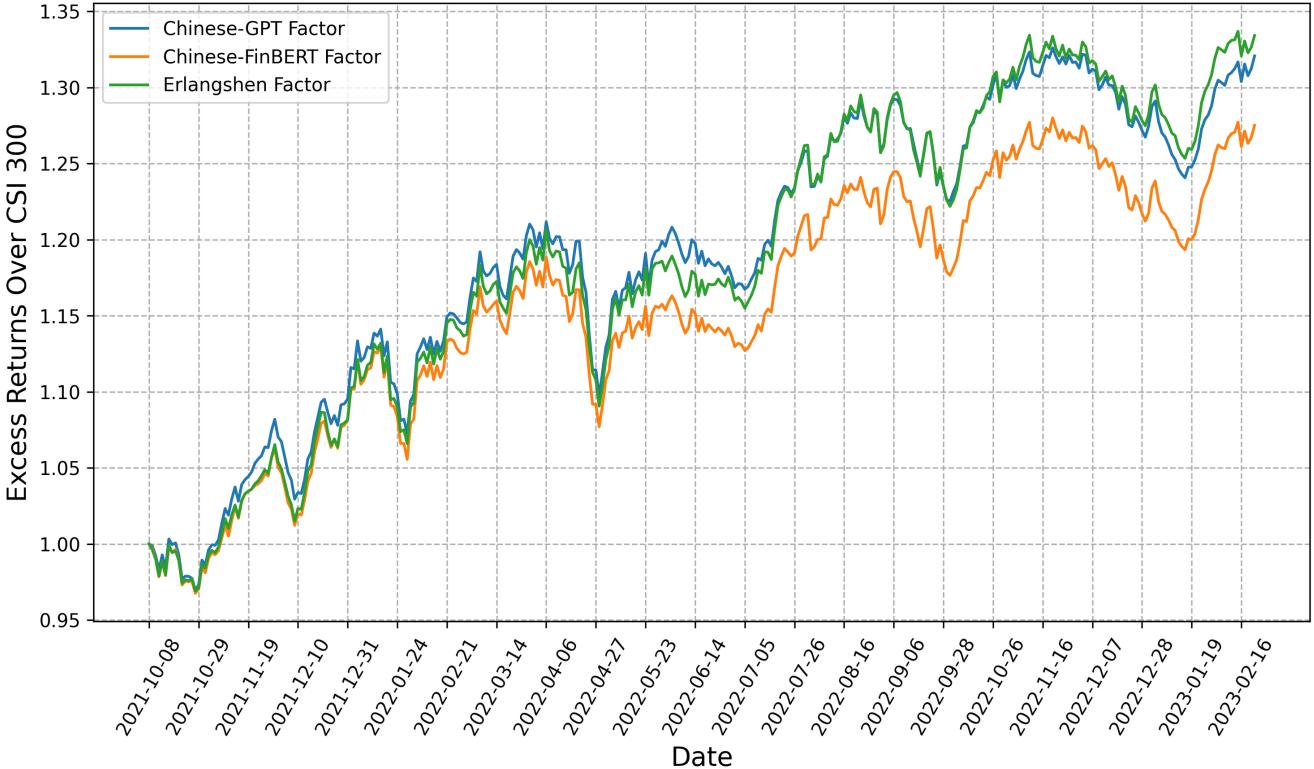


Figure 2: Excess Returns of All Three Sentiment Factors

Table 2: Main Metrics of Back-test

Factor Name	Annual Excess Return (%)	Annual Net Asset Return (%)	Win Rate(%)	Sharpe Ratio	Max Withdrawal Rate
Chinese-GPT	23.17	11.11	57.78	0.6444	0.2315
Chinese-FinBERT	19.88	7.82	57.19	0.4836	0.2397
Erlangshen-110M	24.14	12.08	58.38	0.6841	0.2219

volume of Chinese news summaries about listed companies in China. Through rigorous stock trading simulation back-tests, we evaluate the performance of these sentiment factors in terms of annual return, risk-adjusted return, and excess return and consequently arrive at some perspectives.

The GPT-3.5 iteration of ChatGPT used in this study, which has around 175 billion parameters, contrasts with Erlangshen-110M, which despite its much smaller size of 110 million parameters, outperformed GPT-3.5 across all metrics in this task. This suggests that in the specialized domain of Chinese sentiment factor extraction, a deeper understanding of the nuances in Chinese characters and contextual relationships can lead to more accurate sentiment classification, a capability that cannot necessarily be compensated for by larger models. Additionally, the methodology of sentiment rating differs between the two models. GPT-3.5 was tasked with returning a discrete sentiment rating from three options (-1, 0, 1), whereas Erlangshen-110M outputs a continuous score be-

tween 0 and 1. This continuous scale could potentially offer better differentiation among news items with varying levels of positive sentiment when used to select stocks based on their sentiment ratings. On the other hand, comparing Erlangshen-110M with Chinese-FinBERT reveals that in this task, extensive pre-training on more generalized Chinese texts can be more critical for accurate sentiment modeling than domain-specific training alone. Firstly, the corpus used by Chinese-FinBERT, though domain-specific, is significantly smaller than the Wudao corpus used by Erlangshen-110M, which facilitates a broader understanding of the language. Secondly, effectively capturing sentiment—an inherently human perception—often does not require deep domain-specific knowledge, which may partly explain why Chinese-FinBERT also underperforms compared to GPT-3.5. This insight suggests that a more expansive training dataset on the target language, even one less focused on specific domains, can be more beneficial for tasks involving fundamental human cognitive pro-

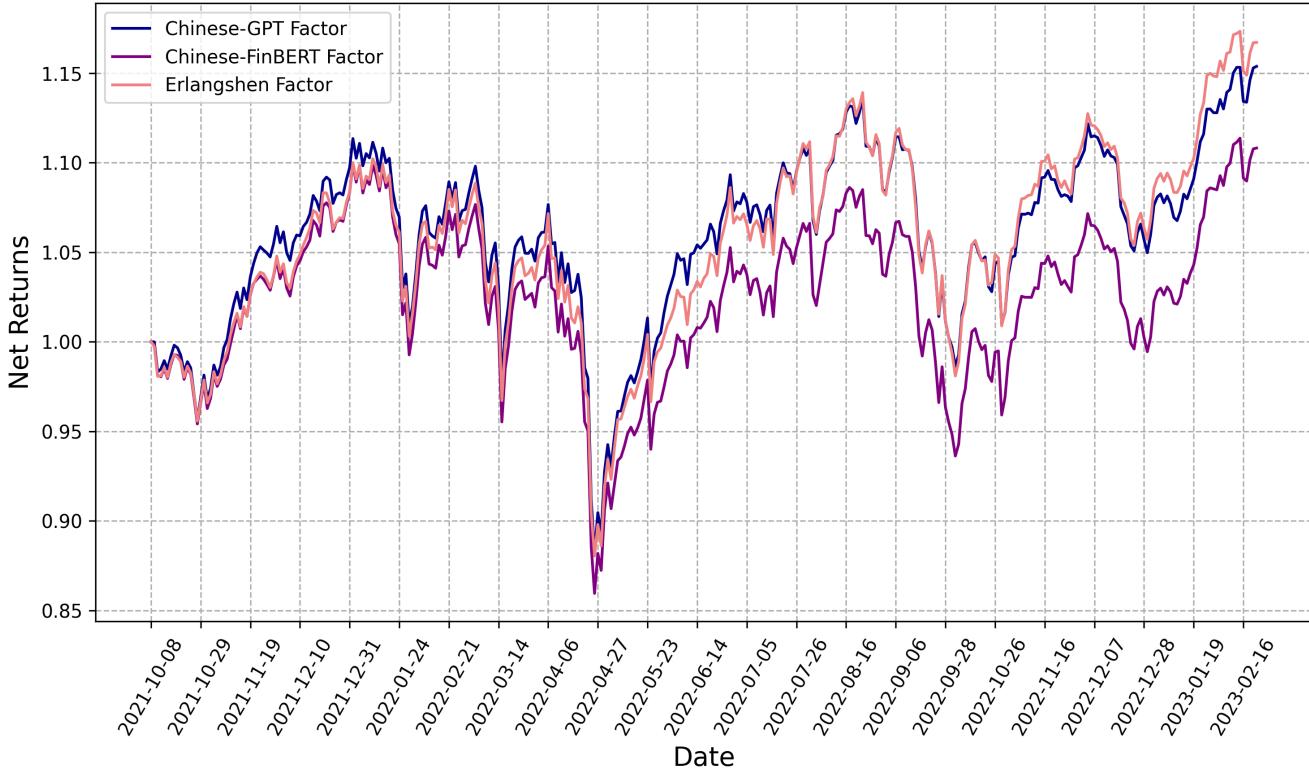


Figure 3: Net Returns of All Three Sentiment Factors

Factor Name	Average Stocks Held per Day	Turn-over Ratio (%)
Chinese-GPT	1684	5.37
Chinese-FinBERT	1788	6.48
Erlangshen-110M	1830	4.17

Table 3: Supplementary Metrics of Back-test

cesses like sentiment analysis.

By providing a comprehensive and standardized procedure, our study contributes to the understanding of LLMs' potential in the specialized domain of sentiment factor extraction from Chinese news text data within the context of quantitative trading. We invite fellow researchers and quantitative finance practitioners to reference our standardized back-testing procedures so that together, we can truly unveil the potential of sentiment.

8 Conflict of Interest and Data Availability Statements

The authors declare that they have conducted this study in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The research was carried out independently and was not influenced by any external organization. The data utilized in this study, derived from Chinese news summaries, were collected through our own efforts and are not publicly available due to the extensive labor and proprietary methods involved in

their aggregation and analysis. Our commitment to scientific integrity ensures that the findings and conclusions presented are the result of unbiased analysis. Should any researchers express interest in establishing joint research in fields related to or as an extension of the work presented in this paper, we do not exclude the possibility of sharing the data for such purposes. We will review such requests on a case-by-case basis, ensuring that any collaboration aligns with our principles of scientific integrity and the potential for mutual advancement in our field.

References

- [Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Kueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark,



Figure 4: Excess Return by Groups of Erlangshen Factor Value

Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Caifuhao, 2023] Caifuhao. <https://caifuhao.eastmoney.com/>, 2023.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL)*, 2019.

[Ding *et al.*, 2023] Yujie Ding, Shuai Jia, Tianyi Ma, Bingcheng Mao, Xiuze Zhou, Liuliu Li, and Dongming Han. Integrating stock features and global information via large language models for enhanced stock return prediction. *arXiv preprint arXiv:2310.05627*, 2023.

[DZZQ, 2023] DZZQ. <http://www.dzzq.com.cn/>, 2023.

[Hexun, 2023] Hexun. <https://www.hexun.com/>, 2023.

[Hithink, 2023] Hithink. <https://www.10jqka.com.cn/>, 2023.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Huang *et al.*, 2023] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

[Jaggi *et al.*, 2021] Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, and Matloob Khushi. Text mining of stocktwits data for predicting stock prices. *arXiv preprint arxiv: 2103.16388*, 2021.

[Koa *et al.*, 2023] Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. *arXiv preprint arxiv: 2402.03659*, 2023.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL)*, 2019.

[Lopez-Lira and Tang, 2023] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

[Netease, 2023] Netease. <https://money.163.com/>, 2023.

- [Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
- [Sharpe, 1966] William F. Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, 1966.
- [Sina, 2023] Sina. <https://finance.sina.com.cn/>, 2023.
- [Steinert and Altmann, 2023] Rick Steinert and Saskia Altmann. Linking microblogging sentiments to stock price movement: An application of gpt-4. *arXiv preprint arxiv:2308.16771*, 2023.
- [Tencent, 2023] Tencent. <https://xw.qq.com/>, 2023.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.
- [Wang *et al.*, 2022] Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970, 2022.
- [Wu *et al.*, 2023] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhjanan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- [Yu *et al.*, 2023] Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm – explainable financial time series forecasting. *arXiv preprint arXiv: 2306.11025*, 2023.
- [Yuan *et al.*, 2021] Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2, 06 2021.

Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis

WENJUN GU¹

Carey Business School, Johns Hopkins University, Baltimore, USA

Yihao Zhong²

Courant Institute of Mathematical Sciences, New York University, New York, USA

Shizun Li³

Independent Researcher, Lynnwood, USA

Changsong Wei⁴

Digital Financial Information Technology Co.LTD, Chengdu, China

Liting Dong⁵

Cox School of Business, Southern Methodist University, Dallas, USA

Zhuoyue Wang⁶

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA

CHAO YAN*

Department of Electrical and Computer Engineering, Northeastern University, Boston, USA

yan.chao@northeastern.edu

The stock market's ascent typically mirrors the flourishing state of the economy, whereas its decline is often an indicator of an economic downturn. Therefore, for a long time, significant correlation elements for predicting trends in financial stock markets have been widely discussed, and people are becoming increasingly interested in the task of financial text mining. The inherent instability of stock prices makes them acutely responsive to fluctuations within the financial markets. In this article, we use deep learning networks, based on the history of stock prices and articles of financial, business, technical news that introduce market information to predict stock prices. We illustrate the enhancement of predictive precision by integrating weighted news categories into the forecasting model. We developed a pre-trained NLP model known as FinBERT, designed to discern the sentiments within financial texts. Subsequently, we advanced this model by incorporating the sophisticated Long Short Term Memory (LSTM) architecture, thus constructing the innovative FinBERT-LSTM model. This model utilizes news categories related to the stock market structure hierarchy, namely market, industry, and stock related news categories, combined with the stock market's stock price situation in the previous week for prediction. We selected NASDAQ-100 index stock data and trained the model on Benzinga news articles, and utilized Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Accuracy as the key metrics for the assessment and comparative analysis of the model's performance. The results indicate that FinBERT-LSTM performs the best, followed by LSTM, and DNN model ranks third in terms of effectiveness.

CCS CONCEPTS • Machine learning • Modeling and simulation

Additional Keywords and Phrases: stock price, FinBERT-LSTM, news sentiment

1 INTRODUCTION

The stock market is a key component of the economic ecosystem [1]. It serves as a conduit through which publicly traded corporations secure financial resources, which provides funding for various research and development projects to create services, products, and employment opportunities that contribute to economic growth. Should a company's performance falter, its share value is likely to experience a sharp decline; Should a company's performance excel, its stock value is likely to experience a dramatic surge. Investors are thus encouraged to delve into the intricacies of the stock market to discern which ventures are poised to yield a return. Anticipating stock price movements is inherently challenging, as it is not governed by a set of rigid mathematical formulas. The market is subject to fluctuation at any given moment, influenced by a myriad of elements including economic inflation and international tensions [2]. Nonetheless, by identifying trends within the data, it is possible to formulate reasonably precise short-term forecasts.

Analyzing current events can significantly contribute to forecasting stock movements, given that the stock market is profoundly swayed by news pertinent to the financial sphere [3]. News articles contain market information, negative articles are related to poor company performance, and positive articles are related to good performance. Therefore, it is possible to understand the trend of stocks by studying news articles, and incisive news analysis can yield substantial advantages by enhancing the accuracy of stock trend forecasts. Over recent years, individuals have examined stock-related news from a variety of angles, yet the potential for extracting valuable insights from financial news archives remains largely untapped. Despite the complexity introduced by a multitude of factors, the task of analyzing news remains a formidable challenge. Sentiment analysis, a subset of textual analysis, is employed to gauge the sentiment polarity of written content. This method evaluates the intrinsic sentiment of a text, categorizing it into positive or negative sentiments [4]. By doing so, it captures the public's emotional response to news stories. For instance, a report highlighting profits and acquisitions can evoke positive sentiments, potentially elevating a company's stock value, whereas a piece detailing layoffs and bankruptcy may incite negative reactions, contributing to a decrease in stock prices. Consequently, discerning the emotional undertones of news articles can offer a more profound insight into a company's operational success and inform predictions regarding its stock market trajectory.

Conversely, technical analysis techniques concentrate on examining the dynamics of stock prices, trading volumes, and the psychological expectations of investors. This approach leverages tools such as K-line charts to scrutinize the trajectory of stock indices for individual equities or the market as a whole, utilizing numerical data to forecast stock prices. Throughout history, a significant portion of initial scholarly work has focused on leveraging data from a particular moment, denoted as time t , to forecast the direction of stock prices at the immediately following moment, time $t+1$. In recent years, there has been a shift in the academic discourse, with some researchers framing the challenge of stock market prediction within the framework of sequential analysis. In this approach, the forecasting model processes a series of data points that span an ongoing timeline, as referenced in the literature [5-7].

2 STATE OF THE ART

Over an extended timeframe, the domain of predicting stock market movements has increasingly relied on the employment of sophisticated machine learning and deep learning methodologies, as highlighted in recent scholarly works [8-9]. Advancements in the field, notably the progression of Recurrent Neural Networks (RNNs), have been pivotal, especially the variants equipped with Long Short-Term Memory (LSTM) features, which have

garnered significant attention in the literature [10], and the incorporation of attention mechanisms, notably self-attention and transformers [11], represent the cutting-edge advancements in deep learning. These sophisticated methods have notably enhanced the precision of tasks that are word-centric [13-14]. The predominant approach in academic literature for news-sensitive stock trend forecasting involves leveraging both stock price data and news articles. Moreover, technical indicators derived from stock price data are also commonly utilized. As highlighted in [15], this study brilliantly showcases the strength of ensemble methods in sentiment analysis, providing a robust framework for future research. Several studies have underscored the efficacy and relevance of integrating multifaceted data into predictive models [16-17]. Furthermore, it has been demonstrated that deep learning frameworks have significantly bolstered feature representation and forecasting accuracy within the financial sector [18-19]. In the work referenced as [20-22], an innovative dual-source stock market forecasting approach was presented, which utilizes numerical attention mechanisms, termed as Numerical Attention-Based (NBA). Subsequently, the study denoted by [23-25] introduced a state-of-the-art autoregressive neural network model that integrates elements of sentiment analysis into its predictive framework. They posit that the integration of predictive factors derived from news articles and Twitter activity can substantially elevate the precision of stock price forecasts. Building on this, [26-28] delves into the realm of stock price prediction through sentiment analysis of news, aiming to anticipate stock market movements.

This paper introduces a hybrid model that integrates sentiment analysis of financial news with stock price trend data, employing both Fin-BERT and Long Short-Term Memory (LSTM) networks. The structure of this paper is delineated as follows: Section 3 elaborates on the selection of pertinent methodologies, predominantly leveraging deep learning techniques, for dataset forecasting and stock market prediction. Section 4 presents and evaluates the efficacy of the Fin-BERT and LSTM models in concert with news headlines and stock price trends, juxtaposed with the performance of LSTM models that solely utilize stock price trends, as well as Deep Neural Network (DNN) models relying exclusively on stock price trends for their predictions. The work culminates in Section 5 with a summary of findings and suggestions for future research avenues.

3 METHODOLOGY

In this section, our work presents an in-depth account of the experimental procedures and evaluation metrics employed to validate the effectiveness of the proposed techniques and models. Our study leverages a comprehensive dataset of historical stock market news from Benzinga, complemented by US stock codes and pricing data sourced from Yahoo Finance. Utilizing the Keras open-source deep learning framework, which is integrated with a TensorFlow backend [10], we have constructed and refined convolutional neural network models. The execution of all experimental trials was conducted on cloud-based servers equipped with NVIDIA Tesla P100 GPU, which provides a unified platform using the NVIDIA Pascal GPU architecture.

3.1 Dataset

The dataset used in this work is a carefully compiled collection of news information, containing a total of 843062 articles, published from February 15, 2009 to June 12, 2020. It comprehensively covers key information from multiple dimensions, aiming to provide a solid data foundation for in-depth insights into media trends, stock market dynamics, and their interrelationships. This dataset not only includes the title of each news article and the URL link to directly access the article content, ensuring the traceability of information and the possibility of instant access, but also accurately records the publisher information of each news article, which is particularly

important for analyzing the reporting tendencies and influence distribution of different media. In addition, the dataset integrates the exact dates of news releases, spanning from February 15, 2009 to June 12, 2020, with a data coverage range of over eleven years, providing valuable time dimensions for time series analysis, trend evolution research, and tracing the influence of specific events on the stock market. Furthermore, the dataset is embedded with news related stock market information and corresponding stock codes, which is particularly prominent because it serves as a bridge for researchers to explore the immediate correlation and long-term effects between news events and stock market volatility. It has significant value in areas such as quantitative investment strategy development, market sentiment analysis, and risk management.

In addition, to ensure equitable representation in the training and validation datasets, it is imperative to employ a methodical strategy for data allocation. The original dataset was applied with simple hierarchical segmentation, with 85% of the data for each stock used in the training set and 15% used for testing. That is, a total of 716603 news information, stock prices, and other data were allocated to the training set, and 126459 news information, stock prices, and other data were designated to the test set. Also, in the training set data, 85% of the data for each stock was split for actual training, and 15% of the data was used for validation. That is, a total of 609113 news information, stock prices, and other data were assigned to the training set, and 107490 news information, stock prices, and other data were assigned to the validation set. as shown in table 1.

Table 1: Numbers of Daily Financial News and stock prices of dataset

data	Daily Financial News and stock prices
Training	609113
validation	107490
testing	126459
Total no. of data	843062

3.2 *Fin-BERT Embedding LSTM Architecture*

In this segment, we delve into the intricacies of the proposed model. The pipeline of this model first applies Fin-BERT word embedding to news data to generate news sentiment analysis indicators with ratings. At the same time, a LSTM is trained by combining historical stock price data to integrate the two models, enabling them to use all the features extracted from the two models (from "digital+news" data) to predict the closing price and reduce errors.

Fin-BERT is a language model based on BERT [23-24]. Fin-BERT represents an evolution in the domain of natural language processing (NLP), specifically tailored for financial sector applications. Its technological leap stems from the adoption of the Transformer architecture's bidirectional training mechanism for language modeling, marking a departure from prior studies [25] that predominantly focused on unidirectional, left-to-right text sequence analysis or a hybrid approach. Empirical evidence from scholarly work suggests that bidirectional training enables language models to achieve a more nuanced comprehension of linguistic contexts and flows. A key element of this architectural design is the integration of an attention mechanism, which is engineered to identify and understand the contextual interconnections between words and sub-word units within textual data. The BERT model stands out by featuring two principal components: an encoder that analyzes the input text and a decoder that generates task-specific predictions. Given BERT's objective to construct a comprehensive language model, it exclusively employs the encoder phase. This selective use of the encoder facilitates the model's ability to deduce word contexts from their immediate surroundings, thereby enhancing its interpretive

capabilities. In our research, we harnessed the capabilities of the pre-trained Fin-BERT model, and its corresponding word segmentation to configure it for sequence classification tasks. The model will classify the input text into several predefined sentiment categories. They are neutral positive、negative。 We split the news headlines related to stocks into tokens and ensure that all input samples have the same length, with any gaps filled in with a special token "PAD". Employ the Softmax transformation to refine the initial logits produced by the model, thereby transforming them into a probabilistic distribution that aligns with the three distinct sentiment inclination percentages. This means that each sentiment category has a score between 0 and 1, and the sum of probabilities for all categories is 1. Calculate the average value of sentiment probability distribution groups. Then separate the tensor from the current computational graph and stop gradient tracking, which is necessary for inference as we do not need to calculate gradients to update model weights. Then, the calculated average emotional score is classified into the date of the news release, and statistical information including each date and its corresponding overall emotional tendency is obtained.

After obtaining overall sentiment statistics including each date and its corresponding stocks, we integrated emotional indicators (indifferent, favorable, unfavorable) with the closing prices from the preceding 8 trading sessions, transforming them into a format that is apt for input into a time series analytical model. The objective is to predict the closing price of the subsequent trading day, and prepared to input it into the LSTM model. We normalize the closing price data to map its value range to (0,1), and vertically concatenate the normalized closing price data with the previous closing price data. We employ a three-dimensional data structure, characterized by dimensions of (sample count, temporal sequence length, feature set size), to serve as the input for our analysis. The number of features includes 3 emotional features plus 1 feature for closing price data. The emotional data is copied to the first 3 feature positions of each time step, and the normalized closing price data is copied to the last feature position of the corresponding time step. We construct an input dataset that includes emotional features and historical closing price features of multiple time steps for each sample, and predict the closing price target for the next day corresponding to each sample. The architecture of the deep learning model is depicted in Figure 1, illustrating a sequential process where each stratum assimilates the preceding layer's output and subsequently forwards its own output to the subsequent stratum. The model will observe data from the past 8 time points as input to predict the output for the next time point. The first layer LSTM has 50 units and will return the output of the entire sequence, so that the next layer LSTM can continue to process this sequence information. The second and third layers of LSTM also have 50 units, which further extract patterns from the sequence. The second layer also returns the sequence output, while the third layer does not return the sequence. It aggregates information into a vector for final prediction. To conclude the model's construction, we incorporate a fully connected layer, commonly referred to as a Dense layer, designed to produce a single output dimension, thereby facilitating the generation of projected stock prices. The model employs a loss function based on the MSE metric, a prevalent choice for regression tasks that quantifies the discrepancy between the estimated and actual outcomes by averaging the squared deviations. Select the Adam optimizer for gradient descent and train at 100 epochs.

3.3 LSTM Architecture

In this method, unlike the data used in the previous section, we did not use the sentiment analysis score of stocks as input. Instead, we directly used the closing prices of stocks from the past 8 days as input data and reshaped the data into a two-dimensional array, where each row represents a sample containing the closing

prices of 8 days, and each column corresponds to the closing prices of a time step. We normalized the data to a range of 0 to 1. Normalization accelerates the model's convergence in the training phase by mitigating the effects of scale disparities among various features. This process ensures the transformation of the initial time series data into a standardized form (mainly the closing prices of the past 8 days and a future day) into a format suitable for model input.

The configuration of the deep learning model demonstrate a cascade of layers where each one receives the preceding layer's result as its input and, in turn, conveys its own result to the following layer. The model will observe data from the past 8 time points as input to predict the output for the next time point. The first layer of LSTM has 50 neurons, which not only output predictions for the last time step, but also predictions for all time steps. The second layer of LSTM also has 50 neurons, continuing to return sequence outputs for all time steps and further extracting time series features. The third layer LSTM still has 50 neurons, but only outputs the result of the last time step. To culminate the model's architecture, a single-neuron fully connected layer is appended, tasked with yielding the ultimate forecast of the stock's closing value. The model opts for the MSE as its loss function, a metric that quantifies the model's predictive accuracy by averaging the squared deviations from the actual closing prices.

3.4 DNN Architecture

In this section, we also use the closing prices of stocks from the past 8 days as input data. The deep learning model's schematic illustrate a sequential arrangement of layers. Each layer integrates the output from its predecessor and subsequently forwards its own output to the next in line. The model initiates with a normalization layer to standardize the data, which is then processed through three fully connected layers containing 256, 128, and 64 neurons respectively. The constituent layers of the model employ a rectified linear unit (ReLU) activation function, augmented with a leaky parameter set to 0.01, addressing the issue of zero gradients for negative inputs. Concluding the sequence, a single-neuron fully connected layer is employed with a linear activation function, which is apt for regression analysis, and the model's objective is quantified by the MSE, a widely accepted metric for gauging the discrepancy between the predicted outcomes and the actual data points.

4 RESULTS

In this chapter, we conducted a detailed analysis and comparison of the predictive effects of three different models and methods on stock prices. Both the training and testing phases are conducted utilizing identical hardware setups. The training process encompasses a total of 100 epochs through the data. Additionally, a checkpoint mechanism is implemented to capture the optimal model parameters at various stages of training, with a focus on monitoring the model's performance based on the validation loss to ensure the most accurate predictions.

4.1 Performance Indicator

Performance metrics serve as critical tools for assessing the efficacy of machine learning and deep learning models. A variety of evaluative measures are at our disposal to gauge model performance. We have conducted a comparative analysis of the methodologies and models employed, focusing on three key metrics:

Mean Absolute Error: The Mean Absolute Error (MAE) represents the mean of the absolute differences between the actual target values and the model's predictions. The MAE is derived through the following formula:

$$MAE = \frac{\sum_{i=1}^n |y_1 - y_2|}{n} \quad (1)$$

Where y_1 denotes the forecasted outcome for the i^{th} training instance, and y_2 corresponds to the actual value for that instance.

Mean Absolute Percentage Error: The Mean Absolute Percentage Error (MAPE) serves as a measure that captures the average proportional discrepancy between forecasted and actual figures. It is derived by averaging the absolute percentage deviations, offering an insight into the model's accuracy in relation to the scale of the variable under consideration. MAPE can be calculated using the following formula:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_1 - y_2}{y_2} \right| \quad (2)$$

Accuracy: Accuracy is obtained by subtracting MAPE from one.

$$Accuracy = 1 - MAPE \quad (3)$$

4.2 Fin-BERT Embedding LSTM Architecture

Our model achieved smaller loss within 77 epochs; Consequently, the model expedites the attainment of high precision. The validation dataset recorded a minimal loss value of 0.00036. The test dataset, as depicted in Figure 1, exhibited a slightly higher loss of 0.00083. Our model's predictions were characterized by a MAE of 173.67 and a MAPE of 0.045, alongside an impressive accuracy score of 0.955. To vividly illustrate the model's forecasting capabilities, we applied it to anticipate eBay's stock price trajectory for the subsequent 100-day period, as shown in Figure 2.

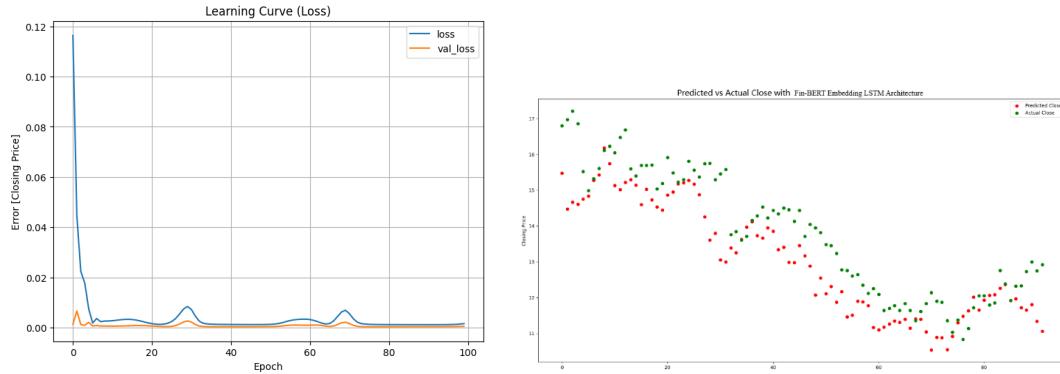


Fig. 1. Loss of Fin-BERT Embedding LSTM Architecture

Fig. 2. Predicted and Actual Close Price Compared under Fin-BERT Embedding LSTM Architecture

4.3 LSTM Architecture

Our model achieved smaller loss within 100 epochs. The validation set obtained a loss of 0.00085. The test dataset, as illustrated in Figure 3, reported a loss value of 0.00092. Our model, upon evaluation, produced a MAE of 183.36 and a MAPE of 0.072, achieving an accuracy level of 0.928. To provide a visual representation

of the model's predictive prowess, we utilized it to forecast eBay's stock price movements for an upcoming 100-day period, as shown in Figure 4.

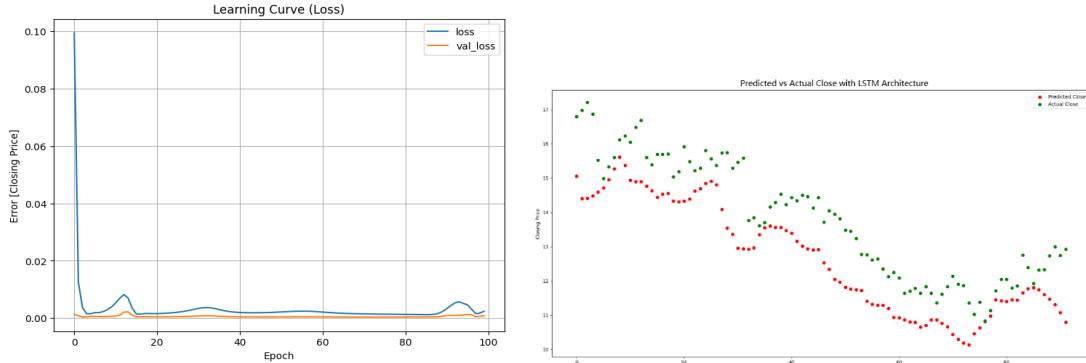


Fig. 3. Loss of LSTM Architecture

Fig. 4. Predicted and Actual Close Price Compared under LSTM Architecture

4.4 DNN Architecture

Our model achieved smaller loss within 100 epochs. The validation set obtained a loss of 0.458. The test set achieved a loss of 21.77, and our proposed model generated a MAE of 489.32 and a MAPE of 0.22, with an accuracy of 0.78. As shown in Figure 5. To visually demonstrate the predictive performance of the model, We deployed the model to project eBay's stock price trends for the forthcoming 100 trading days, as shown in Figure 6.

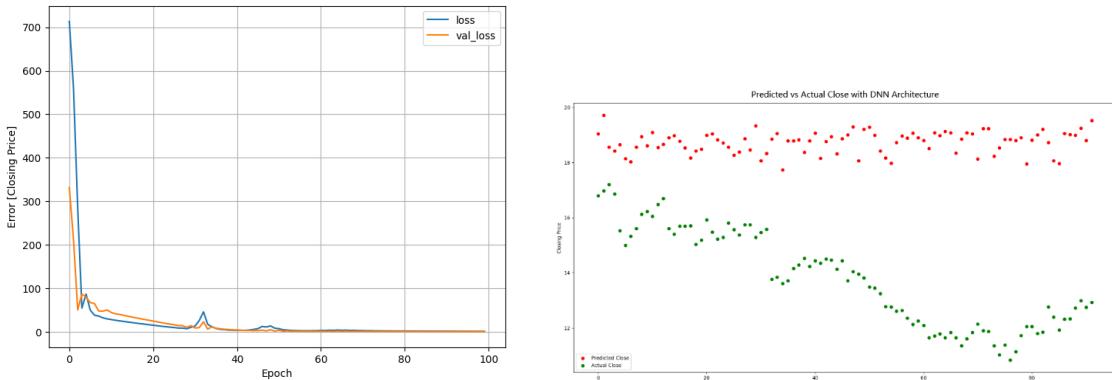


Fig. 5. Loss of DNN Architecture

Fig. 6. Predicted and Actual Close Price Compared under DNN Architecture

In conclusion, Fin-BERT Embedding LSTM Architecture and LSTM Architecture perform relatively well, while DNN Architecture performs the worst. In terms of testing loss, MAE, MAPE, and Accuracy, Fin-BERT Embedded LSTM Architecture performed the best, followed by LSTM Architecture, and DNN Architecture performed the worst. However, it is not difficult to see from the result data that the performance of Fin-BERT Embedding LSTM Architecture and LSTM Architecture is relatively close, while the performance of DNN Architecture is far from satisfactory. As shown in table 2.

Table 2: Comparison of different approach in terms of performance

Approach	Testing loss	MAE	MAPE	Accuracy
Fin-BERT	0.00083	173.67	0.045	0.955
Embedding LSTM				
LSTM	0.00092	183.36	0.072	0.928
DNN	21.77	489.32	0.22	0.78

5 CONCLUSION

Our work proposes three models and methods for predicting financial stock market prices, namely Fin-BERT Embedded LSTM Architecture, LSTM Architecture, and DNN Architecture. The Fin-BERT Embedding LSTM Architecture utilizes stock news content to quantitatively analyze news sentiment. At the same time, combining the past closing price trends of stocks with sequential data, the aim is to comprehensively extract features from the dual perspectives of numerical data and news information, to enhance the precision of forecasting stock closing prices and substantially diminish the predictive error. LSTM Architecture and DNN Architecture, on the other hand, only predict historical stock closing prices. The proposed Fin-BERT Embedding LSTM Architecture method performed the best, with a Testing loss of 0.00083, MAE of 173.67, MAPE of 0.045, and Accuracy of 0.955 at 77 epochs, resulting in better performance with less computation time. LSTM Architecture performed second, while DNN Architecture performed the worst. However, there is not much difference in performance between Fin-BERT Embedding LSTM Architecture and LSTM Architecture. In our future research, our goal will focus on two core directions: firstly, expanding the experimental scope, validating model performance on larger datasets, and deploying it to the current state-of-the-art model framework for in-depth testing. The second is to deepen the emotional analysis of stock market news, and it is expected that this strategy can significantly enhance the stability and reliability of model predictions. In addition, the plan is to include a fake news identification mechanism to distinguish the authenticity of financial reports, which is also expected to improve the robustness of predictions. The research perspective will not only be limited to the stock market, but also cross over to other investment sectors with huge potential, such as precious metals (such as gold), energy (oil), and the real estate industry. The aim is to validate and refine the model's performance across a variety of market conditions.

REFERENCES

- [1] Bosworth, B., Hymans, S., & Modigliani, F. (1975). The Stock Market and the Economy. *Brookings Papers on Economic Activity*, 1975(2), 257–300. <https://doi.org/10.2307/2534104>
- [2] Liu, H., Shen, F., Qin, H., & Gao, F. (2024). Research on Flight Accidents Prediction based Back Propagation Neural Network. *arXiv preprint arXiv:2406.13954*.
- [3] Jin, C., Huang, T., Zhang, Y., Pechenizkiy, M., Liu, S., Liu, S., & Chen, T. (2023). Visual prompting upgrades neural network sparsification: A data-model perspective. *arXiv preprint arXiv:2312.01397*.
- [4] Jin, C., Che, T., Peng, H., Li, Y., & Pavone, M. (2024). Learning from teaching regularization: Generalizable correlations should be easy to imitate. *arXiv preprint arXiv:2402.02769*.
- [5] Xie T, Wan Y, Wang H, Østrøm I, Wang S, He M, Deng R, Wu X, Grazian C, Kit C, Hoex B. Opinion Mining by Convolutional Neural Networks for Maximizing Discoverability of Nanomaterials. *J Chem Inf Model*. 2024 Apr 8;64(7):2746-2759. doi: 10.1021/acs.jcim.3c00746. Epub 2023 Nov 20. PMID: 37982753.
- [6] Hu Z, Liu W, Bian J, Liu X, Liu T-Y, editors. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. Proceedings of the eleventh ACM international conference on web search and data mining; 2018.
- [7] Zhong, Y., Liu, Y., Gao, E., Wei, C., Wang, Z., & Yan, C. (2024). Deep Learning Solutions for Pneumonia Detection: Performance Comparison of Custom and Transfer Learning Models. *medRxiv*. <https://doi.org/10.1101/2024.06.20.24309243>
- [8] Ni, H., Meng, S., Geng, X., Li, P., Li, Z., Chen, X., ... & Zhang, S. (2024). Time Series Modeling for Heart Rate Prediction: From ARIMA

- to Transformers. arXiv preprint arXiv:2406.12199.
- [9] Chong E, Han C, Park FC. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*. 2017 Oct 15; 83:187-205.
- [10] Xu, K., Wu, Y., Li, Z., Zhang, R., & Feng, Z. (2024). Investigating Financial Risk Behavior Prediction Using Deep Learning and Big Data. *International Journal of Innovative Research in Engineering and Management*, 11(3), 77-81.
- [11] Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text sentiment analysis and classification based on bidirectional Gated Recurrent Units (GRUs) model. arXiv preprint arXiv:2404.17123.
- [12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Advances in neural information processing systems*. 2017;30.
- [13] Kalyani J, Bharathi P, Jyothi P. Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958. 2016 Jul 7.
- [14] Thakkar A, Chaudhari K. Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions. *Information Fusion*. 2021; 65:95–107. <https://doi.org/10.1016/j.inffus.2020.08.019> PMID: 32868979
- [15] Lin, Z., Wang, Z., Zhu, Y., Li, Z., & Qin, H. (2024). Text Sentiment Detection and Classification Based on Integrated Learning Algorithm. *Applied Science and Engineering Journal for Advanced Research*, 3(3), 27-33.
- [16] Ni, H., Meng, S., Geng, X., Li, P., Li, Z., Chen, X., ... & Zhang, S. (2024). Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers. arXiv preprint arXiv:2406.12199.
- [17] Liu, R., Xu, X., Shen, Y., Zhu, A., Yu, C., Chen, T., & Zhang, Y. (2024). Enhanced detection classification via clustering svm for various robot collaboration task. arXiv preprint arXiv:2405.03026.
- [18] Shen, Y., Liu, H., Liu, X., Zhou, W., Zhou, C., & Chen, Y. (2024). Localization through particle filter powered neural network estimated monocular camera poses. arXiv preprint arXiv:2404.17685.
- [19] Li X, Cao J, Pan Z. Market impact analysis via deep learned architectures. *Neural Computing and Applications*. 2019; 31(10):5989–6000.
- [20] Ozbayoglu AM, Gudelek MU, Sezer OB. Deep learning for financial applications: A survey. *Applied Soft Computing*. 2020; 93:106384.
- [21] Hu, X., Sun, Z., Nian, Y., Dang, Y., Li, F., Feng, J., ... & Tao, C. (2023). Explainable Graph Neural Network for Alzheimer's Disease And Related Dementias Risk Prediction. arXiv preprint arXiv:2309.06584.
- [22] Ru, J., Yu, H., Liu, H., Liu, J., Zhang, X., & Xu, H. (2022). A bounded near-bottom cruise trajectory planning algorithm for underwater vehicles. *Journal of Marine Science and Engineering*, 11(1), 7.
- [23] Li, X., Chang, J., Li, T., Fan, W., Ma, Y., & Ni, H. (2024). A Vehicle Classification Method Based on Machine Learning.
- [24] Ni, H., Meng, S., Geng, X., Li, P., Li, Z., Chen, X., ... & Zhang, S. (2024). Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers. arXiv preprint arXiv:2406.12199.
- [25] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
- [26] Yu, C., Xu, Y., Cao, J., Zhang, Y., Jin, Y., & Zhu, M. (2024). Credit card fraud detection using advanced transformer model. arXiv preprint arXiv:2406.03733.
- [27] Zheng, Q., Yu, C., Cao, J., Xu, Y., Xing, Q., & Jin, Y. (2024). Advanced Payment Security System: XGBoost, CatBoost and SMOTE Integrated. arXiv preprint arXiv:2406.04658.
- [28] Cao, J., Jiang, Y., Yu, C., Qin, F., & Jiang, Z. (2024). Rough set improved therapy-based metaverse assisting system. arXiv preprint arXiv:2406.04465.

FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs

**Yixuan Liang^{1,2}, Yuncong Liu^{1,2}, Neng Wang¹, Hongyang Yang^{1,2},
Boyu Zhang^{1*}, Christina Dan Wang^{1,3*}**

¹AI4Finance Foundation

²Columbia University

³New York University Shanghai

contact@ai4finance.org

Abstract

Financial sentiment analysis is crucial for understanding the influence of news on stock prices. Recently, large language models (LLMs) have been widely adopted for this purpose due to their advanced text analysis capabilities. However, these models often only consider the news content itself, ignoring its dissemination, which hampers accurate prediction of short-term stock movements. Additionally, current methods often lack sufficient contextual data and explicit instructions in their prompts, limiting LLMs' ability to interpret news. In this paper, we propose a data-driven approach that enhances LLM-powered sentiment-based stock movement predictions by incorporating news dissemination breadth, contextual data, and explicit instructions. We cluster recent company-related news to assess its reach and influence, enriching prompts with more specific data and precise instructions. This data is used to construct an instruction tuning dataset to fine-tune an LLM for predicting short-term stock price movements. Our experimental results show that our approach improves prediction accuracy by 8% compared to existing methods.

Introduction

Financial markets are highly reactive to news, social media, and other public sentiment signals; these affect trading behaviors and, ultimately, stock prices. Understanding these sentiment shifts can provide valuable insights into price movement patterns, making sentiment analysis an essential component of modern financial forecasting.

Traditional sentiment analysis typically categorizes sentiment as positive, negative, or neutral. Advances in natural language processing (NLP) have significantly enhanced our ability to analyze and interpret sentiment data from vast text sources. Most prior research has focused on improving the accuracy of sentiment analysis for individual news items, rarely aggregating them to assess the overall market sentiment or integrating the results into downstream tasks such as stock prediction and risk management.

The emergence of Large Language Models (LLMs) has revolutionized financial sentiment analysis by providing not only sentiment-based classification but also explanations for stock movement predictions (Zhang et al. 2023b,a; Araci 2019; Wu et al. 2023). Recent works like FinRobot (Yang

et al. 2024; Zhou et al. 2024; Han et al. 2024) demonstrate this capability through their “Market Forecaster” tool, which moves beyond single-news analysis to capture more breadth sentiment landscapes. Additionally, we are witnessing a growing body of research that extends beyond individual news analysis (Wang, Izumi, and Sakaji 2024), aiming to offer a more holistic view of stock market dynamics.

Despite the advancements brought by LLMs in financial sentiment analysis, existing methods often rely solely on the news content itself for predictions. This approach neglects the crucial factor of news dissemination, which significantly affects market reactions and stock price movements. Additionally, these methods often lack sufficient contextual data and explicit instructions, limiting the LLMs' ability to interpret news. Our proposed approach addresses these limitations by incorporating the breadth of news dissemination, detailed contextual data, and precise instructions, thereby enhancing the accuracy of short-term stock price movement predictions.

In this paper, we propose a novel approach to enhance LLM-powered sentiment-based stock movement predictions by incorporating news dissemination breadth, contextual data, and precise instructions. Our methodology clusters recent company-related news articles, using cluster attributes to evaluate the news's reach and influence. We operate under two key assumptions: *i*) The centroid article of each cluster encapsulates the most comprehensive information for LLM processing; *ii*) The cluster size indicates the topic's market impact, with larger clusters signifying more significant events. Additionally, we enhance the prompts with daily stock price and return data, along with instructions to consider the short-term or long-term impact of the news. Utilizing this information, we construct an instruction tuning dataset to fine-tune an LLM for short-term stock price predictions.

We summarize by our key contributions:

1. We propose a data-driven clustering-based method to capture the breadth of news dissemination, and incorporate it into the training dataset.
2. By enriching prompts with contextual data and instructions tailored to our proposed data format, we offer a more nuanced approach to financial sentiment analysis.
3. Our experimental results demonstrate that our approach

*Corresponding author

improves prediction accuracy by 8% compared to existing methods, offering a more robust and efficient framework for understanding the impact of news on stock prices.

The remainder of the paper is structured as follows: Section 2 reviews related work in sentiment-based financial prediction. Section 3 presents the problem statement. Section 4 describes our data-centric methodology, including high-granularity and news clustering methods. Section 5 discusses performance metrics, evaluation results, and a case study on Boeing. Section 6 concludes with findings and future directions.

Related Works

NLP for Financial Sentiment Analysis

Sentiment analysis has long been a key application of natural language processing (NLP), especially in finance, where it provides valuable insights into market trends and investor sentiment (Chan and Chong 2017; Atkins, Niranjan, and Gerding 2018). Various models and methodologies have been developed to enhance the accuracy and efficiency of sentiment analysis (Tai and Kao 2013; Hamilton et al. 2016; Day and Lee 2016; Sohangir et al. 2018; Mishev et al. 2020; Rizinski et al. 2024), ranging from lexicon-based techniques to machine learning and deep learning approaches.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding complex natural language, with promising applications in financial sentiment analysis (e.g. FinBERT (Araci 2019), Bloomberggpt (Wu et al. 2023), Fingpt (Wang, Yang, and Wang 2023; Yang, Liu, and Wang 2023)).

While these financial language models excel at individual news sentiment analysis, the systematic integration of related news for stock price prediction remains largely unexplored.

LLMs for Sentiment-based Stock Price Prediction

Leveraging text or news for stock prediction is not new; prior work has used tweets and historical data to forecast prices (Xu and Cohen 2018). Now, with the advent of LLMs, we can achieve more nuanced understanding and interpretation of financial text, allowing these models to capture complex relationships within news data and better inform stock price predictions.

Recent works demonstrate diverse applications of LLMs in stock prediction. LLMFactor (Wang, Izumi, and Sakaji 2024) targets short-term prediction through Sequential Knowledge-Guided Prompting, providing real-time interpretable insights. Similarly, (Elahi and Taghvaei 2024) addresses longer-term predictions by combining financial data and news through retrieval-augmented techniques for 3-6 month horizons.

Our work builds on this line of work, but are independent of the model itself: we focus exclusively on the data preparation process, incorporating the impact of news dissemination on stock price movements and providing LLMs with more precise instructions. In this paper, we follow the

standard framework of instruction tuning LLMs for financial forecasting and used the data organization as the baseline (outlined in (Yang et al. 2024)) to show the significant improvement (Section 4), but we believe that our methodology has a much broader application, having the potential to be applied to all the existing models.

Problem Setting and Overall Framework

Our objective is to predict weekly stock price movements based on news sentiment. The movements are categorized into twelve labels: U1-U5 and U5+ for upward trends (0-1%, 1-2%, 2-3%, 3-4%, 4-5%, over 5%), and D1-D5+ for corresponding downward trends. Predictions are based on previous week’s stock prices, recent news, and company fundamentals (updated quarterly and included three weeks after the quarterly report release). The model also generates reasoning for the prediction by identifying [Positive Developments] and [Potential Concerns]—highlighting the 2-4 most significant factors in each category—as well as providing [Prediction & Analysis].

Our overall framework is illustrated in Figure 1, following a standard framework of fine-tuning LLMs for financial analysis. Our work focuses on the Data Processing part and Prompt Engineering part in this flow.

Methodology

This section outlines our data-driven methodologies. Specifically, we 1) increase stock price granularity and implemented news clustering in the Data Processing part, and 2) incorporate contextual and more targeted instructions in Prompt Engineering part.

Data Processing

High Granularity in Stock Price Information (HG): The baseline method uses only weekly aggregate stock price movements (e.g. 3% weekly gain). To enhance prediction performance, we increase data granularity by incorporating daily closing prices and corresponding returns throughout each week. This granular approach serves two key purposes: 1) it reduces the uncertainty of the calculation within the LLMs by providing explicit daily price movements and 2) enables precise temporal alignment between price changes and news events, providing a basis for the differentiation of short- long-term impacts. We will refer to this method as **HG** in the following part.

News Clustering (HG-NC): Traditional stock market analysis often lacks systematic quantification of news dissemination. A more comprehensive approach requires analyzing the complete news landscape—often exceeding 200 articles weekly for active stocks. It presents significant challenges: redundant information processing, computational inefficiency, and potential token limitations in language models.

To address these challenges, built upon our HG method, we further developed a clustering approach that efficiently organizes high-volume news content while capturing news impact through two key dimensions: reporting frequency

and temporal span. We will refer to this method as **HG-NC** in the following part. Our approach comprises the following steps:

- Data Collection:** We retrieve weekly financial news data, including titles and summaries, from the Finnhub API.
- Topic Clustering:** News articles are transformed into embedding representations using Sentence Transformers, followed by BERTopic-based topic modeling to identify and cluster thematically related news content.
- Cluster Quality Assessment:** We evaluate cluster cohesiveness through pairwise similarity analysis:
 - High-Cohesion Clusters** (average pairwise similarity > 0.6): For these clusters, we select the centroid-proximate article as the cluster representative and preserve the metadata including cluster size and temporal span.
 - Low-Cohesion Clusters** (average pairwise similarity ≤ 0.6): we again select the article closest to the centroid but limit the topic size to 2 and record the time range. This setting reflects lower clustering quality and lower confidence in these less cohesive groups.
- Topic Selection Strategy:** When high-cohesion clusters fall below six, we supplement with low-cohesion clusters (at most 4) to ensure sufficient information coverage. All parameters—similarity thresholds, cluster sizes, and topic quotas—are adjustable based on analysis needs and LLM constraints.

The clustering approach, leveraging BERTopic (Grootendorst 2022) and cosine similarity evaluation, efficiently condenses large volumes of news into representative samples while quantifying news dissemination, thereby enhancing stock movement prediction.

Prompt Engineering: Context-enhanced Instructions

To adapt to our proposed data format where we incorporate daily stock information and quantified news dissemination, context-enhanced instructions are needed for better analysis.

For HG: We instruct LLMs to differentiate between short-term and long-term impacts of news, as daily stock prices and returns can reveal immediate market reactions. This distinction is crucial because the influence of short-term news is often already reflected in stock movements within the same week. For a prompt template, see Figure 4.

For HG-NC: Built upon HG, we construct the news component using selected representative articles and their associated metadata (topic size and temporal coverage). Then, we enhance the instructions by describing the news component and providing guidelines for analyzing the impact of news dissemination on stock movement. For a prompt template, see Figure 5.

Instruction-tuning

The training dataset pairs our structured input prompts (including company introduction, historical stock prices, related news, company fundamentals, and instructions for

utilizing sentiment analysis for prediction) with GPT-4o-generated analysis based on known future movements. Then, removing the ground truth stock price in the prompt, we use this dataset to fine-tune Llama3-8B for weekly stock movement prediction and evaluate both numerical accuracy and reasoning quality (see Appendix A for details.)

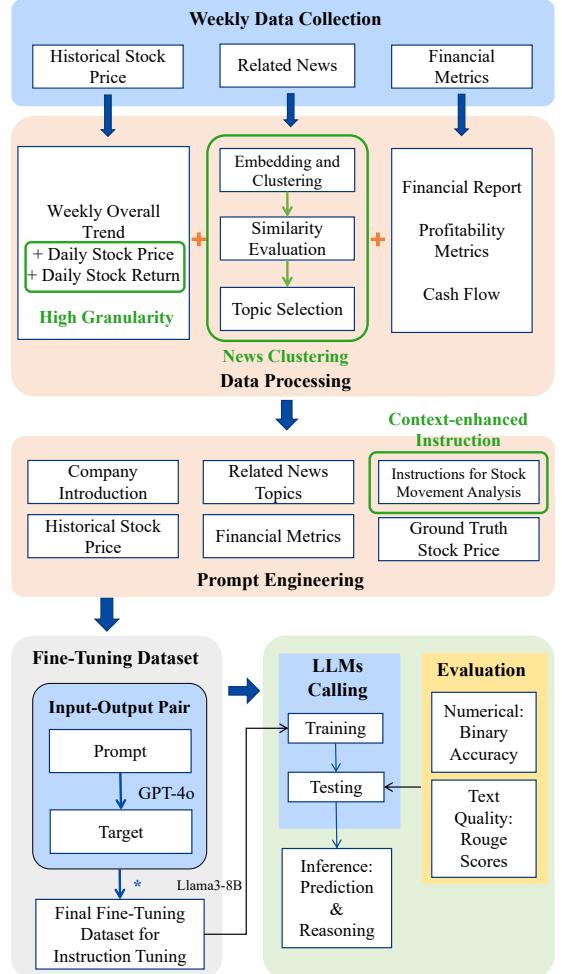


Figure 1: Overall framework. In Data Processing part, we increase granularity in the historical stock price data and employ news data clustering. In Prompt Engineering part, we incorporate context-enhanced instructions for stock movement analysis. The operation we mark with “*” is removing future movement label.

Performance Evaluation

We evaluate our models' performance using two key metrics: binary accuracy for stock prediction and ROUGE scores for reasoning quality. Our analysis compares three progressive methods: the **baseline**, the **HG** method with increased stock price granularity, and the **HG-NC** method.

Binary Accuracy in Stock Movement Prediction

We assess each model’s ability to predict directional stock price movements (up/down) using binary accuracy metrics. Our comprehensive dataset comprises 380 observations across 20 companies, spanning multiple market sectors to ensure robust evaluation. The results demonstrate a consistent improvement pattern across our model iterations:

Method	Avg Acc	Long term	Short term
Baseline	55.0%	15.0%	7.5%
HG	59.2%	69.8%	56.6%
HG-NC	63.0%	58.5%	50.9%

Table 1: Our methods show a significant increase in the binary accuracy. Also, the increase in “long-term” word frequency and “short-term” word frequency provides evidence for attention to temporal aspects, which may account for the accuracy improvement.

With high granularity stock price and targeted instructions, accuracy improves from 0.550 to 0.592. A detailed analysis reveals increased attention to temporal aspects in the **Prediction & Analysis** component, with the frequency of “long-term” rising from 15.0% to 69.8% and “short-term” from 7.5% to 56.6%. This indicates that the LLM effectively follows the instructions to differentiate between short-term and long-term news impacts, balancing them in stock price predictions, leading to improved accuracy.

The further increase to 63% with NC validates our hypothesis that incorporating news clustering results enhances the LLM’s capacity to capture market dynamics and the impact of news dissemination on stock movements.

ROUGE Scores for Reasoning Quality Evaluation

Given the large size of our training and test datasets, obtaining ground-truth sentiment-based analysis for every instance is impractical. Therefore, we rely on automated evaluation metrics to assess the quality of model outputs. We use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores (Lin 2004) to assess the reasoning quality as they measure the overlap of key words and phrases between LLM-generated outputs and reference summaries. Higher ROUGE scores indicate closer alignment with the reference text and broader coverage of news content as the basis for reasoning. However, precise and comprehensive evaluation still requires human judgment.

We evaluate three metrics: ROUGE-1, ROUGE-2, and ROUGE-L (longest common subsequences), with ROUGE-N representing N-gram co-occurrence statistics. Across all metrics, the HG-NC method consistently outperforms both the baseline and HG approaches.

Specifically, we analyze ROUGE scores for the **Prediction & Analysis** component in the output, which integrates positive and negative factors to justify the model’s directional predictions. This critical component highlights the LLM’s ability to weigh competing factors from sentiment analysis and articulate its decision-making process. As shown in Table 2, the HG-NC method better captures and

Method	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	0.450	0.121	0.224
HG	0.469	0.131	0.224
HG-NC	0.472	0.140	0.234

Table 2: ROUGE scores for different methods

articulates the complex interactions between various market factors in its analysis.

Case Study: Boeing Company

We use The Boeing Company (NYSE: BA) as a representative case study and compare the prediction performance of the HG method and HG-NC method. Overall, the HG-NC method has accuracy (63.2%) compared to the HG method (52.63%). For prediction results see Figure 6 in Appendix.

We examine the ratio of news articles in high-coherence clusters (those with average pairwise similarity > 0.6) to the total number of news articles as an indicator of clustering performance. In general, we observe a strong correlation between clustering performance and prediction performance. Specifically, as shown in Figure 2, in 7 instances where our HG-NC method outperforms HG (Case 1), we observe relatively high ratios of high-coherence clusters (mostly exceeding 50%). Conversely, performance declines when this ratio falls below 40% (Case 2), suggesting insufficient capture or preservation of significant market information.

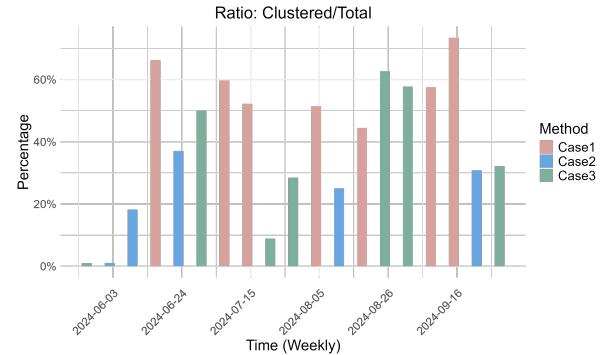


Figure 2: The ratio of news articles in high-coherence clusters to the total number of news articles. We classify prediction comparisons into three cases: Case 1: HG-NC correct vs. NC incorrect; Case 2: HG correct vs. HG-NC incorrect; Case 3: Both methods yield identical predictions.

Conclusion

In this paper, we proposed an approach to enhance LLM-powered sentiment-based stock movement prediction. By increasing granularity for stock price and providing instructions for short-term or long-term analysis, we enhance the contextual understanding of news. Further, we evaluate news dissemination through clustering and incorporate its market impact to improve predictions. We developed an instruction tuning dataset to fine-tune LLMs for more accurate

short-term stock movement predictions. Experimental results validate our approach, achieving 63% binary accuracy compared to the 55% baseline, with better predictions at high clustering ratios ($> 50\%$). These findings highlight the importance of enriched contextual data and dissemination-aware methods in improving prediction accuracy.

References

- Araci, D. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.
- Atkins, A.; Niranjan, M.; and Gerding, E. 2018. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2): 120–137.
- Chan, S. W.; and Chong, M. W. 2017. Sentiment analysis in financial texts. *Decision Support Systems*, 94: 53–64.
- Day, M.-Y.; and Lee, C.-C. 2016. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1127–1134. IEEE.
- Elahi, A.; and Taghvaei, F. 2024. elahi2024combiningCombining Financial Data and News Articles for Stock Price Movement Prediction Using Large Language Models. *arXiv preprint arXiv:2411.01368*.
- Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hamilton, W. L.; Clark, K.; Leskovec, J.; and Jurafsky, D. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing, conference on empirical methods in natural language processing*, volume 2016, 595. NIH Public Access.
- Han, X.; Wang, N.; Che, S.; Yang, H.; Zhang, K.; and Xu, S. X. 2024. Enhancing Investment Analysis: Optimizing AI-Agent Collaboration in Financial Research. In *ICAIF 2024: Proceedings of the 5th ACM International Conference on AI in Finance*, 538–546.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L. T.; and Trajanov, D. 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8: 131662–131682.
- Rizinski, M.; Peshov, H.; Mishev, K.; Jovanovik, M.; and Trajanov, D. 2024. Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). *IEEE Access*.
- Sohangir, S.; Wang, D.; Pomeranets, A.; and Khoshgoftaar, T. M. 2018. Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, 5(1): 1–25.
- Tai, Y.-J.; and Kao, H.-Y. 2013. Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of international conference on information integration and web-based applications & services*, 53–62.
- Wang, M.; Izumi, K.; and Sakaji, H. 2024. LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction. *arXiv preprint arXiv:2406.10811*.
- Wang, N.; Yang, H.; and Wang, C. D. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.
- Wu, S.; Irsoy, O.; Lu, S.; Dabrowski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Xu, Y.; and Cohen, S. B. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1970–1979.
- Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.
- Yang, H.; Zhang, B.; Wang, N.; Guo, C.; Zhang, X.; Lin, L.; Wang, J.; Zhou, T.; Guan, M.; Zhang, R.; et al. 2024. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv preprint arXiv:2405.14767*.
- Zhang, B.; Yang, H.; Zhou, T.; Ali Babar, M.; and Liu, X.-Y. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, 349–356.
- Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.
- Zhou, T.; Wang, P.; Wu, Y.; and Yang, H. 2024. FinRobot: AI Agent for Equity Research and Valuation with Large Language Models. In *ICAIF 2024: The 1st Workshop on Large Language Models and Generative AI for Finance*.

Appendix A: Model Training

The training parameters are given in Table. For our model, we initialize it with LLAMA-3-8B model and perform instruction tuning over 5 epochs. The training process utilizes the AdamW optimizer, with a batch size of 32, an initial learning rate of 1×10^{-5} , and a weight decay of 0.01. In order to be able to input the normal amount of news in a week, we set a maximum input text length of 8000 tokens. We utilize DeepSpeed for the fine-tuning process on A100 (40GB) GPU, resulting in a total training time of 162 minutes.

Parameter	Value
Learning rate	1×10^{-5}
Weight Decay	0.01
Batch size	32
Training epochs	5
Num warmup Steps	0
Max Token Length	8000
GPUs	A100 (40GB)

Table 3: Training parameters.

Appendix B: Prompt Templates

Prompt Template 1: Baseline
[Company Introduction] Apple Inc is a leading entity in the Technology sector... Apple Inc operates primarily in the US, trading under the ticker AAPL ...
[Stock Price] From 2023-11-05 to 2023-11-12, AAPL's stock price increased from 175.75 to 185.70.
[News] News during this period are listed below: [Headline]: Apple Vision Pro Spatial Video arrives for iPhone in iOS 17.2 beta 2 [Summary]: At a glance, there don't appear to be many new features in the second beta of iOS 17.2 — with the exception of Spatial Video recording for Apple Vision Pro. [Headline]: ... [Summary]: ...
[Basic Financials] No basic financial reported.
[Instruction] Based on all the information before 2023-11-12, let's first analyze the positive developments and potential concerns for AAPL. Come up with 2-4 most important factors for each category and keep them concise. Most factors should be inferred from company related news. Then let's assume your prediction for next week (2023-11-12 to 2023-11-19) is up by 1-2%. Provide a summary analysis to support your prediction. The prediction result need to be inferred from your analysis at the end, and thus not appearing as a foundational factor of your analysis.

Figure 3: Baseline Prompt Template. This template contains company introduction, stock price weekly trend, news headlines and summaries, basic financials, and analytical instruction for prediction.

Prompt Template 2: High Granularity
[Company Introduction] Apple Inc is a leading entity in the Technology sector... Apple Inc operates primarily in the US, trading under the ticker AAPL ...
[Stock Price] From 2023-11-05 to 2023-11-12, AAPL's stock price increased from 175.75 to 185.70. The daily stock price during this period is [178.31, 180.89, 181.95, 181.48, 185.69], and the corresponding daily return is given by [0.0146, 0.0144, 0.0058, -0.0026, 0.0232].
[News] News during this period are listed below: [Headline]: Apple Vision Pro Spatial Video arrives for iPhone in iOS 17.2 beta 2 [Summary]: At a glance, there don't appear to be many new features in the second beta of iOS 17.2 — with the exception of Spatial Video recording for Apple Vision Pro. [Headline]: ... [Summary]: ...
[Basic Financials] No basic financial reported.
[Instruction] Based on all the information before 2023-11-12, let's first analyze the positive developments and potential concerns for AAPL. Note that some of the factors only influence the price for the following 1 or 2 days while some others may have longer-term effects. Come up with 2-4 most important positive factors and 2-4 most significant negative effects and keep them concise. Most factors should be inferred from company related news. Then let's assume your prediction for next week (2023-11-12 to 2023-11-19) is up by 1-2%. Provide a summary analysis to support your prediction. The prediction result need to be inferred from your analysis at the end, and thus not appearing as a foundational factor of your analysis.

Figure 4: HG Prompt Template. This template includes daily stock prices with corresponding returns and context-enhanced analytical instructions emphasizing temporal effects.

Prompt Template 3: HG-NC
[Company Introduction] Apple Inc is a leading entity in the Technology sector... Apple Inc operates primarily in the US, trading under the ticker AAPL ...
[Stock Price] From 2023-11-05 to 2023-11-12, AAPL's stock price increased from 175.75 to 185.70. The daily stock price during this period is [178.31, 180.89, 181.95, 181.48, 185.69], and the corresponding daily return is given by [0.0146, 0.0144, 0.0058, -0.0026, 0.0232].
[News] News Topics during this period are listed below: [Topic]: Apple product upgrade. [Duration]: 2023-11-07 to 2023-11-10 [News Headline]: Apple Vision Pro Spatial Video arrives for iPhone in iOS 17.2 beta 2 [News Summary]: At a glance, there don't appear to be many new features in the second beta of iOS 17.2 — with the exception of Spatial Video recording for Apple Vision Pro. [Topic Size]: 12
[Basic Financials] No basic financial reported.
[Instruction] Based on all the information before 2023-11-12 and the following guidelines, let's analyze the positive developments and potential concerns for AAPL. Apart from stock price information and basic financials, several news topics are given and they are derived from news clustering. Each topic contains its duration, a key news headline and summary, and a topic size, indicating the number of related news items within that cluster. First, note that some of the factors only influence the price for the following 1 or 2 days while some others may have longer-term effects. Second, the topics that are closer to 2023-11-12 are likely to have a stronger influence on the upcoming stock price forecast. Third, take into account the topic size: larger clusters represent greater market attention and, consequently, likely exert more influence on the stock price. Following these instructions, please come up with 2-4 most important positive factors and 2-4 most significant negative effects and keep them concise. Most factors should be inferred from company related news. Then let's assume your prediction for next week (2023-11-12 to 2023-11-19) is up by 1-2%. Provide a summary analysis to support your prediction. The prediction result need to be inferred from your analysis at the end, and thus not appearing as a foundational factor of your analysis.

Figure 5: HG-NC Prompt Template. This template enhances the HG format by incorporating clustered news topics with topic size and temporal span, along with guidelines for analyzing news dissemination on price predictions.

Appendix C: Experimental results for Boeing Company

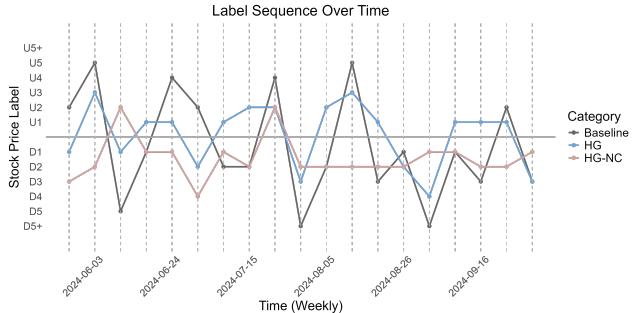


Figure 6: BA prediction labels during test period.

Prediction Result Example: BA, 2024/6/16 to 2024/6/23	
[Positive Developments]:	
1.	Arajet, a carrier expanding international routes, has taken delivery of its 10th Boeing 737 MAX, indicating ongoing demand and successful sales of Boeing's fuel-efficient aircraft.
2.	Despite concerns, there is a sustained strong demand for new planes as travel has surged post-pandemic, which supports Boeing's longer-term sales prospects.
[Potential Concerns]:	
1.	Boeing is facing an investigation into quality issues with undelivered 787 Dreamliners, related to incorrect installation of fasteners, which can hurt the company's reputation and future deliveries.
2.	The FAA is investigating potential issues with the authenticity of titanium documentation in newly manufactured Boeing jets. This adds regulatory scrutiny and could delay production approvals.
3.	Boeing and the FAA are subjected to a congressional probe initiated by Senator Chuck Grassley, further intensifying scrutiny and potentially impacting operational efficiency and decision-making.
[Prediction & Analysis]:	
Prediction:	Down by 0-1%
Analysis:	In light of recent developments, Boeing is confronting multiple headwinds that may overshadow its positive sales developments. The recent investigations, both by the FAA and a congressional probe, could lead to regulatory delays or production setbacks, which could dampen investor confidence in the short term. Furthermore, the concerns regarding quality control in the 787 Dreamliner line emphasize operational challenges that Boeing needs to address swiftly to prevent damage to its reputation and future sales. Despite strong demand for aircraft, these issues are likely to pose a significant risk to Boeing's immediate stock performance. Given the news topics predominantly reflecting negative sentiment recently and the fact that scrutiny often results in market skepticism, I predict a slight decline of 0-1% in Boeing's stock price in the upcoming week (2024-06-16 to 2024-06-23).

Figure 7: The prediction outcome of HG-NC method: Boeing Company, 2024/6/16-2024/6/23.

Start Date	News Count	Clusters	Good Clusters	Clustered News	Ration
5/26	74	2	0	0	0
6/2	76	4	1	3	0.04
6/9	77	5	1	14	0.18
6/16	118	9	5	78	0.66
6/23	135	6	3	50	0.37
6/30	110	6	2	55	0.50
7/7	87	7	4	52	0.60
7/14	69	7	3	36	0.52
7/21	102	2	1	9	0.09
7/28	109	4	2	31	0.28
8/4	70	4	2	36	0.51
8/11	68	4	1	17	0.25
8/18	63	2	1	28	0.44
8/25	67	9	5	42	0.63
9/1	78	3	2	45	0.58
9/8	139	8	5	80	0.58
9/15	124	9	9	91	0.73
9/22	120	7	3	37	0.31
9/29	84	2	1	27	0.32

Table 4: Weekly news clustering statistics (2024/5/26 - 2024/9/29) showing total news count, number of clusters, good (high-cohesion) clusters, clustered news (news in high-cohesion clusters), and the ratio of clustered to total news.

RETuning: Upgrading Inference-Time Scaling for Stock Movement Prediction with Large Language Models

Xueyuan Lin^{1,2,3*} Cehao Yang^{1,2*} Ye Ma³ Ming Li³ Rongjunchen Zhang³
 Yang Ni¹ Xiaojun Wu^{1,2} Chengjin Xu^{2,4} Jian Guo^{2†} Hui Xiong^{1†}

¹ The Hong Kong University of Science and Technology (Guangzhou)

² IDEA Research ³ Hithink RoyalFlush Information Network Co., Ltd ⁴ DataArc Tech Ltd
 {xlin058, cyang289, yni002}@connect.hkust-gz.edu.cn, xionghui@ustc.edu.cn
 {maye, zhangrongjunchen}@myhexin.com, lm92@mail.ustc.edu.cn
 {xuchengjin, wuxiaojun, guojian}@idea.edu.cn

Abstract

Recently, large language models (LLMs) have demonstrated outstanding reasoning capabilities and inference-time scaling on mathematical and coding tasks. However, their application to financial tasks—especially the most fundamental task of stock movement prediction—remains underexplored. We study a three-class classification problem (up, hold, down) and, by analyzing existing reasoning responses, observe that: (1) LLMs are easily swayed by contextual viewpoints, tending to follow analysts’ opinions rather than exhibit a systematic, independent analytical logic in their chain-of-thoughts (CoTs). (2) LLMs often list summaries from different sources without weighing adversarial evidence, yet such counterevidence is crucial for reliable prediction. It shows that the model does not make good use of its reasoning ability to complete the task. To address this, we propose **Reflective Evidence tuning (RETuning)**, a cold-start method prior to reinforcement learning, to enhance prediction ability. While generating CoT, **RETuning** encourages dynamically constructing an analytical framework from diverse information sources, organizing and scoring evidence for price up or down based on that framework—rather than on contextual viewpoints—and finally reflecting to derive the prediction. This approach maximally aligns the model with its learned analytical framework, ensuring independent logical reasoning and reducing undue influence from context. We also build a large-scale dataset spanning all of 2024 for 5,123 A-share stocks, with long contexts (32K tokens) and over 200K samples. In addition to price and news, it incorporates analysts’ opinions, quantitative reports, fundamental data, macroeconomic indicators, and similar stocks. Experiments on this new dataset show that, as a cold-start method, **RETuning** successfully unlocks the model’s reasoning ability in the financial domain. During reinforcement learning, response length steadily increases under the designed curriculum setting. Furthermore, inference-time scaling still works even after 6 months or on out-of-distribution stocks, since the models gain valuable insights about stock movement prediction.

*Equal contribution.

†Corresponding Authors.

1 Introduction

Stock Movement Prediction (SMP) is one of the most fundamental and consequential tasks in finance. It not only directly affects the interests of individual investors but also plays a central role in algorithmic trading (Mahfooz et al., 2022; Ta et al., 2018), financial risk control (Adyatma & Alamsyah, 2022; Vui et al., 2013), and intelligent research platforms (Shi et al., 2020). In recent years, Large Language Models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Bai et al., 2023; Touvron et al., 2023) have demonstrated remarkable reasoning capabilities in domains such as code generation and mathematical problem-solving (DeepSeek-AI et al., 2025; OpenAI et al., 2024; Uesato et al., 2022). This has sparked growing interest in exploring whether such models can also excel in financial tasks. However, it remains an open question whether LLMs’ strength in *reasoning* and *inference-time scaling* can be effectively harnessed for stock price prediction.

In the traditional planning phase of a financial agent, the model has access to a wide range of information sources—news articles, analyst opinions, research reports, quantitative factor analyzes, and more. Despite this richness, making reliable and interpretable predictions remains a major challenge. On one hand, LLMs often exhibit strong prior biases due to the optimistic slant of their training data, which is skewed toward long positions and excludes contrarian views for political or regulatory reasons. On the other hand, these models tend to lack the ability to construct independent reasoning frameworks, reconcile conflicting information, and perform reflective analysis—capabilities that are essential for robust financial decision-making.

To address these challenges, we propose a novel modeling paradigm that treats stock movement prediction as a *generative reasoning task*. By processing all textual information sources end-to-end, this approach aims to simulate the thought process of a human trader, ultimately generating structured and interpretable predictions. Two key innovations underpin this paradigm.

First, we introduce **Reflective Evidence Tuning (RETuning)**, which instills LLMs to dynamically construct reasoning frameworks based on diverse information sources, collect and evaluate evidence for potential price directions (up, down, or hold), and reflect on the evidence before making a final prediction. This structured approach is a cold-start training mechanism prior to reinforcement learning (RL). It enables models to avoid merely summarizing or echoing external viewpoints and instead follow an internally consistent logic, improving both interpretability and accuracy.

Second, we explore the role of **inference-time scalability**, a technique that has shown promise in mathematical and programming tasks (Muennighoff et al., 2025; Li et al., 2025). Specifically, we investigate whether *majority voting* can significantly improve predictive accuracy in financial domain. Although widely successful elsewhere, its efficacy in stock movement prediction has not yet been systematically examined.

To support this research, we construct a large-scale, high-quality dataset that reflects the complexity and information density of real-world financial environments. Covering the full year of 2024 across over 4,000 A-share stocks, this dataset integrates six heterogeneous information sources: news, fundamentals, analyst opinions, quantitative factor reports, macroeconomic context, and stocks of similar trends. With over 200,000 samples and an average input length of up to 32K tokens, it overcomes the limitations of prior datasets that were outdated and lacked information diversity (Xu & Cohen, 2018; Luo et al., 2023; Zhou et al., 2021). The details of the dataset construction are discussed in Appendix C.

Empirical results show that **RETuning** effectively enhances reasoning structure and improves predictive performance over strong baselines. It also generalizes beyond stock movement prediction, yielding significant improvements in other financial tasks, and demonstrates strong performance under inference-time scaling and out-of-distribution settings.

Our contribution can be summarized as follows: (1) We build a large-scale, long-context financial dataset with diverse evidence sources beyond price and news, which fill the gap that existing datasets are outdated and lack information diversity. (2) We introduce **RETuning**, synthesizing cold-start responses that guide LLMs to construct and reflect on an analytical framework for stock movement prediction. It allows significant inference-time scalability of LLMs in the prediction task. (3) We empirically show that **RETuning** unlocks prediction ability and generalizes beyond stock movement prediction. We believe this research lays the groundwork for deploying trustworthy, reasoning-driven LLMs in real-world financial applications.

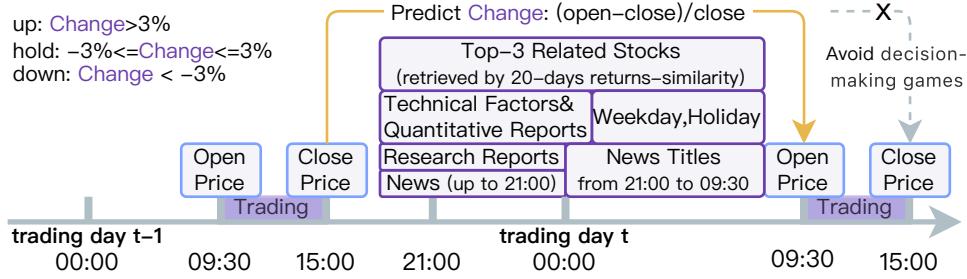


Figure 1: Overview of data sample for one stock at trading day t .

2 Related Work

Stock Movement Prediction with LLMs. Recent work has focused on exploring various types of information sources for stock movement prediction with LLMs. Several studies emphasize the importance of stock-related news in revealing fundamental market insights (Vargas et al., 2018; Li et al., 2021). Meanwhile, other research highlights the significance of understanding the relationships between companies and industries (Feng et al., 2019; Hsu et al., 2021). Recent studies have also provided empirical evidence of the impact of public sentiment on market trends, with researchers working to extract sentiment and keyphrases from news and social media data (Nguyen et al., 2015; Hao et al., 2021). However, these works often focus on a single type of information source, such as news or sentiment, and do not fully leverage the potential of LLMs to integrate multiple heterogeneous sources. In contrast, our work aims to construct a comprehensive dataset that incorporates diverse information sources, including news, fundamentals, analyst opinions, quantitative reports, macroeconomic indicators, and similar stocks, to enhance the predictive capabilities of LLMs in stock movement prediction.

Inference-Time Scaling for LLMs. Inference-time scaling methods (Snell et al., 2024) enhance LLM performance by leveraging additional computation during generation, broadly categorized into three strategies. **Repeated sampling** improves diversity and accuracy via parallel candidate generation, utilizing verification strategies like majority voting (Li et al., 2024a; Lin et al., 2024; Wang et al., 2023; Toh et al., 2024) or best-of-N (BoN) selection with verifiers (Stiennon et al., 2020; Cobbe et al., 2021; Nakano et al., 2022; Li et al., 2023; Liu et al., 2025a), while efficiency optimizations prune low-scoring paths early (Zhang et al., 2024b; Qiu et al., 2024; Sun et al., 2024; Manvi et al., 2024; Ye & Ng, 2024). **Self-correction** iteratively refines outputs using feedback from tools, external models, or self-critique (Shinn et al., 2023; Gou et al., 2024; Li et al., 2024b; Song et al., 2025), though its efficacy depends on feedback reliability (Olausson et al., 2024; Huang et al., 2024; Wang et al., 2024a; Yang et al., 2024). **Tree searching** combines parallel and sequential scaling via algorithms like MCTS or A* (Yao et al., 2023; Xie et al., 2023; Long, 2023; Chari et al., 2025) guided by value functions (Xu, 2023; Hao et al., 2023; Chen et al., 2024; Zhang et al., 2024a). Training techniques distill these scaling benefits into more efficient models (Gao et al., 2023; Hou et al., 2024; Gulcehre et al., 2023; Zhang et al., 2024c). However, these methods have not been systematically applied to financial tasks, particularly stock movement prediction. Our work aims to fill this gap by exploring how inference-time scaling can be effectively utilized in this domain.

3 Preliminaries

3.1 Strict Controlled Dataset for Stock Movement Prediction

We aim to evaluate the ability of LLMs to predict stock movements based on diverse information sources. To this end, we construct a strict controlled dataset named **Fin-2024**, which covers the entire year of 2024, including 5,123 A-share stocks and 209,063 samples. Each sample is designed with a long context window of 32K tokens. Figure 1 illustrates the data sample for one stock at trading day t . LLMs will be trained from January to November, and then evaluated on December. In addition, we also collect data **Fin-2025[June]** on June 2025 for long-horizon prediction evaluation. The dataset construction process is detailed in Appendix C.

Information Sources The dataset consists of diverse information sources that have been proven valuable in machine learning-based quantitative trading research. These include: (1) **News articles** providing real-time market updates and company-specific & sector-specific information, (2) **Fundamental reports** reflecting company financial health and performance, (3) **Analyst opinions** offering professional market insights, (4) **Quantitative reports** containing technical analysis and market indicators, (5) **Macroeconomic indicators** showing broader economic trends, and (6) **Similar stocks information** for comparative analysis. The information are primarily textual, friendly for LLMs.

Prediction Target We define the prediction target as the price movement between the current trading day’s opening price and the previous trading-day’s closing price. This setting is less common in pre-training data compared to closing price-based movements, which helps prevent the model from exploiting memorization. The setting also avoids the model from decision-making games in trading periods, which is hard to capture in the given context. Based on price change, we classify the stock movement into three classes: **up** for change $> 3\%$, **down** for $< -3\%$, and **hold** for else. The three-class classification scheme requires more significant signals for price movements than binary classification. The **hold** class serves as a **decoy**. If the model never learns to distinguish between **up** and **down**, it would indicate a shortage on the model’s ability to make price movement predictions.

Evaluation Protocol During evaluation, we require the model to simultaneously predict both the price change percentage and direction to assess its instruction-following capability and verify the consistency between the two predictions, as the direction should align with the predicted change percentage.

3.2 Observation on Existing Models

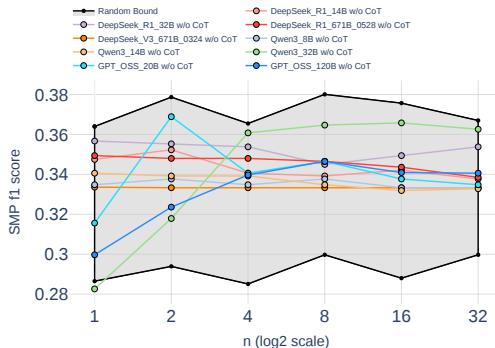


Figure 2: Existing LLMs without RETuning CoT perform no better than random guessing in stock movement prediction, and most (except Qwen3 32B) fail to scale at inference time.

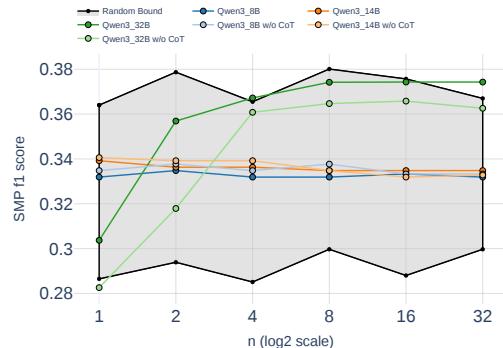


Figure 3: Abalton on CoT prompting. The proposed CoT in Section 4.1 can help improve the prediction performance on Qwen3 32B, but not smaller models. Others lose to random guessing.

To understand the limitations of existing models in stock movement prediction, we analyze their performance on **Fin-2024[December]** in Section 3.1. We use 32 different random seeds to uniformly sample the prediction (hold, up, and down) to construct the random bound (grey area in the figure). Any results that are covered by the random bound will be regarded failing to make trust-worthy prediction. The results reveal two key issues:

Firstly, by investigating existing LLMs’ performance in Figure 2, we observe that current LLMs are almost randomly guessing the prediction result. And the most models cannot scale their ability of prediction at inference time. Secondly, we prompt the model with fine-grained CoT (Section 4.1) to inject knowledge of financial analysis into the reasoning process. We further analyze the responses of the models, as shown in Figure 3. We find that CoT can help improve the model’s prediction performance on Qwen3 32B, but not other models.

To understand why these models fail, we further sample multiple responses for case study (Appendix E.1) and find that: the outputs of these LLMs tend to be vague, detached from the prediction target, and biased toward the **hold** class on label-balanced datasets. Thus, we propose to utilize

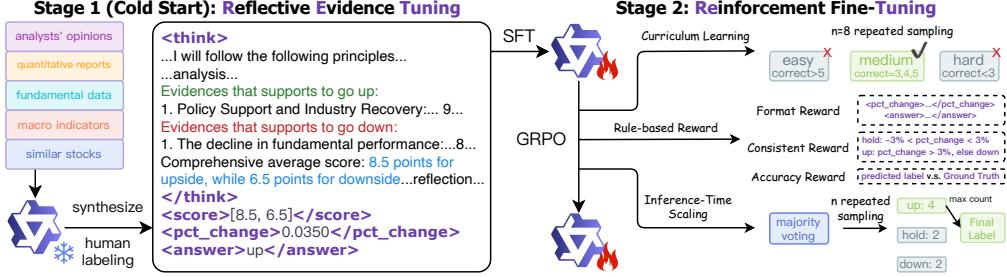


Figure 4: Two-stage stock movement prediction model training framework: **Stage 1** (Cold Start) uses multi-source data with human labeling and synthesis pipeline for **Reflective Evidence Tuning**; **Stage 2** applies **Reinforcement Fine-Tuning** with curriculum learning, reward shaping, and inference-time scaling for final label determination.

supervised fine-tuning to induce coherent, task-specific reasoning to change the output distribution of these models, as discussed in the following section.

4 Reflective Evidence Tuning (REtuning)

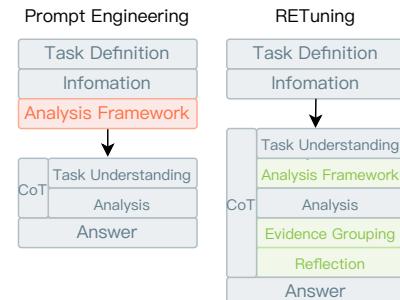
This section introduces **Reflective Evidence Tuning (REtuning)**, a two-stage framework designed to unlock the latent reasoning ability of LLMs in stock movement prediction tasks. As illustrated in Figure 4, RETuning comprises: (1) an SFT stage to *cold-start generative reasoning modeling*, and (2) *rule-based reinforcement learning* for performance refinement and alignment.

4.1 Stock Movement Prediction via Generative Reasoning Modeling

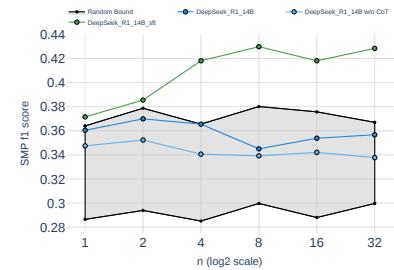
We frame the SMP task as a *generative reasoning problem*, in which the LLM leverage its reasoning ability to make predictions. It must construct an analytical framework, extract and score evidence from heterogeneous sources, and reflect before reaching a conclusion. As is shown in Figure 5a, this contrasts with zero-shot settings, where models superficially summarize and avoid making grounded predictions.

Then we employ supervised fine-tuning (SFT) to instill this reasoning structure into the model. The SFT dataset is constructed through a semi-automated pipeline (Appendix C.2) that uses DeepSeek-R1 (671B) as the backbone model to rejected sampling to synthesize 118 golden cold-start items.

By grounding each prediction in a dynamically built framework, **REtuning** promotes robust and context-aware reasoning. This reduces susceptibility to dominant context bias and improves the model's ability to rationally weigh adversarial evidence—crucial for reliable financial forecasting. By fine-tuning on this structured reasoning process, we find that the model **DeepSeek_R1_14b_SFT** can scale its prediction performance via repeated sampling far beyond random guessing, as shown in Figure 5b. It indicates that the model already possesses a certain level of predictive capability. We can further leverage this weak predictive power to assess the difficulty of predicting samples, thereby enhancing the overall predictive performance of the model more efficiently.



(a) **REtuning** guides the model to generate a principle, collect evidence, and reflect before making a prediction.



(b) DeepSeek_R1_14B_SFT scales prediction performance via repeated sampling.

4.2 Rule-based Reinforcement Learning

To further align model outputs with desired reasoning behavior, we introduce a rule-based reinforcement learning (RL) stage. Rather than relying on simple correctness-based rewards—which are noisy and statistically uninformative in financial prediction—we design more principled signals through **reward shaping** and **curriculum learning**.

Reward Shaping We design a **multi-faceted reward function** to capture both the **structural** and **semantic** correctness of model outputs. First, a **Format Score** ensures that the response adheres to the expected structured format, maintaining clarity and consistency. The **Accuracy Score** focuses on whether the model correctly predicts the directional movement—**up**, **down**, or **hold**. Lastly, the **Consistency Score** encourages logical alignment between the predicted percentage change and the stated directional label. Final score is given by $R = \alpha \cdot \text{Format} + \beta \cdot \text{Accuracy} + \gamma \cdot \text{Consistency}$, where α, β, γ are hyperparameters. This design mitigates the issue of misleading signals from noisy.

Curriculum Learning Not all samples are equally difficult: **hold** predictions are often trivial, while confident **up/down** predictions require strong signal integration. To make training more efficient and targeted, we propose a curriculum learning strategy:

We use the **cold-started model** to generate N ($=8$ in practice) predictions for each training sample. The difficulty of a sample is measured by counting how many of these predictions are incorrect. Based on this difficulty score, we categorize examples into three groups: **easy** (correct $\in [\frac{2}{3}N, N]$), **medium** (correct $\in [\frac{1}{3}N, \frac{2}{3}N]$), and **hard** (correct $\in [0, \frac{1}{3}N]$).

In Figure 6, we observe that a clear correlation between difficulty levels and labels. Low-difficulty samples are mostly dominated by **hold** predictions, which tend to be either spurious or too simple to be informative. High-difficulty samples, on the other hand, often involve **up** or **down** predictions but with weak or noisy signals. In contrast, medium-difficulty samples tend to reflect realistic market complexities and require non-trivial reasoning. To ensure the model focuses on meaningful learning signals, we discard both low and high-difficulty examples and train only on medium-difficulty ones, progressing in order of increasing difficulty.

Inference-Time Scaling We apply **majority voting** on predicted labels over n repeated generations with temperature 0.6. The final decision is given by: $\hat{y} = \arg \max_{y \in \{\text{up}, \text{down}, \text{hold}\}} \sum_{i=1}^n \mathbf{1}[y_i = y]$

5 Experiment and Results

We conduct several experiments to ascertain the effectiveness of **REtuning**, with the aim to gain insights into the following: (1) Basically, can we improve the stock movement prediction performance of LLMs? How does **REtuning** compare with existing methods? What insights does the model learn from the data? (2) Can we scale the prediction performance of LLMs at inference time? (3) What are the key factors that contribute to the success of **REtuning**? (4) Does the enhanced prediction ability contribute to other financial tasks?

5.1 Experiment Setup

Datasets. We use the data from January to November in **Fin-2024** for training, and the December data **Fin-2024[December]** for testing. We also use the **Fin-2025[June]** dataset to evaluate whether the model persists its scaling ability on prediction performance after 6 months. Besides, we evaluate the generalization ability on **BizFinBench** (Lu et al., 2025), a comprehensive financial benchmark covering 10 tasks, including Anomalous Event Attribution (AEA), Financial Time Reasoning (FTR), Financial Tool Usage (FTU), Financial Numerical Computation (FNC), Financial Knowledge QA (FQA), Financial Data Description (FDD), Emotion Recognition (ER), Stock Price Prediction (SP), Financial Named Entity Recognition (FNER). The details of the datasets are shown in Appendix C

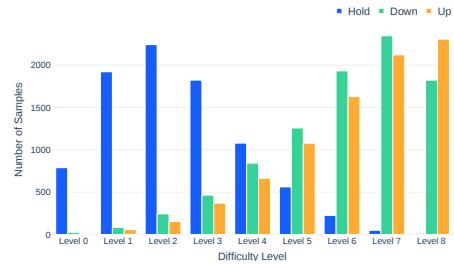


Figure 6: Difficulty distribution given by the cold-started model (DeepSeek_R1_14B_SFT).

Evaluation and Metrics. We care about the generalization ability and consider three types of out-of-distribution (OOD) settings: **OOD_Stock**, **OOD_Date**, and **OOD_Stock&Date**. We choose 50 stocks in random as the OOD stocks and the last month of 2024 as the OOD dates. The 50 stocks are also OOD stocks in **Fin-2025[June]**. We adopt the standard metrics **F1-score** because it balances precision and recall, making it suitable for our multi-class classification task.

Implementation and Baselines. The models are trained on up to 4*8 H100 GPUs. Rollout n is set to 8. Default results are obtained by greedy decoding. For inference-time scaling, we use $k \in \{1, 2, 4, 8, 16, 32\}$ and temperature=0.6. More implementation details are shown in Appendix B. Based on DeepSeek_R1_14B_Instruct (originally DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025)), we apply **REtuning** and get DeepSeek_R1_14B_SFT and DeepSeek_R1_14B_SFT_GRPO, which are after SFT stage and after SFT + GRPO stages, respectively. We also implement DeepSeek_R1_32B_SFT and DeepSeek_R1_32B_SFT_GRPO based on DeepSeek_R1_32B_Instruct. We compare to several strong baselines, including: LLMFactor (Wang et al., 2024b), Fino1 (Qian et al., 2025), Fin-R1 (Liu et al., 2025b), CMIN (Luo et al., 2023) and StockNet (Xu & Cohen, 2018). We also report results of several state-of-the-art open-weight LLMs: DeepSeek (DeepSeek-AI et al., 2025) (R1-7B, R1-14B, R1-32B, R1-671B, V3-671B), Qwen3 (Yang et al., 2025) (8B, 14B, 32B), GPT-OSS (OpenAI et al., 2025) (20B, 120B).

5.2 Results and Analysis

Table 1: Results of different methods on **Fin-2024[December]** benchmarks. w/ CoT means using the CoT prompting in Section 3.2. The relative improvements (%) over the baselines are shown in parentheses. The best results are in **bold**.

Model	F1 Score
<i>Results of Public Models</i>	
Random Guessing	0.3333
LLMFactor (Wang et al., 2024b)	0.3345
Fino1 (Qian et al., 2025)	0.0622
Fin-R1 (Liu et al., 2025b)	0.2543
CMIN (Luo et al., 2023)	0.3275
StockNet (Xu & Cohen, 2018)	0.3081
<i>Results with CoT Ablation</i>	
Qwen3_8B (Yang et al., 2025) w/ CoT	0.3348 0.3319 (-0.87%)
Qwen3_14B (Yang et al., 2025) w/ CoT	0.3406 0.3392 (-0.41%)
Qwen3_32B (Yang et al., 2025) w/ CoT	0.2826 0.3037 (+7.47%)
GPT_OSS_20B (OpenAI et al., 2025) w/ CoT	0.3156 0.3249 (+2.95%)
GPT_OSS_120B (OpenAI et al., 2025) w/ CoT	0.2997 0.3436 (+14.65%)
DeepSeek_R1_671B_0528 w/ CoT	0.3333 0.3494 (+4.83%)
DeepSeek_V3_671B_0324 w/ CoT	0.3336 0.3456 (+3.60%)
<i>Results of Our Models</i>	
DeepSeek_R1_14B_Instruct w/ CoT	0.3475 (baseline) 0.3604 (+3.71%)
DeepSeek_R1_14B_GRPO (Ours)	0.3377 (-2.25%)
DeepSeek_R1_14B_SFT (Ours)	0.3715 (+6.91%)
DeepSeek_R1_14B_SFT_GRPO (Ours)	0.4196 (+20.75%)
DeepSeek_R1_32B_Instruct w/ CoT	0.3567 (baseline) 0.3589 (+0.62%)
DeepSeek_R1_32B_GRPO (Ours)	0.3683 (+3.22%)
DeepSeek_R1_32B_SFT (Ours)	0.3639 (+2.02%)
DeepSeek_R1_32B_SFT_GRPO (Ours)	0.4071 (+14.13%)

REtuning vs baselines. The results of different methods on **Fin-2024[December]** benchmarks are shown in Table 1. We observe that RETuning (SFT + GRPO) significantly outperforms all baselines, including state-of-the-art open-weight LLMs (DeepSeek, Qwen3, GPT-OSS) and public models specifically designed for stock movement prediction (LLMFactor, Fino1, Fin-R1, CMIN,

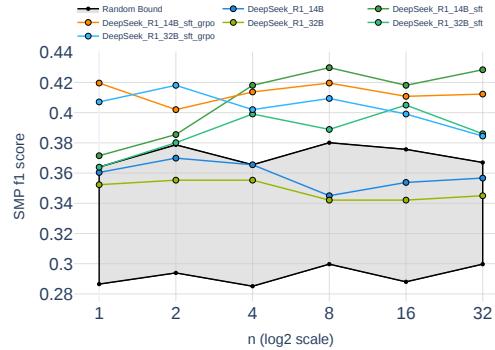


Figure 7: Inference-time scalability results on **Fin-2024[December]**. SFT model already has prediction ability, and GRPO further refines it.

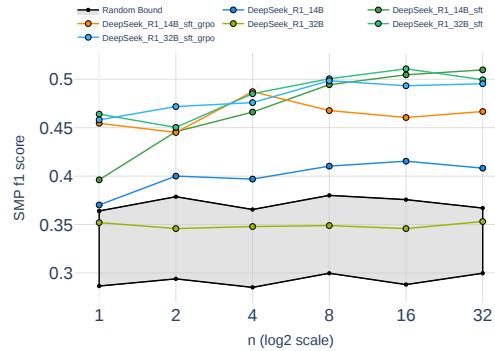


Figure 8: Inference-time scalability results on **Fin-2025[June]**. Finetuned models continue to scale via repeated sampling even after 6 months.

StockNet). For instance, DeepSeek_R1_14B_SFT_GRPO achieves an F1 score of 0.4196, which is a 20.75% relative improvement over its instruct baseline (0.3475) and surpasses the best public model (GPT-OSS-120B w/ CoT at 0.3436) by 22.15%. Similarly, DeepSeek_R1_32B_SFT_GRPO attains an F1 score of 0.4071, marking a 14.13% relative improvement over its instruct baseline (0.3567) and outperforming the best public model by 18.55%.

Can stock movement prediction benefit from inference-time scaling? Yes, but the gains from inference-time scaling are limited. Figure 7 and Figure 8 present the inference-time scalability results on Fin-2024[December] and Fin-2025[June], respectively. We observe monotonic or near-monotonic improvements up to $n \approx 8-16$, after which returns plateau and can even regress for some settings. RL (GRPO) makes test-time scaling less necessary by improving one-sample quality, yet does not increase the peak accuracy.

Can predictive ability generalize to unseen stocks, future dates, or both? Yes. We evaluate out-of-distribution (OOD) robustness along two axes: unseen stocks and forward-in-time generalization. The dataset Fin-2024[December] consists of OOD_Stock, OOD_Date, and OOD_Stock&Date cases, where RETuning maintains or increases F1 score as the number of inference-time samples n grows (Figure 7). On Fin-2025[June] (future dates only), RETuning preserves its advantage and continues to benefit from moderate repeated sampling (Figure 8), indicating strong temporal and cross-ticker generalization.

To further determine how the model scales on different OOD cases, we group the results by OOD split and present in Figure 9|10|11. The scaling is significant in **OOD_Stock**, then is **OOD_Date**. **OOD_Stock&Date** is the hardest cases to scale up, but it still outperforms baselines. We leave detailed analysis in Appendix D.2.

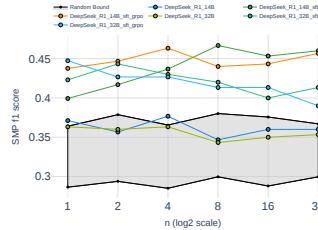


Figure 9: **OOD_Stock** results on Fin-2024[December]

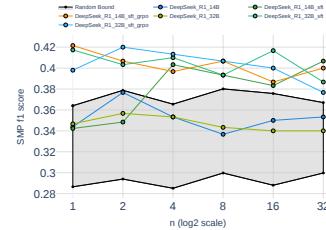


Figure 10: **OOD_Date** results on Fin-2024[December]

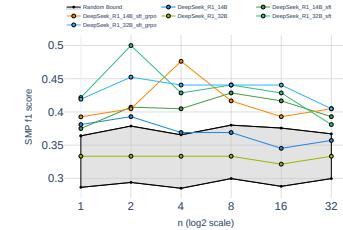


Figure 11: **OOD_Stock&Date** results on Fin-2024[December]

We also explore how the model scales on different ground truth labels. The results are grouped by ground truth label and presented in Figure 12|13|14. The scaling is significant in **hold** cases, and the model performance exceeds the baseline on **up** and **down** cases. We claim that **up** and **down** cases are more challenging, and **REtuning** enhances the model’s performance in these scenarios by enabling it to better leverage its reasoning capabilities, thus to identify factors influencing stock movements more effectively, thereby allowing the model to make more accurate predictions.

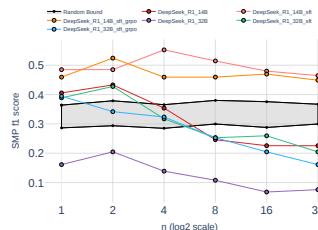


Figure 12: Ground truth **up** results on Fin-2024[December]

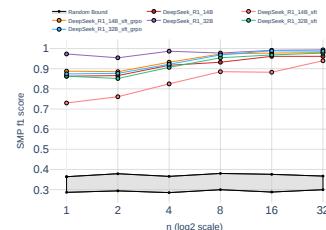


Figure 13: Ground truth **hold** results on Fin-2024[December]

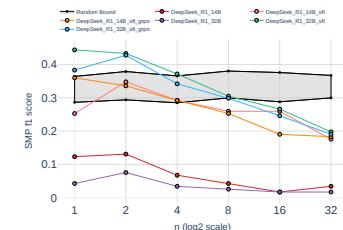


Figure 14: Ground truth **down** results on Fin-2024[December]

Ablation on CoT prompting. We further examine the effect of Chain-of-Thought (CoT) prompting (Table I). It causes slightly negative effects on smaller models (Qwen3-8B, Qwen3-14B), but significantly benefits larger models (Qwen3-32B, GPT-OSS-120B, DeepSeek-R1-671B, DeepSeek-

V3-671B). This suggests that larger models have a greater capacity to leverage CoT prompting effectively. For the 14B model, CoT improves the baseline by **+3.71%**, while the 32B model only gains a marginal **+0.62%**. This indicates that CoT alone yields limited benefits. In contrast, our SFT and SFT+GRPO variants consistently outperform CoT, suggesting that structured fine-tuning and reinforcement optimization are more effective than relying on prompting strategies alone.

Ablation on SFT stage. We compare the effect of applying GRPO directly versus combining it with an SFT stage (*_SFT_GRPO). For the 14B model, GRPO alone underperforms the baseline (0.3377 vs. 0.3475), while SFT followed by GRPO achieves a substantial gain (0.4196, **+20.75%**). A similar trend is observed in the 32B model: GRPO alone yields modest improvement (0.3683, **+3.22%**), whereas SFT+GRPO achieves the best performance (0.4071, **+14.13%**). These results highlight that the SFT stage provides essential initialization, enabling GRPO to realize its full benefit.

Table 2: Performance Comparison of LLMs on BizFinBench (Lu et al., 2025). The colors represent the top three performers for each task: **golden** indicates the top-performing model, **silver** represents the second-best result, and **bronze** denotes the third-best performance.

Model [†]	AEA	FNC	FTR	FTU	FQA	FDD	ER	SP	FNER	Average
Closed-Source LLMs										
ChatGPT-o3	86.23	61.30	75.36	89.15	91.25	98.55	44.48	53.27	65.13	73.86
ChatGPT-o4-mini	85.62	60.10	71.23	74.40	90.27	95.73	47.67	52.32	64.24	71.29
GPT-4o	79.42	56.51	76.20	82.37	87.79	98.84	45.33	54.33	65.37	71.80
Gemini-2.0-Flash	86.94	62.67	73.97	82.55	90.29	98.62	22.17	56.14	54.43	69.75
Claude-3.5-Sonnet	84.68	63.18	42.81	88.05	87.35	96.85	16.67	47.60	63.09	65.59
Open-Weight LLMs										
Qwen3-14B	84.20	58.20	65.80	82.19	84.12	92.91	33.00	52.31	50.70	67.05
Qwen3-32B	83.80	59.60	64.60	85.12	85.43	95.37	39.00	52.26	49.19	68.26
DeepSeek_R1_14B_Instruct ¹	71.33	44.35	50.45	81.96	85.52	92.81	39.50	50.20	52.76	59.49
DeepSeek_R1_32B_Instruct ¹	73.68	51.20	50.86	83.27	87.54	97.81	41.50	53.92	56.80	66.29
Our LLMs										
DeepSeek_R1_14B_SFT ²	80.63	51.67	52.61	83.53	89.05	96.72	36.68	50.43	50.85	65.36
14B Δ_{Instruct} (SFT)	+9.25	+7.28	+2.19	+1.53	+3.42	+3.85	-2.93	+0.24	-1.92	+5.83
DeepSeek_R1_14B_SFT_GRPO ²	81.46	52.41	53.47	83.57	89.02	95.58	36.83	54.06	51.24	66.92
14B Δ_{Instruct} (SFT_GRPO)	+10.03	+8.09	+2.91	+1.64	+3.45	+2.63	-2.74	+3.82	-1.53	+7.46
14B Δ_{SFT} (SFT_GRPO)	+0.82	+0.85	+0.81	+0.06	0.00	-1.23	+0.25	+3.63	+0.42	+1.53
DeepSeek_R1_32B_SFT	80.45	66.42	63.28	86.88	88.43	93.76	46.05	55.27	68.41	70.08
32B Δ_{Instruct} (SFT)	+6.75	+15.23	+12.37	+3.64	+0.83	-4.14	+4.52	+1.25	+11.63	+3.75
DeepSeek_R1_32B_SFT_GRPO	80.67	66.83	64.45	86.79	88.52	91.26	45.68	54.83	67.75	70.44
32B Δ_{Instruct} (SFT_GRPO)	+6.95	+15.62	+13.57	+3.55	+0.93	-6.64	+4.13	+0.85	+10.93	+4.12
32B Δ_{SFT} (SFT_GRPO)	+0.23	+0.42	+1.23	-0.06	+0.13	-2.53	-0.42	-0.43	-0.72	+0.33

[†] Closed-source LLMs results are sourced from Lu et al. (2025). Open-weight LLMs results are reproduced using temperature 0.6.

¹ DeepSeek_R1_14B_Instruct (32B) here is the short of DeepSeek-R1-Distill-Qwen-14B (32B) in the original paper (DeepSeek-AI et al., 2025).

² DeepSeek_R1_14B_SFT and DeepSeek_R1_14B_SFT_GRPO are DeepSeek_R1_14B_Instruct model after RETuning SFT stage and after RETuning SFT + GRPO stages, respectively.

Does the enhanced predictive ability contribute to other financial tasks? Yes. On the financial benchmark BizFinBench (Lu et al., 2025) (Table 2), RETuning generalizes beyond SMP: for 14B, the average score improves from 59.49 (Instruct) to 65.36 (SFT, +5.83) and 66.92 (SFT+GRPO, +7.46); for 32B, it improves from 66.29 to 70.08 (+3.75) and 70.44 (+4.12). The 32B models reach top-3 results on several tasks (e.g., FNC, FTU, ER, SP, FNER), while minor regressions appear on highly saturated dimensions (e.g., FDD after RL: -1.23 for 14B, -2.53 for 32B). Overall, RETuning yields broad, transferable gains with small trade-offs on a few tasks.

What insights does model learn to predict stock movement? Through analyzing the model’s responses detailed in Appendix E.2 we find that the model learns to: 1. Identifying key evidences that influence daily fluctuations in stock prices from multiple information sources. 2. Trending to correctly evaluate the short-term impact of gathered evidences, which benefits from the trade-off ability induced by adversarial scoring. 3. Gradually improving the consistency between the predicted fluctuation and the movement label during reinforcement learning, which we named “vibe prediction”.

6 Conclusion

In this work, we explored the underexamined problem of applying large language models (LLMs) to stock movement prediction. Our analysis revealed that vanilla LLMs tend to rely on contextual viewpoints rather than developing independent analytical reasoning, which limits their predictive reliability. To address this challenge, we introduced **REtuning**, a reflective evidence-based tuning method that encourages models to construct analytical frameworks, weigh adversarial evidence, and refine predictions through reflection. Experiments on our newly constructed large-scale financial dataset demonstrate that RETuning substantially improves predictive performance over strong baselines, enabling more systematic reasoning in the financial domain. Moreover, RETuning generalizes beyond stock movement prediction, yielding gains across diverse financial tasks, and exhibits robustness under inference-time scaling and out-of-distribution settings. Overall, this study highlights the importance of evidence-oriented reasoning in financial LLMs and establishes RETuning as a promising direction for enhancing their reliability and applicability in real-world financial decision-making.

Ethics Statement

To address potential ethical considerations related to our research on large language models (LLMs) for stock movement prediction, we provide the following statement:

First, regarding data ethics: Our large-scale dataset (spanning 2024 for 5,123 A-share stocks) is constructed exclusively from [publicly available sources](#), including market price data, publicly disclosed news, analysts' public opinions, company fundamental reports, official macroeconomic indicators, and publicly accessible information on peer stocks. We strictly comply with China's Data Security Law, Securities Law, and relevant financial regulatory requirements, ensuring no collection or use of private, sensitive personal data, or non-public material information that could violate market fairness.

Second, on potential harm and application boundaries: This research is conducted for [academic purposes only](#) to advance LLM reasoning capabilities in financial tasks. We explicitly emphasize that our model (RETuning) and findings do not constitute financial advice, nor do they endorse or promote real-world investment decisions. Stock market prediction inherently carries high uncertainty, and any practical application of such models for trading could lead to financial risks; we disclaim responsibility for any losses arising from non-academic use of our work.

Third, on bias and fairness: While we designed RETuning to enhance independent logical reasoning (reducing undue reliance on contextual viewpoints) and constructed a diverse dataset to cover a broad range of A-share stocks, we acknowledge potential residual biases (e.g., sector-specific skews in training data or sensitivity to market cycles). Future work will further validate and mitigate such biases to improve the model's fairness across different market scenarios.

Finally, regarding research integrity: Our study involves no human subjects, so institutional review board (IRB) approval is not applicable. We commit to transparency in dataset construction details and methodology implementation (as detailed in the full paper) to enable reproducibility. We adhere to rigorous academic standards to avoid misrepresentation of results or misuse of technical insights.

Reproducibility Statement

We make every effort to ensure that the experiments in this paper are reproducible. Specifically, anonymized source code (training and evaluation scripts), model checkpoints, and processed dataset splits will be released as supplementary material and at the repositories indicated in the Appendix.* The Appendix contains detailed descriptions of data collection and preprocessing (Appendix C), the prompt templates and example inputs/outputs (Figures 17–27), and the exact training and evaluation settings including compute and hardware details (Appendix B, Tables 3 and 4). Evaluation splits used for OOD and long-horizon tests (e.g., **Fin-2024[December]**, **Fin-2025[June]**) and the scripts to compute all reported metrics will also be provided. Where full raw data cannot be released due to third-party licensing or privacy constraints, we describe the access procedure and provide processed, reproducible derivatives in the supplementary materials.

Broader Impact

This work investigates the use of large language models (LLMs) for stock movement prediction, a domain with potentially high economic and societal implications. Our proposed method, RETuning, demonstrates how reflective evidence organization can enhance independent reasoning in financial tasks. On the positive side, this research contributes to the broader effort of making LLMs more reliable in high-stakes applications by encouraging systematic analysis rather than context-driven imitation. Such improvements may benefit both academic research in reasoning and practical applications in financial decision-support systems.

At the same time, we emphasize that financial forecasting is inherently uncertain and subject to market volatility, regulatory constraints, and ethical considerations. Automated prediction systems, if

*<https://github.com/LinXueyuanStdio/RETuning>, <https://huggingface.co/collections/linxy/retuning-68c999be4d9ac2834c64fd00>, <https://huggingface.co/datasets/linxy/Fin-2024>

misused, could amplify risks, encourage speculative behavior, or contribute to unfair advantages for certain market participants. Our dataset and methods are designed for research purposes only, and we strongly discourage their direct use for live trading or investment without rigorous safeguards, human oversight, and compliance with financial regulations.

More broadly, this work highlights both the opportunities and limitations of applying LLMs in sensitive domains. We hope that our findings spur further research into building transparent, evidence-based reasoning frameworks that improve model reliability while also ensuring responsible deployment in practice.

Limitations

Domain specificity. Our study focuses on the Chinese A-share market in 2024, which provides a rich testbed but may limit the generalizability of findings to other markets, such as U.S. equities or emerging markets with different structures, liquidity, and regulatory conditions.

Data coverage. Although our dataset integrates multiple information sources (prices, news, analyst opinions, fundamentals, and macroeconomic indicators), it remains incomplete. Certain high-frequency signals (e.g., intraday order flow, alternative data, or global macro shocks) are not incorporated, potentially constraining predictive accuracy.

Model assumptions. RETuning assumes that LLMs can benefit from explicitly structuring and reflecting on evidence. While this holds in our experiments, the approach may be less effective in domains like healthcare or cryptocurrency, where reliable evidence is scarce, noisy, or difficult to formalize.

Evaluation scope. Our evaluation mainly relies on F1 scores for three-class stock movement prediction and selected benchmarks (e.g., BizFinBench (Lu et al., 2025)). Broader metrics, such as profitability in trading simulations or risk-adjusted returns, are not considered. It takes time to validate real-world trading performance, which is beyond the scope of this paper.

Computation and scalability. Both training (SFT + GRPO) and inference-time scaling are computationally expensive. This may limit accessibility for smaller institutions or researchers without large-scale compute resources, raising questions about cost-efficiency in real-world deployment.

Future Work

Extending to other markets. Future research could examine the effectiveness of RETuning across diverse markets such as U.S. equities, European exchanges, and emerging markets. This would validate whether the approach generalizes under different regulatory regimes, liquidity conditions, and investor behaviors.

Incorporating richer data sources. Enhancing the dataset with high-frequency trading signals, alternative data (e.g., satellite imagery, ESG reports), and global macroeconomic factors could provide a more comprehensive information environment and further strengthen predictive power.

Advancing reasoning frameworks. While RETuning encourages evidence-based reasoning, future work may integrate causal inference, probabilistic reasoning, or game-theoretic perspectives to capture deeper structures behind stock movements and reduce susceptibility to spurious correlations.

Evaluation beyond prediction accuracy. A natural next step is to link model predictions with financial outcomes, such as profitability, portfolio optimization, and risk-adjusted returns. This would bridge the gap between benchmark metrics and real-world decision-making.

Efficiency and accessibility. Research on lightweight RETuning variants, parameter-efficient fine-tuning, and inference-time acceleration could reduce computational costs, making the method more accessible to practitioners and researchers with limited resources.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simon Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Maiteusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrew Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantiliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023. URL <https://api.semanticscholar.org/CorpusID:257532815>.

Farhan Adyatma and Andry Alamsyah. The indonesia stock exchange composite prediction based on macroeconomic indicators using arima, lstm, and ann. *2022 8th International Conference on Science and Technology (ICST)*, 1:1–5, 2022. URL <https://api.semanticscholar.org/CorpusID:258992891>.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhong Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu,

Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *ArXiv*, abs/2309.16609, 2023. URL <https://api.semanticscholar.org/CorpusID:263134555>.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Anirudh Chari, Aditya Tiwari, Richard Lian, Suraj Reddy, and Brian Zhou. Pheromone-based learning of optimal reasoning paths, 2025. URL <https://arxiv.org/abs/2501.19278>.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamat almost zero: Process supervision without process, 2024. URL <https://arxiv.org/abs/2405.03553>.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. <https://github.com/InternLM/xtuner>, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jia Shi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghai Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv: 2501.12948*, 2025.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)*, 37(2): 1–30, 2019. URL <https://doi.org/10.1145/3309547>.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth*

International Conference on Learning Representations, 2024. URL <https://openreview.net/forum?id=Sx038qxjek>.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. URL <https://arxiv.org/abs/2308.08998>.

Pei-Yi Hao, Chien-Feng Kung, Chun-Yang Chang, and Jen-Bing Ou. Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane. *Applied Soft Computing*, 98:106806, 2021. URL <https://www.sciencedirect.com/science/article/pii/S1568494620307444>.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL <https://aclanthology.org/2023.emnlp-main.507>.

Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. Does rlhf scale? exploring the impacts from data, model, and method, 2024. URL <https://arxiv.org/abs/2412.06000>.

Yi-Ling Hsu, Yu-Che Tsai, and Cheng-Te Li. Fingat: Financial graph attention networks for recommending top- k profitable stocks. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):469–481, 2021. doi: 10.1109/TKDE.2021.3079496.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=IkmD3fKBPQ>.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1611835114>.

Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. S*: Test time scaling for code generation. *arXiv preprint arXiv:2502.14382*, 2025.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need, 2024a. URL <https://arxiv.org/abs/2402.05120>.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models, 2024b. URL <https://arxiv.org/abs/2402.12563>.

Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. Modeling the stock relation with graph network for overnight stock movement prediction. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pp. 4541–4547, 2021. URL <https://doi.org/10.24963/ijcai.2020/626>.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL <https://aclanthology.org/2023.acl-long.291>.

Lei Lin, Jiayi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3829–3852, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.230. URL <https://aclanthology.org/2024.findings-acl.230>.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Pairwise rm: Perform best-of-n sampling with knockout tournament. *arXiv preprint arXiv:2501.13007*, 2025a.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. Fin-rl: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv: 2503.16252*, 2025b.

Jieyi Long. Large language model guided tree-of-thought, 2023. URL <https://arxiv.org/abs/2305.08291>.

Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. Bizfinbench: A business-driven real-world financial benchmark for evaluating llms. *arXiv preprint arXiv: 2505.19457*, 2025.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12164–12176, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.679. URL <https://aclanthology.org/2023.acl-long.679>.

S. Z. Mahfooz, Iftikhar Ali, and Muhammad Navaid Khan. Improving stock trend prediction using lstm neural network trained on a complex trading strategy. *International Journal for Research in Applied Science and Engineering Technology*, 2022. URL <https://api.semanticscholar.org/CorpusID:251176332>.

Rohin Manvi, Anikait Singh, and Stefano Ermon. Adaptive inference-time compute: Llms can predict if they can do better, even mid-generation, 2024. URL <https://arxiv.org/abs/2410.02725>.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv: 2501.19393*, 2025.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL <https://arxiv.org/abs/2112.09332>.

Thien Hai Nguyen, Kyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015. URL <https://www.sciencedirect.com/science/article/pii/S0957417415005126>.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=y0GJXRungR>.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi,

Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilabert, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card. *arXiv preprint arXiv: 2412.16720*, 2024.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezzcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv: 2508.10925*, 2025.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. Fino1: On the transferability of reasoning enhanced llms to finance. *ArXiv*, abs/2502.08127, 2025. URL <https://api.semanticscholar.org/CorpusID:276287200>.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling, 2024. URL <https://arxiv.org/abs/2410.16033>.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL <https://doi.org/10.1145/3394486.3406703>.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL <https://arxiv.org/abs/2402.03300>.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL <https://doi.org/10.1145/3689031.3696075>.

Yong Shi, Yuanchun Zheng, Kun Guo, and Xinyue Ren. Stock movement prediction with sentiment analysis based on deep learning networks. *Concurrency and Computation: Practice and Experience*, 33, 2020. URL <https://api.semanticscholar.org/CorpusID:228848908>.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL <https://arxiv.org/abs/2303.11366>.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv: 2408.03314*, 2024.

Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, and Bo Zheng. Progco: Program helps self-correction of large language models, 2025. URL <https://arxiv.org/abs/2501.01264>.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=348hfcprUs>.

Van-Dai Ta, Chuan-Ming Liu, and Direselign Addis. Prediction and portfolio optimization in quantitative trading using machine learning techniques. *Proceedings of the 9th International Symposium on Information and Communication Technology*, 2018. URL <https://api.semanticscholar.org/CorpusID:56450817>.

Vernon Y. H. Toh, Deepanway Ghosal, and Soujanya Poria. Not all votes count! programs as verifiers improve self-consistency of language models for math reasoning, 2024. URL <https://arxiv.org/abs/2410.12608>.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *Arxiv*, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv: 2211.14275*, 2022.

Manuel R Vargas, Carlos EM Dos Anjos, Gustavo LG Bichara, and Alexandre G Evsukoff. Deep learning for stock market prediction using technical indicators and financial news articles. In *2018 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2018. doi: 10.1109/IJCNN.2018.8489208.

Chang Sim Vui, Kim Soon Gan, Chin Kim On, R. Alfred, and Patricia Anthony. A review of stock market prediction with artificial neural network (ann). *2013 IEEE International Conference on Control System, Computing and Engineering*, pp. 477–482, 2013. URL <https://api.semanticscholar.org/CorpusID:9567658>.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies, 2024a. URL <https://arxiv.org/abs/2406.06461>

Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3120–3131, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.185. URL <https://aclanthology.org/2024.findings-acl.185/>

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Bw82hwg5Q3>.

Haotian Xu. No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function, 2023. URL <https://arxiv.org/abs/2309.03224>

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1183. URL <https://aclanthology.org/P18-1183/>

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. Confidence v.s. critique: A decomposition of self-correction capability for llms, 2024. URL <https://arxiv.org/abs/2412.19513>

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

Hai Ye and Hwee Tou Ng. Preference-guided reflective sampling for aligning language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21646–21668, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1206. URL <https://aclanthology.org/2024.emnlp-main.1206>.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang, and Dongzhan Zhou. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning, 2024a. URL <https://arxiv.org/abs/2410.02884>.

Ruiqi Zhang, Momin Haider, Ming Yin, Jiahao Qiu, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Accelerating best-of-n via speculative rejection. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024b. URL <https://openreview.net/forum?id=dRp8tAIPhj>.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding, 2024c. URL <https://arxiv.org/abs/2412.00154>.

Zhihan Zhou, Liqian Ma, and Han Liu. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2114–2124, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.186. URL <https://aclanthology.org/2021.findings-acl.186>.

Appendix

A The Usage of Large Language Models (LLMs)

This section details the specific role of Large Language Models (LLMs) in this paper.

We employ LLMs *GPT-5* (developed by OpenAI) and *Douba* (developed by ByteDance) to enhance the clarity, coherence, and overall quality of our manuscript. The LLM assistance primarily focus on language polishing (refining structure, terminology consistency, grammar) and formatting adjustments, ensuring that the paper meets high standards of academic writing. No other LLMs were used for research ideation or image generation.

All reviewed/approved by authors. All authors bear *full responsibility* for the final paper. All content (including LLM-generated/polished text) was verified: factual claims cross-checked against datasets/literature, and the manuscript screened to avoid unintended plagiarism.

B Implementation Details

B.1 Open Source, Open Weights, and Open Data

The source code is available at GitHub[†]. The model weights are available at HuggingFace[‡]. The training and evaluation datasets are available at HuggingFace[§].

B.2 Model Training

The models are trained on up to 4*8 H100 GPUs. Rollout n is set to 8. The training epochs are set to 3 for SFT and 1 for RL. The details of the data synthesis workflow for building SFT dataset are shown in Appendix C.2. With the help of the SFT model to determine the prediction difficulty, we further apply GRPO (Shao et al., 2024) with curriculum learning and reward shaping. The objective is

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) &= \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \\ &\left\{ \min \left[\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL} [\pi_\theta || \pi_{ref}] \right\}, \end{aligned} \quad (1)$$

where advantage $\hat{A}_{i,t} = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$, G is the group size, β is the coefficient of KL penalty, and $q = (x, \{y_i\}_{i=1}^n)$ with prompt x and n generations y .

B.3 Hyperparameters

We report the detailed hyperparameters for Supervised Fine-Tuning (SFT) in Table 3 and for GRPO in Table 4. We use up to 4*8 H100 80GB GPUs for experiments.

B.4 Training Frameworks

We use Xtuner (Contributors, 2023) to SFT with DeepSpeed (Rasley et al., 2020) to accelerate training and ZeRO-3 to reduce memory usage, and verl (Sheng et al., 2025) to implement GRPO.

[†]<https://github.com/LinXueyuanStdio/RETuning>

[‡]<https://huggingface.co/collections/linxy/retuning-68c999be4d9ac2834c64fd00>

[§]<https://huggingface.co/datasets/linxy/Fin-2024>

Table 3: The training hyperparameters for Supervised Fine-Tuning (SFT). 32B, 14B, and 7B denote models based on DeepSeek-R1-Distill-Qwen with 32B, 14B, and 7B parameters respectively.

Hyperparameter Category	32B	14B	7B
1. Data Configuration			
Train Micro-Batch Size per GPU	1	1	1
Gradient Accumulation Steps	128	128	128
Total Effective Batch Size	1024	1024	1024
($1 \times 128 \times 8$)	($1 \times 128 \times 8$)	($1 \times 128 \times 8$)	
Pack Sequences to Max Length	False	False	False
Data Shuffling Before Packing	True	True	True
2. Model & LoRA Configuration			
LLM Torch Dtype	torch.float16	torch.float16	torch.float16
LoRA Rank (r)	32	32	32
LoRA Alpha (α)	64	64	64
LoRA Dropout	0.1	0.1	0.1
LoRA Bias Type	none	none	none
LoRA Task Type	CAUSAL_LM	CAUSAL_LM	CAUSAL_LM
Variable-Length Attention	False	False	False
3. Optimizer & LR Scheduler			
Optimizer Type	torch.optim.AdamW		
Learning Rate (LR)	2×10^{-4}	2×10^{-4}	2×10^{-4}
AdamW Betas (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Weight Decay	0	0	0
Gradient Clipping Max Norm	1	1	1
LR Scheduler Type	Linear Warmup + Cosine Annealing		
Warmup Ratio (Warmup Epochs / Total Epochs)	0.03 (0.09/3)	0.03 (0.09/3)	0.03 (0.09/3)
Warmup Start Factor	1×10^{-5}	1×10^{-5}	1×10^{-5}
Cosine Annealing Final LR (η_{min})	0.0	0.0	0.0
4. Training Strategy & Distributed Config			
Training Strategy Type	DeepSpeedStrategy		
DeepSpeed Zero Optimization Stage	3	3	3
BF16 Precision Enabled	True	True	True
FP16 Precision Enabled	False	False	False
Sequence Parallel Size	8	8	8
Sampler Shuffling	True	True	True
6. Environment & Misc Config			
Launcher Type	pytorch		
Distributed Backend	nccl	nccl	nccl
Multiprocessing Start Method	fork	fork	fork
Deterministic Training	False	False	False

C Dataset Details

C.1 Fin-2024 and Fin-2025

We consider studying the stock movement prediction task based on data from the Chinese A-share market. Naturally, the collected data are in Chinese, and consequently, the associated prompts and synthetic data are also in Chinese. This consistency within a single language allows LLMs to achieve better understanding and more coherent reasoning over the data. Thus, the language model can fully leverage its pre-trained knowledge in Chinese to analyze the stock market. Therefore, in this paper, we choose Qwen (Bai et al., 2023) and DeepSeek (DeepSeek-AI et al., 2025) as the backbone models, both of which are strong Chinese LLMs. We apologize for any inconvenience the language gap may cause to readers who are not native Chinese speakers, and we hope that this gap will not hinder the understanding of our work.

The reason why we decide to construct a new dataset to study the stock movement prediction task is twofold. First, existing datasets (StockNet (Xu & Cohen, 2018), CMIN-US (Luo et al., 2023), CMIN-

Table 4: The training hyperparameters for GRPO. 32B SFT, 14B SFT, and 7B SFT denote models based on DeepSeek R1 with 32B, 14B, and 7B parameters after SFT stage respectively.

Base Model	32B SFT	14B SFT	7B SFT
1. Data Configuration			
Training Batch Size	256	256	256
Validation Batch Size	256	256	256
Max Prompt Length	32768	32768	32768
Max Response Length	4096	4096	4096
Data Shuffling	True	True	True
2. Algorithm Configuration			
Advantage Estimator	GRPO	GRPO	GRPO
Gamma (γ , Discount Factor)	1.0	1.0	1.0
Lambda (λ , Advantage Smoothing)	1.0	1.0	1.0
KL Coefficient	0.001	0.001	0.001
Target KL Divergence	0.1	0.1	0.1
Normalize Advantage by Std	True	True	True
3. Actor & Ref Configuration			
Learning Rate	3×10^{-7}	3×10^{-7}	3×10^{-7}
Weight Decay	0.01	0.01	0.01
Clip Ratio	0.2	0.2	0.2
Entropy Coefficient	0.0	0.0	0.0
PPO Epochs	1	1	1
Log Prob Micro Batch Size	8	4	4
Tensor Parallel Size	4	4	4
4. Rollout Configuration			
Rollout Count (n)	8	8	8
Rollout Mode	sync	sync	sync
Engine Name	vllm	vllm	vllm
Data Type (dtype)	bfloat16	bfloat16	bfloat16
Temperature	0.6	0.6	0.6
Max Num Batched Tokens	36864	36864	36864
Tensor Parallel Size	4	4	4
Enable Chunked Prefill	True	True	True
5. Trainer Configuration			
Number of Nodes	4	2	1
GPUs per Node	8	8	8
Total Epochs	1	1	1
Checkpoint Save Frequency (steps)	10	10	10
Validation Frequency (steps)	10	10	10
Validate Before Training	True	True	True

CN (Luo et al., 2023), EDT (Zhou et al., 2021)) are outdated and do not reflect the current market conditions. Financial markets are dynamic and constantly evolving, with new trends, regulations, and events shaping the landscape. Using outdated datasets may lead to models that are not well-suited for current market scenarios. Second, existing datasets lack diversity in data sources. Relying solely on price and news data may not capture the full complexity of stock movements. Incorporating additional data sources such as analyst reports, macroeconomic indicators, and quantitative reports can provide a more comprehensive view of the market and improve prediction accuracy. Rich enough data sources are crucial for reliable forecasting in the high-signal-noise ratio of financial markets.

Then, we build a new dataset **Fin-2024** covering January to December 2024, with a test split set **Fin-2024[December]** and an additional long-horizon evaluation set **Fin-2025[June]**. We collect data from multiple sources, including stock prices, financial news, analyst reports, macroeconomic indicators, and quantitative reports. We process the raw data into a structured format suitable for LLMs, including entity recognition, sentiment analysis, event extraction, and traditional time-series analysis (for generating quantitative reports). The dataset contains 209,063 data points across 5,123 A-share stocks from various sectors. Each data point includes a timestamp, stock identifier, historical

prices (open, close, high, low, volume), relevant news articles, analyst reports, macroeconomic indicators, quantitative reports, and the corresponding stock movement label (up/down/hold), which is based on the change_pct between the open price of the next trading day and the close price of the current trading day.

In order to ensure data quality, we apply several filtering steps. We remove data points with missing or incomplete information, filter out stocks with low trading volume or insufficient historical data, and balance the label distribution in the dataset to avoid bias towards any particular class. The data processing pipeline is shown in Figure 15, and the prompt length distribution is shown in Figure 16.

The final dataset is split into training (90%) and out-of-distribution (OOD) (10%) sets. The training set is used for model fine-tuning and reinforcement learning, and the OOD set for final evaluation. The long-horizon evaluation set **Fin-2025[June]** contains data from June 2025 to assess model performance in a future market scenario. Please refer to Figure 15 for detailed numbers and splits.

We present an example of the prompt template in Figure 17, which consists of multiple parts: stock news (Figure 18 and Figure 19), stock price information of the current stock and top-3 similar stocks (Figure 20), macroeconomic indicators report (Figure 21), stock fundamentals report (Figure 22 and Figure 23), stock basic information (Figure 24), stock quantitative reports (Figure 25), model response (Figure 26), and model response grading (Figure 27).

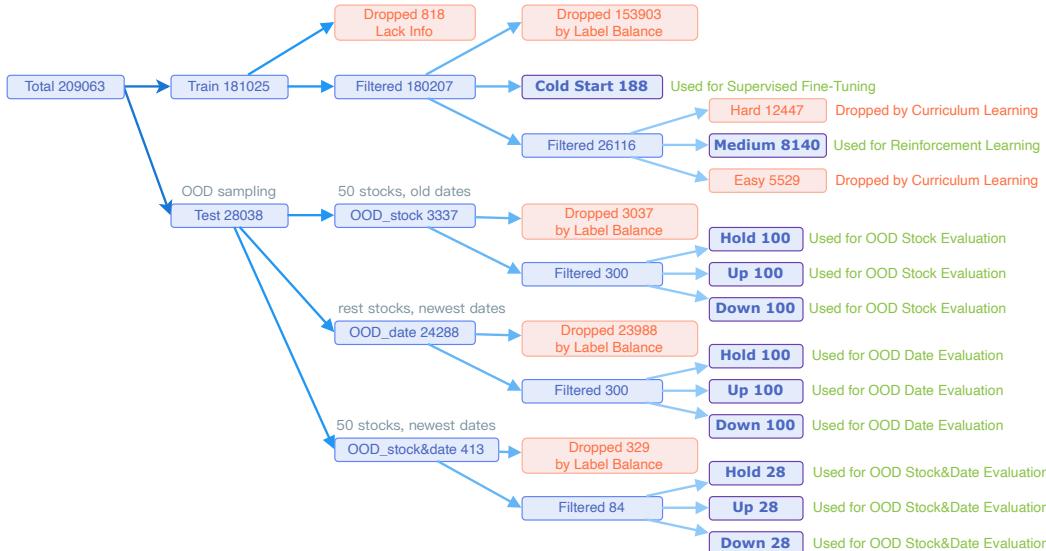


Figure 15: Data processing pipeline showing total 209063 points split into train/test, filtered via multiple steps (lack info, label balance, curriculum learning), and categorized for SFT, RL, OOD evaluations.

C.2 Fin-2024-SFT

The dataset **Fin-2024-SFT** contains 188 cold-start items and 10K general reasoning samples. Firstly, we selected 10K items from <https://huggingface.co/datasets/GeneralReasoning/GeneralThought-323K>, which aims to avoid catastrophic forgetting (Kirkpatrick et al., 2017) and even strengthen reasoning ability when fine-tuning the model. These data points are related to math, code, common sense, chatting, role play, writing, etc. However, they are not allowed to be related to finance to avoid bias, so that we can determine the model’s performance in a more controlled setting.

Then, the samples for cold-starting are constructed using a workflow with DeepSeek-R1 (671B) as the backbone model and the polish prompt (Figure 28) to generate high-quality, diverse training samples that follow the proposed thinking schema presented in Section 4.1. These synthesized samples are further filtered by a reward function, which checks format, validates prediction (<score>, <change_pct>, <answer>) against the ground truth, and ensures adherence to the desired output style. At this stage, 300 samples passed the validation. From these, we cherry-picked 188 samples

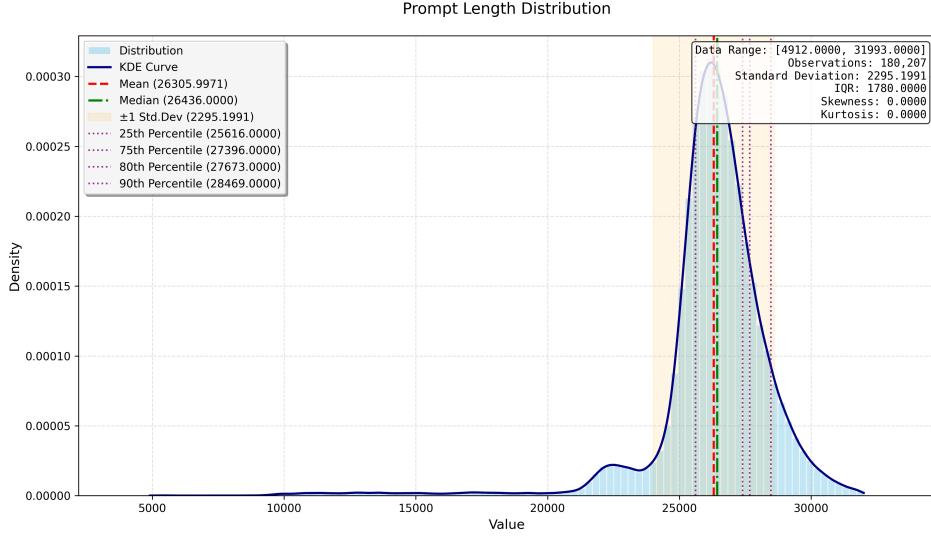


Figure 16: Prompt length distribution across the dataset, illustrating the varying lengths of input prompts used for training and evaluation.

in total, which required approximately six hours of expert annotation (around 100 USD) for final curation.

Notably, the financial data we collected spans the entire year of 2024 and June 2025. Among these, the data from January to November 2024 were used for training. From this subset of training, we extracted 300 instances (100 labeled as *up*, 100 as *hold*, and 100 as *down*) to construct the cold-start dataset for synthesis.

RETuning guides the model through a multistep process:

1. **Understand the Task.** The model first internalizes the task—predicting the direction of price movement from the previous close to the next open, with a $\pm 3\%$ threshold—improving alignment and reducing hallucination.
2. **Establish Analytical Principles.** It constructs a dynamic analytical framework (e.g., fundamentals, news trends, macro signals), independent of analyst commentary.
3. **Extract Evidence.** The model collects multiple context-based signals and assesses their support for each directional hypothesis.
4. **Group and Score.** Extracted evidence is grouped and scored by directional leaning, forming a soft-evidence pool.
5. **Reflect and Reconcile.** Averaging directional scores, the model enters a reflection phase to resolve conflicting evidence.
6. **Produce Structured Output.** The model generates a final decision in a consistent, interpretable format, specifying both direction and percentage change.

In detail, we employed the following workflow:

- **Step 1: Multi-Source Data Aggregation.** We systematically collected a diverse range of financial data covering the period from January 2024 to June 2025. This included quantitative data, such as historical stock prices (open, high, low, close) and trading volumes, alongside qualitative data from relevant financial news articles. For training purposes, we focused on the January–November 2024 subset, from which 300 balanced instances (*up*, *hold*, *down*) were selected for cold-start data synthesis. To ensure the dataset’s breadth and representativeness, we selected a varied portfolio of stocks from multiple market sectors.
- **Step 2: Structured Prompt Formulation.** We engineered a series of structured prompts designed to elicit detailed, context-aware analytical reasoning from the language model.

现在要求你预测股价涨跌：给出支持上涨和下跌的证据评分，并预测下一个交易日的开盘价相对于当前交易日的收盘价的涨跌幅，最后给出涨跌分类结果。请先在脑海中思考推理过程，然后向用户提供最终答案。
请回忆一下你在金融领域的知识和经验，结合当前的市场环境，给出一个合理的预测。

你需要考虑：做好这个预测任务所需的关键条件；整理自己的状态；明确自己的分析范式；

为了避免被材料中的情绪误导，请遵循以下思维范式：

1. 理解所要分析的股票的个性，如蓝筹股、成长股、ST股等，不同类型股票的分析方法不同
2. 理解所要预测的时间特征，关注节假日、周末、月末等，不同时间特征的分析方法不同
3. 查看所提供的市场状态包含哪方面的信息，不同的信息覆盖的维度不同，因此分析方法也不同
4. 初步动态构建分析方法逻辑
5. 按分析方法逻辑分析各维度信息
6. 开始整理这些信息，按支持涨的、跌的进行分类，每一类都要对每一条证据进行评分，10分制
7. 进行假设检验，市场模拟，未来推演，反事实假设等，对这些证据进行反思，直到你确信你已经考虑了所有可能的情况。
8. 综合平均这些评分，给出最终支持分数。如 `<score>[a, b]</score>` 表示支持涨的a分，支持跌的b分；a和b的范围在[0, 10]之间
9. 给出最终的涨跌幅预测和方向预测

预测目标：

1. 证据评分：请给出支持上涨和下跌的证据评分，范围在[0, 10]之间。请在 `<score>[a, b]</score>` 中填写你的评分，即支持涨的a分，支持跌的b分；a和b的范围在[0, 10]之间。
2. 涨跌幅预测：你需要预测下一个交易日的开盘价相对于当前交易日的收盘价的涨跌幅，范围在[-1, 1]之间，精确到小数点后四位。请在 `<pct_change>0.xxxx</pct_change>` 中填写你的答案。
3. 方向预测：必须且只能从以下三个选项中选择一个
 - 所预测的涨跌幅大于 3%（显著上涨）：`<answer>up</answer>`
 - 所预测的涨跌幅小于 -3%（显著下跌）：`<answer>down</answer>`
 - 所预测的涨跌幅在 -3% 和 3% 之间（震荡）：`<answer>hold</answer>`

推理过程、证据评分、涨跌幅预测、答案分别包含在 `<think></think>`、`<score></score>`、`<pct_change></pct_change>` 和 `<answer></answer>` 标签中，即 `<think>` 此处为推理过程 `</think>` `<score>` 此处为证据评分 `</score>` `<pct_change>` 此处为涨跌幅预测 `</pct_change>` `<answer>` 此处为答案 `</answer>`。

当前环境：

```
Current Trading Date: {{ example.cur_trading_date }}{{ get_weekday(example.cur_trading_date) }}
Next Trading Date: {{ example.next_trading_date }}{{ get_weekday(example.next_trading_date) }}
Stock: {{ example.code }}

// region 基础信息:
{{ example.base_info }}
// endregion

// region 截止到 {{ example.cur_trading_date }} 晚上 9:00 的新闻信息:
{{ example.news }}
// endregion

// region 技术面数据:
{{ example.price }}
// endregion

// region 宏观环境:
{{ example.macro }}
// endregion

// region 基本面数据:
{{ example.fundamental }}
// endregion

{%- if example.olhcv_price %}
// region 价格:
{{ example.olhcv_price }}
// endregion
{%- endif %}

{%- if example.stock_news %}
// region 截止到 {{ example.next_trading_date }} 上午 9:30 开盘的新闻信息:
{%- for item in stock_news %}
编号: {{ loop.index0 }}
{{ item }}
{%- endfor %}
// endregion
{%- endif %}
```

Figure 17: The prompt template for stock movement prediction.

Current Trading Date: 2025-05-30(星期五)
Next Trading Date: 2025-06-03(星期二)
Stock: 000555.SZ

//region 截止到 2025-06-03 上午 9:30 开盘的新闻信息:
编号: 0

Title: 【热点直击】多重利好加持 数字货币含金量还在提升, Content: <p> 5月30日，数字货币概念股开盘即启动集体大涨模式，海联金汇、翠微股份竞价涨停，并最终以涨停报收，雄帝科技、四方精创盘中最大涨幅双双超过10%，恒宝股份、神州信息、天阳科技、汇金股份、中科金财、金证股份、京北方等概念股跟涨。</p><p> 香港推行稳定币出现重大进展</p><p> 数字货币概念强力启动得到多个利好消息刺激。5月21日，香港特别行政区立法会通过《稳定币条例草案》，以在香港设立法币稳定币发行人的发牌制度，完善对虚拟资产活动在香港的监管框架。以该条例为基础，香港将正式设立法币稳定币发行人的发牌制度。此前，京东币链科技于2024年7月入选香港金管局公布的首批“稳定币沙盒”参与者，京东进入币圈的大动作引起市场高度关注。</p><p> 花旗银行报告指出，稳定币是去中心化金融的入口，跟踪稳定币发行增长有助于确定整体数字资产环境健康状况和增长情况。稳定币可以被视为没有原生代币固有波动性的价值储存手段，可用于支付和跨境交易。作为交换媒介，稳定币支付比例正在增长，随着监管态度的明确，支付市场空间有望进一步打开。</p><p> 国海证券认为，稳定币交易相对欧美传统银行交易有着更高的效率，尤其是跨境支付方面，稳定币的转账费用较低，费用取决于网络情况（以USDT为例，通常只有几美元），而一些支付系统的手续费按金额比例收费，且费率较高。</p><p> 比特币突破11万美元</p><p> 从全球数字货币市场表现看，近期受资金面回暖、机构持续加仓等多重因素影响，以比特币为代表的加密货币整体呈现上涨趋势。5月22日，比特币价格一度突破11万美元，超过今年1月创下的高点。</p><p> 2024年12月5日，比特币价格有史以来首次突破10万美元，后在今年1月美国总统特朗普的就职创下历史新高，价格突破10.9万美元。但是，随着特朗普的关税政策影响，加密货币整体暴跌，比特币价格在4月初曾短暂跌破8万美元。</p><p> 比特币的强劲表现对包括稳定币在内的数字货币，产生了正向刺激作用。</p><p> 香港稳定币概念股：众安在线</p><p> 众安在线：去年7月，众安银行成为香港首家为稳定币发行方提供储备银行服务的数字银行，并与金管局首批沙盒参与者之一的圆币创新科技合作，其将成为首家利用众安银行储备银行服务的稳定币发行人。</p><p> 连连数字：连连数字通过多层全资子公司间接持有连连国际100%股权。连连国际与圆币科技合作稳定币在跨境支付的场景应用项目。</p><p> 渣打集团：渣打银行（香港）联合安拟集团和香港电讯成立合资公司，专注于港元稳定币发行，目前处于沙盒测试后期，目标降低大湾区汇兑成本。</p><p> 京东集团：京东币链科技已进入沙盒测试第二阶段，开发与数字人民币衔接的稳定币（JD-HKD），重点测试跨境支付和供应链金融应用。</p><p> A股稳定币概念股：海联金汇、翠微股份、雄帝科技、四方精创、恒宝股份、神州信息、天阳科技、汇金股份、中科金财、金证股份、京北方。</p><p> 青岛财经日报/首新闻记者 荣晓敏</p>

编号: 1

Title: 神州信息王永利：推动大模型金融应用需解决数据壁垒问题, Content: <p> 新浪科技讯 5月31日午间消息，近日，由神州信息主办的“数云原力2025·数智金融论坛”于西安召开。神州信息联席董事长王永利在致辞中表示，AI大模型的广泛应用正成为金融业实现新发展的重大推动力。回顾金融科技的发展脉络，一次次关键技术的革新，促使金融行业一步一步的持续发展、健康发展。今天，人工智能作为引领未来的战略性技术，已成为重塑金融行业核心竞争力的关键要素，促使行业迈入金融与科技领域深度融合，积极探索发展新机的关键时点。</p><p> 王永利称，推动大模型金融应用，需要着重解决“数据壁垒、投入产出和安全保障”三方面问题。金融科技的高质量发展，需要全行业生态协同和跨领域的资源整合。实现技术的突破，以及与场景的深度融合，需要汇聚政产学研多方智慧，开放交流共解AI转型难题，以生态合力加速技术普惠。</p><p> 论坛现场，针对金融安全标准、金融大模型联合研发、金融大模型产业生态、国产云金融新核心等内容，神州信息与产业生态伙伴发布系列成果。神州信息正式发布“乾坤”企业级数智底座白皮书。作为行业首个基于“数云融合”理念打造的企业级数智底座，“乾坤”平台具有“云原生、数字原生、AI原生”三大技术特点，可针对“企业级开发平台、企业级工艺及架构治理、企业级云原生平台和企业级AI平台”四大典型建设场景，提供一站式解决方案。</p><p> 此外，神州数码宣布自研的神州问学由AI原生赋能平台升级为企业级Agent中台，为AI规模化落地提供全栈解决方案；神州信息与青岛银行成立金融科技大模型联合实验室；由神州信息牵头，金融机构标准——“金融机构信息安全控制措施指南标准”研究正式启动。</p><p> 同时，神州信息与华为、腾讯云、阿里云、中科海光和昆仑芯等行业科技伙伴举行多项成果发布：神州信息与华为正式对外发布“金融知识问答”联合智能体；神州信息与腾讯云签署大数据产品合作框架协议；神州信息发起行业首个国产云金融核心系统联盟；神州信息发起成立行业首个AIGC大模型金融生态体系。</p>

编号: 2

Title: 数云原力2025·数智金融论坛在西安召开, Content: <p> 上证报中国证券网讯（记者 张问之）5月29日，由神州信息主办的“数云原力2025·数智金融论坛”于西安召开。论坛以“信启原力·智启未来”为主题，围绕金融AI应用、金融场景创新、金融生态协同等数智化发展关键问题展开交流。</p><p> 本次论坛邀请全国近200家商业银行、300多位科技和相关业务负责人参会。来自工商银行、民生银行、恒丰银行、北京银行、陕西农信等机构负责人分享数智化转型建设经验。多家银行相关负责人围绕“金融大模型应用”和“银行转型新机遇”等话题展开讨论。</p><p> 陕西省决策咨询委员会委员、省委金融办原一级巡视员张春明</p><p> 陕西省决策咨询委员会委员、省委金融办原一级巡视员张春明在致辞中表示，陕西依托科教优势，紧扣中央金融工作会议关于数字金融的指导思想，在数智金融领域持续探索创新，形成了一系列特色实践与显著成果。通过深化科技赋能应用、优化金融服务模式，在缓解中小企业融资难题，助力乡村振兴、绿色发展等重点领域积极作为，为全省经济高质量发展注入了强劲动力。</p><p> 陕西省委金融办、金融工委将一如既往地高度重视数智金融发展，持续完善政策支持体系，加强数字金融基础设施建设，为各类金融机构和企业在陕发展提供全方位、高质量的服务。陕西将以更加开放的格局，深化与国内外金融机构、科研院所及企业的协同合作，汇聚全球资源与创新智慧，携手打造具有国际竞争力的数智金融产业集群，抢占数字经济时代的产业制高点。</p><p> 神州信息联席董事长王永利在致辞中表示，AI大模型的广泛应用正成为金融业实现新发展的重大推动力。回顾金融科技的发展脉络，一次次关键技术的革新，促使金融行业一步一步的持续发展、健康发展。今天，人工智能作为引领未来的战略性技术，已成为重塑金融行业核心竞争力的关键要素，促使行业迈入金融与科技领域深度融合，积极探索发展新机的关键时点。推动大模型金融应用，需要着重解决“数据壁垒、投入产出和安全保障”三方面问题。</p><p> 金融科技的高质量发展，需要全行业生态协同和跨领域的资源整合。实现技术的突破，以及与场景的深度融合，需要汇聚政产学研多方智慧，开放交流共解AI转型难题，以生态合力加速技术普惠。“神州信息愿以本次论坛为契机，以AI技术为笔，以数据要素为墨，共同书写金融数字化转型的新篇章，共筑行业发展的新生态。”王永利表示。</p><p> 论坛同期，神州信息与中国金融标准研究院、中国信通院、华为、腾讯云、阿里云、中科海光、昆仑芯等产业侧相关机构达成合作，并发布“金融安全标准、区域银行新核心建设、全栈国产云金融核心、AIGC金融生态体系、大模型金融一体机、智能大数据合作”等多项行业重磅合作成果。</p>

Figure 18: Example. Part 1.1: Stock News.

编号：3

Title: 神州信息发布“乾坤”企业级数智底座白皮书, Content: <p> 上证报中国证券网讯（记者 张问之）近日，在神州信息主办的“数云原力2025·数智金融论坛”现场，针对金融安全标准、金融大模型联合研发、金融大模型产业生态、国产云金融新核心等内容，神州信息与产业生态伙伴发布系列成果。</p> <p> 其中，神州信息正式发布“乾坤”企业级数智底座白皮书。作为行业首个基于“数云融合”理念打造的企业级数智底座，“乾坤”平台具有“云原生、数字原生、AI原生”三大技术特点，可针对“企业级开发平台、企业级工艺及架构治理、企业级云原生平台和企业级AI平台”四大典型建设场景，提供一站式解决方案。</p> <p> 发布现场，神州信息副总裁徐启昌表示，神州信息“乾坤”数智化底座可以帮助金融机构实现“基础设施资源利用率提升一倍、研发效率提升30%至50%、数据利用提升60%以上、AI模型迭代速度提升10倍以上”。“乾坤”数智化底座，已经成为金融企业的新质生产力平台。</p> <p> 神州数码副总裁、CTO李刚正式宣布神州数码自研的神州问学由AI原生赋能平台升级为企业级Agent中台，为AI规模化落地提供全栈解决方案。</p> <p> 同时，神州信息与青岛银行成立金融科技大模型联合实验室。论坛现场，青岛银行信息技术部总经理韩朝丽与神州信息大行BU总经理李拥军代表双方共同签署合作协议。联合实验室的成立，将成为产业协同合作的标杆。通过技术与业务深度融合，双方将共同加速银行业大模型技术的场景化创新与落地实践。同时，实验室正式启动首个重点研究课题——对公多智能体业务助手研发，将为银行对公业务带来全新的智能化服务体验。</p> <p> 由神州信息牵头，金融国际标准——“金融机构信息安全控制措施指南标准”研究也正式启动。神州信息与邮储银行、兴业银行、中国银联、国泰海通、北京国家金融标准化研究院、中金金融认证中心、新华三集团等单位共同启动该标准研制。该标准研究将围绕金融服务机构在内部运营和外部企业交易中的特殊需求，针对金融服务领域的流程控制系统，提供基于信息安全管理的实施指导规范，旨在将ISO/IEC 27002:2022的标准内容扩展应用到金融服务领域的流程控制系统以及自动化技术的使用范围。</p>

编号：4

Title: 神州信息与华为发布“金融知识问答”联合智能体, Content: <p> 上证报中国证券网讯（记者 张问之）近日，在神州信息主办的“数云原力2025·数智金融论坛”现场，神州信息与华为、腾讯云、阿里云、中科海光和昆仑芯等行业科技伙伴举行多项成果发布。</p> <p> 神州信息与华为正式对外发布“金融知识问答”联合智能体。以神州信息FinancialMaster知识问答智能体为核心应用，基于华为昇腾一体机，金融知识问答联合智能体可为金融机构提供高效的资料查询与信息提炼能力，有效改善用户体验。</p> <p> 神州信息与腾讯云签署大数据产品合作框架协议。面对金融行业云转型建设趋势和金融大模型应用热潮，金融行业对国产数据库应用需求大幅提升。双方将加大在国产分布式数据库及相关数据产品的市场推广力度。通过联合推广、联合技术合作等方式，实现双方产业生态的深度发展，满足金融机构各类大数据建设需求。</p> <p> 同时，神州信息发起行业首个国产云金融核心系统联盟。面对金融行业新一代核心系统建设，为了满足金融机构对更高效、稳定、安全的核心应用系统建设需求，神州信息与华为云、腾讯云和阿里云，联合成立国产云金融核心联盟，通过新一代云原生金融核心+全栈国产云，为金融机构数智化夯实数字底座。</p> <p> 神州信息亦发起成立行业首个AIGC大模型金融生态体系，旨在面对金融数智化转型热潮，为金融机构提供低成本、高效率、高安全和高可用的金融大模型建设服务。神州信息与华为、腾讯云、阿里云、中科海光、昆仑芯合作，围绕核心应用、国产云等多层面，构建从基础硬件到上层应用软件，全层面的AIGC生态体系。</p>

编号：5

Title: 神州信息发起成立行业首个AIGC大模型金融生态体系, Content: <p> 本报讯(记者桂小笋)6月2日，《证券日报》记者从神州数码信息服务集团股份有限公司(以下简称“神州信息”)处获悉，在日前由神州信息主办的“数云原力2025·数智金融论坛”中，神州信息正式对外发布“乾坤”企业级数智底座白皮书。作为行业首个基于“数云融合”理念打造的企业级数智底座，“乾坤”平台具有“云原生、数字原生、AI原生”三大技术特点，可针对“企业级开发平台、企业级工艺及架构治理、企业级云原生平台和企业级AI平台”四大典型建设场景，提供一站式解决方案。</p> <p> 神州信息副总裁徐启昌表示，神州信息“乾坤”数智化底座可以帮助金融机构实现“基础设施资源利用率提升一倍、研发效率提升30%至50%、数据利用提升60%以上、AI模型迭代速度提升10倍以上”。“乾坤”数智化底座，已经成为金融企业的新质生产力平台。</p> <p> 神州信息联席董事长王永利表示：“AI大模型的广泛应用正成为金融业实现新发展的重大推动力。推动大模型金融应用，需要着重解决‘数据壁垒、投入产出和安全保障’三方面问题。”</p> <p> 王永利分析，金融科技的高质量发展，需要全行业生态协同和跨领域的资源整合。实现技术的突破，以及与场景的深度融合，需要汇聚政产学研多方智慧，开放交流共解AI转型难题，以生态合力加速技术普惠。</p> <p> 与此同时，神州信息还发起行业首个国产云金融核心系统联盟。面对金融行业新一代核心系统建设，为了满足金融机构对更高效、稳定、安全的核心应用系统建设需求，神州信息与华为云、腾讯云和阿里云，联合成立行业首个国产云金融核心联盟，通过新一代云原生金融核心+全栈国产云，为金融机构数智化夯实数字底座。同时，神州信息发起成立行业首个AIGC大模型金融生态体系。旨在面对金融数智化转型热潮，为金融机构提供低成本、高效率、高安全和高可用的金融大模型建设服务。神州信息与华为、腾讯云、阿里云、中科海光、昆仑芯合作，围绕核心应用、国产云、国产芯等多层面，构建从基础硬件到上层应用软件，全层面的AIGC生态体系。</p>

// endregion

//region 截止到 2025-05-30 晚上 9:00 的新闻信息:

2025-05-30新闻: 2025-05-30每日信息总结:

实时新闻: <神州信息在5月29日获得融资买入6044.86万元，占当日流入资金比例为12.65%。当前融资余额为5.15亿元，占流通市值的4.39%，超过历史60%分位水平。融券方面，5月29日融券偿还0股，融券卖出1.95万股，融券余额为174.75万元，低于历史50%分位水平。>

分析师观点: <融资买入的增加表明市场对神州信息的信心有所提升，投资者情绪偏向买方，可能反映出对公司未来发展的乐观预期。尽管融券卖出量较小，但融券余额的下降显示出市场对该股的谨慎态度。>/n2025-05-31新闻: 2025-05-31每日信息总结:

实时新闻:

1. 神州信息在“数云原力2025·数智金融论坛”上发布了“乾坤”企业级数智底座白皮书，强调其在金融机构基础设施和研发效率方面的提升。
2. 神州信息与青岛银行成立金融科技大模型联合实验室，旨在推动银行业大模型技术的应用。
3. 神州信息牵头启动金融国际标准“金融机构信息安全控制措施指南标准”的研究，涉及多家金融机构的合作。
4. 数字货币概念股集体上涨，神州信息作为其中一员，受益于香港稳定币政策的推进。
5. 神州信息与华为等科技伙伴发布“金融知识问答”联合智能体，提升金融机构的信息查询能力。

分析师观点:

1. AI大模型的应用被视为金融行业发展的重要推动力，需解决数据壁垒等问题以实现高质量发展。
2. 数字货币市场的积极动态和政策支持可能进一步提升相关概念股的市场表现，神州信息在此领域的布局将增强其竞争力。

Figure 19: Example. Part 1.2: Stock News.

```

//region 价格:
### 000555.SZ price infos
|index|date|open|high|low|close|volume|turnover_rate|pct_change|lowerband|middleband|upperband|
|---|---|---|---|---|---|---|---|---|---|---|---|
|77|2025-04-30|10.6500|10.8200|10.5800|10.7000|13307055.0000|1.3688|0.0113|9.9833|10.7405|11.4977|
|78|2025-05-06|10.8800|11.2800|10.8600|11.2700|27991423.0000|2.8792|0.0533|10.0506|10.7195|11.3884|
|79|2025-05-07|11.4500|11.5200|11.0400|11.1300|25053400.0000|2.5770|-0.0124|10.0850|10.7055|11.3260|
|80|2025-05-08|11.1000|11.3800|11.0900|11.3500|21471050.0000|2.2030|0.0198|10.1126|10.7595|11.4064|
|81|2025-05-09|11.3500|11.5200|11.1600|11.2200|2180221.0000|2.3843|-0.0115|10.2202|10.8165|11.4128|
|82|2025-05-12|11.3200|11.4100|11.2300|11.3700|19581446.0000|2.0142|0.0134|10.2767|10.8700|11.4633|
|83|2025-05-13|11.4900|11.5400|11.1800|11.2300|18089400.0000|1.8607|-0.0123|10.3005|10.9010|11.5015|
|84|2025-05-14|11.2000|11.4200|11.1100|11.3100|17626700.0000|1.8131|0.0071|10.3064|10.9285|11.5506|
|85|2025-05-15|11.2500|11.3100|10.9600|10.9900|16474300.0000|1.6946|-0.0283|10.3132|10.9350|11.5568|
|86|2025-05-16|10.9200|11.0700|10.8900|10.9500|10325300.0000|1.0621|-0.0036|10.3301|10.9455|11.5609|
|87|2025-05-19|11.1000|11.2200|10.8500|11.1100|16296246.0000|1.6762|0.0146|10.3923|10.9760|11.5597|
|88|2025-05-20|11.1100|11.4200|11.0000|11.2400|18187500.0000|1.8708|0.0117|10.4795|11.0165|11.5535|
|89|2025-05-21|11.2400|11.2400|11.0400|11.1600|2196250.0000|1.2545|-0.0071|10.5595|11.0485|11.5375|
|90|2025-05-22|11.0900|11.1800|10.9200|10.9200|13373250.0000|1.3756|-0.0215|10.5639|11.0505|11.5371|
|91|2025-05-23|11.0200|11.5500|10.9600|11.0400|43082619.0000|4.4315|0.0101|10.5671|11.0455|11.5303|
|92|2025-05-26|10.9300|11.1800|10.9300|11.1000|23766388.0000|2.4446|0.0054|10.5598|10.9430|11.5262|
|93|2025-05-27|11.1400|11.1500|10.9100|11.0700|16139100.0000|1.6601|-0.0027|10.5880|11.0570|11.5260|
|94|2025-05-28|11.0200|11.1200|10.9200|10.9600|14931917.0000|1.5359|-0.0099|10.5994|11.0620|11.5246|
|95|2025-05-29|10.9600|12.0600|10.9200|12.0600|69115467.0000|7.1093|0.1004|10.5587|11.1380|11.7173|
|96|2025-05-30|12.7900|12.8600|12.0100|12.0900|103139544.0000|10.6090|0.0025|10.5564|11.2135|11.8706|
### 300079.SZ price infos (rank 1 similar to 000555.SZ)
|index|date|open|high|low|close|volume|turnover_rate|pct_change|lowerband|middleband|upperband|
|---|---|---|---|---|---|---|---|---|---|---|---|
|77|2025-04-30|5.2400|5.3400|5.2200|5.2600|38795267.0000|3.0274|0.0057|4.4328|5.0760|5.7192|
|78|2025-05-06|5.3200|5.4900|5.3100|5.4900|57199858.0000|4.4636|0.0437|4.4271|5.0805|5.7339|
|79|2025-05-07|5.5500|5.6000|5.3700|5.4300|58119984.0000|4.5354|-0.0109|4.4247|5.0830|5.7413|
|80|2025-05-08|5.4200|5.4900|5.3900|5.4700|39975485.0000|3.1195|0.0074|4.5629|5.1405|5.7181|
|81|2025-05-09|5.4500|5.4600|5.3100|5.3400|32286401.0000|2.5975|-0.0238|4.6617|5.1795|5.6973|
|82|2025-05-12|5.3800|5.4100|5.3400|5.4002|26762700.0000|2.0884|0.0124|4.7279|5.2125|5.6971|
|83|2025-05-13|5.4800|5.5100|5.3400|5.3600|28349215.0000|2.2122|-0.0074|4.7730|5.2365|5.7000|
|84|2025-05-14|5.3300|5.4300|5.3000|5.3900|30913111.0000|2.4123|0.0056|4.8200|5.2610|5.7020|
|85|2025-05-15|5.3600|5.3800|5.2300|5.2500|24995500.0000|1.9505|-0.0260|4.8526|5.2745|5.6964|
|86|2025-05-16|5.2400|5.2900|5.2000|5.2500|17641898.0000|1.3767|0.0004|4.8660|5.2875|5.6890|
|87|2025-05-19|5.2400|5.3000|5.1700|5.2800|21042500.0000|1.6420|0.0057|4.9491|5.3070|5.6649|
|88|2025-05-20|5.2700|5.3800|5.2200|5.3800|25037225.0000|1.9538|0.0189|5.0248|5.3310|5.6372|
|89|2025-05-21|5.4400|5.5300|5.3400|5.3900|56098830.0000|4.3768|0.0019|5.0974|5.3525|5.6076|
|90|2025-05-22|5.3600|5.4200|5.2600|5.2600|35300931.0000|2.7554|-0.0241|5.1235|5.3595|5.5957|
|91|2025-05-23|5.3400|5.4000|5.1600|5.1600|35569924.0000|2.7757|-0.0190|5.0959|5.3510|5.6025|
|92|2025-05-26|5.1400|5.2500|5.1400|5.2400|23263219.0000|1.8153|0.0155|5.1486|5.3255|5.5024|
|93|2025-05-27|5.2700|5.2700|5.1700|5.2400|19213460.0000|1.4993|0.0000|5.1399|5.3180|5.4961|
|94|2025-05-28|5.2500|5.2800|5.1900|5.2300|18887298.0000|1.4739|-0.0019|5.1319|5.3140|5.4961|
|95|2025-05-29|5.2300|5.4900|5.2100|5.4900|60921610.0000|4.7544|0.0497|5.1339|5.3270|5.5201|
|96|2025-05-30|5.4500|5.5300|5.3400|5.3800|38312377.0000|2.9897|-0.0200|5.1455|5.3345|5.5235|
### 300479.SZ price infos (rank 2 similar to 000555.SZ)
|index|date|open|high|low|close|volume|turnover_rate|pct_change|lowerband|middleband|upperband|
|---|---|---|---|---|---|---|---|---|---|---|---|
|77|2025-04-30|19.3100|19.8500|19.1000|19.6600|6985000.0000|3.5470|0.0299|17.6005|19.3575|21.1145|
|78|2025-05-06|19.9900|20.2500|19.8100|20.1400|7816845.0000|3.9694|0.0244|17.6925|19.3115|20.9305|
|79|2025-05-07|20.4000|20.5500|19.9300|20.1500|7743299.0000|3.9320|0.0005|17.7607|19.2780|20.7953|
|80|2025-05-08|19.9800|20.4800|19.9800|20.4500|6243600.0000|3.1705|0.0149|18.2142|19.4455|20.6768|
|81|2025-05-09|20.3500|20.3600|19.9200|19.9600|5139300.0000|2.6097|-0.0240|18.6053|19.5585|20.5117|
|82|2025-05-12|20.1500|20.3700|20.0400|20.3300|4720400.0000|2.3970|0.0185|18.7457|19.6455|20.5453|
|83|2025-05-13|20.6600|20.6700|19.9000|19.9500|5478600.0000|2.7820|-0.0187|18.8154|19.6880|20.5606|
|84|2025-05-14|19.9600|20.3300|19.7000|20.1600|6406300.0000|3.2531|0.0105|18.8529|19.7305|20.6081|
|85|2025-05-15|20.3300|20.3500|19.3500|19.3800|6546000.0000|3.3241|-0.0387|18.8361|19.7230|20.6099|
|86|2025-05-16|19.3800|19.5800|19.1300|19.4000|4054599.0000|2.0589|0.010|18.8337|19.7220|20.6103|
|87|2025-05-19|19.4000|19.7000|19.1200|19.6500|40404000.0000|2.0517|0.0129|18.8463|19.7295|20.6127|
|88|2025-05-20|19.5900|19.9800|19.3500|19.9100|4695700.0000|2.3845|0.0132|18.8979|19.7605|20.6231|
|89|2025-05-21|19.8100|20.5200|19.4500|20.1000|11138149.0000|5.6559|0.0050|18.9740|19.8015|20.6290|
|90|2025-05-22|19.9500|20.4300|19.6300|20.3600|11830800.0000|6.0077|0.0175|18.9654|19.8280|20.6906|
|91|2025-05-23|20.2400|20.2600|19.3000|19.3600|11955500.0000|7.0710|-0.0491|18.9142|19.7975|20.6808|
|92|2025-05-26|19.5900|20.4800|19.4500|19.6200|9397557.0000|4.7721|0.0134|18.8996|19.7715|20.6434|
|93|2025-05-27|19.4500|19.6800|19.1300|19.2000|7293700.0000|3.7037|-0.0214|18.8587|19.7575|20.6563|
|94|2025-05-28|19.2500|19.6200|19.0400|19.1900|6843854.0000|3.4753|-0.0005|18.8058|19.7365|20.6672|
|95|2025-05-29|19.2500|20.3800|19.1100|20.3400|13791300.0000|7.0032|0.0599|18.9650|19.8155|20.6660|
|96|2025-05-30|20.2100|20.5800|19.9600|20.0900|10359145.0000|5.2604|-0.0123|19.0761|19.8655|20.6549|
### 300541.SZ price infos (rank 3 similar to 000555.SZ)
|index|date|open|high|low|close|volume|turnover_rate|pct_change|lowerband|middleband|upperband|
|---|---|---|---|---|---|---|---|---|---|---|---|
|77|2025-04-30|13.1000|13.6000|12.9700|13.4000|48074617.0000|12.4359|0.0276|10.7158|12.1235|13.5312|
|78|2025-05-06|13.6900|14.2600|13.5300|14.2400|4478106.0000|16.6792|0.0627|10.5460|12.1800|13.8140|
|79|2025-05-07|14.3000|14.3800|13.6500|13.8700|57774792.0000|14.9452|-0.0260|10.4577|12.2340|14.0103|
|80|2025-05-08|13.8100|14.0200|13.7900|13.8600|34989228.0000|9.0510|-0.0007|10.6733|12.4010|14.1287|
|81|2025-05-09|13.8000|14.2500|13.6800|13.7800|40699099.0000|10.5280|-0.0058|10.8431|12.5410|14.2389|
|82|2025-05-12|13.9500|14.1000|13.7400|14.0200|30871359.0000|7.9858|0.0174|10.9308|12.6690|14.4072|
|83|2025-05-13|14.1300|14.2800|13.8200|13.8500|27313896.0000|7.0655|-0.0121|10.9834|12.7630|14.5426|
|84|2025-05-14|14.2800|13.8700|14.2600|40205397.0000|10.4003|0.0296|11.0265|12.8785|14.7305|
|85|2025-05-15|14.0800|14.1100|13.5800|13.6800|32079108.0000|8.2982|-0.0407|11.1103|12.9575|14.8047|
|86|2025-05-16|13.6000|13.8200|13.4900|13.6900|1874531.0000|4.6755|0.0007|11.2250|13.0425|14.8600|
|87|2025-05-19|13.6500|13.8200|13.5500|13.7400|15513988.0000|4.0132|0.0037|11.4001|13.1415|14.8829|
|88|2025-05-20|13.6600|13.7500|13.4100|13.6200|19220998.0000|4.9721|-0.0087|11.6418|13.2425|14.8432|
|89|2025-05-21|13.7000|14.1300|13.5500|13.8800|3533270.0000|9.1413|0.0191|11.8765|13.3480|14.8195|
|90|2025-05-22|13.7100|13.8700|13.4300|13.4700|25255268.0000|6.5330|-0.0295|12.0608|13.4165|14.7722|
|91|2025-05-23|13.4000|13.5200|12.9300|12.9300|26857256.0000|6.9474|-0.0401|12.1381|13.4405|14.7429|
|92|2025-05-26|12.9900|13.2500|12.9500|13.2400|15712590.0000|4.0645|0.0240|12.2439|13.4775|14.7111|
|93|2025-05-27|13.2000|13.2800|13.0600|13.1600|11762310.0000|3.0427|-0.0060|12.4804|13.5340|14.5876|
|94|2025-05-28|13.2200|13.3900|13.0600|13.1700|1641706.0000|4.2531|0.0008|12.8171|13.5960|14.3749|
|95|2025-05-29|13.1700|14.2600|13.1300|14.2600|62903024.0000|16.2717|0.0828|12.8750|13.6580|14.4410|
|96|2025-05-30|13.8100|14.0300|13.7100|13.8600|40900299.0000|10.5801|-0.0261|12.9654|13.6990|14.4326|
// endregion

```

Figure 20: Example. Part 2: Stock Price Info of Current Stock and Top-3 Similar Stocks.

```

//region 宏观环境:
# A股市场分析报告

## 1. 市场态势分析

### 资金流向与强度分析
近期A股市场表现出资金流入的积极态势，尤其是医疗健康和科技板块。根据龙虎榜数据，资金净流入前十的板块中，仿制药一致性评价和创新药板块分别吸引了4.33亿元和1.77亿元的资金流入，显示出市场对医药行业的强烈关注。

### 市场参与主体行为特征
从龙虎榜数据来看，游资和机构的交易行为活跃，尤其是在医疗和科技领域。游资在多个个股中表现出明显的买入倾向，尤其是华森制药（002907.SZ），其连续涨停的表现吸引了大量资金关注。

### 市场情绪指标与波动特征
市场情绪指标显示出乐观情绪，尤其是在医疗健康板块。根据市场情绪指数，当前市场情绪处于高位，反映出投资者对短期市场的信心增强。

### 近期市场主要矛盾和焦点
当前市场的主要矛盾集中在经济复苏与外部环境的不确定性之间。尽管国内政策支持力度加大，但全球经济形势和中美贸易关系仍然是市场关注的焦点。

## 2. 板块轮动分析

### 强势板块梳理与未来持续性判断
近期表现强势的板块包括仿制药一致性评价、创新药和流感板块。仿制药一致性评价板块在过去三日内涨幅达到1.16%，资金净流入4.33亿元，显示出其未来的持续性。

### 资金布局方向和规模
资金布局主要集中在医疗健康和科技板块，尤其是仿制药和创新药。资金流入的规模和强度表明市场对这些板块的信心。

### 板块轮动节奏与特征
板块轮动节奏较快，医疗健康板块在短期内表现突出，而科技板块则在中长期内有望继续吸引资金。市场情绪的变化可能导致板块间的快速切换。

### 龙头个股表现分析
华森制药（002907.SZ）作为仿制药一致性评价板块的龙头个股，近期表现强劲，连续涨停，吸引了大量资金流入。其未来表现值得关注。

## 3. 市场机会与风险分析

### 短期市场机会识别
短期内，医疗健康板块仍然是市场的主要机会，尤其是仿制药和创新药相关个股。建议投资者关注相关龙头个股的表现。

### 潜在风险因素预警
潜在风险主要来自于外部经济环境的不确定性，尤其是中美贸易关系的变化可能对市场情绪产生影响。此外，市场情绪的波动也可能导致短期内的剧烈调整。

### 交易时机与规模建议
建议投资者在医疗健康板块的回调时适度加仓，关注龙头个股的表现。交易规模应根据市场情绪和资金流向进行动态调整。

### 止盈止损参考位
建议设定止盈位在近期高点附近，止损位则可设定在个股支撑位下方，以控制风险。

## 4. 外围市场联动性分析

### 全球主要市场表现
近期全球主要市场表现分化，欧美市场在经济数据改善的背景下有所反弹，但仍面临通胀压力和利率上升的挑战。

### 外围市场风险溢出效应
美国加息预期和通胀压力可能对全球市场产生溢出效应，影响A股市场的资金流动和投资者情绪。

### 全球流动性状况
全球流动性仍然宽松，但美联储的加息政策可能导致流动性收紧，需密切关注其对市场的影响。

### 重要经济体政策动向
美国的减税政策和关税政策仍在不断调整，可能对全球经济和市场产生深远影响。投资者需关注相关政策的变化及其对市场的影响。
---  

本报告基于多维度数据源的分析，旨在为短线交易决策提供参考。建议投资者密切关注市场动态，灵活调整投资策略。  

// endregion

```

Figure 21: Example. Part 3: Macroeconomic Indicators Report.

```

//region 基本面数据:
#### 一、公司概述

1. **所属行业与主营业务**
   - 神州信息(股票代码: 000555.SZ)主要从事技术服务、农业信息化、应用软件开发、金融专用设备相关业务及集成解决方案。公司在量子科技、数字货币、网络安全等多个领域具有业务布局。

2. **行业地位**
   - 神州信息在量子通信、数字货币及网络安全等领域处于行业前列，尤其在数字金融和信息安全方面，积极参与国家信息安全战略，具备较强的市场竞争力。

3. **业务特点**
   - 公司业务涵盖多个高科技领域，尤其在量子通信和数字货币方面具有较强的技术积累和市场应用。与华为的合作进一步增强了其在金融科技和物联网领域的竞争力。

#### 二、最新经营态势

1. **收入利润分析**
   - 截至2024年9月30日，神州信息的营业总收入为66.80亿元，较上年同期的66.18亿元略有增长(0.94%)。然而，净利润为-1.80亿元，显示出公司在盈利能力方面面临较大压力。

2. **盈利能力分析**
   - 销售毛利率为13.61%，销售净利率为-2.70%，表明公司在成本控制和盈利能力方面存在问题。营业总成本占营业总收入的比例高达103.83%，显示出公司在运营效率上存在较大挑战。

3. **经营效率分析**
   - 经营活动产生的现金流量净额为-7.45亿元，表明公司在经营活动中现金流出大于流入，反映出经营效率低下。应收账款周转率为2.67次，存货周转率为2.36次，显示出公司在资产管理方面的效率较低。

4. **变动趋势分析**
   - 归属母公司股东的净利润同比下降236.41%，显示出公司在盈利能力方面的显著恶化。预计2024年将出现净利润亏损，主要由于行业客户数字化转型进程放缓及市场竞争加剧。

#### 三、财务状况评估

1. **资产负债分析**
   - 截至2024年9月30日，公司的总资产为126.69亿元，负债合计为65.90亿元，资产负债率为52.02%。股东权益合计为60.79亿元，表明公司在财务结构上相对稳健，但负债水平仍需关注。

2. **现金流分析**
   - 经营活动产生的现金流量净额为-7.45亿元，显示出公司在现金流管理方面的压力。尽管销售商品、提供劳务收到的现金为78.45亿元，但经营活动现金流出大于流入，反映出公司在运营中的现金流问题。

3. **资本结构分析**
   - 公司的流动比率为1.53，速动比率为1.01，表明公司在短期偿债能力上相对良好。然而，保守速动比率仅为0.65，显示出在流动性管理上存在一定风险。

4. **风险指标分析**
   - 公司的已获利息倍数为-6.94，表明公司在偿还利息方面面临较大压力，可能影响未来的融资能力和财务稳定性。

#### 四、估值水平分析

1. **当前估值水平**
   - 神州信息的市盈率为-67.6942，市净率为1.7637，表明市场对公司未来盈利能力的预期较低。每股净资产为6.1858元，反映出公司在资产方面的相对价值。

2. **估值合理性分析**
   - 由于公司预计在2024年出现净利润亏损，当前的估值水平可能反映出市场对其未来盈利能力的悲观预期。市净率相对较高，可能意味着市场对公司资产的价值认可度不足。

#### 五、综合结论

1. **主要经营优势**
   - 神州信息在量子科技、数字货币和网络安全等领域具有技术积累和市场应用优势，尤其与华为的合作为其提供了强大的市场支持。

2. **潜在风险提示**
   - 公司面临的主要风险包括市场竞争加剧、客户数字化转型进程放缓、现金流管理不善及高负债水平等，这些因素可能影响公司的持续盈利能力及财务稳定性。

3. **基本面趋势判断**
   - 由于公司预计在2024年将出现净利润亏损，且经营效率和现金流状况不佳，基本面趋势显示出一定的下行压力。未来需关注公司在业务结构转型和市场开拓方面的进展，以期改善经营状况。

```

Figure 22: Example. Part 4.1: Stock Fundamentals Report.

最新估值：神州信息(股票代码：000555.SZ)最新财务估值信息
企业估值：

当前，公司总股本为9.76亿，发行总股本为0.74亿，流通股本为9.72亿，每股净资产为5.6886，ps市销率为1.366，pcf市现率为-1.144，市净率为2.1253，市盈率为-31.7323。

截至报告期2025-03-31，公司财务分析指标如下：

{'归属母公司股东的净利润/报告期末总股本': '-0.10元/股', '归属于普通股股东的扣除非经常性损益后的净利': '-0.10元/股', '股东权益周转率': '0.39%', '人力投入回报率rop': '-14.76%', '每股营业总收入': '2.21元/股', '每股资产公允价': '2.39元/股', '每股盈余公积': '0.05元/股', '每股未分配利润': '2.32元/股', '每股留存收益': '2.37元/股', '每股现金流量净额': '-0.48元/股', '每股息税前利润': '-0.13元/股', '每股企业自由现金流流量': '0.15元/股', '每股股东自由现金流量': '2.39元/股', '平均净资产收益率': '-1.66%', '扣除非经常损益后的平均净资产收益率': '-1.79%', '净资产收益率roe—增发条件': '-1.79%', '总资产净利率roa': '-0.89%', '总资产报酬率roa': '-1.03%', '投入资本回报率roi': '-1.30%', '年化净资产收益率': '-6.64%', '年化总资产报酬率': '-4.12%', '年化总资产净利率': '-3.58%', '销售毛利率': '12.76%', '销售净利率': '-5.28%', '销售成本率': '87.24%', '期间费用率': '12.97%', '净利润/营业总收入': '-5.28%', '营业利润/营业总收入': '-6.48%', '息税前利润/营业总收入': '-6.08%', '营业总成本/营业总收入': '107.03%', '营业费用/营业总收入': '3.32%', '管理费用/营业总收入': '9.32%', '(管理费用+研发费用)/营业总收入': '0.33%', '资产减值损失/营业总收入': '2.16%', '经营活动净收益/利润总额': '0%', '价值变动净收益/利润总额': '0%', '营业外收支净额/利润总额': '0%', '所得税/利润总额': '0%', '扣除非经常损益后的净利润/净利润': '107.88%', '销售商品提供劳务收到的现金/营业收入': '94.41%', '经营活动产生的现金流量净额/营业收入': '-119.37%', '经营活动产生的现金流量净额/经营活动净收益': '0%', '资本支出/折旧和摊销': '0%', '资产负债率': '58.38%', '权益乘数': '2.40%', '流动资产比率(流动资产/总资产)': '81.80%', '非流动资产/总资产': '18.20%', '有形资产/总资产': '30.30%', '归属母公司股东的权益/全部投入资本': '67.82%', '带息债务/全部投入资本': '32.18%', '流动负债比率(流动负债/负债合计)': '99.07%', '非流动负债比率(非流动负债/负债合计)': '0.93%', '流动比率': '1.41', '速动比率': '0.85', '保守速动比率': '0.53', '产权比率(负债合计/归属母公司股东的权益)': '1.42', '归属母公司股东的权益/负债合计': '0.70', '归属母公司的股东权益/带息债务': '2.11', '有形资产/负债合计': '0.52', '有形资产/带息债务': '1.55', '有形资产/净债务': '3.84', '息税折旧摊销前利润/负债合计': '-0.02', '经营活动产生的现金流量净额/负债合计': '-0.33', '经营活动产生的现金流量净额/带息债务': '-0.98', '经营活动产生的现金流量净额/流动负债': '-0.33', '经营活动产生的现金流量净额/净债务': '-2.42', '已获利息倍数(ebit/利息费用)': '18.32', '长期债务与营运资金比率': '0.02', '应收账款及应收票据周周转率': '1.04次', '无形资产周转率': '12.63次', '营业周期': '241.50天', '存货周转天数': '156.82天', '应收账款周转天数': '84.68天', '存货周转率': '0.57次', '应收账款周转率': '1.06次', '流动资产周转率': '0.21次', '固定资产周转率': '5.53次', '总资产周转率': '0.17次', '基本每股收益同比增长率': '-65.30%', '稀释每股收益同比增长率': '-65.30%', '每股经营活动产生的现金流量净额同比增长率': '-80.93%', '营业总收入同比增长率': '22.13%', '营业收入同比增长率': '22.13%', '投资收益同比增长率': '9.04%', '营业利润同比增长率': '-73.33%', '利润总额同比增长率': '-64.90%', '归属母公司股东的净利润同比增长率': '-64.63%', '扣除非经常损益后的归属母公司股东的净利润同比增长率': '-74.68%', '经营活动产生的现金流量净额同比增长率': '-80.93%', '全面摊薄净资产收益率同比增长率': '-0.76%', '每股净资产相对年初增长率': '-1.65%', '资产总计相对年初增长率': '12.75%', '归属母公司的股东权益相对年初增长率': '-1.65%', '营业成本同比增长率': '26.13%', '毛利同比增长率': '0.36%', '应付账款周转率': '0.90次', '应付账款周转天数': '100.27天', '现金循环周期': '141.24天', '营运资金周转率': '0.64次', '营运资金周转天数': '140.32天', '净资产周转率': '0.38', '净资产周转天数': '236.66', '扣除非经常性损益后的基本每股收益同比增长率': '-75.29%', '已动用资本回报率roc'e': '-0.02', '研发费用率': '6.28%', '管理费用率': '3.04%', '销售费用率': '3.32%'}

对于公司报告期2025-03-31的财务审计情况如下

{'归属母公司股东的净利润/报告期期末总股本': '-0.10', '归属于普通股股东的扣除非经常性损益后的净利': '-0.10', '股东权益周转率': '0.39', '人力投入回报率rop': '-14.76', '每股营业总收入': '2.21', '每股营业收入': '2.21', '每股资本公积': '2.39', '每股盈余公积': '0.05', '每股未分配利润': '2.32', '每股留存收益': '2.37', '每股市现金流量净额': '-0.48', '每股息税前利润': '-0.13', '每股企业自由现金流量': '0.15', '每股股东自由现金流量': '2.39', '平均净资产收益率': '-1.66', '扣除非经常损益后的平均净资产收益率': '-1.79', '净资产收益率roe—增长率': '-1.79', '总资产净利率roa': '-0.89', '总资产报酬率roa': '-1.03', '投入资本回报率roic': '-1.30', '年化净资产收益率': '-6.64', '年化总资产报酬率': '-4.12', '年化总资产净利率': '-3.58', '销售毛利率': '12.76', '销售净利率': '-5.28', '销售成本率': '87.24', '期间费用率': '12.97', '净利润/营业总收入': '-5.28', '营业利润/营业总收入': '-6.48', '息税前利润/营业总收入': '-6.08', '营业总成本/营业总收入': '107.03', '营业费用/营业总收入': '3.32', '管理费用/营业总收入': '9.32', '(管理费用和研发费用)/营业总收入': '0.33', '资产减值损失/营业总收入': '2.16', '扣除非经常损益后的净利润/净利润': '107.88', '销售商品提供劳务收到的现金/营业收入': '94.41', '经营活动产生的现金流量净额/营业收入': '-119.37', '资产负债率': '58.38', '权益乘数': '2.40', '流动资产比率(流动资产/总资产)': '81.80', '非流动资产/总资产': '18.20', '有形资产/总资产': '30.30', '归属母公司股东的权益/全部投入资本': '67.82', '带息债务/全部投入资本': '32.18', '流动负债比率(流动负债/负债合计)': '99.07', '非流动负债比率(非流动负债/负债合计)': '0.93', '流动比率': '1.41', '速动比率': '0.85', '保守速动比率': '0.53', '产权比率(负债合计/归属母公司股东的权益)': '1.42', '归属母公司股东的权益/负债合计': '0.70', '归属母公司的股东权益/带息债务': '2.11', '有形资产/负债合计': '0.52', '有形资产/带息债务': '1.55', '有形资产/净债务': '3.84', '息税折旧摊销前利润/负债合计': '-0.02', '经营活动产生的现金流量净额/负债合计': '-0.33', '经营活动产生的现金流量净额/带息债务': '-0.98', '经营活动产生的现金流量净额/流动负债': '-0.33', '经营活动产生的现金流量净额/净债务': '-2.42', '已获利息倍数(ebit/利息费用)': '-18.32', '长期债务与营运资金比率': '0.02', '应收账款及应收票据周转率': '1.04', '无形资产周转率': '12.63', '营业周期': '241.50', '存货周转天数': '156.82', '应收账款周转天数': '84.68', '存货周转率': '0.57', '应收账款周转率': '1.06', '流动资产周转率': '0.21', '固定资产周转率': '5.53', '总资产周转率': '0.17', '基本每股收益同比增长率': '-65.30', '稀释每股收益同比增长率': '-65.30', '每股经营活动产生的现金流量净额同比增长率': '-80.93', '营业总收入同比增长率': '22.13', '营业收入同比增长率': '22.13', '投资收益同比增长率': '9.04', '营业利润同比增长率': '-73.33', '利润总额同比增长率': '-64.90', '归属母公司的股东的净利润同比增长率': '-64.63', '扣除非经常损益后的归属母公司的股东的净利润': '-74.68', '经营活动产生的现金流量净额同比增长率': '-80.93', '全面摊薄净资产收益率同比增长率': '-0.76', '每股净资产相对年初增长率': '-1.65', '资产总计相对年初增长率': '12.75', '归属母公司的股东权益相对年初增长率': '-1.65', '营业成本同比增长率': '26.13', '毛利同比增长率': '0.36', '应付账款周转率': '0.90', '应付账款周转天数': '100.27', '现金循环周期': '141.24', '营运资金周转率': '0.64', '营运资金周转天数': '140.32', '净资产周转率': '0.38', '净资产周转天数': '236.66', '扣除非经常性损益后的基本每股收益同比增长率': '-75.29', '已动用资本回报率roce': '-0.02', '研发费用率': '6.28', '管理费用率': '3.04', '销售费用率': '3.32'}

```
// endregion
```

Figure 23: Example. Part 4.2: Stock Fundamentals Report

```

//region 基础信息:
神州信息(股票代码: 000555.SZ)的基本信息如下:
一、主营业务: 技术服务、农业信息化、应用软件开发、金融专用设备相关业务及集成解决方案。
二、所属行业: ['元件 III', '终端设备', 'IT服务 III', 'IT服务']
三、所属概念(按当前相关性排序, 从高往低): 1.数字货币: 公司领先市场发布区块链平台Sm@rtGAS和数字货币 (DCEP) 解决方案。数字钱包系统作为承接数字货币、发展数字货币相关业务的重要系统, 公司目前已在建设银行、广发银行、北京银行实现落地。
2.跨境支付(CIPS): 根据2025年1月15日互动易: 公司参与中国现代化支付体系建设多年, 在支付业务方面, 公司主要涉及大额实时支付系统、小额批量支付系统等支付清算业务, 有上百家银行案例。公司拥有CIPS相关技术储备, 在跨境、外币相关系统方面已有项目落地, 具备提供人民银行现代化支付体系相关系统完整的解决方案能力。
3.互联网金融: 2023年9月18日互动易: 公司深耕数字金融领域, 以未来银行整体架构规划ModelB@ank为指引, 形成了以“核心应用、云计算、数据智能、智能银行、数字金融、信贷管理、风险管理及科技监管”为矩阵的八大产品族, 并持续探索创新。公司助力金融客户建设手机银行、网上银行、视频银行、移动展业平台等数字化渠道, 利用积分、权益、商城等数字化营销手段实现获客和活客, 并通过互联网金融中台和开放平台输出金融能力下沉各场景。
4.财税数字化: 2024年9月10日互动易, 2024年上半年, 在财税数字化领域, 公司基于多年参与金税工程的业务和技术优势, 深度参与金税三期和四期项目建设, 在区块链新技术探索、国家信息工程网络安全防护响应、混合云应用研究、大数据技术应用、不同部委数据共享等多个方面, 持续推动财税数字化建设。上半年, 公司中标签约国家税务总局多边税务数据服务平台升级完善及运维项目, 参与到北京、浙江、江苏、山西等21个省级税务局的数据交换共享、管理决策支持核算与分析系统、智慧税务办公平台建设等多个项目中, 同时积极推进多个地市级税务局的信息系统建设。
5.华为鲲鹏
6.区块链: 2019年10月28日公司互动平台回复公司在2016年开始在区块链方面进行跟踪和研究, 公司还专门成立区块链研究院, 拥有专门的团队研究底层技术, 并已经掌握了区块链的核心技术, 坚持自主创新。
7.信创
8.鸿蒙概念: 2024年5月11日互动易: 公司与HUAWEI合作覆盖金融科技、智能服务以及物联网等多领域的行业应用解决方案及运维服务, 已经成为HUAWEI在中国行业解决方案领域最大的合作伙伴之一。2024年4月, 公司成功完成“fPaaS全渠道技术开发平台”的HarmonyOS NEXT原生基线版本的适配开发工作, 并正式对外发布。作为HUAWEI的战略合作伙伴, 公司不仅是HUAWEI鸿蒙首批生态伙伴, 更是率先完成面向金融机构的鸿蒙原生迁移应用平台——“fPaaS全渠道技术开发平台”, 帮助金融机构完成面向掌银、网银、移动信贷、移动展业等多场景的应用搭建。
9.网络安全: 公司拥有“信息安全”相关业务, 是国家信息安全战略的重要参与者和推动者。公司具有“国家安全可靠计算机信息系统集成重点企业”资质, 联合成立安全可靠产业联盟“龙安联盟”, 积极推动我国各个行业的信息安全水平。
10.数据要素: 2025年4月1日公告: 公司以数据要素为驱动, 构建了全域数据管理平台, 为金融客户提供全面数据体系咨询和方案落地。形成了数据汇聚、数据治理、数据资产、数据应用、数据服务等数据解决方案。
11.AI智能体: 2025年02月18日官微: 作为国内领先的金融科技企业, 神州信息完成旗下重要集成服务解决方案“神州灵境”与DeepSeek大模型R1&V3的接入, 基于DeepSeek大模型在多任务、多模态和多语言以及知识问答领域的优势, 结合多行业场景智能化升级和应用调优, 率先推出“运维智能体、办公智能体和业务管理智能体”, 通过一站式应用部署, 让客户直观感受到大模型带来的智能变化。
12.ERP概念
13.数字乡村: 主营技术服务、农业信息化相关业务及集成解决方案, 为我国金融、农业等国民经济重点行业提供技术服务、应用软件开发以及行业云建设及运营等产品和服务;
14.腾讯概念: 2024年11月7日互动易: 公司作为腾讯生态合作伙伴受邀参会, 基于腾讯云在云计算、数据库、大数据平台等领域基础产品能力, 叠加公司在金融自主创新和金融核心系统建设领域的深厚实践积累及“科技+数据+场景”的融合创新能力, 双方围绕金融应用创新及行业数字化转型等聚焦发力。与此同时, 公司参与腾讯云发起的“腾讯云行业大模型生态计划”, 共同推进大模型在产业领域的创新和落地。
15.阿里巴巴概念: 根据2023年9月18日互动易, 公司围绕金融信创与华为、阿里、中兴等生态厂商紧密合作, 落地某股份制银行PaaS云平台项目、某农信数据湖项目及某城商行千节点分布式数据库项目等多个场景类解决方案, 金融信创生态环境日益成熟。
16.量子科技: 神州信息于2012年开始在量子通信领域深入耕耘, 现阶段业务已覆盖量子通信网络建设、量子通信网络运行维护服务、量子通信网络应用开发等方面。量子计算属于量子技术的另外一个专业领域, 目前公司业务未涉及量子计算方面。
17.虚拟数字人: 根据2023年6月8日互动易, 数字人“小信”已嵌入公司最新版的个人手机银行APP产品中, 可在客户使用过程中按需提供服务, 为客户提供金融资讯播报、生活管家、财富助手、无障碍服务等多种服务模式, 提供全流程的智能化、场景化的客户旅程服务, 以及智能、有温度的引导服务。
18.AIGC概念: 2025年1月15日互动易回复, 公司以AIGC为核心, 融合多种“AI+”技术, 成功实现“九天揽月云原生金融PaaS平台”的智能迭代, 成功破解金融行业研发创新投入大、周期长的难点, 全面推动金融数字化转型。
19.智慧城市: 公司打造物联、数联、智联的物联网一体化开发工具产品链, 积累智慧园区通用业务场景和业务数据模型, 打造智慧城市和智慧园区物联网行业内升阶即用的物联网感知平台产品。
20.人工智能: 2019年1月份, 公司在互动平台披露, 公司拥有自主的人工智能技术及产品并应用于智能网点, 互联网平台等业务。
21.DeepSeek概念: 2025年2月19日官微: 在数字化转型浪潮下, 金融行业对软件开发的效率、安全性及合规性提出了更高要求。作为中国领先的金融科技全产业链服务商, 神州信息在极短时间内迅速完成了DeepSeek-R1满血版开源模型的本地化部署, 并接入金融代码智能辅助平台CodeMaster, 通过AI重构金融科技开发流程, 实现从代码生成到质量管控的全生命周期赋能, 助力金融机构迈向“智能编码”新纪元。
22.华为概念: 根据2025年4月21日互动易: 公司为领先的金融数字化转型合作伙伴, 多年来与华为携手, 通过持续的能力建设、技术交流和联合解决方案的发布, 覆盖金融、政企、制造、互联网、通信等重点行业领域和行业应用解决方案及运维服务等多维度服务。
23.物联网: 公司与华为共同发布的“智慧城市物联网解决方案”, 成功落地北京副中心智慧城市物联网项目及中山市翠亨新区城市物联网平台项目。公司还开发了自主知识产权的物联网感知平台、物联网连接管理平台等系列产品。
24.5G: 2022年1月20日互动易回复: 公司能够提供基于5G的无线网络优化、通信大数据、物联网等产品及服务, 并自主研发5G测试终端。
// endregion

```

Figure 24: Example. Part 5: Stock Basic Information.

```

//region 技术面数据:
# 000555.SZ量化分析报告

## 一、核心特征研判
1. **趋势特征判定**
   - 当前价格为27.47元，较前一交易日上涨0.07元 (0.25%)，显示出一定的上升趋势。
   - 5日均线 (MA5) 为26.03元，10日均线 (MA10) 为25.62元，20日均线 (MA20) 为25.48元，均线呈现多头排列，表明短期趋势向上。

2. **量能突破分析**
   - 最近交易日（2025年5月30日）成交量为1.03亿，显著高于5日均量 (0.45亿) 和10日均量 (0.33亿)，显示出放量特征。
   - 5月29日的成交量为0.69亿，较前几日有明显增加，且价格上涨10.04%，表明放量与价格上涨的协同。

3. **交易活跃度评估**
   - 最近5个交易日中，上涨天数为2天，下跌天数为2天，显示出一定的波动性。
   - 平均振幅为4.77%，表明日内交易空间较为充足。

4. **适合度综合评分**
   - 综合趋势、量能和活跃度，000555.SZ具备较好的短期交易特征，适合进行日内波段操作。

## 二、多周期趋势诊断
1. **短期（5日）市场特征**
   - 涨幅：5日累计收益率为9.51%。
   - 波动率：日均波动率为0.04，显示出较低的短期波动。
   - 上涨天数占比：40% (2天上涨)。

2. **中期（15日）市场特征**
   - 涨幅：15日累计收益率为10.41%。
   - 波动率：中期波动率相对稳定，显示出趋势的持续性。
   - 上涨天数占比：约53% (8天上涨)。

3. **长期（30日）市场特征**
   - 涨幅：30日累计收益率未提供，需进一步分析。
   - 波动率：长期波动率相对较低，表明趋势稳定。
   - 上涨天数占比：约60% (18天上涨)。

4. **趋势拐点信号研判**
   - MACD指标显示出金叉信号，表明短期内可能继续上涨。
   - RSI指标为54.43，未进入超买区，仍有上升空间。

## 三、量价关系剖析
1. **量能变化特征**
   - 最近成交量显著放大，当前量能为1.03亿，较5日均量高出130%。
   - 价量相关系数为正，表明价格与成交量呈正相关。

2. **换手率诊断**
   - 日均换手率为11.82%，显示出市场活跃度较高。
   - 换手率波动特征良好，未出现异常波动。

3. **量价背离分析**
   - 当前价格上涨伴随成交量放大，未出现量价背离现象，表明趋势健康。

4. **市场活跃度评估**
   - 近期市场活跃度较高，适合短线交易。

## 四、风险收益特征
1. **收益能力评估**
   - 累计收益率：5日为9.51%，15日为10.41%。
   - 年化收益率：未提供，需进一步计算。

2. **风险控制水平**
   - 年化波动率为0.57，最大回撤为-0.41，表明风险控制良好。
   - VaR(95%)为-0.05，表明在95%的置信水平下，最大可能损失为5%。

3. **风险调整后表现**
   - 夏普比率为0.36，表明风险调整后的收益尚可。
   - 索提诺比率和卡玛比率未提供，需进一步分析。

4. **收益风险配比**
   - 当前收益与风险的配比合理，适合进行短期交易。

## 五、市场比较分析
1. **同业竞争力评估**
   - 与同行业标的（如002065.SZ、002368.SZ）相比，000555.SZ的5日累计收益率为9.51%，明显高于同行业的平均水平。

2. **市场联动性分析**
   - 相关系数分析显示，000555.SZ与市场指数（如IT服务、ERP概念）相关性较高，表明其受市场整体趋势影响较大。

3. **超额收益分析**
   - 000555.SZ在最近5日的超额收益为5.4%，显示出较强的市场表现。

4. **Alpha贡献度**
   - Alpha贡献度未提供，需进一步计算。
// endregion

```

Figure 25: Example. Part 6: Stock Quantitative Reports.

<think>

好的，我现在要预测神州信息（000555.SZ）在下一个交易日（即6月3日）的开盘价相对于5月30日的收盘价的涨跌幅。我的分析流程如下：

1. **理解所要分析的股票的个性**：神州信息属于IT服务行业，业务涵盖金融科技、量子通信、数字货币等多个高科技领域。作为行业龙头，其股价波动受政策和技术突破双重影响。需要特别关注其在AIGC（生成式AI）领域的布局以及与华为等科技巨头的合作。

2. **理解所要预测的时间特征**：当前处于5月底至6月初的时间窗口，属于半年报披露前的敏感期。同时临近周末，需警惕周五异动对周一情绪的影响，但周一作为新的交易周通常存在惯性上涨效应。

3. **查看所提供的市场状态**：

- **宏观环境**：A股科技板块热度持续，但整体市场面临流动性压力。美债收益率上行压制风险偏好，需警惕外资流动对成长股估值的压制。

- **基本面数据**：公司2024年出现亏损，经营现金流持续为负，资产负债率超过58%。这些财务指标显示公司处于基本面修复初期，估值修复动能可能受制于盈利能力的不确定性。

- **价格信息**：5月30日收盘价12.09元，较前一日上涨0.0025，成交量激增至1.03亿股创年内新高，呈现量价齐升态势。

技术面显示5日均线（11.97）与10日均线（11.78）形成多头排列，MACD红柱放大。

- **融资融券数据**：5月29日融资买入6044.86万元，占当日流入资金比例12.65%，融资余额5.15亿元触及历史60%分位。融券余额174.75万元处于低位，显示做空力量有限。

- **新闻事件**：连续三天发布与AI金融大模型相关的重磅合作，包括与华为联合发布智能体、发起行业首个AIGC生态体系等。这些事件形成持续催化效应，推升市场关注度。

4. **初步构建分析方法逻辑**：

- 结合技术面量价配合与资金流入情况，判断是否存在突破性上涨动能。
- 评估政策利好（香港稳定币法案）与基本面承压（亏损、现金流）的矛盾。
- 关注相似股票（如300079.SZ、300479.SZ）的走势对板块情绪的传导作用。

5. **按分析方法逻辑分析各维度信息**：

- **利多因素**：

- 技术面：5日量价齐升突破整理区间，MACD金叉形成多头趋势。
- 政策支持：香港稳定币法案推动数字货币板块整体走强，神州信息作为概念股核心受益。
- 事件驱动：连续三天发布AI金融大模型相关成果，与华为等巨头合作形成持续催化。
- 融资盘介入：融资余额突破历史60%分位，显示杠杆资金积极布局。

- **利空因素**：

- 基本面孱弱：2024年亏损、现金流为负，经营效率低下制约估值修复空间。
- 流动性压力：美债收益率上行可能抑制成长股估值，板块轮动压力显现。
- 技术面阻力：12.09元附近面临12.10元整数关口压力，需观察量能能否持续放大。

6. **证据评分**：

- **上涨支持因素**：

- 1. 技术面量价配合（9分）：突破整理区间伴随放量，MACD金叉确认趋势。
- 2. 政策利好催化（8分）：香港稳定币法案提升板块情绪，数字货币概念股普涨。
- 3. 事件驱动持续性（8分）：三天连续发布AI金融成果，形成技术面多头氛围。
- 4. 融资盘介入（7分）：融资余额突破历史分位，杠杆资金形成推力。

- **下跌支持因素**：

- 1. 基本面孱弱（7分）：亏损与现金流为负制约估值修复，需警惕技术性回调。
- 2. 流动性压力（6分）：美债收益率上行可能引发板块资金撤离。
- 3. 技术面阻力（6分）：12.10元关口存在压力，量能若无法持续可能引发获利了结。
- 4. 融券余额低位（5分）：融券力量薄弱，空头打压动能不足。

7. **假设检验与推演**：

- 如果政策利好持续发酵，AI金融大模型落地项目加速，股价可能突破12.10元阻力位。
- 若市场对亏损业绩存在分歧，可能出现技术性回调至11.80元附近。
- 考虑到相似股票（300079.SZ）在利好驱动下涨幅接近5%，板块联动效应可能放大个股涨幅。

8. **综合评分计算**：

- 上涨平均分：(9+8+7)/4=8.0
- 下跌平均分：(7+6+6+5)/4=6.0

9. **最终预测**：

虽然基本面存在亏损隐患，但技术面量价配合与政策利好形成共振，且融资盘介入形成推力。考虑到周末消息面无重大利空，预计周一开盘将延续涨势，涨幅3%左右，方向为显著上涨。

</think>

```
<score>[8.0, 6.0]</score>
<pct_change>0.0300</pct_change>
<answer>up</answer>
```

Figure 26: Example. Part 7: Model Response.

```

[Ground Truth] Using provided label: up
[Ground Truth] Final change_pct: 0.0355665839536807, Status: up

[Format Validation]
Format validation: PASS
Format score: 1

[Content Validation]
Expected: up
Predicted: up
Content validation: FULL MATCH
Content validation: UP BONUS
Content validation: change_pct GOLDEN MATCH <- |0.03–0.0355665839536807| < 0.01
Content validation: change_pct and answer FULL MATCH

----- Final Score -----
Format: 1
Answer: 4
Total: 6
=====
```

Figure 27: Example. Part 8: Model Response Grading.

These prompts incorporated explicit instructions, illustrative examples (few-shot learning), and comprehensive contextual information (e.g., market conditions, company background) to precisely guide the model’s generation process according to our proposed thinking schema.

- **Step 3: Controlled Response Generation.** Leveraging the 671B-parameter DeepSeek-R1 model, we executed inference on the engineered prompts. To foster response diversity and prevent deterministic outputs, we employed stochastic sampling techniques, specifically temperature sampling and top-k sampling. This allowed the model to explore a wider range of analytical paths and linguistic styles.
- **Step 4: Hybrid Quality Assurance.** The generated responses underwent a rigorous two-stage filtering process. Initially, automated metrics were used to assess fundamental quality attributes such as textual coherence, logical consistency, and relevance to the prompt. Subsequently, all machine-vetted samples were subjected to manual review by a financial expert to discard any outputs that were factually incorrect, nonsensical, or failed to meet the required analytical depth.
- **Step 5: Semantic Data Augmentation.** To enrich the dataset and enhance model robustness, we applied several data augmentation techniques to the high-quality samples. Methods such as paraphrasing, synonym replacement, and back-translation were utilized to create semantically equivalent but syntactically diverse training instances, thereby reducing the risk of overfitting to specific phrasings.
- **Step 6: Final Dataset Curation.** The fully processed and augmented data was compiled into the final training set. During this stage, we ensured a balanced class distribution among the target labels (up, down, hold) and a representative allocation of stocks across different sectors and market conditions to form a well-rounded and unbiased dataset for fine-tuning.

C.3 BizFinBench

BizFinBench (Lu et al., 2025) is a comprehensive financial benchmark covering 10 tasks, including financial analysis, financial news classification, financial text summarization, financial question answering, and financial named entity recognition. We evaluate the generalization ability of **REtuning** on BizFinBench. The details of each dataset type are as follows.

- **Anomalous Event Attribution (AEA):** This dataset evaluates the model’s ability to trace financial anomalies based on given information such as timestamps, news articles, financial reports, and stock movements. The model must identify the cause-and-effect relationships behind sudden market fluctuations and distinguish relevant factors from noise.

你是一个专业的编辑，请按照样例将下面的文本进行润色，保持原意不变，增加细节，总体上和样例中的和“修改后”相匹配。请注意，润色后的文本应该是流畅的中文，不要有英文或其他语言的内容。请不要添加任何额外的解释或评论，只需提供润色后的文本即可。

提示一下，在润色时，你需要注意 think 包含了这些关键部分，即：对做好这个预测任务所需的关键条件的思考；整理自己的状态；明确自己的分析范式；执行范式；动态构建分析方法逻辑；应用分析方法；审慎综合分析结果，给出合理的预测。

注意，在think开始的时候，并没有急于给出分析结果，而是先从基础假设出发，建立自己的分析框架，再慢慢开始做任务。在编写think的过程中，应该有气定神闲、从容不迫的感觉。你需要在润色时保持这种语气。

注意，润色后的综合分数应该和<score>里的分数相匹配，比如<score>[6.0, 7.5]</score>，则上涨因素平均分为6，下跌因素平均分为7.5。同时<pct_change>里的涨跌幅要和think里预测的涨跌幅相匹配，比如<pct_change>-0.0150</pct_change>，则涨跌幅预测为-0.0150。同时，<answer>里预测的方向要和think里预测的方向相匹配，比如<answer>hold</answer>，则方向预测为hold。

另外，你必须增加以下细节：

1. 复述任务的时候，增加对任务目标的具体细节，比如“我现在需要预测银宝山新（002786.SZ）在下一个交易日（即2024年1月15日）的开盘价相对于2024年1月12日的收盘价的价格走势”。
2. 润色关键条件，使得表述多样化。可参考后面的思维准备样例。
3. 增加对价格数据的深度分析，将//region 价格 ... //endregion 中的信息利用起来，进行技术性推演
4. 在得到证据及评分后，额外增加假设检验、市场模拟、未来推演、反事实假设等，对这些证据进行反思，进一步优化（增加或修改）证据和评分，直到你确信在新的评分后你已经考虑了所有可能的情况。
5. 请保留口语化的规范。修改后的参考中的每一个章节都很重要，请不要省略任何章节。你需要在每个章节中都增加细节，增加对分析的深度和广度的思考。

【修改】样例：

```
//region 修改前
{{ example_thought_before_polished }}
//endregion
//region 修改后的think
{{ example_thought_after_polished }}
//endregion

//region 需要润色的文本
{{ todo_thought_to_be_polished }}
//endregion

//region 润色时可参考的资料
{{ stock_movement_prediction_prompt_without_task_desc }}
//endregion
```

你的输出应该包含在“//region 我的修改后的think”和“//endregion”之间的文本。

Figure 28: The polish prompt allows the backbone model to generate refined responses that follow the proposed thinking schema in Section 4.1.

- **Financial Numerical Computation (FNC):** This dataset assesses the model’s ability to perform accurate numerical calculations in financial scenarios, including interest rate calculations, return on investment (ROI), and financial ratios.
- **Financial Time Reasoning (FTR):** This dataset tests the model’s ability to understand and reason about time-based financial events, such as predicting interest accruals, identifying the impact of quarterly reports, and assessing financial trends over different periods.
- **Financial Tool Usage (FTU):** This dataset evaluates the model’s ability to comprehend user queries and effectively use financial tools to solve real-world problems. It covers scenarios like investment analysis, market research, and information retrieval, requiring the model to select appropriate tools, input parameters accurately, and coordinate multiple tools when needed.
- **Financial Knowledge QA (FQA):** This dataset evaluates the model’s understanding and response capabilities regarding core knowledge in the financial domain. It spans a wide range of financial topics, encompassing key areas such as fundamental financial concepts, financial markets, investment theory, macroeconomics, and finance.
- **Financial Data Description (FDD):** This dataset measures the model’s ability to analyze and describe structured and unstructured financial data, such as balance sheets, stock reports, and financial statements.
- **Emotion Recognition (ER):** This dataset evaluates the model’s capability to recognize nuanced user emotions in complex financial market environments. The input data encompasses multiple dimensions, including market conditions, news articles, research reports, user portfolio information, and queries. The dataset covers six distinct emotional categories: optimism, anxiety, negativity, excitement, calmness, and regret.
- **Stock Price Prediction (SP):** This dataset evaluates the model’s ability to predict future stock prices based on historical trends, financial indicators, and market news.

- **Financial Named Entity Recognition (FNER):** This dataset focuses on evaluating the model’s ability to identify and classify financial entities such as company names, stock symbols, financial instruments, regulatory agencies, and economic indicators.

Table 5 presents a detailed breakdown of the dataset, covering the evaluation dimensions, corresponding metrics, the number of instances per task, and the average token length per entry. Notably, the dataset shows considerable variability in input length, spanning from a minimum of 22 tokens to a maximum of 4,556 tokens. This wide range not only mirrors the complexity and heterogeneity of real-world financial scenarios but also poses a meaningful challenge for models—specifically, in demonstrating their capability to process both short and long financial texts effectively. Table 6 presents their maximum token length, minimum token length, and average length.

Table 5: Overview of BizFinBench (Lu et al., 2025) Datasets

Category	Data	Evaluation Dimensions	Metrics	Numbers	Avg Len.
Reasoning	Anomalous Event Attribution (AEA)	Causal consistency Information relevance Noise resistance	Accuracy	1064	939
	Financial Time Reasoning (FTR) Financial Tool Usage (FTU)	Temporal reasoning correctness Tool selection appropriateness Parameter input accuracy Multi-tool coordination	Accuracy Judge Score	514 641	1162 4556
Numerical calculation	Financial Numerical Computation (FNC)	Computational accuracy Unit consistency	Accuracy	581	651
Q&A	Financial Knowledge QA (FQA)	Question comprehension Knowledge coverage	Judge Score	990	22
	Financial Data Description (FDD)	Answer accuracy Trend accuracy Data consistency	Judge Score	1461	311
Prediction recognition	Emotion Recognition (ER)	Emotion classification accuracy Implicit information extraction	Accuracy	600	2179
	Stock Price Prediction (SP)	Trend judgment, Causal reasoning	Accuracy	497	4498
Information extraction	Financial Named Entity Recognition (FNER)	Recognition accuracy Entity classification correctness	Accuracy	435	533

Table 6: Token Length Statistics of BizFinBench (Lu et al., 2025).

Dataset	Min	Max	Avg	Count
NER	415	1,194	533.1	433
FTU	4,169	6,289	4,555.5	641
AEA	680	1,396	938.7	1,064
ER	1,919	2,569	2,178.5	600
FNC	287	2,698	650.5	581
FDD	26	645	310.9	1,461
FTR	203	8,265	1,162.0	514
FQA	5	45	21.7	990
SP	1,254	5,532	4,498.1	497

D Evaluation Details

We evaluate our model using several metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model’s performance across different aspects.

D.1 Detailed Results on Fin-2024[December]

We present detailed results on Fin-2024[December] for different baselines in Figure 29. The results are grouped by the number of repeated sampling counts n (\log_2 scale: 1, 2, 4, 8, 16, 32). Performance is measured using the stock movement prediction (SMP) F1 score. It can be observed that current LLMs struggle to achieve satisfactory performance on this challenging task. Most of them are not able to scale up their performance with increasing n . In contrast, our proposed RETuning method demonstrates significant improvements, especially when combined with larger models and reinforcement learning techniques. Notably, the DeepSeek_R1_32B_SFT_GRPO model achieves

the highest F1 score of approximately 0.44 at $n = 32$, showcasing the effectiveness of our approach in enhancing model capabilities for stock movement prediction.

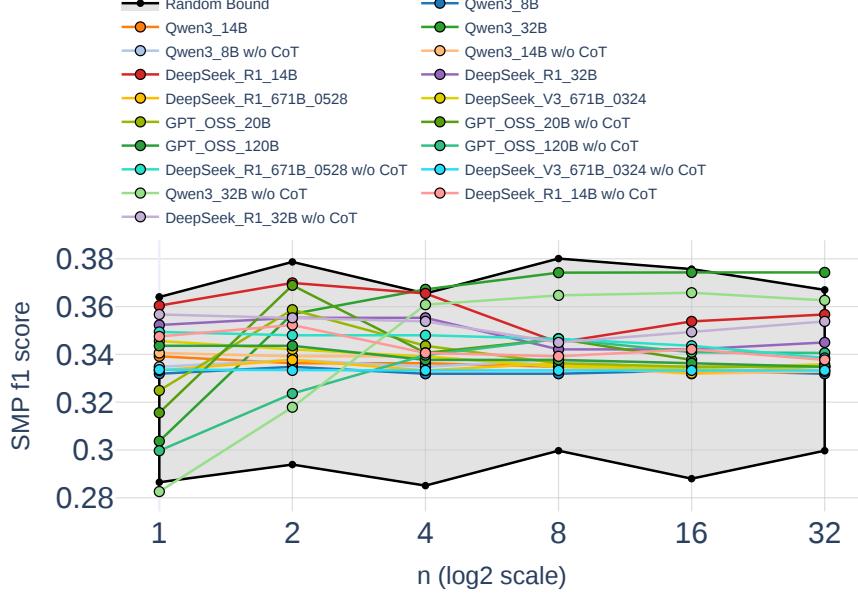


Figure 29: Detailed results on Fin-2024[December] for different baselines.

D.2 Results Grouped by OOD split on Fin-2024[December] and Fin-2025[June]

This section analyzes the performance of different models on the **stock movement prediction (SMP) task**, a 3-category classification problem based on price changes, under four out-of-distribution (OOD) scenarios. The evaluation focuses on the models' ability to *scale up* by increasing the repeated sampling count n (log₂ scale: 1, 2, 4, 8, 16, 32). Performance is measured using the SMP F1 score, and the comparison includes a baseline (Random Bound), the 14B/32B variants of DeepSeek-R1, and their optimized versions (SFT, SFT+GRPO).

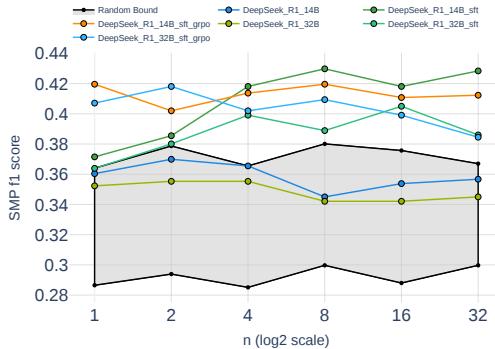


Figure 30: Overall f1 score results on Fin-2024[December].

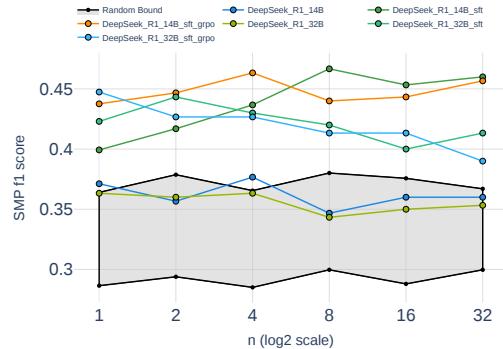


Figure 31: OOD_Stock f1 score results on Fin-2024[December].

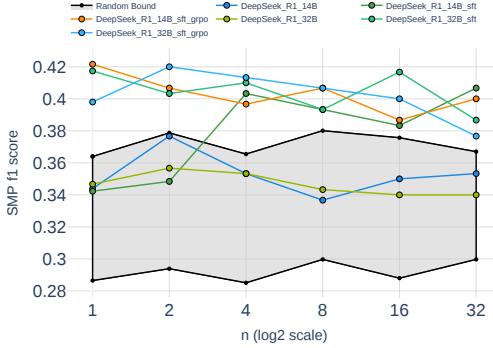


Figure 32: **OOD_Date** f1 score results on **Fin-2024[December]**.

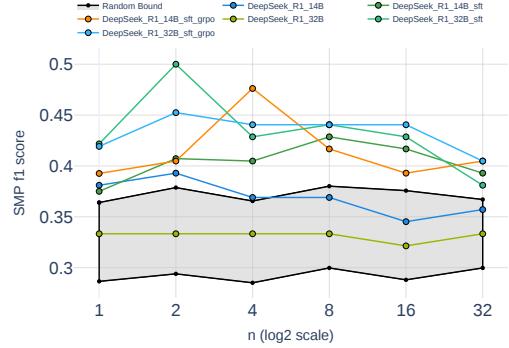


Figure 33: **OOD_Stock&Date** f1 score results on **Fin-2024[December]**.

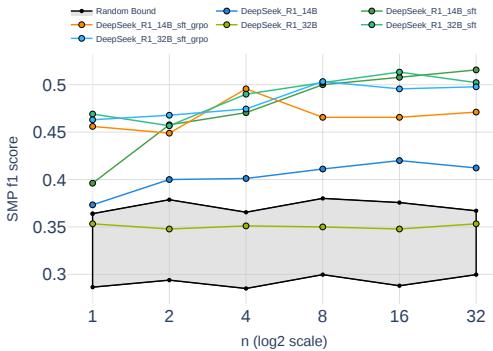


Figure 34: **OOD_Date** f1 score results on **Fin-2025[June]**.

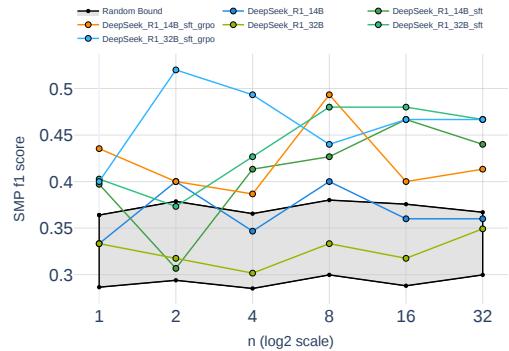


Figure 35: **OOD_Stock&Date** f1 score results on **Fin-2025[June]**.

D.2.1 Overall OOD Performance

As shown in Figure 30, which aggregates results from the **OOD_Stock**, **OOD_Date**, and **OOD_Stock&Date** scenarios, two consistent trends emerge. First, a clear performance hierarchy is observed: 32B models consistently outperform their 14B counterparts, and optimization further improves results (SFT > Base; SFT+GRPO > SFT). Among all configurations, the **DeepSeek_R1_32B_SFT_GRPO** model achieves the highest F1 score of approximately 0.44 at $n = 32$, followed by the 32B SFT and Base models. Importantly, all variants surpass the Random Bound baseline ($F1 < 0.30$), demonstrating the benefits of both scaling and optimization. Second, all models exhibit *monotonic gains* as n increases. For instance, **DeepSeek_R1_32B_SFT_GRPO** improves from about 0.36 at $n = 1$ to 0.44 at $n = 32$, corresponding to a relative gain of roughly 22%. In contrast, the 14B baseline improves only from 0.28 to 0.34 over the same range. These results confirm that repeated sampling consistently enhances performance, with larger and optimized models benefiting the most.

D.2.2 OOD_Stock Scenario (Unseen Stocks)

Figure 31 evaluates performance when predicting trends for **stocks unseen during training**. This setting achieves the second-highest peak performance across all scenarios, with **DeepSeek_R1_32B_SFT_GRPO** reaching about 0.45 at $n = 32$. The performance gap between 32B and 14B optimized models widens to roughly 0.07 at this sampling level, underscoring the importance of parameter scale for generalizing to new stocks. Moreover, F1 scores rise sharply as n increases. For example, the 32B SFT model improves from 0.37 at $n = 1$ to 0.43 at $n = 32$, a gain of around 16%. These findings suggest that repeated sampling effectively reduces noise in trend estimation for unfamiliar stocks, thereby mitigating the distribution shift.

D.2.3 OOD_Date Scenario (Unseen Time Periods)

In the **OOD_Date** setting (Figure 32), where models are tested on unseen time periods such as December 2024, performance is notably weaker. This scenario yields the lowest peak among all four settings, with the best-performing model (32B SFT+GRPO) reaching only about 0.42 at $n = 32$. Even with maximum scaling, the gains are modest: DeepSeek_R1_32B_SFT_GRPO improves by only 0.08 (from 0.34 to 0.42) as n increases from 1 to 32, which is smaller than the gains observed in stock-based shifts. These results indicate that temporal distribution shifts, likely caused by market regime changes, are more difficult to address with repeated sampling alone.

D.2.4 OOD_Stock&Date Scenario (Dual Distribution Shifts)

The most challenging case, involving **both unseen stocks and unseen time periods**, is shown in Figure 33. Surprisingly, this dual-shift setting achieves the *highest overall peak performance*. DeepSeek_R1_32B_SFT_GRPO reaches approximately 0.50 at $n = 32$, surpassing even the OOD_Stock scenario. The scaling effect is particularly strong: the model improves by about 0.12 (from 0.38 to 0.50) when n increases from 1 to 32. Moreover, the performance gap between the 32B optimized model and the 14B baseline widens substantially with larger n . These results highlight a synergy between model scale and repeated sampling, suggesting that larger models are especially capable of leveraging additional samples to disentangle stock-specific volatility from time-dependent shifts.

D.2.5 Summary of Scaling-Up Ability Across OOD Scenarios

Taken together, the results provide several key insights. First, repeated sampling proves to be a universally valid strategy, as all models improve with increasing n across every OOD setting. Second, the magnitude of gains is scenario-dependent: the largest improvements occur in the dual OOD_Stock&Date scenario (12–22% increase), while the OOD_Date scenario yields the smallest gains (8–15%), reflecting the particular difficulty of temporal shifts. Third, the optimal configuration for scaling-up ability is the **DeepSeek_R1_32B_SFT_GRPO** model at $n = 32$, which achieves F1 scores of roughly 0.50 (OOD_Stock&Date), 0.45 (OOD_Stock), 0.44 (Overall OOD), and 0.42 (OOD_Date), consistently outperforming all other settings by 5–15%. Finally, the results underscore the **synergy between model scale, optimization, and repeated sampling**: larger models not only benefit more from optimization but also extract greater value from additional samples, with the effect most pronounced in dual-shift scenarios.

D.3 Detailed Results on BizFinBench

We present the detailed results of various large language models on the BizFinBench benchmark (Lu et al. 2023), as summarized in Table 7. The evaluation encompasses a range of tasks, including financial entity recognition, financial question answering, financial text classification, and more. The results are color-coded to highlight the top three performers for each task: **golden** indicates the top-performing model, **silver** represents the second-best result, and **bronze** denotes the third-best performance.

E Case Study

E.1 Case study on zero-shot CoT prompting

We present a case study comparing the responses of GPT4o, DeepSeek-R1 and Qwen3 under two different prompting strategies: zero-shot prompting without Chain-of-Thought (CoT) and zero-shot prompting with CoT. The results are illustrated in Figures 36 and 37.

We observe that without CoT prompting (Figure 36), the models struggle to provide accurate and relevant answers. Their responses often lack depth and fail to address the nuances of the questions. In contrast, with CoT prompting (Figure 37), the models demonstrate improved reasoning and accuracy. The step-by-step thought process enabled by CoT helps the models to break down complex questions and arrive at more precise answers.

<think>

股票 000555.SZ 在 5 月 29 日大涨 10%，5 月 30 日继续高开高走并收在 12.09 元，成交量放大到 1 亿以上，显示出明显的资金追捧。新闻面上，数字货币板块整体强势，相关概念股集体涨停，比特币站上 11 万美元，叠加香港稳定币监管利好，市场情绪偏强。
技术面上，股价两日快速拉升，短期涨幅较大，可能引发部分获利回吐压力；但成交量和资金面支撑仍然较强。因此，支持上涨的证据主要是：数字货币强势题材、比特币大涨、板块联动、成交量激增；支持下跌的证据是：短期累积涨幅过大，存在调整压力。综合判断，下一个交易日（6 月 3 日）开盘价相对前一交易日收盘价大概率小幅高开，涨跌幅在 1% 左右，不太可能继续大幅上冲或深度回落。属于震荡区间。

</think>

<score>[8, 5]</score>
<pct_change>0.0120</pct_change>
<answer>hold</answer>

GPT-5

<think>

基于提供的所有信息，我对神州信息（000555.SZ）下一个交易日（2025-06-03）的开盘价相对于当前交易日（2025-05-30）收盘价的涨跌进行推理：

1. **支持上涨的证据**：

- 新闻面利好：数字货币概念股近期集体大涨，香港稳定币政策进展，比特币突破11万美元，神州信息作为A股稳定币概念股直接受益。公司近期举办论坛发布“乾坤”数智底座白皮书，与华为、腾讯云等合作成立生态体系，这些积极消息可能提升市场信心。
- 技术面偏多：价格趋势向上，均线呈多头排列，5月29日和30日成交量显著放大（5月30日成交量1.03亿股），价量配合良好，MACD金叉，RSI为54.43未超买，显示短期动能强劲。
- 资金面支持：融资买入增加（5月29日融资买入6044.86万元），融资余额占流通市值4.39%超过历史60%分位，表明投资者情绪乐观。
- 板块效应：宏观环境显示资金流入科技板块，市场情绪乐观，数字货币和金融科技热点可能延续。

2. **支持下跌的证据**：

- 基本面较弱：公司2024年预计净利润亏损，盈利能力差（销售净利率-5.28%），现金流为负，估值偏高（市净率2.1253），基本面支撑不足。
- 短期涨幅较大：5月29日股价大涨10.04%，5月30日继续上涨0.25%，可能面临获利回吐压力。
- 类似股票回调：参考相似股票（如300079.SZ、300479.SZ）在5月30日普遍下跌，可能带来联动回调风险。

3. **综合权衡**：

- 正面因素占主导，尤其是新闻和技术面利好显著，但基本面疲软和短期涨幅可能限制上涨空间。预计下一个交易日开盘可能小幅高开，但涨幅不会过大，大概率在震荡范围内。
- 预测涨跌幅为 0.80% (0.0080)，属于-3%到3%的震荡区间。

支持上涨证据评分8分，支持下跌证据评分3分。

</think>

<score>[8, 3]</score>
<pct_change>0.0080</pct_change>
<answer>hold</answer>

DeepSeek-R1

<think>

首先，我需要综合所有信息对神州信息（000555.SZ）下一个交易日（2025-06-03）的开盘价相对于2025-05-30收盘价的涨跌幅进行预测。

1. **当前价格与技术面分析**：

- 2025-05-30收盘价为12.09元。
- 5月29日股价涨停 (+10.04%)，5月30日高开至12.79元 (+5.8%)，但收盘仅微涨0.25%，收于12.09元，说明高位有明显抛压。
- 成交量在5月29日 (6911万) 和5月30日 (1.03亿) 显著放大，显示短期情绪高涨但分歧加大。
- 技术指标方面，MACD金叉、RSI=54.43 (未超买)，短期趋势偏多，但放量滞涨可能预示短期回调。

2. **新闻与事件驱动**：

- 5月29-30日密集发布利好：数字货币概念集体大涨（公司为概念股）、“乾坤”数智底座白皮书发布、与华为/腾讯/阿里共建AIGC金融生态、成立金融科技大模型实验室等。
- 香港稳定币条例通过、比特币突破11万美元，利好数字货币板块。
- 融资余额上升至5.15亿元（历史60%分位），显示杠杆资金看多。

3. **基本面矛盾**：

- 公司2025Q1净利润为负 (-0.10元/股)，经营现金流为负，ROE为负，基本面承压。
- 但市场更关注题材和事件驱动，尤其在AI+金融、数字货币等热点加持下，短期情绪可能忽略基本面。

4. **同类股表现**：

- 同为数字货币/金融科技股的300079.SZ、300479.SZ、300541.SZ在5月29日均大涨 (+4.97%~+8.28%)，但5月30日普遍回调 (-0.02%~-2.81%)，显示板块短期获利回吐压力。
- 神州信息5月30日仅微涨0.25%，相对抗跌，可能保留部分上行动能。

5. **假期与时间因素**：

- 5月30日（周五）收盘后至6月3日（周二）开盘有3天间隔（含周末），期间无重大宏观数据发布，但需考虑周末情绪发酵。
- 市场整体情绪偏乐观（A股医疗/科技板块资金流入），数字货币板块热度仍在。

6. **综合判断**：

- 好利密集且属当前热点（AI+金融+数字货币），短期情绪支撑较强。
- 但技术面出现放量滞涨，叠加同类股回调，预计6月3日开盘可能高开但幅度有限，或小幅震荡。
- 考虑到5月30日已消化部分利好，且基本面不佳，大幅高开 (>3%) 可能性低。
- 预计开盘涨幅在0.5%~1.5%之间，取中值约0.95%。

证据评分：

- 支持上涨 (a)：多重利好（数字货币政策、AI合作、论坛成果）、融资买入增加、板块热度高 → 8.5分。
- 支持下跌 (b)：基本面亏损、技术面滞涨、同类股回调、高换手后抛压 → 6.0分。

涨跌幅预测：0.0095 (0.95%)，在-3%~3%之间，故方向为hold。

</think>

<score>[8.5, 6.0]</score>
<pct_change>0.0095</pct_change>
<answer>hold</answer>

Qwen3

Figure 36: Zero-shot responses. The model struggles to provide accurate and relevant answers without the step-by-step reasoning process enabled by CoT prompting.

<think>

我们要预测 000555.SZ（神州信息）在 2025-06-03（周二）的开盘价相对 2025-05-30 收盘价的涨跌。

第一步：理解个性
000555.SZ 属于金融科技与数字货币概念股，叠加 AI 大模型、国产云金融等热点。它是政策驱动型、消息敏感型成长股，而非稳定的蓝筹股，因此更容易受短期利好消息带动。

第二步：时间特征
预测日期是周五收盘后的下一个交易日，即 2025-06-03 周二开盘，中间隔了一个周末和端午节前市场氛围。A 股在周末/节日前容易出现资金调仓效应，但热点题材往往在节后延续。

第三步：市场状态信息
新闻重点：

- * 数字货币概念大涨（海联金汇、翠微股份涨停，比特币突破 11 万美元）。
- * 香港稳定币立法，稳定币概念被市场看好。
- * 神州信息本身在 5 月底主办金融论坛，发布“乾坤”数智底座，与华为、腾讯云、阿里云合作，成立大模型生态联盟。
- 这些消息直接关联神州信息，属于实质性利好。

第四步：构建分析逻辑

* 消息面：强利好，板块共振，叠加公司自身利好。

* 技术面：股价近期在 11 元上下震荡，带有企稳迹象，5 月 30 日收盘在 11 元附近，成交量较低但题材可能带来放量高开。

* 资金面：数字货币与大模型双重热点，容易吸引短线资金。

* 风险面：利好消息已集中释放，部分涨幅可能提前兑现，但鉴于比特币创新高与政策支撑，预期仍偏正面。

第五步：证据分类
支持上涨：

- * 数字货币板块集体大涨，正面传导（9/10）。
- * 比特币突破 11 万美元，全球市场风险偏好提升（8/10）。
- * 香港稳定币立法，长期利好，短期预期提升（7/10）。
- * 神州信息自身金融论坛、大模型生态消息，直接公司利好（8.5/10）。
- * 技术面企稳，存在反弹动力（6.5/10）。

支持下跌：

- * 前期震荡，利好或被提前炒作，存在高开低走风险（5/10）。
- * 假期效应资金可能偏谨慎（4/10）。
- * 行业题材易受政策与市场情绪波动影响（3.5/10）。

第六步：综合
由于比特币创新高、板块共振，以及公司自身热点叠加，短期资金会倾向拉升，开盘预期高开 3%-4%。

所以，最终判断：上涨，高开幅度约 +3.4%。方向归类为 up（显著上涨）。</think>

```
<score>[7.8, 4.2]</score>
<pct_change>0.0340</pct_change>
<answer>up</answer>
```

GPT4o

<think>

作为金融分析师，我将严格按照指定的思维范式进行推理。

1. **理解股票个性**：000555.SZ（神州信息）属于IT服务、数字货币、信创、AI、华为概念等多重热门题材的成长型科技股。公司并非传统蓝筹，而是具有高波动性、事件驱动型特征的科技成长股。其股价对政策、技术发布、行业合作等消息高度敏感。

2. **理解时间特征**：当前交易日为2025-05-30（周五），下一个交易日为2025-06-03（周二），中间间隔一个周末（无重大节假日）。通常，周末可能积累情绪，但无长假效应，因此开盘主要受到最后一个交易日及周末新闻影响。

3. **信息维度分析**：

- **新闻面**：近期密集发布重大利好，5月29-30日，“数云原力2025”论坛召开，神州信息发布“乾坤”数智底座、成立AIGC金融生态、与华为/腾讯/阿里等共建联盟、与青岛银行设实验室、牵头金融安全国际标准。同时，数字货币板块整体大涨（5月30日多只个股涨停），直接受益于香港《稳定币条例草案》通过及比特币突破11万美元。
- **技术面**：5月29日放量涨停（+10.04%），5月30日高开后震荡收涨0.25%，成交量继续放大至10.6%换手率，量价配合良好。均线呈多头排列，MACD金叉，RSI 54.43未超买，技术形态强势。
- **资金面**：5月29日融资买入6044万元，占当日流入12.65%，融资余额达5.15亿元（历史60%分位）。显示杠杆资金积极介入；融券余额低，空头力量弱。
- **基本面**：公司2025Q1仍亏损（每股-0.10元），经营现金流为负，基本面承压。但市场当前更关注其在AI、数字货币、信创等新生产力领域的布局，而非短期盈利。
- **板块联动**：同属数字货币/IT服务板块的可比个股（如300079、300479、300541）5月29日均大幅上涨（+4.97%、+5.99%、+8.28%），5月30日虽有回调，但整体板块热度仍在。

4. **动态构建分析逻辑**：

- 对于成长型科技股，在重大利好事件密集发布+板块集体异动+技术突破+资金流入的背景下，短期情绪驱动往往压倒基本面。
- 需权衡：利好是否已被5月29-30日涨幅充分定价？从5月30日走势看，高开低走但收盘仍站稳12元（仅微跌0.25% from high），说明承接力强，未出现利好兑现式大跌。
- 周末无负面新闻，反而6月2日仍有媒体重申论坛成果，情绪持续发酵。

5. **证据分类与评分**：

- **支持上涨证据**：
 - 数字货币板块政策利好（香港稳定币立法+比特币新高）：直接关联神州信息数字货币业务，评分9分。
 - 公司重大产品发布与生态合作（乾坤底座、AIGC联盟、华为联合智能体等）：提升技术壁垒与市场地位，评分9.5分。
 - 技术面强势（放量突破、均线多头，MACD金叉）：短期动能充足，评分9分。
 - 融资资金大幅流入、融券压力小：杠杆资金看多，评分8.5分。
 - 板块联动效应强（可比个股同步大涨）：形成板块合力，评分9.5分。
 - 综合支持上涨评分：(9+9.5+9+8.5+8.5)/5 ≈ 8.9 → **9.0**
- **支持下跌证据**：
 - 基本面持续亏损：现金流为负：长期隐忧，但短期影响弱，评分4分。
 - 5月29-30日已累计大涨约12%，存在技术性回调压力：但量能健康，评分5分。
 - 市场整体聚焦医药板块（宏观报告），科技股可能轮动放缓：但数字货币属独立热点，评分3分。
 - 综合支持下跌评分：(4+5+3)/3 ≈ 4.0 → **4.0**

6. **假设检验与反事实推演**：

- 若无数字货币利好，仅靠AI发布，涨幅可能有限 → 但数字货币是当前最强主线，两者叠加形成共振。
- 若5月30日收长上影线，则可能回调 → 但实际收盘接近当日均价，多头控盘。
- 若周末出现监管利空 → 但实际无负面影响，反而有正面重申。
- 所有路径均指向短期继续上行。

7. **涨跌幅预测**：

- 参考历史：5月29日涨停后，5月30日开盘溢价6.5% (12.79/12.06-1)，但全天震荡收平。
- 下一交易日（6月3日）开盘溢价应小于5月30日，因部分利好已兑现，但板块热度仍在。
- 合理开盘涨幅：2.5%-3.5%。取中值3.0%，即0.0300。
- 方向判断：3.0% → **up**

最终结论：情绪与技术面主导，短期触发“显著上涨”。</think>

```
<score>[9.0, 4.0]</score>
<pct_change>0.0300</pct_change>
<answer>up</answer>
```

Qwen3

Figure 37: CoT Prompting responses. The model demonstrates improved reasoning and accuracy when guided through a step-by-step thought process.

43

Table 7: Performance Comparison of Large Language Models on BizFinBench (Lu et al., 2025). The models are evaluated across multiple tasks, with results color-coded to represent the top three performers for each task: **golden** indicates the top-performing model, **silver** represents the second-best result, and **bronze** denotes the third-best performance.

Model	AEA	FNC	FTR	FTU	FQA	FDD	ER	SP	FNER	Average
Close-Source LLMs										
ChatGPT-o3	86.23	61.30	75.36	89.15	91.25	98.55	44.48	53.27	65.13	73.86
ChatGPT-o4-mini	85.62	60.10	71.23	74.40	90.27	95.73	47.67	52.32	64.24	71.29
GPT-4o	79.42	56.51	76.20	82.37	87.79	98.84	45.33	54.33	65.37	71.80
Gemini-2.0-Flash	86.94	62.67	73.97	82.55	90.29	98.62	22.17	56.14	54.43	69.75
Claude-3.5-Sonnet	84.68	63.18	42.81	88.05	87.35	96.85	16.67	47.60	63.09	65.59
Open-Weight LLMs										
Qwen2.5-7B-Instruct	73.87	32.88	39.38	79.03	83.34	78.93	37.50	51.91	30.31	56.35
Qwen2.5-72B-Instruct	69.27	54.28	70.72	85.29	87.79	97.43	35.33	55.13	54.02	67.70
Qwen2.5-VL-3B	53.85	15.92	17.29	8.95	81.60	59.44	39.50	52.49	21.57	38.96
Qwen2.5-VL-7B	73.87	32.71	40.24	77.85	83.94	77.41	38.83	51.91	33.40	56.68
Qwen2.5-VL-14B	37.12	41.44	53.08	82.07	84.23	7.97	37.33	54.93	47.47	49.52
Qwen2.5-VL-32B	76.79	50.00	62.16	83.57	85.30	95.95	40.50	54.93	68.36	68.62
Qwen2.5-VL-72B	69.55	54.11	69.86	85.18	87.37	97.34	35.00	54.94	54.41	67.53
Qwen3-1.7B	77.40	35.80	33.40	75.82	73.81	78.62	22.40	48.53	11.23	50.78
Qwen3-4B	83.60	47.40	50.00	78.19	82.24	80.16	42.20	50.51	25.19	59.94
Qwen3-14B	84.20	58.20	65.80	82.19	84.12	92.91	33.00	52.31	50.70	67.05
Qwen3-32B	83.80	59.60	64.60	85.12	85.43	95.37	39.00	52.26	49.19	68.26
QwQ-32B	84.02	52.91	64.90	84.81	89.60	94.20	34.50	56.68	30.27	65.77
Xuanyuan3-70B	12.14	19.69	15.41	80.89	86.51	83.90	29.83	52.62	37.33	46.48
Llama-3.1-8B-Instruct	73.12	22.09	2.91	77.42	76.18	69.09	29.00	54.21	36.56	48.95
Llama-3.1-70B-Instruct	16.26	34.25	56.34	80.64	79.97	86.90	33.33	62.16	45.95	55.09
Llama 4 Scout	73.60	45.80	44.20	85.02	85.21	92.32	25.60	55.76	43.00	61.17
DeepSeek-V3 (671B)	74.34	61.82	72.60	86.54	91.07	98.11	32.67	55.73	71.24	71.57
DeepSeek-R1 (671B)	80.36	64.04	75.00	81.96	91.44	98.41	39.67	55.13	71.46	73.05
DeepSeek_R1_14B_Instruct	71.33	44.35	50.45	81.96	85.52	92.81	39.50	50.20	52.76	59.49
DeepSeek_R1_32B_Instruct	73.68	51.20	50.86	83.27	87.54	97.81	41.50	53.92	56.80	66.29
Our LLMs										
DeepSeek_R1_14B_SFT	80.63	51.67	52.61	83.53	89.05	96.72	36.68	50.43	50.85	65.36
14B $\Delta_{\text{Instruct}}(\text{SFT})$	+9.25	+7.28	+2.19	+1.53	+3.42	+3.85	-2.93	+0.24	-1.92	+5.83
DeepSeek_R1_14B_SFT_GRPO	81.46	52.41	53.47	83.57	89.02	95.58	36.83	54.06	51.24	66.92
14B $\Delta_{\text{Instruct}}(\text{SFT_GRPO})$	+10.03	+8.09	+2.91	+1.64	+3.45	+2.63	-2.74	+3.82	-1.53	+7.46
14B $\Delta_{\text{SFT}}(\text{SFT_GRPO})$	+0.82	+0.85	+0.81	+0.06	0.00	-1.23	+0.25	+3.63	+0.42	+1.53
DeepSeek_R1_32B_SFT	80.45	66.42	63.28	86.88	88.43	93.76	46.05	55.27	68.41	70.08
32B $\Delta_{\text{Instruct}}(\text{SFT})$	+6.75	+15.23	+12.37	+3.64	+0.83	-4.14	+4.52	+1.25	+11.63	+3.75
DeepSeek_R1_32B_SFT_GRPO	80.67	66.83	64.45	86.79	88.52	91.26	45.68	54.83	67.75	70.44
32B $\Delta_{\text{Instruct}}(\text{SFT_GRPO})$	+6.95	+15.62	+13.57	+3.55	+0.93	-6.64	+4.13	+0.85	+10.93	+4.12
32B $\Delta_{\text{SFT}}(\text{SFT_GRPO})$	+0.23	+0.42	+1.23	-0.06	+0.13	-2.53	-0.42	-0.43	-0.72	+0.33

E.2 Case study on cold-started responses

We present a case study comparing the original response and the cold-started response generated by DeepSeek_R1_14B_Instruct and DeepSeek_R1_14B_SFT. The results are illustrated in Figure 38. Before applying cold-starting techniques, the original response tends to be verbose and includes unnecessary elaboration. After cold-starting, the response becomes more concise and is much longer, which indicates the cold-started model is leveraging its reasoning capabilities to provide a more comprehensive answer. This demonstrates the effectiveness of cold-starting in enhancing the clarity and relevance of model-generated responses.

```

<think>
### 推理过程

1. **市场环境分析**:
    - **宏观环境**: A股市场近期资金流入积极，尤其是医疗健康和科技板块。市场情绪乐观，但外部经济环境存在不确定性。
    - **行业动态**: 数字货币概念股表现强劲，受政策利好推动，稳定币在香港的进展和比特币价格上涨对市场有正向刺激作用。
    - **公司新闻**: 神州信息发布了“乾坤”企业级数智底座白皮书，与华为等合作推出金融知识问答联合智能体，并成立金融科技大模型联合实验室。这些进展显示公司在金融科技领域的领先地位和创新能力。

2. **技术面分析**:
    - **趋势特征**: 均线多头排列，显示短期趋势向上。
    - **量能分析**: 成交量显著放大，价量关系正相关，显示资金流入积极。
    - **指标信号**: MACD金叉，RSI未超买，趋势健康。

3. **基本面分析**:
    - **财务状况**: 尽管净利润亏损，但公司在技术研发和市场拓展方面有积极进展，未来盈利能力有望提升。
    - **估值水平**: 当前估值较低，市盈率为负，可能反映市场对未来的悲观预期，但随着业务改善，估值有上升空间。

4. **综合评估**:
    - **利好因素**: 政策支持、行业领先、技术突破、资金流入。
    - **风险因素**: 历史亏损、高负债、现金流压力。

### 证据评分

- **支持上涨的证据**:
    1. **行业利好政策**: 数字货币和AI大模型的应用推动行业发展，评分: 9/10。
    2. **技术突破和新品发布**: 发布“乾坤”平台和AIGC生态体系，提升竞争力，评分: 8/10。
    3. **资金流入和市场情绪**: 融资买入增加，市场情绪乐观，评分: 8/10。
    4. **量价配合**: 成交量放大，价量正相关，评分: 7/10。
    5. **均线多头排列**: 显示短期趋势向上，评分: 7/10。

- **支持下跌的证据**:
    1. **历史亏损**: 公司过去净利润亏损，评分: 6/10。
    2. **高负债和现金流问题**: 负债率较高，现金流净额为负，评分: 7/10。
    3. **行业竞争加剧**: 科技和金融板块竞争激烈，评分: 6/10。
    4. **政策不确定性**: 全球经济和贸易政策的不确定性，评分: 6/10。

### 涨跌幅预测

综合考虑利好因素和技术面的积极信号，预计下一个交易日的开盘价将上涨约1.5%。

### 方向预测

预计涨幅在3%以内，属于震荡区间。
</think>
<score>[10.0, 6.0]</score>
<pct_change>0.0150</pct_change>
<answer>hold</answer>
```

```

<think>
好的，我现在要预测神州信息（000555.SZ）在下一个交易日（即6月3日）的开盘价相对于5月30日的收盘价的涨跌幅。我的分析流程如下：

1. **理解所要分析的股票的个性**: 神州信息属于IT服务行业，业务涵盖金融科技、量子通信、数字货币等多个高科技领域。作为行业龙头，其股价波动受政策和技术突破双重影响。需要特别关注其在AIGC（生成式AI）领域的布局以及与华为等科技巨头的合作。
2. **理解所要预测的时间特征**: 当前处于5月底至6月初的时间窗口，属于半年报披露前的敏感期。同时临近周末，需警惕周五异动对周初情绪的影响，但周一作为新的交易周通常存在惯性上涨效应。
3. **查看所提供的市场状态**:
    - **宏观环境**: A股科技板块热度持续，但整体市场面临流动性压力。美债收益率上行压制风险偏好，需警惕外资流动对成长股估值的压制。
    - **基本面数据**: 公司2024年出现亏损，经营现金流持续为负，资产负债率超过58%。这些财务指标显示公司处于基本面修复初期，估值修复动能可能受制于盈利能力的不确定性。
    - **价格信息**: 5月30日收盘价12.09元，较前一日上涨0.0025。成交量激增至103亿股创年内新高，呈现量价齐升态势。技术面上显示5日均线（11.97）与10日均线（11.78）形成多头排列，MACD红柱放大。
    - **融资融券数据**: 5月29日融资买入6044.86万元，占当日流入资金比例12.65%，融资余额5.15亿元触及历史60%分位。融券余额124.75万元处于低位，显示做空力量有限。
    - **新闻事件**: 连续三天发布与AI金融大模型相关的重磅合作，包括与华为联合发布智能体、发起行业首个AIGC生态体系等。这些事件形成持续催化效应，推升市场关注度。
4. **初步构建分析方法逻辑**:
    - 结合技术面量价配合与资金流入情况，判断是否存在突破性上涨动能。
    - 评估政策利好（香港稳定币法案）与基本面承压（亏损、现金流）的矛盾。
    - 关注相似股票（如300079.SZ、300479.SZ）的走势对板块情绪的传导作用。
5. **按分析方法逻辑分析各维度信息**:
    - **利多因素**:
        - 技术面：5日量价齐升突破整理区间，MACD金叉形成多头趋势。
        - 政策支持：香港稳定币法案推动数字货币板块整体走强，神州信息作为概念股核心受益。
        - 事件驱动：连续三天发布AI金融大模型相关成果，与华为等巨头合作形成持续催化。
        - 融资盘介入：融资余额突破历史60%分位，显示杠杆资金积极布局。
    - **利空因素**:
        - 基本面弱弱：2024年亏损、现金流为负，经营效率低下制约估值修复空间。
        - 流动性压力：美债收益率上行可能抑制成长股估值，板块轮动压力显现。
        - 技术面阻力：12.09元附近面临12.10元整数关口压力，需观察量能否持续放大。
6. **证据评分**:
    - **上涨支持因素**:
        1. 技术面量价配合（9分）：突破整理区间伴随放量，MACD金叉确认趋势。
        2. 政策利好催化（8分）：香港稳定币法案提升板块情绪，数字货币概念股普涨。
        3. 事件驱动持续性（8分）：三天连续发布AI金融成果，形成技术面多头氛围。
        4. 融资盘介入（7分）：融资余额突破历史分位，杠杆资金形成推力。
    - **下跌支撑因素**:
        1. 基本面弱弱（7分）：亏损与现金流为负制约估值修复，需警惕技术性回调。
        2. 流动性压力（6分）：美债收益率上行可能引发板块资金撤离。
        3. 技术面阻力（6分）：12.10元关口存在压力，量能若无法持续可能引发获利了结。
        4. 融券余额低位（5分）：融券力量薄弱，空头打压动能不足。
7. **假设检验与推演**:
    - 如果政策利好持续发酵，AI金融大模型落地项目加速，股价可能突破12.10元阻力位。
    - 若市场对亏损业绩存在分歧，可能出现技术性回调至11.80元附近。
    - 考虑到相似股票（300079.SZ）在利好驱动下涨幅接近5%，板块联动效应可能放大个股涨幅。
8. **综合评分计算**:
    - 上涨平均分:  $(9+8+8+7)/4=8.0$ 
    - 下跌平均分:  $(7+6+6+5)/4=6.0$ 
9. **最终预测**:
    虽然基本面存在亏损隐患，但技术面量价配合与政策利好形成共振，且融资盘介入形成推力。考虑到周末消息面无重大利空，预计周一开盘将延续涨势，涨幅3%左右，方向为显著上涨。
</think>

<score>[8.0, 6.0]</score>
<pct_change>0.0300</pct_change>
<answer>up</answer>
```

Figure 38: Case study on cold-started responses. Left is the original response, and right is the cold-started response. The cold-started response is more concise and to the point, avoiding unnecessary elaboration.

WORKING PAPER

Measuring Retail Trade Using Card Transactional Data

Diego Bodas / Juan R. García / Juan Murillo / Matías Pacce / Tomasa Rodrigo / Pep Ruiz de Aguirre / Camilo Ulloa / Juan de Dios Romero and Heribert Valero



Measuring Retail Trade Using Card Transactional Data¹

**Diego Bodas¹ / Juan R. García² / Juan Murillo¹ / Matías Pacce³ / Tomasa Rodrigo² /
Pep Ruiz de Aguirre² / Camilo Ulloa² / Juan de Dios Romero¹ and Heribert Valero¹**

1: BBVA Data & Analytics, Madrid, Spain

2: BBVA Research, Madrid, Spain

3: Banco de España, Madrid, Spain*

Abstract

In this paper we present a high-dimensionality Retail Trade Index (RTI) constructed to nowcast the retail trade sector economic performance in Spain, using Big Data sources and techniques. The data are the footprints of BBVA clients from their credit or debit card transactions at Spanish point of sale (PoS) terminals. The resulting indexes have been found to be robust when compared with the Spanish RTI, regional RTI (Spain's autonomous regions), and RTI by retailer type (distribution classes) published by the National Statistics Institute (INE). We also went one step further, computing the monthly indexes for the provinces and sectors of activity and the daily general index, by obtaining timely, detailed information on retail sales. Finally, we analyzed the high-frequency consumption dynamics using BBVA retailer behavior and a structural time series model.

Key words: retail sales, Big Data, electronic payments, consumption, structural time series model

JEL classification: C32; C55; C81; E21

1: We are grateful to Gonzalo de Cadenas-Santiago for his contribution in the first stages of the project, Miguel Cardoso, Alvaro Ortiz and for their comments, as well as all of those who give us feedback, which greatly improved the quality of this paper.

The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Banco de España.

1. Introduction

Recent improvements in data storage, management, and processing have led to an exponential increase in the amount and quality of the information available for economic analysis, both from an individual and a macroeconomic perspective. In particular, the latest developments in Big Data technologies permit a ‘quasi-real-time’ analysis of the information emerging from citizens, governments, and firms in all interactions that generate a digital footprint. New data sources, such as those originating from social networks and search engines, have been proven to help the forecasting of economic variables, such as employment, consumption, and tourism flows (see (Chamberlain, 2010); McLaren 2011; (Choi & Varian, 2012); (Camacho & Pacce, 2017); among many others). More recently, (Cavallo A. , 2016) and (Cavallo & Rigobon, 2016) have shown that the prices of goods and services offered over the Internet can be used to estimate high-frequency price indices. The authors used web scraping tools for the data compilation and proposed a common methodology for generating a price index for a large sample of countries, to allow for the international comparability of prices. We wish to contribute to this line of research by proposing an alternative method for measuring the business evolution of the retail trade sector based on data from credit and debit card transactions.

The retail trade index (RTI) has traditionally been measured by National Statistics Institutes (NSIs) using surveys conducted with a limited sample of retailers, resulting in the collection of relevant information based on data from the supply side. In this paper, we propose a different focus and show that, as could be expected, data emerging from the demand side can also offer similar measures to the official statistics. In particular, we replicated the evolution of the Spanish RTI released by the Spanish National Institute of Statistics (INE), using information obtained from retail transactions by credit and debit card holders of BBVA (one of the largest banks in Spain). The possibility of studying aggregate economic patterns from individual economic transactions using card transaction data was demonstrated by (Sobolevsky, et al., 2015), who were able to obtain regional socioeconomic signals in Spain using this type of information. Also, electronic payment data has already been shown to be helpful in forecasting the evolution of economic aggregates (see Tkacz, 2013; Galbraith and Tkacz, 2016 or (Duarte, Paulo, & Rua, 2017)). However, to the best of our knowledge, this is the first time that the evolution of an official index published by an NSI has been replicated using information from credit and debit card transactions.

Having accurate estimates of the evolution of retail trade sector activity is of great importance given that this is a key indicator of the current economic situation. In general, its dynamic drives the evolution of aggregate consumption (see Figure 1.1), which in turn represents a high proportion of the gross domestic product (GDP). In this sense, it is not surprising that most short-term macroeconomic forecasting models used by Central Banks or private agencies around the world include the aggregate RTI as an important input. One clear example comes from Stock and Watson ((Stock & Watson, 1989), (Sims, Stock, & Watson, 1990), (Stock & Watson, 1991), who included the RTI as one of the four economic indicators needed to construct a coincident indicator index for the evolution of the US GDP. Besides this important feature, the RTI is also important as an indicator for studying the evolution of the retail sector itself, as the possible disaggregation published by the NSIs (by sectors or regions) is key for a detailed analysis.

[Figure 1.1]

The results of this paper show that developments in Big Data analysis have the potential to replicate the evolution of relevant macroeconomic indicators. In particular, we reproduced the dynamics not only of the aggregate Spanish RTI but also the regional RTI (of the Spanish autonomous regions) and retailer type (distribution classes). A number of benefits emerged from our proposed methodology, which were related to the quasi-real-time availability of the data, the higher frequency with which the index can be computed, and the greater geographical and sectoral disaggregation. In this sense, we were able to construct a RTI for the 50 Spanish provinces, a geographical detail that is not published by the INE, and even a daily aggregate Spanish RTI. In addition, based on this daily index, we were able to analyze consumption dynamics by using a structural time series model like the one proposed by (Harvey, Koopman, & Riani, 1997). We found regular, significant patterns that displayed strong intra-weekly, intra-monthly and intra-yearly seasonalities, which were also affected by holiday effects.

The remainder of this paper is organized as follows: Section 2 describes the methodology followed to replicate the Spanish RTI and the data used, while Section 3 shows how alternative indices are a good way to replicate the dynamics of the official ones. Section 4 describes the daily model used to study regular consumption pattern and Section 5 presents the conclusions.

2. An Alternative Way to Compute Retail Trade Indexes

NSIs around the world use the same pillars to estimate the evolution of business in the retail trade sector, which is in general summarized in the retail trade index (RTI). This index reflects the total gross sales of retailers during a fixed period of time (generally a month) and is constructed by conducting surveys directed to a limited number of companies selected using random sampling techniques². In other words, the relevant data is collected using information obtained from the supply side. Alternatively, it is possible to consider getting the same information from the demand side by using surveys asking the retailers' customers about their expenditure. Even though the latter was never a real option for statistical offices, a major breakthrough has occurred in view of recent developments in Big Data technologies. In particular, the increase in payments using credit and debit cards makes it possible to use the information recorded whenever a credit or debit card is used for a retail transaction to obtain similar measures to the ones given in the RTI, but using real data on consumption instead of data from surveys. Taking this hypothesis, we propose an alternative measure to the Spanish RTI that is based on the information obtained from retail transactions made by credit and debit card holders of the BBVA's Spanish bank.

² In the case of Spain, information is obtained from a sample that covers between 20% and 25% of the 12,500 companies registered in the Central Companies Directory (CCD), which provide data by completing a monthly questionnaire over the telephone, email, fax, or the web.

2.1 The Data Sources

We analyzed a complete set of Point of Sale (PoS) purchases or transactions performed in Spanish retail stores between 2013/01 and 2016/12 by clients of BBVA Spain³. For the purposes of this paper, we focused on information⁴ relating to the amount of each transaction, the geo-localization and principal activity of each PoS, as well as the company that owns it and the exact time the transaction took place. Following the definition given by INE, retail trade here did not include expenditure on motor vehicles and motorcycles, food service, the hospitality industry, or financial services, while sales at gas stations were taken into account. In other words, retail activities refer to Section G, category 47 of the National Classification of Economic Activities (CNAE-2009). Filtering for that specific category is possible because the dataset on card transactions includes the main activity of each PoS.

Following INE, we also grouped purchases into 5 distribution classes based on the following categories of retail store⁵:

1. Gas stations.
2. Department stores: premises⁶ with 2,500 m^2 or more.
3. Large chain stores: chain stores with 25 premises or more.
4. Small chain stores: chain stores with more than one premises and less than 25.
5. Single retail stores: only one premises.

We only considered transactions below €30,000 (the maximum credit limit for BBVA Gold Cards). Values over that threshold were considered to be outliers. The sample contained approximately 12,500 stores and more than 200 million annual transactions by over 4 million cardholders. Columnar databases were used to deal with this huge amount of information.

2.2 Methodology

In order to build the Spanish BBVA-RTI based on card transactions it was necessary to create a data engine capable of regularly delivering the index pursued. With this aim, a number of steps were followed during the process of building the data engine (see Figure 2.1).

[Figure 2.1]

We started selecting data sources and useful variables to meet the goals of the project. Big data queries are not a trivial task when using columnar databases and cluster solutions, as they need to be optimized in order to avoid

³ Both face-to-face and online purchases have been analyzed for this project.

⁴ The transactions database has been anonymized and aggregated before analyzing it.

⁵ Since information on the number of employees in each company was not available, we did not include the restriction of having more than 50 employees or more in the definition of a “large chain” or “small chain”

⁶ According to the INE, a premises is “any structurally separate and independent building that is not dedicated exclusively to family housing, and in which economic activities dependent on a company are carried out, and in which one or more persons work for the company”.

cluster failure or malfunction. We obtained daily, weekly, and monthly aggregate data on the total number and total amount of the transactions. At the same time, we queried this information at different levels of granularity, getting data for the entire country, data for each of the 17 autonomous regions of Spain, and the 50 provinces⁷. Ensuring data quality required that the data be cleaned and formatted. During this process, outliers were deleted before the data were standardized⁸. In the final stage, the data were tested to check whether these data sources and variables were useful for the project's goals. Finally, the process was automated by implementing a code library.

2.3 Strengths and Weakness of the BBVA-RTI

As we are proposing an alternative way of computing the Spanish RTI (which could potentially be translated to other countries), it is important to point out the advantages and disadvantages of using card transaction data rather than the classical method of estimation used by official statistical offices. The comparison is summarized in Table 19.

Table 2.1 Comparison between RTI Data Sources

	Card Transaction Data (BBVA)	Survey Data (INE)
Cost per observation	Marginally Low	High
Data Frequency	Daily	Monthly
Disaggregation by activity	High	Low
Geographical disaggregation	High	Low
Real-time availability	Yes	No
Retailer sample	12,500	≈ 3,000
Payment methods covered	BBVA's clients credit and debit cards	All
Possible bias of technological trends	Yes	No

The first advantage of using card transaction data is related to the cost of obtaining each extra observation. Even though storing huge amounts of information is not cheap, the economic scale of digital information storage means the cost of marginal observation is close to zero, allowing for an obvious gain as compared to conducting 50 parallel regional monthly surveys (one for each INE provincial delegation) to obtain the relevant information.

A second advantage is related to the frequency of data collection, which allows for a deeper analysis of the behavior of retailers' customers than the one than can be performed when using monthly data. Section 5 shows an example of this applicability.

⁷ Spain's autonomous regions and provinces correspond, respectively, to NUTS-2 and NUTS-3 in EUROSTAT nomenclature.

⁸ We took natural logs or the first difference of logs, subtracted the mean, and divided by standard deviations.

⁹ This table is comparable to Table 1 of (Cavallo & Rigobon, 2016), which describes the advantages and disadvantages of using online data prices to construct an alternative consumer price index.

Thirdly, card transaction data include information on each PoS's main activity, allowing for greater economic activity disaggregation, and not only for each of the 5 groups published by INE. As an example, in Section 4, we show the median expenditure by sector at the end of 2017.

Fourthly, the geographical disaggregation that can be obtained is greater than that published by INE. In particular, with data on the geo-localization of each PoS, it is potentially possible to generate an RTI for a city or even a single postcode. In the present work, we computed the RTI at the provincial level, a disaggregation that is not available for the INE data.

Fifthly, INE publishes its data with a one-month delay, while card transaction data is available almost in real time. This would allow policymakers to access the latest information without any kind of delay.

Lastly, in regard to the sample, the card transaction data include almost the entire sample of companies registered in the CCD (12,500 companies), while INE's data are based on a sample that covers only some 25% of the companies registered.

On the other hand, some disadvantages can be found when using card transaction data. Firstly, the total amount billed only refers to expenditure made using BBVA client credit and debit cards. This means that we excluded all transactions made using cash or non-BBVA cards. Nonetheless, given BBVA's high market share (13.8%) in Spain, we assumed that the sample we were using was sufficiently representative. A second disadvantage was related to the potential bias that technological trends could generate if they affect preferences for using credit or debit cards.

3. The Spanish Retail Trade Index

In this section, we show that information obtained from card transaction data can replicate the dynamics of the official RTI for Spain, not only for the national aggregate, but also for all five sub-divisions of the national index and for all 17 official retail trade indices published for each of the autonomous regions.

3.1 Similarities with the Official Aggregate Indices

By comparing both the official Spanish RTI and the BBVA-RTI, it is easy to see how the two are closely related. In Figure 3.1, we plot the monthly BBVA-RTI next to the nominal non-seasonally adjusted official RTI. As can be seen, even though the dynamic of both indices appears to be similar, the BBVA-RTI shows a steeper trend than the official index. As previously said, this may be associated with the existence of some kind of technological trend affecting consumer behavior and the intensity of use of credit and debit cards. As an example of this pattern, in Figure 3.2 we plot the evolution of the BBVA-RTI's average transaction amount. As shown, the median transaction amount decreased from €45.70 in December 2013 to €40.50 in June 2017. This result, together with the upward trend in the BBVA-RTI, can be interpreted not only as the fact that people are increasingly using credit and debit cards but also that there is a higher number of lower amount transactions. Additionally, the BBVA data may be affected by the addition of new clients or the loss of old ones. This is particularly relevant in the case of mergers and acquisitions, like

BBVA's absorption of UNIM and CatalunyaCaixa in the second quarter of 2013 and the last quarter of 2016¹⁰, respectively.

[Figures 3.1 and 3.2]

Even though Figure 3.1 does not clearly show the official RTI as being a non-stationary series, it does become clear when the index is plotted for a larger sample period¹¹ (see Figure A1 in the Annex). As both series show a non-stationary pattern (even though with different trends), working with growth rates is a possible solution to continuing with the comparison. In Figure 3.3, we show how the similarities between the indices become stronger when expressed in monthly standardized¹² growth rates, giving strong support to the BBVA-RTI as a very close approximation to the official index. This became even more evident when we analyzed the 5 distribution channels for which the INE gives retail trade indices, besides the aggregate one. In Figure 3.4, it is possible to see that those similarities are high enough for all 5 disaggregations, even though with some heterogeneity between them. The case of "department stores" is the one that shows most proximity between the indices, even though the agreement is also very high for "large chain stores". When analyzing the cases of "small chain stores" and "single retail stores", it can be seen that the dynamics of the series (in monthly growth rates) are similar, but the indices built on card transaction data appear to be more volatile. In contrast, "gas stations" is where greater differences appear. One possible explanation for the greater similarities found for the indices relating to larger retailers could be explained by a more intense use of credit and debit cards in these kinds of stores. The differences that emerge for "gas stations" could be due to inflows of cash payments. In the left pane of table 3.1, we show the R-squared from the linear regression between the BBVA-RTI and the official indices. In the right pane of the table, we show the Hansen stability test p values for these regressions. The results indicate that, although the correlation between the series levels is high, the relationship is statistically stable for all the distribution channels only when growth rates are taken into account. Altogether, the results clearly show that co-movement between the series is robust enough to reinforce the idea that card transaction data provide suitable information for correctly replicating the official RTI.

[Figure 3.3]

[Figure 3.4]

¹⁰ BBVA Research estimates indicate that the relationship between BBVA-RTI's growth rates and the official RTI deviate by 0.47 and 0.59 sd in September and October 2016 as a result of the absorption of CatalunyaCaixa. No statistically significant effects were found for national transactions after the UNIMUnnim takeover.

¹¹ The official data start in 1995/01.

¹² We rescale each of the month to month growth rate series to have a mean of zero and a standard deviation of one.

Table 3.1 INE-RTI and BBVA-RTI

	R-squared		Hansen stability test P-Value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Total	0.89	0.94	0.22	0.90
Department Stores	0.89	0.95	0.03	0.99
Large chain stores	0.73	0.91	0.32	0.93
Small chain stores	0.48	0.91	0.30	0.67
Single Retail stores	0.57	0.92	0.21	0.85
Service stations	0.05	0.79	0.01	0.18

Note: sample: 2013.01-2017.12 period

3.2 Similarities with Regional Indices

As mentioned above, the INE publishes an RTI for each of Spain's 17 autonomous regions. Taking advantage of the high geographical disaggregation permitted by the card transaction data, we constructed each of these 17 indices based on that information. Figure 3.5 shows the dynamic of all the regional INE-RTI and BBVA-RTI, once again expressing the indices as monthly growth rates for better comparability. The figures show the great similarities in the dynamics of the indices, which are also reflected in the high R-squared and Hansen stability test p values (see Table 3.2).

[Figure 3.5]

One of the bi-products of using card transaction data for computing regional RTIs is the possibility of obtaining 5 distribution groups for each of the autonomous regions, which are not publicly available from INE13.

¹³ All figures regarding the 5 distribution groups by regions are available upon request

Table 3.2 INE-RTI and BBVA-RTI by Autonomous Region (R-squared)

	R-squared		Hansen stability test P-value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Andalusia	0.72	0.96	0.14	0.32
Aragon	0.52	0.91	0.05	0.24
Asturias	0.76	0.94	0.20	0.70
Balearic Islands	0.72	0.90	0.19	0.96
Canary Islands	0.87	0.94	0.16	0.21
Cantabria	0.87	0.96	0.10	0.87
Castile and León	0.70	0.94	0.05	0.96
Castilla-La Mancha	0.65	0.93	0.09	0.86
Catalonia	0.38	0.83	0.37	0.23
Valencian Community	0.81	0.94	0.06	0.69
Extremadura	0.54	0.94	0.06	0.48
Galicia	0.82	0.95	0.07	0.51
Community of Madrid	0.79	0.96	0.09	0.88
Region of Murcia	0.65	0.91	0.27	0.15
Navarre	0.78	0.91	0.08	0.27
La Rioja	0.78	0.95	0.05	0.82
Basque Country	0.82	0.95	0.47	0.38

Note: sample period: 2013/01-2017/12

4. Higher Dimensionality: Granular Data by Time Span, Geography and Further Dimensions

After checking that the constructed BBVA-RTI replicated the official figures published by the INE at all levels in which is available, we went one step further in taking advantage of the BBVA quasi-real-time transaction data, getting insights into retail sales at higher frequencies (e.g., daily) with greater geographical detail (i.e., at the provincial level), as well as exploiting new dimensions that the INE-RTI does not provide, both on the supply side (e.g., sector of activity) and the demand side (e.g., socioeconomic characteristics of consumers, such as sex, age, and income level).

The high frequency of the BBVA-RTI (Figure 4.1) covered the one-month lag in publication by INE, providing timely answers on retail sales for particular events. It also permitted a deeper analysis of the retailers' customers' behavior,

uncovering the aggregate consumption dynamic using a structural time series model like the one proposed by (Harvey, Koopman, & Riani, 1997) (an example is shown in Section 5).

[Figure 4.1]

The geo-located information from the PoSs gave a higher geographical disaggregation to the BBVA-RTI, providing information on the evolution of retail sales that is not published by the INE. We had the RTI for the 50 provinces (NUTS 3 geographical division in the EUROSTAT nomenclature), but it would potentially be possible to generate an RTI at the city or postcode level. Figure 4.2 shows the evolution of the RTI in December 2017 in each province as compared to December 2016 (yoY levels). Although it was not feasible to compare the dynamics of these series with the INE statistics (the INE-RTI information is published by autonomous region), the consistency of the index at the national level and by autonomous region, as well as the correspondence between the INE-RTI and the BBVA-RTI for autonomous regions with only one province, brings a high likelihood to the rest of the provincial indices.

[Figure 4.2]

The transaction data included information on the main activity of each PoS, allowing for more economic activity disaggregation than the information published by INE. The analysis of the BBVA-RTI by sector of activity showed that the lowest median ticket in December 2017 was for healthcare, other services, and the bar and restaurant sectors, respectively. In contrast, technology, sports and toys, and automotive were among the sectors in which we found the highest median expenditure. (Figure 4.3).

[Figure 4.3]

5. Working with Daily Data

One of the most important features of working with card transaction data is the possibility of studying aggregate consumption patterns. In other words, given that high frequency data is available, it is feasible to study the actors' decisions regarding daily, or even hourly, expenditure. The daily BBVA-RTI displays weekly, monthly and annual seasonalities, plus some calendar effects. Even though the figure shows a very volatile pattern, it is clear that Saturdays are in general the day on which people consume the most, while on Sundays they consume the least. In addition, it seems that, within a year, December is the month with the highest consumption, followed by July, while calendar effects relating to public holidays or Easter Week can be found where the blue line becomes thicker. Modeling all those patterns into one single model that operates at a daily frequency was not an easy task and several issues had to be taken into account. Not only the numbers of days within a month or within a year¹⁴ change, but also the position of the date on a specific day of the week¹⁵ or for holidays like Easter is not the same from year to year. As mentioned in Cabrero et al. (2007), two major approaches exist for dealing with those and other problems in the context of daily time series: the ARIMA model suggested by (Bell & Hillmer, 1984) and the structural time series (STS)

14: The number of days in a year depends on its being a leap year or not.

15: e.g., January 1st is not always on a Monday.

approach of (Harvey, Koopman, & Riani, 1997). In this paper, we used the second approach, which includes *periodic cubic splines* to model some of the seasonal components exhibited by the daily BBVA-RTI data¹⁶.

[Figure 5.1]

Using the Harvey et al. (1997) notation, the basic STS model can be described for a univariate time series (y_t) where

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$

where μ_t , γ_t and ε_t are, respectively, the stochastic trend, the stochastic or deterministic seasonal components, and the irregular component, while $t = 1, \dots, T$. The dynamic of the stochastic trend is defined by,

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim NID(0, \sigma_\eta^2)$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim NID(0, \sigma_\zeta^2)$$

where the level and the slope of the trend are given by μ_t and β_t , while η_t and ζ_t are mutually independent processes. The seasonal component is characterized as the sum of the day of the week effect (γ_t^d), the intra-monthly effects (γ_t^m), the intra-yearly effects (γ_t^y), and moving and fixed holidays (γ_t^h)¹⁷:

$$\gamma_t = \gamma_t^d + \gamma_t^m + \gamma_t^y + \gamma_t^h \quad (3.1)$$

Each of the seasonal components is described by its own dynamics. In particular, to model the day of the week effect, we rely on stochastic dummies of the form:

$$\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d + \omega_t, \quad \omega_t \sim NID(0, \sigma_\omega^2) \quad (3.2)$$

where $s = 7$ is the number of days in a week¹⁸. By imposing $\sigma_\omega^2 = 0$, the seasonality becomes deterministic (the main results remain unchanged when this is done).

Intra-monthly and intra-yearly effects are both modeled by using time-varying cubic splines. For setting a spline it is necessary to choose h^i knots in the range of $[0, N^i]$, where N^i is the number of the days in a month or in a year ($i = \{m, y\}$). Once again, following Harvey et al. notation, we define

$$\gamma_d^i = \mathbf{z}_d^i \gamma_t^i \quad d = 1, \dots, N^i ; \quad i = \{m, y\} \quad (3.3)$$

where \mathbf{z}_d^i is vector of dimension $(h^i - 1) \times 1$, which depends on the number and positioning of the knots and should be defined in a way that guarantees continuity from one period to the next. By letting the vector γ_t^i follow a random

16: Harvey et al. (1997) indicate that, by giving more scope for a parsimonious parameterization, this approach better captures periods with sharp peaks, like the one that can be observed surrounding Christmas.

17: For notational simplicity, we have included the fixed and moving festivals as seasonal components, even though both are calendar effects.

18: Dummies (z_{jt}) in equation 3.1 are introduced as $\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d z_{jt} + \omega_t$ for $t = 1, \dots, T$ where for $t = i, i+s, i+2s, \dots$ and $i = 1, \dots, s-1$ the variable z_{jt} is one for $j = i$ and zero for $j \neq i$, while for $t = s, 2s, 3s, \dots$ and $j = 1, \dots, s-1$ the value of z_{jt} is equal to minus one.

path, γ_d^i becomes stochastic. For a detailed explanation on the modeling of periodic cubic splines, see Harvey et al. (1997).

In order to specify the length N^i , it was necessary to take into account the fact that not all months or all years have the same number of days. To deal with this problem, we followed the strategy of Cabrero et al. (2009) and set $N^m = 31$ for all months and $N^y = 366$ for all years. The days that do not exist (February 29th when the year is not a leap year or days like April 31st) were considered to be missing values and were easily handled, given that our estimation strategy relied on Kalman filter iterations. When using periodic cubic splines, a second issue to be taken into account was related to the number and position of the knots. Moreover, to obtain periodicity, the value of the first and the last knot within a period should be equal¹⁹. Therefore, two consecutive days with similar characteristics should be chosen for placing the starting and final knots. Cabrero et al. (2009) correctly highlight that setting the first and last knot as January 1st and December 31st for intra-yearly seasonality gives the particularities of those days, while for intra-monthly patterns the first and last days of the month are less likely to be similar than two days in the middle of the month²⁰. The dates finally chosen for the placement of the first and last knot for the annual periodicity were February 18th and 19th, while the 22nd and 23rd of each month were selected for monthly splines. The decision on the final number and position of the knots was based on the analysis of residual correlograms, goodness-of-fit performance, and visual observation. In particular, after trying many different specifications, we decided to include 7 knots for intra-month patterns and 18 knots for intra-year seasonality. As in (Harvey, Koopman, & Riani, 1997), when dealing with annual seasonality, we needed to impose a relatively larger number of knots in the short period of time surrounding Christmas, while fewer knots were needed when seasonal patterns changed slowly. The knots for the intra-yearly spline were placed at 1, 5, 9, 13, 17, 21, 25 and 28²¹ and for the intra-monthly spline at 1, 25, 50, 75, 125, 150, 200, 225, 251, 286, 296, 301, 306, 313, 320, 330, 345, and 365²².

To model holidays effects (γ_t^h), we used a deterministic approach and included dummy variables for each of those specific days²³. It should be noted that $\gamma_t^h = \sum_{i=1}^I \gamma_t^{h,i}$ where $i = 1, \dots, I$ is an indicator for each holiday. Under this notation, the holidays effect was modeled as

$$\gamma_t^{h,i} = w_i(B)h(\tau_i, t) \quad (3.4)$$

where $w_i(B)$ is a polynomial lag operator and $h(\tau_i, t)$ is an indicator function that takes the value 1 when $\tau_i = t$ and zero otherwise. The presence of a polynomial lag operator is related to the fact that days surrounding a holiday could

19: As mentioned in Cabrero et al. (2009), this is strictly true only for the case of deterministic periodic cubic splines.

20: It should also be remembered that in 5 out of the 12 months of the year, the last day of the month does not really exist (e.g., April 31st) and it is considered as a missing value for estimation purposes. Also, the end of the month displays a sharp trend given the monthly seasonal pattern.

21: Notice that knots 1 and 365 correspond, respectively, to February 18th and 19th

22: Knots 1 and 28 correspond, respectively, to the 22nd and 23rd day of the month.

23: In Spain, there are three classes of public holidays: national holidays, holidays specific to each autonomous region and municipal holidays. As the daily model will be applied to the national aggregate RTI, only national holidays were taken into account.

also show some peculiarities (e.g., people going to the supermarket the day before a holiday)²⁴ (see table 5.1 to see the polynomial lag operator set for each holiday). When a National holiday falls on a Sunday, we opted not to include a dummy for that day.

Table 5.1 Fixed Holiday Lag Polynomials

Card Transaction Data (BBVA)	
Good Friday	$(w_0 + w_1B + \dots + w_{12}B^{12})B^{-6}$
New Year, Epiphany (Jan 6th), St. Joseph (Mar 19th), Labor Day (May 1st), Assumption (Aug 15th), Spain's National Holiday (Oct 12th) and All Saints' Day	$(w_0 + w_1B + \dots + w_4B^4)B^{-1}$
Immaculate Conception (Dec 8th)	$(w_0 + w_1B + \dots + w_8B^8)B^{-3}$

Since the whole model described in 3.1-3.4 can be written in state space form, maximum likelihood estimation in combination with a Kalman filter and Kalman smoothed could be used. The main results are summarized in Figures 3.2 to 3.7.

Figure 5.1 shows the intra-weekly effects for a week of the year²⁵. As highlighted by the raw data (Figure 4.1), Sundays are the days of the week with the lowest consumption while Saturdays have the highest. This behavior is not surprising in a country like Spain, where Sunday is a day for the family and rest, meaning that most retail shops are closed. On the other hand, it seems that Saturdays are used for doing the shopping that is harder to do on weekdays, maybe because of restrictions caused by working hours. For weekdays, the consumption pattern looks to be very similar between Monday and Thursday, while it rises on Fridays.

The intra-monthly effects are displayed in Figure 5.2. The results show that consumption is higher during the first two weeks and the last three days of a month, suggesting a consumption pattern linked to salary payment²⁶. Working with statistics on the daily banknotes in circulation in Europe, Cabrero et al. (2009) found a similar intra-monthly behavior. This result is in line with Stephens (2003, 2006) (Shapiro, 2005), (Mastrobuoni & Weinberg, 2009) and (Aguila, Kapteyn, & Pérez-Arce, 2017), who found monthly increase in consumption during the week of and the week after payroll. Alternative explanations for this kind of consumption behavior rely on the existence of credit restrictions, liquidity constraints, myopia, or the existence of hyperbolic discounting in the actor's preferences.

Figure 5.3 shows the intra-yearly seasonality. As can be observed, there is a sharp peak starting in the first few days of December and ending around January 10th. This period is related to the Christmas holidays, when the increase in

24: To model holidays, we also took into account the fact that, to be treated as a seasonal effect, the holidays effects plus the non-holiday factor should be null (the dummy variables were altered to get this kind of effect).

25: As we are working with stochastic dummies, the consumption pattern is not exactly similar for every week of the year. Figure 3 shows the intra-weekly behavior for the first week of June, but results are very similar for any week of the year.

26: In Spain, wages are paid monthly, normally on the first or last day of the month.

retail sales may be associated with purchasing Christmas gifts. Another period of high retail consumption is during July, which may be related to the summer sales period. The rest of the year displays a very similar pattern, although February and March appear to be the months with the lowest sales. The same kind of intra-yearly seasonality is the one that the INE found in the monthly data for the RTI.

The holiday effects are shown in Figure 5.4. As expected, all national holidays have a negative effect on retail sales, which is obviously related to the fact that most retail stores are closed on those days²⁷. Also, the days before and after a holiday show a positive pattern. This could be explained by a distribution of consumption around the holiday date if it coincides with a working day. December 25th and January 1st and 6th are the holidays with the highest negative effects on retail sales followed by May 1st.

Given the importance of Easter, Figure 5.5 shows consumption during a period of two weeks around those days. As can be seen, consumption increases during the week prior to the Easter weekend and falls on Good Friday. The fall observed on Easter Monday is not surprising as it is a holiday in some of the biggest autonomous regions (e.g., Catalonia). Finally, Figure 5.6 shows the estimated stochastic daily trend. As expected from dynamic observed in Figure 3.1, a positive trend was found during the period in which the model was estimated.

The results obtained using the daily STS model should be considered carefully. Since we only had four years of data, the intra-yearly and fixed holiday effects were mostly indicative. As time passes and we amass more data, a better estimate will be possible.

6. Conclusions

The new digital era, together with the development of data infrastructure, technologies, and data science techniques, presents a chance for economic research to take advantage of unprecedented amounts of data. In this paper, we developed an alternative way of measuring the retail trade in Spain using high dimensional data collected from the digital footprint of BBVA clients using their credit or debit card transactions at a Spanish point of sale (PoS) terminal.

The results of this paper show that card transaction data replicate with great precision the evolution of the Spanish RTI, an important macroeconomic indicator showing the evolution of aggregate consumption and, therefore, of economic activity. The RTI indicator developed replicated the dynamics of the aggregate Spanish RTI, the RTI by region (Spain's autonomous regions) and the RTI by retailer type (distribution classes). In addition, the high granularity of the data allowed us to reproduce the evolution of daily retail sales, with timely answers on the impact of any retail sales event, great geographical detail (by province or even by postcode) and information on further dimensions (such as the sector of activity).

We also investigated the behavior of retailers' customers to analyze the high frequency consumption dynamics using a structural time series model. We found regular, significant patterns that displayed strong intra-weekly (Sundays are

27: Note the absence of some holidays during the period s plotted (October 12th, for example). As previously explained, we did not include dummies for a holiday when it fell on a Sunday.

the days of the week with the lowest consumption while Saturdays are the ones with the highest), intra-monthly (consumption is higher during the first two weeks and last three days of a month) and intra-yearly seasonalities (we found a sharp peak in retail sales starting in the first few days of December and ending around January 10th, and also in July), which are also affected by holiday effects.

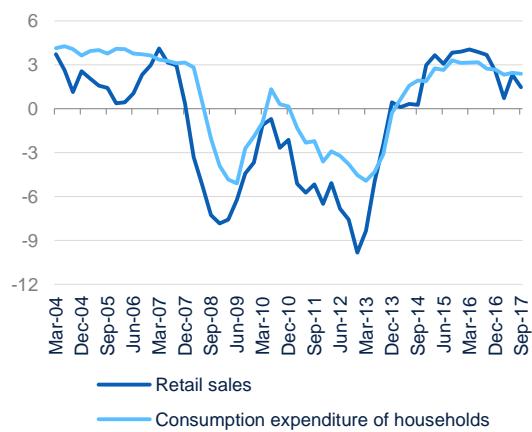
This line of research could be extended to exploit further dimensions offered by the data, such as the credit consumption behavior of BBVA clients or the socioeconomic features of online versus offline payments. Deseasonalizing the index to work with real values instead of nominal ones, and testing its predictive power at nowcasting, is left for further research.

References

- Aguila, E., Kapteyn, A., & Pérez-Arce, F. (2017). Consumption Smoothing and Frequency of Benefit Payments of Cash Transfer Programs. *American Economic Review*, 107(5), 430-35.
- Bell, W., & Hillmer, S. (1984). Issues Involved with the Seasonal Adjustment of Economic Time Series. *Journal of Business & Economic Statistics*, 2(4).
- Camacho, M., & Pacce, M. (2017). Forecasting Travellers in Spain with Google's search volumen indices. *Tourism Economics*.
- Cavallo, A. (2016). Scrapped Data and Sticky Prices. *Review of Economics and Statistics*.
- Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, 30(2), 151-78.
- Chamberlain, G. (2010). Binary Response Models for Panel Data: Identification and Information. *Econometrica*, 78(2), 159-168.
- Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2-9.
- Duarte, C., Paulo, M., & Rua, A. (2017). A Mixed Frequency Approach to the Forecasting of Private Consumption with ATM/POS Data. *International Journal of Forecasting*, 61-75.
- Harvey, A., Koopman, S., & Riani, M. (1997). The Modelling and Seasonal Adjustment of Weekly Observations. *Journal of Business & Economic Statistics*, 15(3), 354-68.
- Mastrobuoni, G., & Weinberg, M. (2009). Heterogeneity in Intra-monthly Consumption Patterns, Self-Control, and Savings at Retirement. *American Economic Journal: Economic Policy*, 1(2), 163-89.
- Shapiro, J. (2005). Is there a daily discount rate? Evidence from the food stamp nutrition cycle. *Journal of Public Economics*, 89(2-3), 303-325.
- Sims, C. A., Stock, J., & Watson, M. (1990). Inference in Linear Time Series Models with some Unit Roots. *Econometrica*, 58(1), 113-144.
- Sobolevsky, S., Bojic, I., Belyi, A., Sitko, I., Hawelka, B., Murillo Arias, J., & Ratti, C. (2015). Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity. *IEEE International Congress on Big Data, Proceedings*, pp. 600-07.
- Stock, J., & Watson, M. (1989). New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual*, 4, 351-409.
- Stock, J., & Watson, M. (1991). A probability Model of the Coincident Economic Indicators. In G. Moore, & K. Lahiri, *The Leading Economic Indicators: New Approaches and Forecasting Records* (pp. 63-90). Cambridge University Press.
- Tkacz, G., (2013). Predicting Recessions in Real-time: Mining Google Trends and Electronic Payments Data for Clues. Working Paper

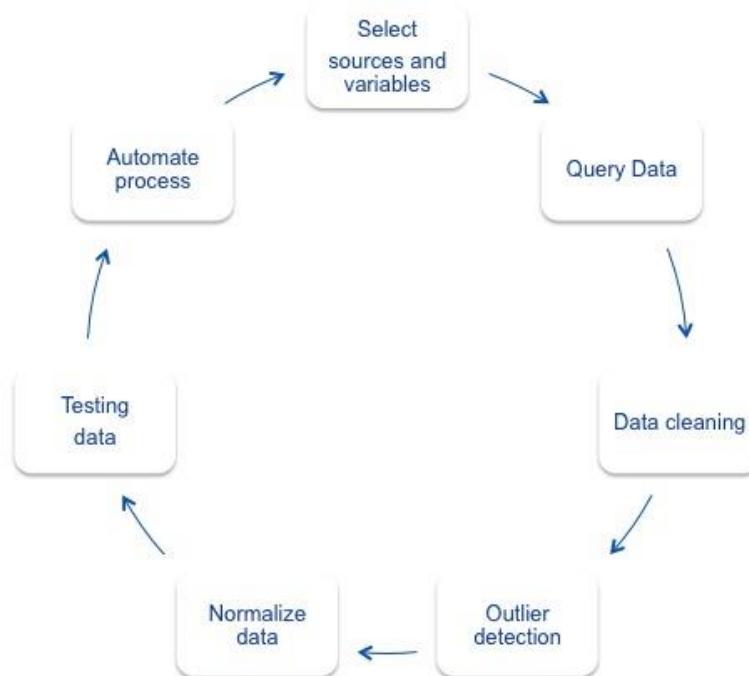
Figures

Figure 1.1 Spain: Retail Sales vs. Household Consumption Expenditure (%), YoY



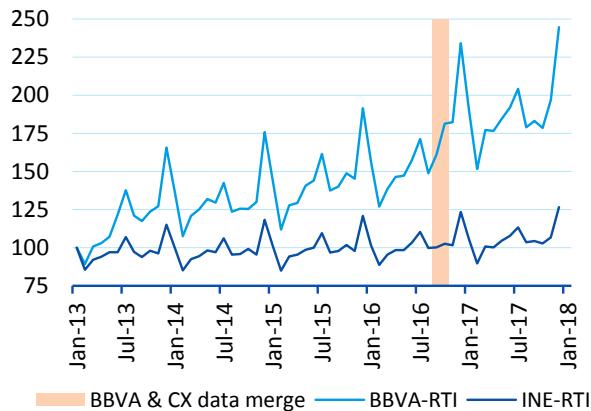
Source: BBVA based on INE

Figure 2.1 Data Engine Building Process. Data extraction, cleansing, and transformation



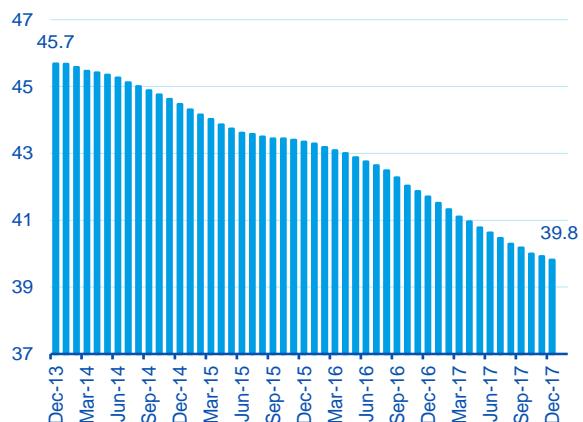
Source: BBVA

Figure 3.1 Aggregate Retail Trade Indices
(Jan-13 = 100)



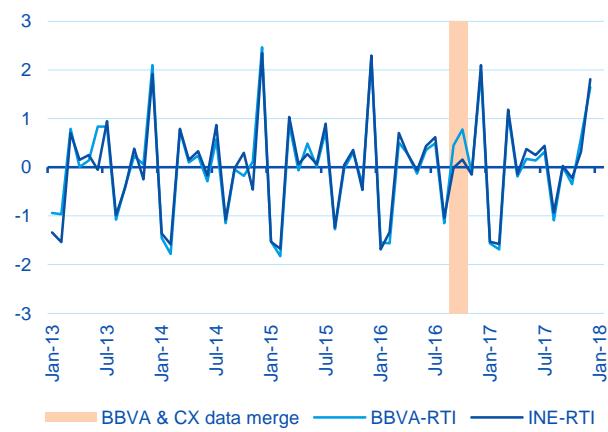
Source: BBVA based on INE

Figure 3.2 BBVA-RTI, average transaction amounts
(euros, 12 months moving average)



Source: BBVA

Figure 3.3 Aggregate Retail Trade Indices (standardized monthly growth rate)



Source: BBVA based on INE

Figure 3.4 RTI by Distribution Classe

Figure 3.4a Department Store RTI
(standardized monthly growth rate)



Figure 3.4c Small Chain Store RTI
(standardized monthly growth rate)

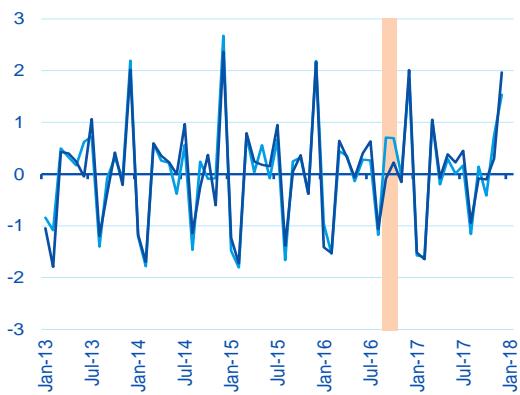
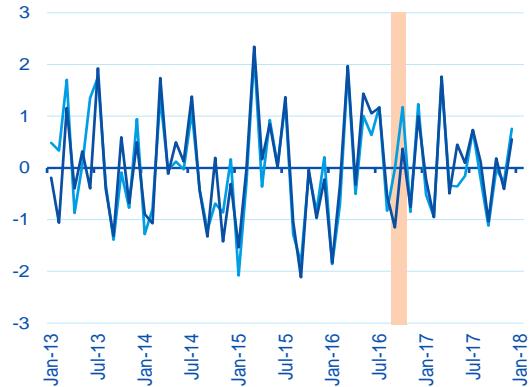


Figure 3.4e Gas Station RTI
(standardized monthly growth rate)



■ BBVA & CX data merge ■ BBVA-RTI ■ INE-RTI

Source: BBVA based on INE

Figure 3.4b Large Chain Store RTI
(standardized monthly growth rate)

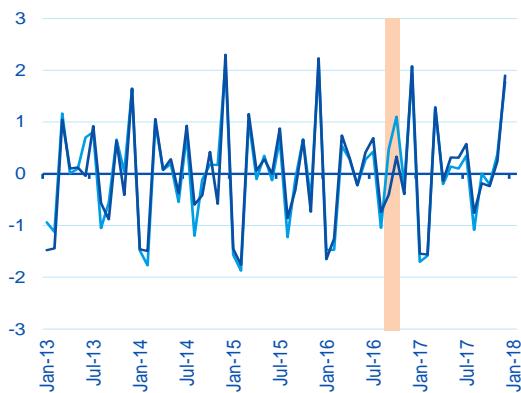


Figure 3.4d Single Retail Store RTI
(standardized monthly growth rate)



Figure 3.5 RTI by Autonomous Region (standardized monthly growth, %) rate) (1/2)

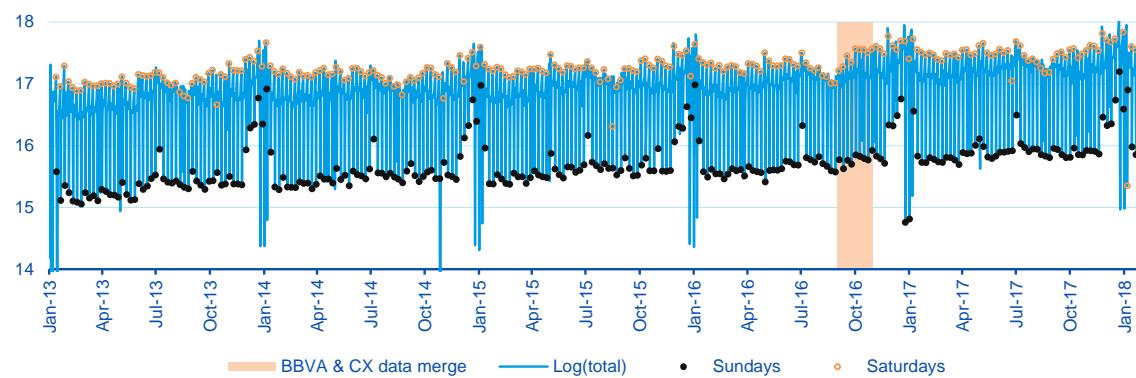


Figure 3.5 RTI by Autonomous Region (standardized monthly growth, %) rate) (2/2)



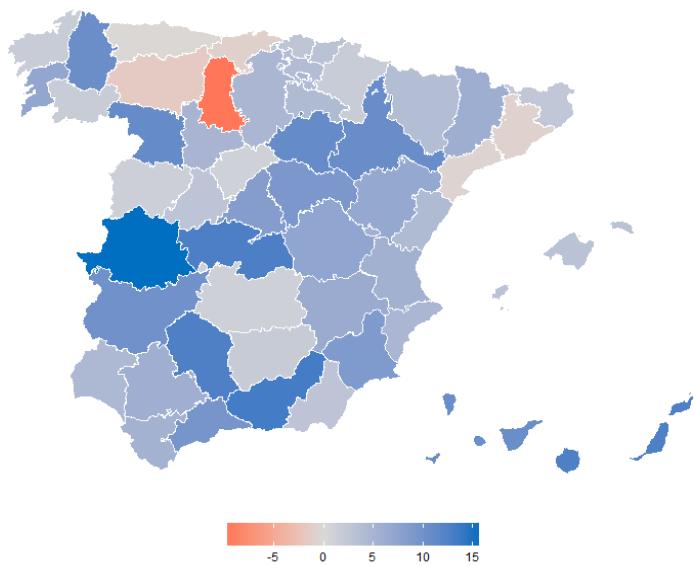
Source: BBVA based on INE

Figure 4.1 Aggregate Retail Trade - Daily Frequency (logarithms)



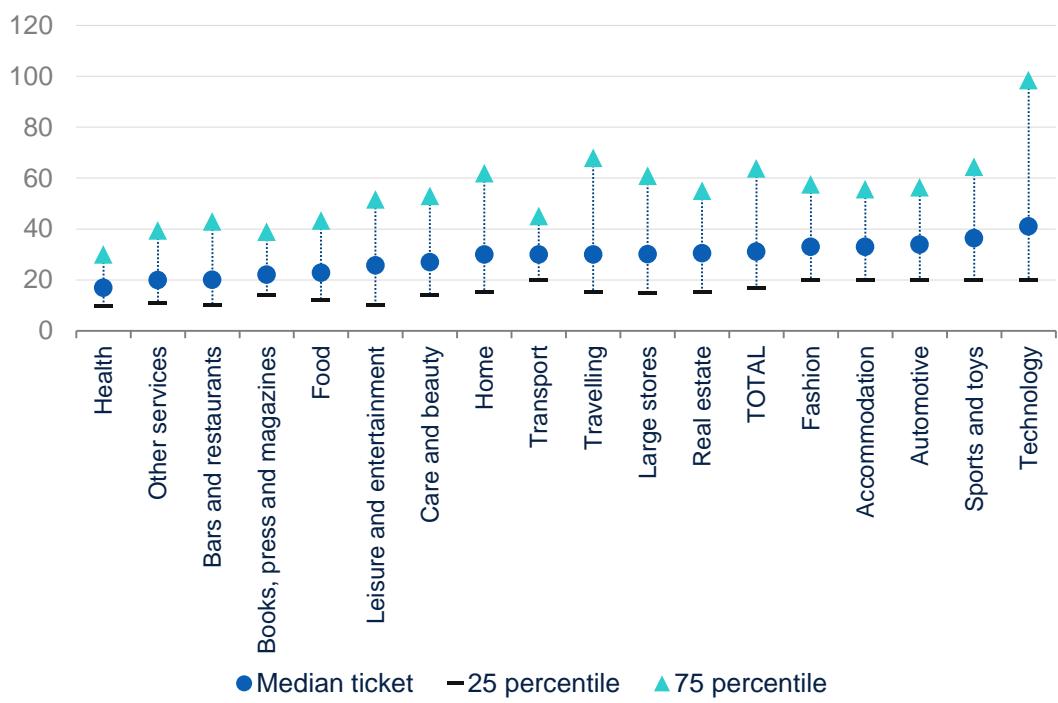
Source: BBVA

Figure 4.2 BBVA-RTI by Province in Dec-17 (% YoY)



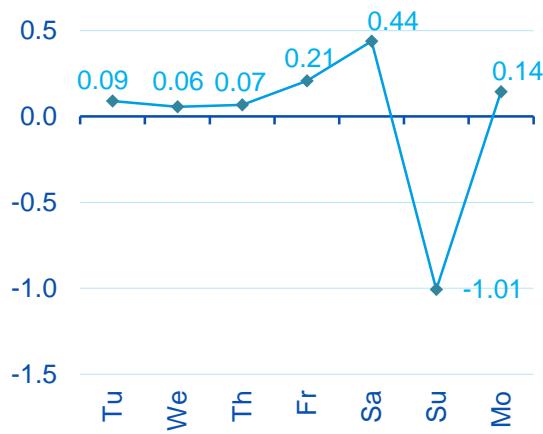
Source: BBVA

Figure 4.3 BBVA-RTI: Median Transaction by Sector of Activity in Dec-17 (Euros)



Source: BBVA

Figure 5.1 Intra-Weekly Effects (γ_t^d)



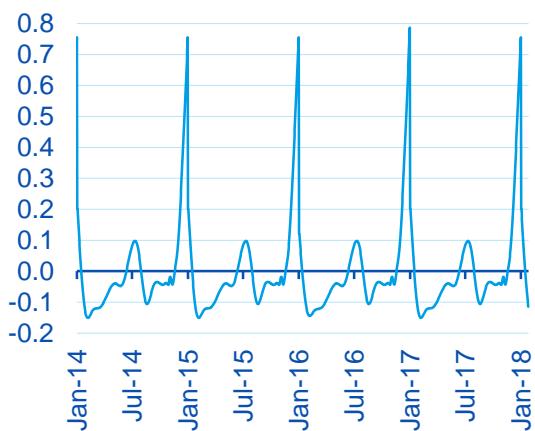
Note: Parameters as estimated for the second week of January 2018.
Source: BBVA

Figure 5.2 Intra-Monthly Effect (γ_t^m)



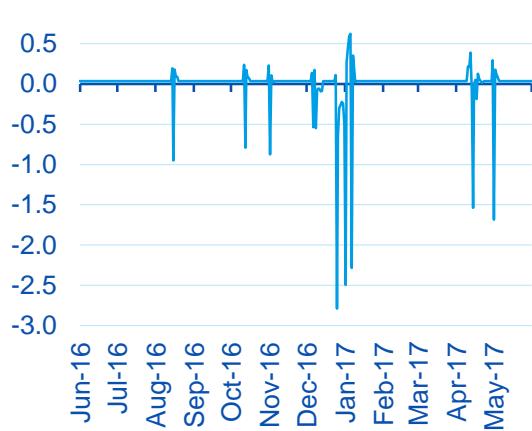
Note: Seasonal pattern for October 2017.
Source: BBVA

Figure 5.3 Intra-Yearly Effect (γ_t^y)



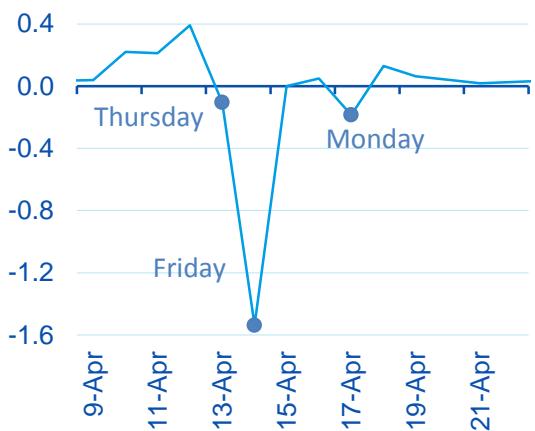
Source: BBVA

Figure 5.4 Holiday Effect (γ_t^h)



Source: BBVA

Figure 5.5 Easter 2017



Source: BBVA

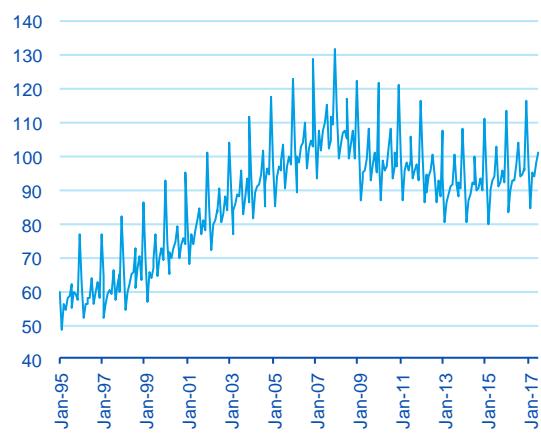
Figure 5.6 Trend (μ_t)



Source: BBVA

Annex

Figure A.1 INE-RTI (nominal and non-seasonal adjusted, base 2010=100)



Working Papers

2018

18/03 **Diego Bodas, Juan R. García López, Juan Murillo Arias, Matías Pacce, Tomasa Rodrigo López, Pep Ruiz de Aguirre, Camilo Ulloa, Juan de Dios Romero Palop and Heribert Valero Lapaz:** Measuring Retail Trade Using Card Transactional Data

18/02 **Máximo Camacho and Fernando Soto:** Consumer confidence's boom and bust in Latin America.

18/01 **Ana I. Segovia Domingo and Álvaro Martín Enríquez:** Digital Identity: the current state of affairs.

2017

17/24 **Joaquín Iglesias, Álvaro Ortiz and Tomasa Rodrigo:** How Do the Emerging Markets Central Bank Talk? A Big Data Approach to the Central Bank of Turkey.

17/23 **Ángel de la Fuente:** Series largas de algunos agregados económicos y demográficos regionales: Actualización de RegData hasta 2016.

17/22 **Ángel de la Fuente:** Series enlazadas de algunos agregados económicos regionales, 1955-2014. Parte II: Otras variables de empleo, rentas del trabajo y paro.

17/21 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2015.

17/20 **Maximo Camacho, Matias Pacce and Camilo Ulloa:** Business cycle phases in Spain.

17/19 **Ángel de la Fuente:** La liquidación de 2015 del sistema de financiación de las comunidades autónomas de régimen común.

17/18 **Víctor Adame y David Tuesta:** The labyrinth of the informal economy: measurement strategies and impacts.

17/17 **Víctor Adame y David Tuesta:** El laberinto de la economía informal: estrategias de medición e impactos.

17/16 **Liliana Rojas-Suárez y Lucía Pacheco:** Índice de prácticas regulatorias para la inclusión financiera en Latinoamérica: Facilitadores, Promotores y Obstaculizadores.

17/15 **Liliana Rojas-Suárez y Lucía Pacheco:** An Index of Regulatory Practices for Financial Inclusion in Latin America: Enablers, Promoters and Preventers.

17/14 **Ángel de la Fuente:** Las finanzas autonómicas en 2016 y entre 2003 y 2016.

17/13 **Carlos Casanova, Joaquín Iglesias, Álvaro Ortiz, Tomasa Rodrigo y Le Xia:** Tracking Chinese Vulnerability in Real Time Using Big Data.

17/12 **José E. Boscá, Rafael Doménech, Javier Ferri y José R. García:** Los Desplazamientos de la Curva de Beveridge en España y sus Efectos Macroeconómicos.

17/11 **Rafael Doménech y José Manuel González-Páramo:** Budgetary stability and structural reforms in Spain: lessons from the recession and options for the future.

17/10 **Ángel de la Fuente:** Series enlazadas de algunos agregados económicos regionales, 1955-2014. Parte I: Metodología, VAB, PIB y puestos de trabajo.

17/09 **José Félix Izquierdo:** Modelos para los flujos de nuevo crédito en España.

17/08 **José María Álvarez, Cristina Deblas, José Félix Izquierdo, Ana Rubio y Jaime Zurita:** The impact of European banking consolidation on credit prices.

17/07 **Víctor Adame García, Javier Alonso Meseguer, Luisa Pérez Ortiz, David Tuesta:** Infrastructure and economic growth from a meta-analysis approach: do all roads lead to Rome?

17/06 **Víctor Adame García, Javier Alonso Meseguer, Luisa Pérez Ortiz, David Tuesta:** Infraestructuras y crecimiento: un ejercicio de meta-análisis.

17/05 **Olga Cerqueira Gouveia, Enestor Dos Santos, Santiago Fernández de Lis, Alejandro Neut y Javier Sebastián:** Monedas digitales emitidas por los bancos centrales: adopción y repercusiones.

17/04 **Olga Cerqueira Gouveia, Enestor Dos Santos, Santiago Fernández de Lis, Alejandro Neut and Javier Sebastián:** Central Bank Digital Currencies: assessing implementation possibilities and impacts.

17/03 **Juan Antolín Díaz and Juan F. Rubio-Ramírez:** Narrative Sign Restrictions for SVARs.

17/02 **Luis Fernández Lafuerza and Gonzalo de Cadenas:** The Network View: applications to international trade and bank exposures.

17/01 **José Félix Izquierdo, Santiago Muñoz, Ana Rubio and Camilo Ulloa:** Impact of capital regulation on SMEs credit.

2016

16/21 **Javier Sebastián Cermeño:** Blockchain in financial services: Regulatory landscape and future challenges for its commercial application

16/20 **Máximo Camacho and Matías Pacce:** Forecasting travelers in Spain with Google queries.

16/19 **Javier Alonso, Alfonso Arellano, David Tuesta:** Factors that impact on pension fund investments in infrastructure under the current global financial regulation.

16/18 **Ángel de la Fuente:** La financiación regional en Alemania y en España: una perspectiva comparada.

16/17 **R. Doménech, J.R. García and C. Ulloa:** The Effects of Wage Flexibility on Activity and Employment in the Spanish Economy.

16/16 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2014.

16/15 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común: Adenda.

16/14 **Alicia García-Herrero, Eric Girardin and Hermann González:** Analyzing the impact of monetary policy on financial markets in Chile.

16/13 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común.

16/12 **Kan Chen, Mario Crucini:** Trends and Cycles in Small Open Economies: Making The Case For A General Equilibrium Approach.

16/11 **José Félix Izquierdo de la Cruz:** Determinantes de los tipos de interés de las carteras de crédito en la Eurozona.

16/10 **Alfonso Ugarte Ruiz:** Long run and short run components in explanatory variables and differences in Panel Data estimators.

16/09 **Carlos Casanova, Alicia García-Herrero:** Africa's rising commodity export dependency on China.

16/08 **Ángel de la Fuente:** Las finanzas autonómicas en 2015 y entre 2003 y 2015.

16/07 **Ángel de la Fuente:** Series largas de algunos agregados demográficos regionales, 1950-2015.

16/06 **Ángel de la Fuente:** Series enlazadas de Contabilidad Regional para España, 1980-2014.

16/05 **Rafael Doménech, Juan Ramón García, Camilo Ulloa:** Los efectos de la flexibilidad salarial sobre el crecimiento y el empleo.

16/04 **Angel de la Fuente, Michael Thöne, Christian Kastrop:** Regional Financing in Germany and Spain: Comparative Reform Perspectives.

16/03 **Antonio Cortina, Santiago Fernández de Lis:** El modelo de negocio de los bancos españoles en América Latina.

16/02 **Javier Andrés, Ángel de la Fuente, Rafael Doménech:** Notas para una política fiscal en la salida de la crisis.

16/01 **Ángel de la Fuente:** Series enlazadas de PIB y otros agregados de Contabilidad Nacional para España, 1955-2014.

[Click here to Access the Working Paper published](#)

[Spanish](#)

and [English](#)

The analysis, opinions, and conclusions included in this document are the property of the author of the report and are not necessarily property of the BBVA Group.

BBVA Research's publications can be viewed on the following website: <http://www.bbvareresearch.com>

Contact details:

BBVA Research
Azul Street, 4
La Vela Building - 4th and 5th floors
28050 Madrid (Spain)

Tel.: +34 91 374 60 00 and +34 91 537 70 00

Fax: +34 91 374 30 25

bbvaresearch@bbva.com

www.bbvaresearch.com



OPEN ACCESS

EDITED BY

Aparna Gupta,
Rensselaer Polytechnic Institute, United States

REVIEWED BY

Dragos Bozdog,
Stevens Institute of Technology, United States
Sotiris Kotsiantis,
University of Patras, Greece
Peggy Lindner,
University of Houston, United States

*CORRESPONDENCE

Vishal Mirza
✉ vishal.mirza@nyu.edu

RECEIVED 08 April 2025

ACCEPTED 18 July 2025

PUBLISHED 27 August 2025

CITATION

Jadhav A and Mirza V (2025) Large Language Models in equity markets: applications, techniques, and insights.
Front. Artif. Intell. 8:1608365.
doi: 10.3389/frai.2025.1608365

COPYRIGHT

© 2025 Jadhav and Mirza. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Large Language Models in equity markets: applications, techniques, and insights

Aakanksha Jadhav and Vishal Mirza*

Independent Researcher, New York, NY, United States

Recent breakthroughs in Large Language Models (LLMs) have the potential to disrupt equity investing by enabling sophisticated data analysis, market prediction, and automated trading. This paper presents a comprehensive review of 84 research studies conducted between 2022 and early 2025, synthesizing the state of LLM applications in stock investing. We provide a dual-layered categorization: first, by financial applications such as stock price forecasting, sentiment analysis, portfolio management, and algorithmic trading; second, by technical methodologies, including prompting, fine-tuning, multi-agent frameworks, reinforcement learning, and custom architectures. Additionally, we consolidate findings on the datasets used, ranging from financial statements to multimodal data (news, market trends, earnings transcripts, social media), and systematically compare general-purpose vs. finance-specialized LLMs used in research. Our analysis identifies key research trends, commonalities, and divergences across studies, evaluating both their empirical contributions and methodological innovations. We highlight the strengths of existing research, such as improved sentiment extraction and the use of reinforcement learning to factor market feedback, alongside critical gaps in scalability, interpretability, and real-world validation. Finally, we propose directions for future research, emphasizing hybrid modeling approaches, architectures that factor reasoning and large context windows, and robust evaluation frameworks to advance AI-driven financial strategies. By mapping the intersection of LLMs and equity markets, this review provides a foundation and roadmap for future research and practical implementation in the financial sector.

KEYWORDS

Large Language Models, LLMS, finance, NLP, stock, equity, investing, algorithmic trading

1 Introduction

Historically, investment strategies largely relied on structured data, fundamental and technical analysis, and human interpretation of financial reports, often resulting in slow decision-making and limited market adaptability. The advent of Large Language Models (LLMs) marks a transformative era in equity investing, shifting the paradigm from traditional, manual analysis to automated, real-time insights. LLMs now enable the rapid processing and integration of vast datasets, combining structured financial metrics with unstructured sources like news, earnings call transcripts, and social media sentiment. This integration uncovers market trends and signals with unprecedented precision, converting previously overlooked textual data into actionable trading signals. Moreover, LLMs are facilitating the evolution from static, rule-based models to dynamic, self-learning systems powered by reinforcement learning and multi-agent frameworks. This evolution enhances market responsiveness, improves risk management, and boosts alpha generation by identifying complex market narratives and emerging shifts. However, the integration of LLMs into equity investing is not without its

challenges, including data reliability, potential biases, regulatory considerations, and the interpretability of AI-driven recommendations.

This review paper analyzes current research on the application of Large Language Models (LLMs) in equity markets. It focuses on the following research questions: (1) What are the major trends in how LLMs are being applied within equity markets? (2) What are the primary technical innovations and methodological approaches employed in LLM-driven equity research? (3) What are the significant limitations, challenges, and research gaps that have been identified in the literature?

1.1 Key considerations for LLM usage in stock investing

1.1.1 Data complexity

Multimodal data: The financial stock investing landscape is characterized by an increasingly complex and voluminous collection of multi-modal data, presenting significant analytical challenges. Investors must process both structured and unstructured data sets, each requiring distinct computational approaches. Structured data, including financial statements, earnings reports, and quantitative market metrics demands rigorous statistical and analytical modeling. In contrast, unstructured data, such as financial news, social media sentiment, and analyst reports, necessitates advanced natural language processing (NLP) techniques to extract meaningful insights. Additionally, the integration of visual elements such as price charts, technical indicators, and graphical financial summaries, along with audio or text data from earnings calls and investor briefings, further complicates the analytical landscape.

Large datasets: The size of datasets used to analyze a single stock, or a portfolio is large and complex with inclusion of multimodal data—multiple unstructured and structured datasets and real-time data streams. Additionally, the size of financial earning reports (quarterly, annual, 10Q, 10 K) and industry report pdf files usually spans 100+ pages, while earning transcripts contains large audio/text files.

Why LLMs can address data complexity in large datasets: Large Language Models (LLMs) are uniquely suited to handle the data complexity inherent in financial stock investing due to their ability to process and synthesize large volumes of heterogeneous data. Unlike traditional models that require separate preprocessing pipelines for each data type, LLMs are trained on vast corpora of multimodal information, enabling them to natively handle unstructured data such as text from earnings reports, analyst notes, and news articles. Their contextual understanding allows them to extract relevant insights from lengthy documents (e.g., 100+ page 10-Ks), and their ability to summarize, infer sentiment, and answer questions from natural language inputs makes them ideal for navigating and distilling large, complex datasets. Moreover, LLMs can be extended or paired with vision and speech models (e.g., via multimodal architectures like GPT-4V or Gemini) to interpret visual data (charts, tables) and audio transcripts (earnings calls), thereby providing a unified framework for holistic financial analysis.

1.1.2 Time sensitivity and real time analysis

Financial markets are inherently time-sensitive, with investment decisions often hinging on the rapid processing and analysis of information. Latency and response times play a pivotal role, particularly in short-term trading strategies such as day trading.

The ability to react swiftly to market fluctuations can significantly impact investment outcomes, emphasizing the need for real-time or near-real-time analytical capabilities. This consideration is especially pertinent when evaluating the suitability of Large Language Models (LLMs) for equity investing, as their effectiveness depends on their ability to process vast amounts of financial data with minimal delay.

1.1.3 Diverse investment strategies and asset classes

Investment strategies in equity markets vary widely, from long-term value and growth investing to short-term momentum and high-frequency trading, each requiring distinct analytical methods. Additionally, this complexity extends across asset classes, including stocks, ETFs, and derivatives like options and futures, each with unique characteristics and risk factors. While LLMs hold promise for financial investing, effectively adapting them to diverse asset classes and investment styles remains a critical challenge. Additionally, financial analysis encompasses a wide range of approaches, including fundamental and technical analysis. Fundamental analysis focuses on the intrinsic value of assets, while technical analysis examines historical price and volume data.

1.2 Review scope

The application of LLMs to stock and equity investing has seen a significant surge in research, particularly in the 2 years following the late 2022 launch of ChatGPT, during which we identified approximately 84 relevant studies. Despite the growing interest in applying LLMs to finance and stock investment, the rapid proliferation of research in this domain has created a fragmented landscape. Some studies focus on multi-agent trading frameworks, others explore time series forecasting, while still others develop domain-specific LLM architectures. This review aims to consolidate these disparate efforts by synthesizing findings from 84 recent studies. Our goal is to provide a comprehensive overview of the research on how LLMs are being applied to transform stock and equity investing, while also highlighting key challenges and gaps in the field.

We selected 84 research papers from a comprehensive set of 100+ research papers sourced from Google Scholar and arXiv. The selection process involved applying keyword filters such as “LLM for Stock Investing” and “Large Language Models for Equity Investing”. Papers that primarily focused on macroeconomic analysis or general financial risk modeling, broader finance topics without explicit application of LLMs to equity markets, were excluded.

To provide a holistic analysis, we adopt a two-fold classification approach:

- 1 Applications of LLMs in Finance-Equity Investing (Section 2): This section categorizes the practical goals (why) and real-world relevance (what) of LLM applications in equity investing.
- 2 LLM Technical Innovations and Approaches (Section 3): This section examines the method (how)—detailing the specific techniques and methodologies used in LLM applications.

This dual approach enables a comprehensive evaluation of both the “why, what” and “how” aspects of research on LLM usage in equity investing. By exploring the key observations and insights

gained from existing research, we seek to highlight both the potential and the limitations of LLMs in this dynamic and challenging field. As investors increasingly integrate LLMs and other reinforcement learning AI techniques into their decision-making processes, understanding the potential and limitations of LLM-driven strategies becomes crucial. This paper serves as both a roadmap and a call to action for researchers and practitioners, paving the way toward more transparent, efficient, and reliable applications of LLMs in stock and equity investing. For simplicity, brevity, and readability we refer to each research paper by ‘Paper No.’ as indexed in the References section and avoid the use of complete paper title and author names unlike the format of traditional literature reviews.

2 Applications of LLMs in finance-equity investing

An effective method for categorizing the research landscape of LLM applications in stock investing is by end use or financial application, as it reflects how both practitioners and researchers typically frame their objectives. By grouping the 84 papers into categories such as stock trend prediction, sentiment analysis, portfolio management, and others we highlight each study’s practical goals and

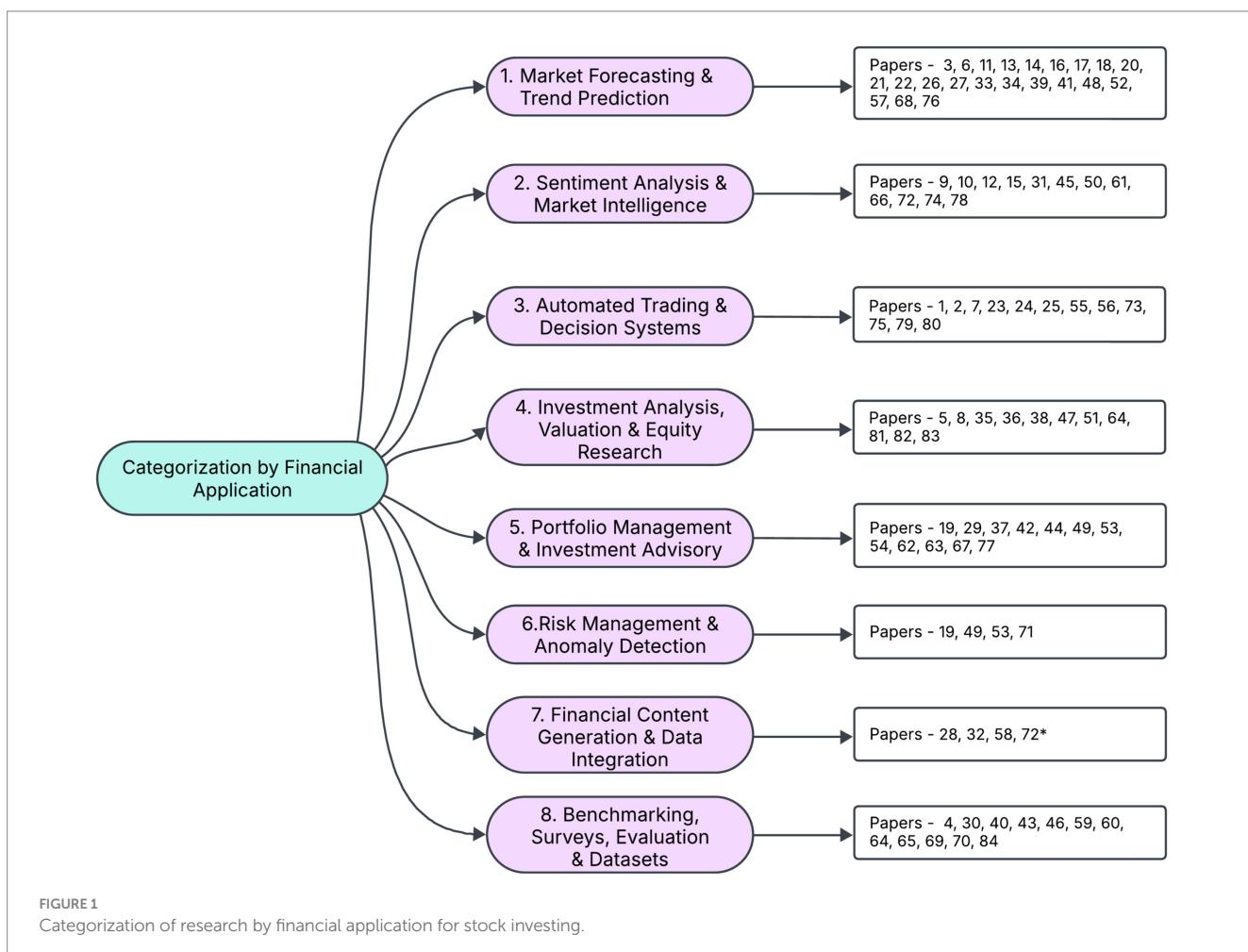
real-world relevance (Figure 1). This application-centric perspective not only spotlights potential synergies among similar works but also reveals gaps in coverage—for example, relatively few studies address risk management tasks. Consequently, categorizing by end use provides a clear, pragmatic lens through which to assess the collective impact and future directions of LLM research in equity and stock investing.

2.1 Stock price forecasting and market trends

A substantial body of research on LLMs in equity investing centers on forecasting stock prices, predicting market returns, and analyzing broader market trends. Typically, these studies combine textual data (e.g., news, press releases, earnings reports, analyst reports, social media content, etc.) with quantitative time-series data to enhance predictive accuracy beyond traditional methods.

2.1.1 Integration of qualitative data in forecasting

Leveraging language models to extract and interpret market signals from unstructured data has proven effective in complementing numerical datasets for forecasting models. This subcategory explores how LLMs integrate unstructured textual data with structured



financial indicators to improve stock prediction models and generate actionable market signals.

2.1.1.1 Stock prediction using multiple data sets

- [Shi and Hollifield \(2024\)](#) compares GPT-4 and BERT for stock return prediction using historical economic indicators, showing the strength of LLMs in processing structured economic data.
- [Ni et al. \(2024\)](#) forecasts post-earnings stock performance by combining textual earnings reports and transcripts with key financial metrics.
- [Lee et al. \(n.d.\)](#) transforms qualitative textual insights into quantitative market signals, offering a direct mapping approach for LLM-based forecasting.
- [Ding et al. \(2023\)](#) fuses LLMs with time-series modeling to improve stock return predictions in the Chinese market, illustrating LLM adaptability to local market contexts.
- [Swamy et al. \(n.d.\)](#) blends LLM-generated “priority indices” with traditional quantitative factors to enhance market trend prediction.
- [Tong et al. \(2024\)](#) introduces Ploutos, which integrates textual and numerical data to generate interpretable forecasts of price movements, focusing on explainability.
- [Cheng and Tang \(2023\)](#) demonstrates GPT-4’s ability to design high-performing investment factors with strong returns and Sharpe ratios.
- [Fatemi and Hu \(2024\)](#) proposes a multi-agent, multimodal system for stock prediction, emphasizing complex agent-based modeling over single-model forecasting.
- [Deng et al. \(2024\)](#) introduces a “denoising-then-voting” method to improve few-shot stock forecasting in noisy and data-scarce environments.
- [Di et al. \(2024\)](#) and [Guo and Shum \(2024\)](#) extend forecasting to broader market indices using LLMs combined with multi-source datasets and knowledge graphs, respectively.

2.1.1.2 Stock prediction using news data sets

- [Tandon \(2024\)](#) compares BERT and FinBERT for general sentiment extraction from news headlines, establishing baseline model differences.
- [Bhat \(n.d.\)](#) focuses on extracting emotional tone from news headlines to forecast stock movements, highlighting finer-grained sentiment modeling.
- [Liang et al. \(2024\)](#) introduces FinGPT, which incorporates contextual nuances from financial news to improve prediction accuracy.
- [Chen et al. \(2022\)](#) shows that complex linguistic features captured by LLMs outperform traditional models in global stock return forecasting.
- [Guo and Hauptmann \(2024\)](#) fine-tunes LLMs using Mistral embeddings on financial news, demonstrating performance gains in portfolio returns from domain-specific training.
- [Vidal \(2024\)](#) reports mixed results with certain LLMs, cautioning against overreliance without robust evaluation.
- [Lopez-Lira and Tang \(2023\)](#) evaluates ChatGPT’s raw capability to predict stock prices from headlines without financial fine-tuning, benchmarking general-purpose LLMs against domain-trained ones.

These studies collectively demonstrate the value of combining LLMs with structured and unstructured data for market prediction, showcasing methodological diversity and a trend toward interpretability and domain adaptation.

2.1.2 Time series specialization

This subcategory focuses on adapting LLMs to handle temporal dependencies and structured price signals for enhanced time-series forecasting in financial markets.

- [Valeyre and Aboura \(2024\)](#) evaluates LLMs on a dataset of major U.S. equities, demonstrating their alpha-generating potential in traditional time-series forecasting tasks.
- [Wang et al. \(2024\)](#) proposes a novel architecture that fuses textual inputs with time-series signals to improve predictive accuracy in stock price movements.
- [Chen et al. \(2024\)](#) investigates the role of historical return patterns in informing future price predictions, using LLMs to extract and model temporal patterns.
- [Voigt et al. \(2024\)](#) explores the methodological convergence between NLP and time-series analysis, applying LLMs to structured forecasting problems in quantitative finance.
- [Wang et al. \(2024\)](#) introduces Stock time—a bespoke architecture specifically designed for financial time-series prediction, optimizing LLM capabilities for sequential modeling tasks.

These studies highlight how LLMs are being repurposed or architecturally enhanced to address the unique demands of time-dependent financial data.

2.2 Sentiment analysis and market intelligence

Research in this domain focuses on extracting, quantifying, and interpreting sentiment from diverse textual sources—such as news articles, social media posts, analyst reports, and press releases—to generate actionable intelligence for market analysis. The key themes include:

2.2.1 Text mining and natural language processing

This category focuses on extracting insights from financial text data using LLMs, with a strong emphasis on sentiment analysis across various domains and data types.

- [Deng et al. \(2022\)](#) applies LLMs to Reddit data to extract investor sentiment, showcasing the utility of social media as an alternative sentiment source.
- [Das et al. \(2024\)](#) evaluates LLMs for single-stock trading, integrating news sentiment and price movement signals to inform trading decisions.
- [Wu \(2024\)](#) investigates the relationship between market sentiment from news sources and resulting stock price fluctuations.
- [Zhao and Welsch \(2024\)](#) introduces an adaptive LLM framework for sentiment analysis, integrating instruction tuning and real-time market feedback to improve adaptability.

- Aparicio et al. (2024) presents BioFinBERT, a domain-specific LLM fine-tuned for biotechnology stock sentiment, highlighting the value of industry-specific modeling.
- Liu et al. (2024) explores the correlation between news sentiment and Bitcoin prices, extending sentiment analysis to cryptocurrency markets.
- Xing (2024) proposes a multi-agent LLM system that enhances sentiment classification accuracy through collaborative agent dynamics.
- Dolphin et al. (2024) develops an LLM-driven system to process unstructured financial news and tickers, converting them into structured data formats for sentiment tracking.
- Yu (2023) conducts a case study on US equity market news, examining the variability and reliability of LLM-generated sentiment labels.
- Kirtac and Germano (2024) compares general-purpose and fine-tuned LLMs for financial sentiment analysis, analyzing performance differences across model types.

These studies reflect a broad spectrum of sentiment analysis techniques using LLMs—from basic extraction to adaptive and agent-based frameworks—demonstrating their growing sophistication and importance in financial modeling.

2.2.2 Sentiment scoring

This subcategory focuses on converting qualitative sentiment from financial text into quantitative scores that can directly inform trading and investment strategies.

- Bond et al. (2023) uses ChatGPT to generate a sentiment-based market indicator from daily news summaries, demonstrating superior performance compared to traditional sentiment analysis methods.
- Lefort et al. (2024) applies ChatGPT to financial news headlines to derive sentiment scores for NASDAQ index predictions, integrating these scores into an NLP-driven investment strategy.

Both studies illustrate how LLMs can translate textual sentiment into actionable numeric signals, with Paper 31 emphasizing broader market trends and Paper 61 focusing on NASDAQ-specific movements.

2.3 Automated trading and decision systems

This category highlights research dedicated to building systems capable of autonomously making trading decisions or crafting strategies based on LLM outputs. These approaches range from multi-agent frameworks to fully functional trading bots, operating in both simulated and real-world environments.

2.3.1 Algorithmic trading/automated trading decision systems (AI agents)

- Yu et al. (2024) introduces a multi-agent framework FinCon designed to handle complex financial tasks, including trading and portfolio management.

- Kou et al. (2024) presents a multi-agent methodology for quantitative stock investing, combining LLMs with established quantitative techniques to enhance performance and stability.
- Zhang et al. (2024) simulates investor behavior through an LLM-driven multi-agent system (Stock Agent) that adapts to live market conditions.
- Xiao et al. (2024) deploys specialized LLM-based agents (Trading Agents) within a structure modeled on real-world trading firms, demonstrating improved performance against standard benchmarks.
- Li et al. (2023) introduces TradingGPT, a multi-agent framework with layered memories and distinct agent characters, aiming to emulate human cognitive processes for improved trading efficiency and accuracy, emphasizing the hierarchical nature of human memory.
- Yu et al. (2023), FinMem, addresses the need for a novel LLM agent architecture to effectively transition from question-answering to purpose-driven financial trading, focusing on multi-source information processing, reasoning chains, and task prioritization.
- Wang et al. (2024) investigates LLM reasoning processes for trading decisions based on trend observations in crypto trading, revealing that less sophisticated LLMs can outperform more sophisticated LLMs, offering a contrast to the trend of increasing model complexity.
- Li et al. (2024) develops an LLM-based trading agent, CryptoTrade, that integrates diverse data (on chain and off-chain data) for cryptocurrency trading, showcasing LLM versatility beyond traditional stock markets.

2.3.2 Sentiment analysis for trading/portfolio management

- Konstantinidis et al. (2024) proposes a sentiment-analysis framework (FinLlama) for algorithmic trading that elevates portfolio returns, even in volatile markets.
- Chen et al. (2024) explores social media sentiment to inform trading strategies, linking shifts in investor sentiment to returns and herding behavior within AI-driven trading ecosystems. This study demonstrates the potential of factoring investor sentiment to inform trading decisions.

2.3.3 Adaptive trading/reinforcement learning

- Saqr and Rudzicz (2024) introduces Reinforcement Learning from Market Feedback (RLMF), enabling LLMs to adapt continuously to evolving market dynamics.

2.3.4 AI agent platforms

- Yang et al. (2024) provides an open-source AI agent platform, FinRobot, broadening access to specialized LLM-driven tools for both researchers and practitioners.

Integrating sentiment analysis (2.3.2) into algorithmic systems (2.3.1) combines quantitative and qualitative data. Reinforcement

learning (2.3.3) enhances agent adaptability. Open-source platforms (2.3.4) broaden access. These advances demonstrate LLMs' transformative potential in automated trading.

2.4 Investment analysis, valuation and equity research

Large Language Models (LLMs) offer considerable potential for streamlining and enhancing traditional equity research. By automating tasks such as stock ratings, identifying new investment opportunities, and assisting in the interpretation of complex financial documents, LLMs can significantly improve analysts' efficiency and insights. Research in this domain can be grouped into several key themes.

2.4.1 Equity research automation

- [Papasotiriou et al. \(2024\)](#) presents a method for automating and improving equity stock ratings by combining GPT-4 with multimodal financial data.
- [Li et al. \(2023\)](#) leverages generative AI (Llama2 and GPT-3.5) to automate fundamental investment research, with a focus on data summarization and ideation.
- [Zhou et al. \(2024\)](#) introduces FinRobot, an open-source AI agent designed for sell-side analysts seeking to automate equity research processes.
- [Yue and Au \(2023\)](#) describes GPTQuant, a conversational chatbot that simplifies investment research by generating and executing Python code.

While all papers focus on automation, they differ in their approach. ([Papasotiriou et al., 2024](#)) focuses on ratings, ([Li et al., 2023](#)) on idea generation, ([Zhou et al., 2024](#)) on open-source tools, and ([Yue and Au, 2023](#)) on user interaction, highlighting the diversity of automation strategies.

2.4.2 Investment research/analysis

- [Kim and Oh \(n.d.\)](#) presents a novel approach that combines LLMs, NLP, and dynamic data retrieval for in-depth stock market analysis.

2.4.3 Stock selection/portfolio management

- [Fatourou et al. \(2024\)](#) proposes MarketSenseAI, a GPT-4-based framework that supports stock selection.
- [Fatourou et al. \(2025\)](#) extends this approach with MarketSenseAI 2.0, an enhanced LLM-driven system that integrates various financial datasets and Retrieval-Augmented Generation to optimize portfolio performance.

Both papers focus on stock selection, but Paper 47 expands on Paper 8 by incorporating more diverse data and advanced generation techniques to improve portfolio optimization.

2.4.4 Executive/corporate communication analysis

- [Chiang et al. \(2025\)](#) explores how LLMs can evaluate Q&A segments in earnings calls to assess the transparency and responsiveness of corporate executives, providing valuable insights for investment decision-making.

2.4.5 Modeling

- [Wang et al. \(2023\)](#) introduces a new paradigm for alpha mining in quantitative investment, addressing the challenge of translating quant researchers' ideas into effective trading strategies.
- [Wang et al. \(2024\)](#) focuses on using Large Language Models (LLMs) to help understand and model investor decision-making, especially when investors are influenced by "herd behavior" (following the crowd).

[Wang et al. \(2023\)](#) focuses on modeling quant researcher ideas, while ([Wang et al., 2024](#)) model's investor behavior, showing two different modeling approaches. Building upon the modeling of financial data, researchers are also looking at how to improve the explainability of LLM models.

2.4.6 Explainability/interpretability

The focus on explainability is important for the real-world application of the models.

- [Koa et al. \(2024\)](#) develops a self-learning framework to enhance the interpretability of stock predictions by generating human-readable explanations, addressing a critical challenge for both traditional models and LLMs.
- Additionally, [Lopez-Lira and Tang \(2023\)](#), [Tong et al. \(2024\)](#), [Abdelsamie and Wang \(2024\)](#), [Zhao \(2024\)](#) cover model interpretability. [Tong et al. \(2024\)](#) and [Lopez-Lira and Tang \(2023\)](#), introduce novel frameworks and [Zhao \(2024\)](#) introduces a novel algorithm for enhancing interpretability for equity application of LLMs, while ([Abdelsamie and Wang, 2024](#)) compares interpretability across general purpose LLMs.

2.5 Portfolio management and investment advisory

Research in this domain applies Large Language Models (LLMs) to portfolio construction, wealth management, and personalized financial advice. The overarching aim is to optimize asset allocations and provide actionable recommendations for investors. Key themes include:

2.5.1 Portfolio construction/optimization

- [Lu et al. \(2023\)](#) demonstrates high-alpha portfolio generation by incorporating insights from news and policy announcements.
- [Romanko et al. \(2023\)](#) uses ChatGPT for stock selection in portfolio construction, integrating it with traditional optimization methods.

- Ko and Lee (2024) explores how ChatGPT can assist in asset class selection and enhance portfolio diversification.
- Gu et al. (2024) proposes an adaptive portfolio management framework that leverages LLMs and Reinforcement Learning for dynamic long–short positions.
- Huang et al. (2024) introduces a novel LLM-based, algorithm-driven system for stock selection and portfolio optimization.
- Zhao (2024) to revolutionize portfolio management by overcoming limitations in traditional approaches, this research develops a framework integrating advanced NLP, LLMs, and DRL (Deep Reinforcement learning) for enhanced return predictions, sentiment extraction, and optimized trading strategies.
- Perlin et al. (n.d.) evaluates Google's Gemini 1.5 Flash LLM's investment performance using extensive U.S. market data, finding it does not consistently outperform basic benchmarks, and its risk-adjusted returns decline with longer investment horizons. It covers a large-scale simulation of investment decisions using different data inputs and time horizons.

While (Lu et al., 2023; Romanko et al., 2023; Ko and Lee, 2024; Gu et al., 2024; Huang et al., 2024; Zhao, 2024) provide frameworks for portfolio construction, Perlin et al. (n.d.) differentiates itself by evaluating an existing and widely utilized LLM, showing the difference between theoretical implementations and real-world evaluation. Also, the complexity of frameworks increases for (Gu et al., 2024; Huang et al., 2024; Zhao, 2024) including a reinforcement learning approach. Some papers utilize hybrid approaches, while other papers rely solely on LLMs.

While the research conducted within the sphere of portfolio construction and optimization can, lead to, and improve the quality of Robo-advisory platforms, we cover that in as separate section below.

2.5.2 Robo-advisory/investor education

- Fieberg et al. (2024): Illustrates how LLMs can generate financial advice tailored to individual investor profiles.
- Lu (2025): Examines the impact of varying levels of financial literacy (alpha/beta) on investor behavior within robo-advisory platforms.

2.6 Risk management and anomaly detection

While much of the literature focuses on improving predictive accuracy, few studies address the robustness and reliability of LLMs in equity investing. Three key themes emerge:

2.6.1 Bias assessment

- Glasserman and Lin (2023) investigates two potential biases that can arise when LLMs use news sentiment for stock predictions: Look-Ahead Bias—when models inadvertently incorporate future returns into current forecasts. Distraction Effect—where extraneous company information skews the sentiment assessment.

2.6.2 Anomaly detection

This category of research examines methods for identifying unusual market conditions or portfolio crashes.

- Park (2024) introduces an LLM-driven multi-agent framework designed to automate anomaly detection in financial markets, reducing the burden of manual alert validation. Similarly,
- Koa et al. (2024) proposes a framework called “Temporal Relational Reasoning (TRR),” which utilizes LLMs to detect portfolio crashes by applying human-like temporal reasoning.
- Yang et al. (2025) proposes TwinMarket, a multi-agent framework that simulates complex human behavior. Simulated stock market experiments show how individual actions lead to emergent group behaviors, including financial bubbles and recessions.

Park (2024) and Yang et al. (2025) both utilize multi-agent frameworks, but (Park, 2024) focuses on automating alert validation, whereas (Yang et al., 2025) simulates human behavior to model emergent market phenomena. Koa et al. (2024) takes a different approach by focusing on temporal relational reasoning, which emulates human-like temporal analysis. Therefore, the papers vary in their approach, from multi-agent systems to temporal reasoning.

The early detection of anomalies is a critical step in risk mitigation, allowing for proactive measures to be taken before significant market disruptions occur.

2.6.3 Risk mitigation

Currently, there are no papers found that directly fit this category.

2.7 Financial content generation and data integration

This section covers research that utilizes LLMs for generating financial content or integrating large-scale datasets. The overarching aim is to synthesize information from diverse sources into coherent, actionable outputs that aid in decision-making and provide market insights. Key areas of focus include:

2.7.1 Report generation/automation

- Nishida and Utsuro (2025) demonstrates automated production of financial news articles covering stock price fluctuations.
- Pop et al. (2024) discusses how LLMs streamline equity research reporting by automating significant portions of the writing process.

While both papers focus on automating financial content, Nishida and Utsuro (2025) is geared toward generating news articles with a focus on timeliness, whereas (Pop et al., 2024) focuses on the automation of more in-depth equity research reports.

2.7.2 Financial LLM development and democratization

- Liu et al. (2023) introduces an open-source framework (FinGPT) designed to democratize financial LLMs. It covers data collection,

fine-tuning, and various adaptation strategies, supporting a wide range of downstream applications.

2.8 Benchmarking, surveys, evaluation and datasets

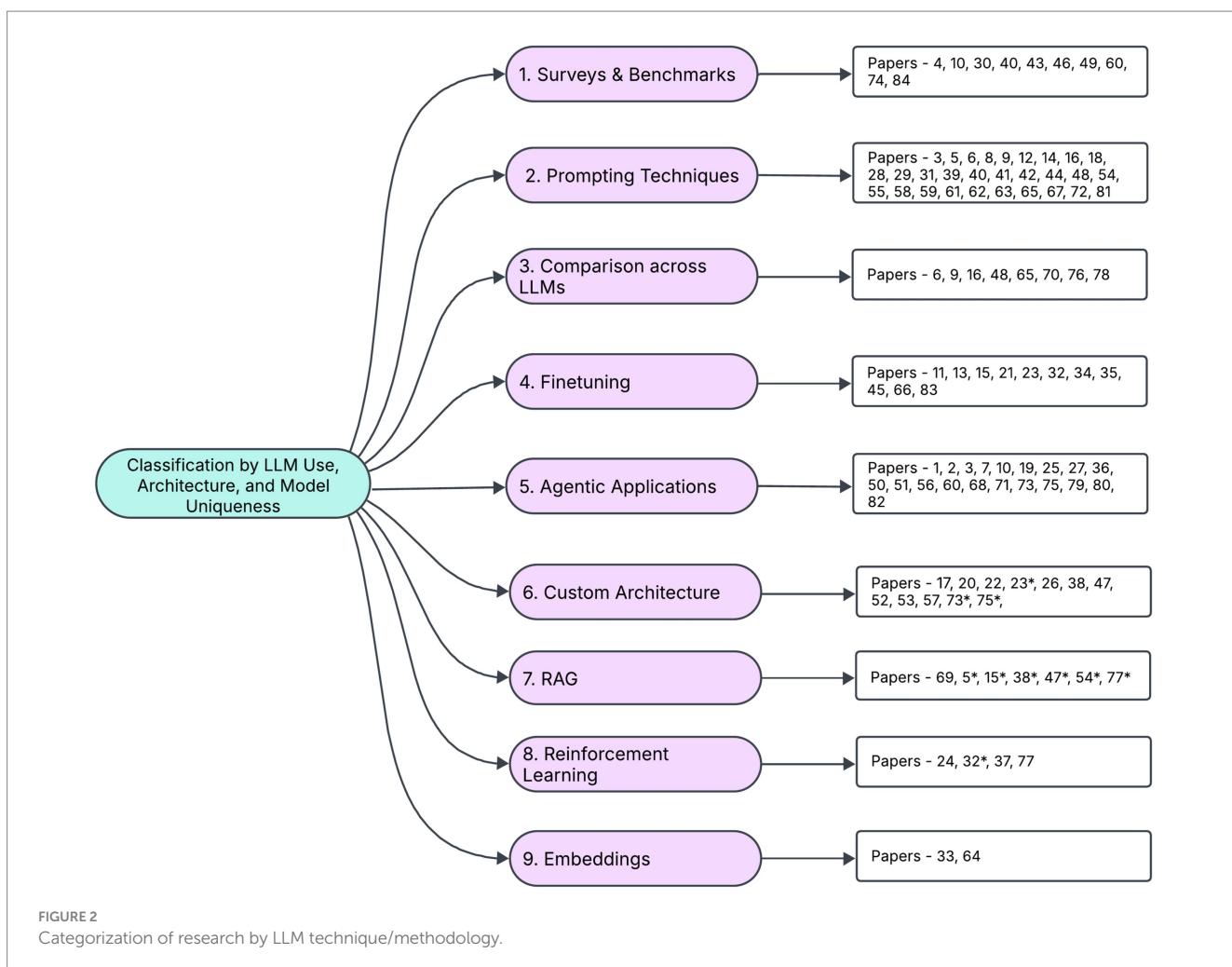
This final category highlights studies that assess and benchmark the performance of LLMs in financial contexts. These works often involve surveys, comparative analyses, or the introduction of new evaluation frameworks—serving as foundational references for future research.

- Surveys and reviews: Several papers (Zhao et al., 2024; Nie et al., 2024; Kong et al., 2024; Ding et al., 2024; Kong et al., 2024) offer broad overviews of LLM applications, ranging from automating financial reports to deploying trading agents.
- Benchmarks: Papers (Li et al., 2024; Mateega et al., 2025; Krause, 2023) focus on the development of new benchmarks (e.g., InvestorBench, FinanceQA), enabling more systematic comparisons of LLM performance. Li et al. (2024) focuses on creating a benchmark/standardized way to evaluate how well LLMs perform in financial analysis tasks.

- Comparisons: Abdelsamie and Wang (2024) benchmarks specialized financial LLMs, like Quantum, against human analysts and general-purpose LLMs to evaluate their market prediction accuracy and efficiency.
- Limitations: Papers (Bi et al., 2024; Chiang et al., 2025) delve into the practical potential and limitations of LLMs for financial forecasting and strategic decision-making, underscoring the need for ongoing evaluation in this rapidly evolving field.

3 LLM technical innovations, approaches

This section presents the key technical innovations underpinning recent research in the application of Large Language Models (LLMs) to Financial Stock Investing. We offer a comprehensive overview of the methodologies employed across these studies, highlighting the diverse techniques and nuanced approaches that drive progress in this field. Figure 2 provides a summary of the categorization of recent research by different approaches for LLMs.



3.1 Surveys and benchmarks

This category includes research that systematically evaluates the performance of LLM-based financial applications through benchmarks, surveys, and evaluation frameworks, providing critical insights into both the capabilities and limitations of current approaches.

3.1.1 Broad overview

- Zhao et al. (2024) and Kong et al. (2024) provides a comprehensive overview of how LLMs are being applied in finance.
- Bi et al. (2024) surveys how AI—especially ChatGPT—can transform financial forecasting, addressing key challenges such as privacy and ethics.
- Kong et al. (2024) reviews the transformation of finance by LLMs, categorizing applications and presenting datasets, benchmarks, and methodologies for future research.

While all three papers provide broad overviews, Zhao et al. (2024) and Kong et al. (2024) offer a general view, Bi et al. (2024) focuses on the potential of a specific LLM, and Kong et al. (2024) provides a structured categorization of the field.

Note that the categorization is not completely mutually exclusive, some papers that fit in two categories are represented in the one category and by an asterisk in the next most relevant category.

3.1.2 Application-specific overviews

- Sentiment analysis: Liu et al. (2024) provides a deep dive into the application of LLMs for financial sentiment analysis, using detailed datasets and case studies.
- Trading agents: Ding et al. (2024) reviews the landscape of LLM-based trading agents.

The application-specific overviews provide a foundation for more detailed technical evaluation.

3.1.3 Technical overview

- Nie et al. (2024) offers an in-depth review of various approaches, detailing progress, prospects, and challenges while classifying applications into linguistic, sentiment, financial, and agent-based tasks.

3.1.4 Benchmarking and bias evaluation

- Li et al. (2024) introduces InvestorBench, a standardized benchmark for evaluating LLM-based agents in financial decision-making across equities, cryptocurrencies, and ETFs, assessing their reasoning and decision-making capabilities using 13 different LLM backbones.
- Glasserman and Lin (2023) evaluates both “Look Ahead Bias” and the “distraction effect” by comparing sentiment-driven trading strategies based on original versus anonymized news

headlines, using in-sample and out-of-sample tests to isolate and measure these biases.

- Yu (2023) presents a case study on US equity market news sentiment analysis revealing significant variability in LLM sentiment classification, highlighting inherent output volatility. This could act as a benchmark reference for future research.

Li et al. (2024) provides a broad benchmark for financial agents, while (Glasserman and Lin, 2023; Yu, 2023) focus on specific aspects of LLM performance, such as bias and output volatility. Glasserman and Lin (2023) and Yu (2023) both use real world data to evaluate LLMs, whereas (Li et al., 2024) builds a benchmark.

The findings from these benchmarking and bias evaluations inform the broader overviews and technical analyses.

3.2 Prompting techniques

Prompting techniques are pivotal in harnessing the capabilities of LLMs for financial applications, enabling precise control over model outputs and facilitating complex reasoning. This section explores various prompting methodologies, demonstrating their impact across diverse financial equity investing tasks, from forecasting to portfolio management.

3.2.1 Zero-shot and few-shot prompting for financial analysis and forecasting

Paper 5: Uses GPT-4-32k with a diverse set of financial datasets (covering fundamental, market, and news data) and a Retrieval Augmented Generation (RAG)-like strategy to generate multi-horizon equity stock ratings. It employs both zero-shot and few-shot prompting for effective data integration.

- Nishida and Utsuro (2025) generates explanatory financial news articles about stock price movements by implementing few-shot learning and contrasting its performance with zero-shot methods.
- Pop et al. (2024) enhances stock trend prediction by applying a “denoising-then-voting” technique that combines few-shot learning with in-context prompt engineering.
- Mateega et al. (2025) automates the extraction of key elements from equity research reports by combining zero-shot/few-shot prompting with information retrieval.
- Chen et al. (2024) investigates ChatGPT’s forecasting tendencies by analyzing its extrapolation behavior, calibration, and inherent biases in predicting historical stock returns using a prompt-based approach.
- Lee et al. (n.d.) converts qualitative textual insights into quantitative market prediction scores by leveraging crafted prompts with dynamic few-shot examples.
- Swamy et al. (n.d.) extracts quantitative signals from qualitative data—such as moving averages and options volume—using innovative prompting strategies and integrates these signals with traditional quantitative features.
- Lefort et al. (2024) classifies financial news sentiment and de-noises aggregated outputs through zero-shot and few-shot prompting, supporting trading decisions on the NASDAQ.

- Chen et al. (2024) applies zero-shot classification to analyze over 77 million investor social-media posts, categorizing them into technical versus fundamental analysis and bullish versus bearish sentiment.
- Romanko et al. (2023) implements a two-step process where prompting (via zero-shot/few-shot methods) gathers stock suggestions, which are then refined using traditional portfolio optimization techniques (e.g., Markowitz mean-variance optimization).

These papers showcase the versatility of zero-shot and few-shot prompting, varying in their complexity and application. Papasotiriou et al. (2024) and Romanko et al. (2023) utilize RAG and optimization techniques, respectively, to enhance the prompting process, while other papers focus on direct application to sentiment analysis, forecasting, and data extraction.

The effectiveness of these prompting techniques extends to portfolio management, where tailored prompts can guide LLMs in generating and evaluating investment strategies.

3.2.2 Prompting techniques in portfolio management

- Li et al. (2024) leverages zero-shot prompting to generate portfolios and assess asset diversification, comparing the performance of these portfolios against randomly selected ones.
- Ko and Lee (2024) uses ChatGPT in a zero-shot setting to produce asset class selections, which are quantitatively evaluated for diversity and performance.
- Fieberg et al. (2024) employs structured prompts across 32 different LLMs, each reflecting diverse investor profiles, to generate financial advice that is then evaluated for suitability, performance, and potential bias.
- Lu (2025) integrates a ChatGPT-based system—utilizing both zero-shot and few-shot prompting—into a robo-advisor framework to enhance investor understanding of alpha and beta, leading to improved portfolio choices.
- Bond et al. (2023) utilizes ChatGPT's zero-shot NLP capabilities to analyze daily U.S. news summaries and generate a market sentiment indicator for short-term stock return forecasting.
- Wang et al. (2023) also covers the development of Alpha-GPT, an interactive system that uses prompt engineering and LLMs to generate creative trading signals, validated through alpha mining experiments.

These papers demonstrate the application of prompting for portfolio management, with variations in the use of zero-shot versus few-shot prompting, as well as the use of LLMs to generate investment advice, and trading signals. Fieberg et al. (2024) and Lu (2025) incorporate the analysis of investor profiles, and educational aspects, whereas the rest of the papers focus on pure portfolio creation and analysis.

3.2.3 Chain of thought prompting

- Fatouros et al. (2024) uses GPT-4 as both a predictor and signal evaluator for stock selection by analyzing diverse financial data.

It employs chain-of-thought and in-context learning techniques to enhance signal accuracy.

- Deng et al. (2022) adopts a semi-supervised approach by prompting an LLM with chain-of-thought techniques to generate weak sentiment labels from Reddit posts, which are then distilled into a smaller model using regression loss—achieving performance comparable to supervised methods.

These papers show how chain of thought prompting allows LLMs to create better outputs, by improving the reasoning process. Fatouros et al. (2024) uses it to improve signal accuracy, whereas (Deng et al., 2022) uses it to create better sentiment labels.

3.2.4 Instruction prompting and context prompting

- Perlin et al. (n.d.) evaluates Google's Gemini 1.5 Flash LLM for investment decision-making using anonymized U.S. market data. It queries the LLM with prompts that specify investment horizons and relevant financial inputs.
- Huang et al. (2024) implements scenario-based iterative prompt engineering to generate stock suggestions. These suggestions are subsequently screened using additional financial algorithms (NBESOA) for optimized portfolio construction.
- Lu et al. (2023) uses a prompt-based approach with data feeds from the Wall Street Journal and China State policy datasets to assess the efficacy of ChatGPT in providing financial stock investing recommendations.
- Dolphin et al. (2024) by combining LLM generative capabilities with advanced prompting and a validation framework using string similarity, the system extracts granular, per-company insights from news articles, demonstrating high accuracy and providing a live API and dataset for further research.

These papers showcase the use of instruction and context prompting across various applications, from evaluating LLM performance to generating stock suggestions and extracting market insights. Perlin et al. (n.d.) and Lu et al. (2023) focus on evaluation, whereas (Huang et al., 2024; Dolphin et al., 2024) focus on the creation of outputs.

3.2.5 Knowledge generation prompting

- Cheng and Tang (2023) leverages GPT-4 to autonomously generate equity investment factors through knowledge inference without direct data input. It employs prompting strategies that guide GPT-4 in reasoning and generating these factors.

3.3 Comparison across LLM models

Comparative analysis of LLMs is crucial for understanding their relative strengths and weaknesses in financial equity applications. This section explores studies that benchmark different models across various tasks, highlighting the nuances in their performance and capabilities.

- Krause (2023) compares the performance of ChatGPT, Bard, and Bing on various financial tasks—including report/text generation, decision support, summarization, and general NLP—to assess their suitability for financial analysis.
- Wu (2024) evaluates and contrasts multiple LLMs (e.g., ChatGPT, Tongyi Qianwen, and Baichuan Intelligence) in generating news-based stock scores and predicting stock market returns.
- Shi and Hollifield (2024) assesses the predictive capabilities of GPT against traditional transformer models like BERT using economic data from the Federal Reserve Beige Book, to determine their effectiveness in financial market prediction.
- Tandon (2024) systematically reviews and compares various LLM techniques (including BERT and FinBERT) for stock market prediction by analyzing financial news headlines using historical datasets from Kaggle and Yahoo Finance.
- Voigt et al. (2024) adapts NLP-centric, Transformer-based architectures for stock price forecasting by leveraging the structural parallels between text sequences and time-series data, with the expectation that models excelling in language processing can capture temporal dependencies in financial markets.
- Abdelsamie and Wang (2024) Compares the performance of Quantum, an advanced LLM specialized for financial forecasting, with other general purpose LLMs—GPT-3, GPT-4, FinGPT, and FinBERT and human analysts. Employing a dataset of historical financial data, news headlines, and social media sentiment, the research systematically assesses predictive accuracy, response efficiency, and interpretability across models.
- Lopez-Lira and Tang (2023) compares the performance of various GPT models, BERT and Finbert to confirm accuracy and reliability in predicting stock prices based on new headlines while also suggesting an interpretability framework
- Kirtac and Germano (2024) covers a comparative analysis of LLMs, including GPT-3-based OPT model, BERT, and FinBERT, alongside traditional methods for financial news sentiment analysis, demonstrating that OPT significantly outperforms others in predicting stock market returns and generating substantial trading profits.

These papers vary in their scope and focus, from comparing general-purpose LLMs across diverse tasks (Krause, 2023; Wu, 2024) to evaluating specialized financial LLMs against human analysts and traditional models (Abdelsamie and Wang, 2024; Kirtac and Germano, 2024). Shi and Hollifield (2024), Tandon (2024) and Voigt et al. (2024) focus on comparing LLMs against traditional transformer models. Lopez-Lira and Tang (2023) adds the dimension of interpretability. The evaluations vary in the data used, and the metrics used.

The findings from these comparative studies inform the development of more effective prompting techniques and the selection of appropriate models for specific financial investing tasks.

3.4 Finetuning

Finetuning plays a critical role in adapting Large Language Models (LLMs) to the specific demands of financial stock investing tasks, enabling them to capture nuanced patterns and generate accurate predictions. This section reviews various finetuning methods used to

adapt Large Language Models to financial tasks, ranging from instruction tuning to parameter-efficient finetuning and knowledge distillation.

3.4.1 Finetuning/domain adaptation

- Guo and Hauptmann (2024) fine-tunes LLMs for stock return forecasting, integrating and comparing token-level embeddings from different LLM architectures.
- Valeyre and Aboura (2024) evaluates the Chronos model for time series prediction in financial markets by testing both its pre-trained and fine-tuned configurations using supervised forecasting.
- Aparicio et al. (2024) adapts BioBERT to the financial domain by fine-tuning it on curated financial textual databases, enabling sentiment analysis of press releases and financial texts around key biotech stock inflection points.
- Wang et al. (2024) this research developed InvestAlign, a method that constructs supervised fine-tuning training datasets for LLMs using theoretical solutions from simplified investment problems, rather than costly and privacy-sensitive real-user data, to better align LLM investment decisions with human investor behavior.

These papers demonstrate various approaches to domain adaptation, ranging from fine-tuning for specific prediction tasks (Guo and Hauptmann, 2024; Valeyre and Aboura, 2024) to adapting models from other domains (Aparicio et al., 2024) and creating synthetic datasets for finetuning (Wang et al., 2024). The data used, and the model architectures adapted vary significantly.

Instruction finetuning builds upon domain adaptation by incorporating specific instructions to guide the model's learning.

3.4.2 Instruction finetuning and instruction prompting

- Liang et al. (2024) constructs an instruction tuning dataset using a multi-step process—clustering company-related news to capture dissemination influence, enriching prompts with context and explicit instructions, and then fine-tuning the LLM—to enhance sentiment-based stock movement prediction.

3.4.3 Parameter efficient finetuning (PEFT)

PEFT aims to reduce the computational and memory costs associated with finetuning large models

- Konstantinidis et al. (2024) implements LoRA fine-tuning on Llama 2 7B within a generator-classifier architecture for efficient financial sentiment analysis. [Custom Architecture]
- Liu et al. (2023) introduces FinGPT, an open-sourced framework that automates real-time financial data collection and adapts general-purpose LLMs for applications like robo-advising and sentiment analysis using LoRA/QLoRA fine-tuning, coupled with reinforcement learning (RLSP). [Reinforcement Learning]

Konstantinidis et al. (2024) and Liu et al. (2023) both utilize LoRA finetuning, but they differ in their architectures and applications. Konstantinidis et al. (2024) uses LoRA within a custom architecture,

whereas (Liu et al., 2023) uses LoRA within its open-source framework, and also in conjunction with reinforcement learning.

3.4.4 Combined instruction and parameter efficient finetuning

Combining instruction finetuning with parameter-efficient techniques allows for both task-specific adaptation and computational efficiency.

- Ni et al. (2024) employs a QLoRA-enhanced instruction finetuning strategy that combines base and external factors for improved stock prediction following earnings reports.
- Li et al. (2023) leverages both unsupervised and supervised LoRA on Llama2 alongside instruction fine-tuning on GPT3.5 to automate summarization and idea generation from diverse financial texts, thereby enhancing fundamental investment research.

3.4.5 Instruction tuning with reinforcement learning

Reinforcement learning can be integrated with instruction finetuning to further refine model performance based on real-world feedback

- Zhao and Welsch (2024) includes finetuning of LLaMA 2 model with instruction tuning to incorporate human instructions. Additionally, reinforcement learning (RL) is used to incorporate stock market feedback by dynamically adjusting knowledge source weights within a RAG module, improving financial sentiment analysis and stock price movement prediction.

Finetuning can also be used to enhance embedding generation for predictive modeling.

3.4.6 Fine tuning for embedding generation and predictive modeling

- Chen et al. (2022) utilizes LLMs (e.g., ChatGPT and LLaMA) to extract contextualized representations from news text for predicting expected stock returns. The study demonstrates that LLM-based predictions, which capture broader article context and complex narratives, significantly outperform traditional technical signals across multiple global equity markets.

3.4.7 Finetuning and knowledge transfer

- Das et al. (2024) combines financial textual data (news and sentiment) with price signals by leveraging large pretrained models (e.g., LLaMA-2-13B, Mistral, Gemma). These models are further fine-tuned or prompted in a zero-/few-shot manner to generate automated trading actions, facilitating effective knowledge transfer.

3.5 Agentic frameworks

Agent-based systems have the potential to transform financial stock and equity investing by automating complex decision-making

processes. This section reviews various agentic applications, from single-agent solutions to sophisticated multi-agent frameworks, highlighting their diverse approaches and impacts. Research on agent-based frameworks in financial stock and equity investing is still emerging, with only 20 of the 84 reviewed papers focusing on agent systems. (Note: Ding et al. (2024) is a survey and is covered in Section 3.1.1.).

3.5.1 Single-agent applications

- Yue and Au (2023) GPTQuant uses prompt templates and LangChain to create a conversational AI agent that generates Python code for investment research, streamlining the analysis process.
- Koa et al. (2024) covers the development of a Summarize-Explain-Predict (SEP) framework, which utilizes a self-reflective agent and Proximal Policy Optimization (PPO) to train an LLM to autonomously generate explainable stock predictions, achieving superior performance in both prediction accuracy and portfolio construction.

3.5.2 Multi-agent systems using chain-of-thought (CoT)

Multi-agent systems expand on these single-agent capabilities by coordinating multiple agents to perform complex tasks

- Zhou et al. (2024) utilizes a multi-agent CoT system to automate equity research and valuation by combining quantitative and qualitative analysis. It dynamically updates its data using three specialized agents (Data CoT, Concept CoT, and Thesis CoT) that emulate human analyst reasoning, resulting in high-quality, timely research.
- Wang et al. (2024) to understand LLM's reasoning approach for trading decision, study introduce FS-ReasoningAgent, a multi-agent framework that separates reasoning into factual and subjective components, demonstrating enhanced LLM trading performance and showing that subjective news drives returns in bull markets, while factual data performs better in bear markets.

Zhou et al. (2024) and Wang et al. (2024) both employ multi-agent CoT systems, but they focus on different aspects of financial investing. Zhou et al. (2024) automates equity research, while (Wang et al., 2024) analyzes the impact of factual versus subjective news on trading decisions.

3.5.3 Multi-agent systems with coordinated networks

Coordinated networks further enhance multi-agent systems by enabling complex interactions and adaptive decision-making.

- Kou et al. (2024) proposes a multi-step framework where LLMs extract alpha factors from multimodal financial data. These factors are integrated into a multi-agent system with dynamic weight-gating to produce an adaptive composite alpha formula for enhanced portfolio management.
- Zhang et al. (2024) introduces a multi-agent system, "StockAgent," which simulates investor trading behavior in a realistic market environment. Agents make trading decisions based on various external factors.

- Park (2024) presents a multi-agent framework for anomaly detection in financial markets. It employs specialized agents for data conversion, expert web analysis, institutional knowledge application, cross-checking, and report consolidation, thereby automating anomaly alert validation in the S&P 500, reducing manual verification.
- Xiao et al. (2024) deploys a multi-agent framework (TradingAgents) that simulates a trading firm by assigning roles such as fundamental, sentiment, and technical analysts, along with risk management agents. These agents collaboratively debate and synthesize diverse analyses and historical data to inform trading decisions, outperforming baseline models on key metrics
- Fatemi and Hu (2024) develops a multi-agent system (FinVision) that analyzes diverse multimodal financial data (including text and charts) using a “reflection module” that reviews past trading signals—especially visual cues—to improve stock market predictions.
- Xing (2024) proposes a heterogeneous LLM agent framework for financial sentiment analysis, where specialized agents collaboratively discuss identified error types to improve accuracy without additional fine-tuning.
- Yang et al. (2024) presents an open-source multi-agent platform (FinRobot) for financial tasks that employs a layered architecture. It leverages Financial Chain-of-Thought to decompose complex problems, dynamically selects LLM strategies, and integrates diverse models through LLMOps and DataOps.
- Yang et al. (2025) presents TwinMarket (a multi agent framework) that simulates individual behaviors and interactions to demonstrate how they lead to collective dynamics and emergent phenomena, such as financial bubbles and recessions, within a simulated stock market.
- Li et al. (2023) TradingGPT (single and multi-agent system) utilizes three memory layers with custom decay mechanisms, inter-agent debate, and individualized trading traits to enable agents to effectively integrate historical data and real-time market signals for enhanced trading decisions. Note the layered memory also fits in custom architecture

These papers showcase diverse approaches to multi-agent systems with coordinated networks, ranging from alpha factor extraction (Kou et al., 2024) to anomaly detection (Park, 2024) and trading simulation (Zhang et al., 2024; Xiao et al., 2024; Fatemi and Hu, 2024; Yang et al., 2025; Li et al., 2023). The complexity of the systems, and the tasks performed vary significantly. Xing (2024) and Yang et al. (2024) focus on frameworks, whereas the rest focus on applications.

3.5.4 Multi-agent systems with reinforcement learning

Reinforcement learning can further enhance multi-agent systems by enabling agents to learn from experience and adapt to changing market conditions.

- Yu et al. (2024) introduces FinCon, a hierarchical multi-agent system modeled after real-world investment firms. It employs a “conceptual verbal reinforcement” mechanism where agents self-critique and update their investment beliefs to guide future

actions, thereby improving performance and reducing unnecessary communication.

- Yu et al. (2023) presents FinMem a multi-agent architecture that incorporates three core modules: Profiling for agent customization, a layered Memory module that emulates human trader cognition for efficient hierarchical data assimilation, and Decision-making to translate insights into investment actions; this design allows the agent to self-evolve, adapt to market cues, and surpass human perceptual limits, resulting in improved trading performance.
- Li et al. (2024) CryptoTrade, incorporates a reflective mechanism to analyze prior trading outcomes and refine daily decisions, demonstrating superior performance compared to traditional strategies and time-series baselines across various cryptocurrencies and market conditions.

These papers demonstrate the use of reinforcement learning in multi-agent systems, with variations in the reinforcement mechanisms and applications. Yu et al. (2024) uses a conceptual verbal reinforcement mechanism, whereas (Yu et al., 2023) focuses on a multi-module architecture, and (Li et al., 2024) uses a reflective mechanism.

3.6 Custom architecture

Custom architectures are pivotal in pushing the boundaries of LLM applications in finance, enabling the development of specialized models that address unique challenges. This section explores a variety of innovative architectural approaches, from knowledge distillation to hybrid and multi-modal frameworks. Please note multi-agent architectures are covered explicitly in the section 3.5—Agentic Applications.

3.6.1 Knowledge distillations

Knowledge distillation includes the transfer knowledge from a large “teacher” model to a smaller “student” model.

- Bhat (n.d.) employs a computationally efficient distilled LLM to extract emotional tone and intensity from financial news headlines. The distilled model’s outputs are fed into classification algorithms for predicting stock price direction, demonstrating that emotion analysis alone can rival the performance of traditional financial data methods.

3.6.2 Transfer learning and foundation models

Transfer learning and foundation models extend this concept by leveraging pre-trained knowledge for broader applications. Guo and Shum (2024) demonstrates the power of transfer learning and foundation models in capturing universal market patterns.

- Guo and Shum (2024) utilizes a novel LLM structure designed for large-scale investment applications. The model, termed Large Investment Model (LIM), employs a “foundation model” approach by training on vast financial datasets to learn universal market patterns, which are then transferred via transfer learning to develop specialized and efficient investment strategies for various financial tasks.

3.6.3 Retrieval-augmented and agent-based architectures

Retrieval-augmented and agent-based architectures further enhance LLM capabilities by integrating external knowledge and dynamic interactions.

- [Fatouros et al. \(2025\)](#) introduces MarketSenseAI 2.0, which combines Retrieval-Augmented Generation (RAG) with LLM agents for comprehensive stock analysis and selection. This framework processes diverse data types—including financial news, historical prices, company fundamentals, and macroeconomic indicators—with RAG handling SEC filings, earnings calls, and institutional reports. Empirical results on S&P 100 stocks (2023–2024) indicate cumulative returns of 125.9% versus an index return of 73.5%.
- [Kim and Oh \(n.d.\)](#) presents an integrated system that fuses RAG with LangChain to dynamically retrieve and synthesize external financial data, producing real-time, contextually enriched stock analysis reports.

[Fatouros et al. \(2025\)](#) and [Kim and Oh \(n.d.\)](#) both use RAG, but [\(Fatouros et al., 2025\)](#) uses RAG in combination with LLM agents.

3.6.4 Time series and temporal reasoning architectures

Time series and temporal reasoning architectures complement other forms by focusing on the temporal dynamics of financial data.

- [Wang et al. \(2024\)](#) proposes StockTime, a specialized LLM architecture designed explicitly for stock price time series data. By treating stock prices as consecutive tokens, StockTime extracts textual information (e.g., correlations, trends, timestamps) and integrates this with time series data into an embedding space. This multimodal fusion yields more accurate predictions while reducing memory usage and runtime costs.
- [Koa et al. \(2024\)](#) introduces Temporal Relational Reasoning (TRR), a framework that combines LLM-based zero-shot text interpretation with cognitively inspired components (memory, attention, reasoning) to track and aggregate news impacts over time. This structured approach improves detection of impending portfolio crashes by modeling temporal relationships among events and stocks.

[Wang et al. \(2024\)](#) and [Koa et al. \(2024\)](#) both address temporal reasoning, but they differ in their approaches. [Wang et al. \(2024\)](#) focuses on treating time series data as tokens, while [\(Koa et al., 2024\)](#) integrates cognitive components.

3.6.5 Hybrid and multi-modal frameworks

Hybrid and multi-modal frameworks integrate diverse data types and models to capture complex market relationships.

- [Di et al. \(2024\)](#) proposes a custom architecture for securities index prediction that integrates LLM-driven knowledge extraction with a heterogeneous graph and a Graph Neural Network (GNN) to capture complex market relationships.
- [Tong et al. \(2024\)](#) presents a novel, integrated two-part framework (Ploutos) that combines specialized experts for

multi-modal data analysis with tailored prompting and dynamic token weighting to enhance interpretability in stock movement prediction.

- [Ding et al. \(2023\)](#) introduces a novel framework that combines a Local–Global model for integrating stock features with LLM-derived semantic information and employs self-correlated reinforcement learning to align these embeddings within a shared semantic space.

These papers showcase diverse hybrid and multi-modal frameworks, combining various techniques to capture complex market relationships. [Di et al. \(2024\)](#), [Tong et al. \(2024\)](#) and [Ding et al. \(2023\)](#) all use different methods to combine different data types and models.

3.7 Retrieval-augmented generation (RAG)

Retrieval-Augmented Generation (RAG) in stock investing enables LLMs to provide more accurate and contextually relevant financial insights by dynamically retrieving and incorporating up-to-date information from diverse sources like news articles, financial reports, and market data.

We observe that only a few papers out of the 84 research papers incorporate RAG in some form, with just two using it as the primary technique.

- [Li et al. \(2024\)](#) introduces AlphaFin, a data set combining traditional research datasets, real-time financial data, and handwritten chain-of-thought (CoT) data to address the limitation of limited financial training dataset availability. Further they introduce a RAG based, two phased framework (StockChain) for financial analysis.

Several papers incorporate RAG with diverse techniques. Papers ([Papasotiriou et al., 2024](#); [Romanko et al., 2023](#)) combine RAG with zero-shot and few-shot prompting; Papers ([Kim and Oh, n.d.](#); [Fatouros et al., 2025](#)) utilize RAG with agents and ([Zhao, 2024](#)) integrates RAG with reinforcement learning while ([Zhao and Welsch, 2024](#)) uses RAG in combination with reinforcement learning and instruction tuning.

3.8 Reinforcement learning

This section examines the integration of Reinforcement Learning (RL) with Large Language Models (LLMs) to advance financial trading. Recent studies demonstrate RL's efficacy in enabling LLM-driven adaptive strategies, particularly in achieving regime-adaptive execution, enhancing the explainability of trading decisions, and improving portfolio performance across volatile market conditions.

3.8.1 Regime adaptation via reinforcement learning

- [Saqr and Rudzicz \(2024\)](#) Leverages Reinforcement Learning from Market Feedback (RLMF), a regime-adaptive market execution method, to dynamically adjust LLM behavior in real-time. This model-agnostic approach, demonstrated with Llama-2

7B, utilizes intrinsic market rewards and a teacher-student dual-phase pipeline (iterative train and execute cycles) to improve predictive accuracy by 15% over models like GPT-4o, effectively circumventing the need for human-labeled data. This highlights a trend in utilizing RL to reduce reliance on extensive human-labeled datasets.

3.8.2 Adaptive and explainable trading systems

- [Gu et al. \(2024\)](#) presents a framework that fuses LLMs with Reinforcement Learning for margin trading. The LLM system analyzes diverse financial data to produce market forecasts with explainable reasoning, which are then integrated with RL to dynamically adjust trading positions, significantly boosting returns and Sharpe ratios while enhancing transparency in portfolio management.
- [Zhao \(2024\)](#) introduces the Hierarchical Reinforced Trader (HRT), an adaptive Retrieval-Augmented Generation (RAG) framework for LLMs, employing bi-level Deep Reinforcement Learning (DRL) and an enhanced Univariate Flagging Algorithm (UFA) for model interpretability. This framework demonstrates significant improvements in portfolio performance and risk management across diverse market conditions. Portfolio comparisons to the S&P 500 reveal favorable results in both bull and bear/volatile market scenarios.

3.9 Embedding based methods

Embedding-based methods leverage the semantic understanding capabilities of Large Language Models (LLMs) to transform textual financial data into dense vector representations. This section explores recent applications of LLM embeddings in financial analysis, focusing on their use in stock prediction and communication evaluation.

- [Chen et al. \(2022\)](#) utilizes state-of-the-art LLMs (e.g., ChatGPT, LLaMA) to extract contextualized embeddings from news articles. These embeddings capture nuanced language features—such as negation and complex narratives—and are used as input features to predict stock returns, outperforming traditional technical signals and simpler NLP methods across diverse global equity markets.
- [Chiang et al. \(2025\)](#) leverages LLM-based vector embeddings derived from 192,000 earnings call transcripts to quantify the semantic alignment and relevance of Q&A segments, providing a novel metric for evaluating investor communication effectiveness. This method allows for a deeper understanding of the relationship between the questions asked, and the answers given, in earnings calls.

4 Data sets and models

For a comprehensive view on data sets used across the 84 different research is presented [Figure 3](#).

Majority of the datasets used in the 84 research studies are public data sets available through SEC, Edgar, Yahoo Finance, News,

Company websites, Social media. Only a few limited studies use specialized datasets—Simulated Trading Environment Dataset ([Zhang et al., 2024](#); [Xiao et al., 2024](#)), Anonymized Portfolio Dataset ([Perlin et al., n.d.](#)), INVESTORBENCH Dataset ([Li et al., 2024](#)), FinLLM Challenge Dataset ([Das et al., 2024](#)), Boardroom Q&A Dataset ([Chiang et al., 2025](#)), FinanceQA Dataset (for question–answer evaluation) ([Mateega et al., 2025](#)).

A central theme across the 84 reviewed papers is the diversity of LLMs employed for financial tasks. Some studies leverage generic transformer models (e.g., GPT, LLaMA) with minimal modifications, while others adopt domain-specific architectures fine-tuned on large corpora of financial text. [Figure 4](#) provides the distribution of LLMs used in research so far. Out of the 84 research papers evaluated, 50 papers provide specific reference to LLM names with a majority around 49 papers mentioning the use of general purpose LLMs such as GPT, Llama, BERT, others. A few studies present domain specific LLMs. [Table 1](#) provides an overview of financial LLMs, including details on base models, parameters and key focus areas.

Overall, the proliferation of finance-specific LLMs underscores a broader industry trend toward domain adaptation, multi-agent frameworks, and hybrid modeling approaches that harness the unique strengths of LLMs and insights from financial data.

5 Discussion

5.1 Strengths of existing research

The reviewed literature highlights several strengths in using LLMs for stock investing:

- **Comprehensive Data Integration:** Many studies successfully integrate structured (e.g., financial statements, historical prices) and unstructured (e.g., news articles, earnings call transcripts, social media) data to enhance predictive accuracy.
- **Breadth of Coverage (Application & LLM Techniques):** Research on use of LLMs in stock investing so far covers a broad set of financial end-use case applications (sentiment analysis, equity research, stock prediction, portfolio management, algorithmic trading, others) and diverse set of LLM techniques ranging from simple prompt based methods, to fine tuning, the use of LLM agents for automated workflows and decisions and the proposals of novel custom architectures.
- **Usage of General Purpose and Domain Specific LLM Models:** Majority of the research studies conducted so far use general purpose LLMs such as GPT 3, 4, LlaMA, BERT and others. Only a few studies use or present fine-tuned versions of general purpose LLMs such as (FinGPT, FinLlama, BioFinBERT) for improved performance on financial investing tasks.
- **Advancements in Model Architectures:** Several studies introduce novel frameworks such as Ploutos (for integrating numerical and textual data) and using a combination of generator and classifier model, StockTime (for time-series adaptation), and MarketSenseAI (for multi-modal analysis using RAG). These innovations improve LLM adaptability to financial markets.
- **Validation of potential to disrupt Sentiment Analysis and Investment Research:** Several research studies demonstrated superior performance of LLMs when used in sentiment analysis

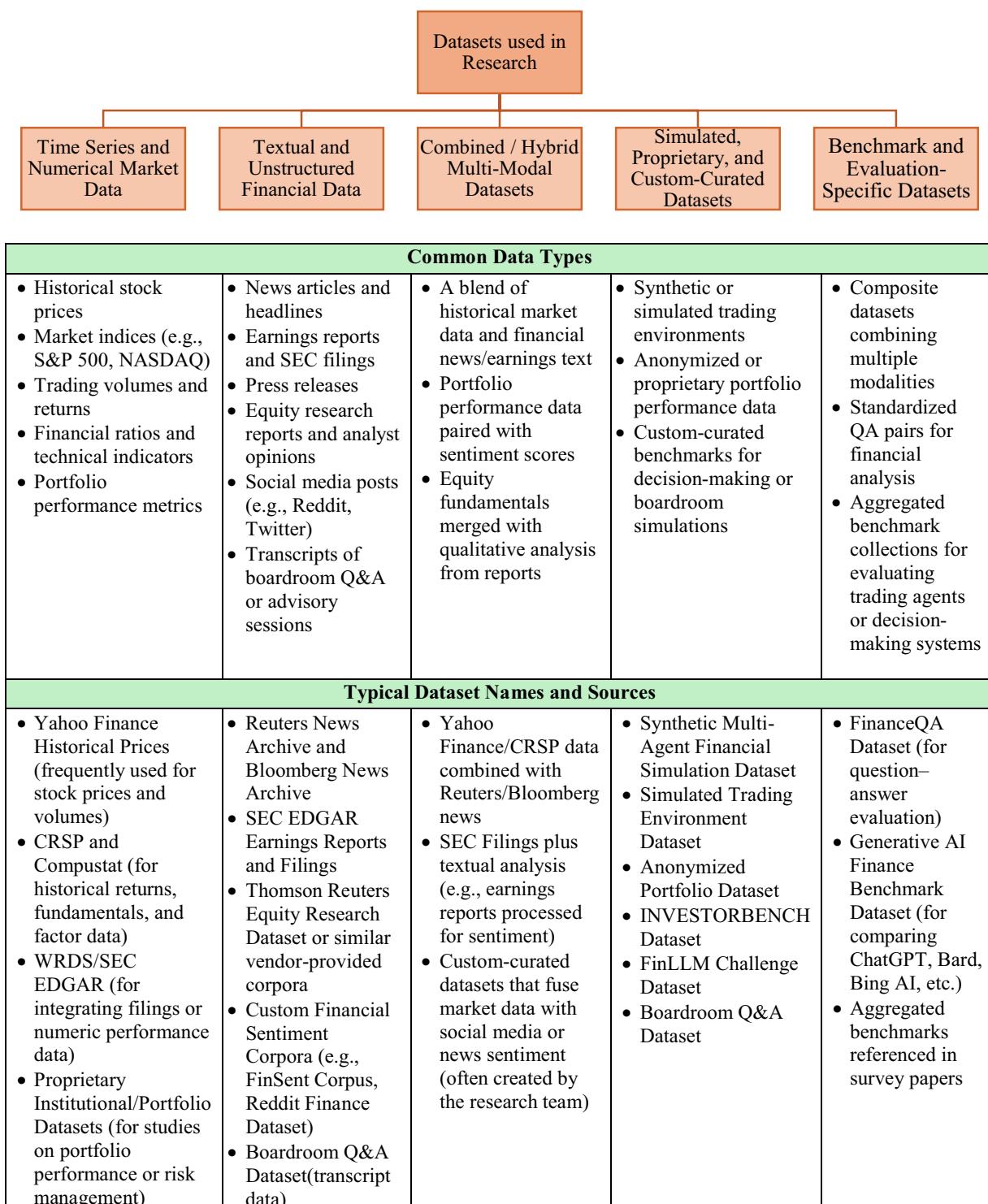


FIGURE 3
Datasets used across 84 different research papers.

to capture signals from large data sets (news, analyst reports and social media) compared to traditional methods. Similarly, research confirmed LLMs streamlining equity research (e.g., GPTQuant for report generation, FinRobot for sell-side research automation) and support algorithmic trading through real-time processing and decision-making.

- **Emerging Multi-Agent and Reinforcement Learning Systems:** Research in AI-driven trading strategies has progressed beyond rule-based models to reinforcement learning and multi-agent frameworks that dynamically adjust to market conditions, improving adaptability and efficacy of predictions, risk management.

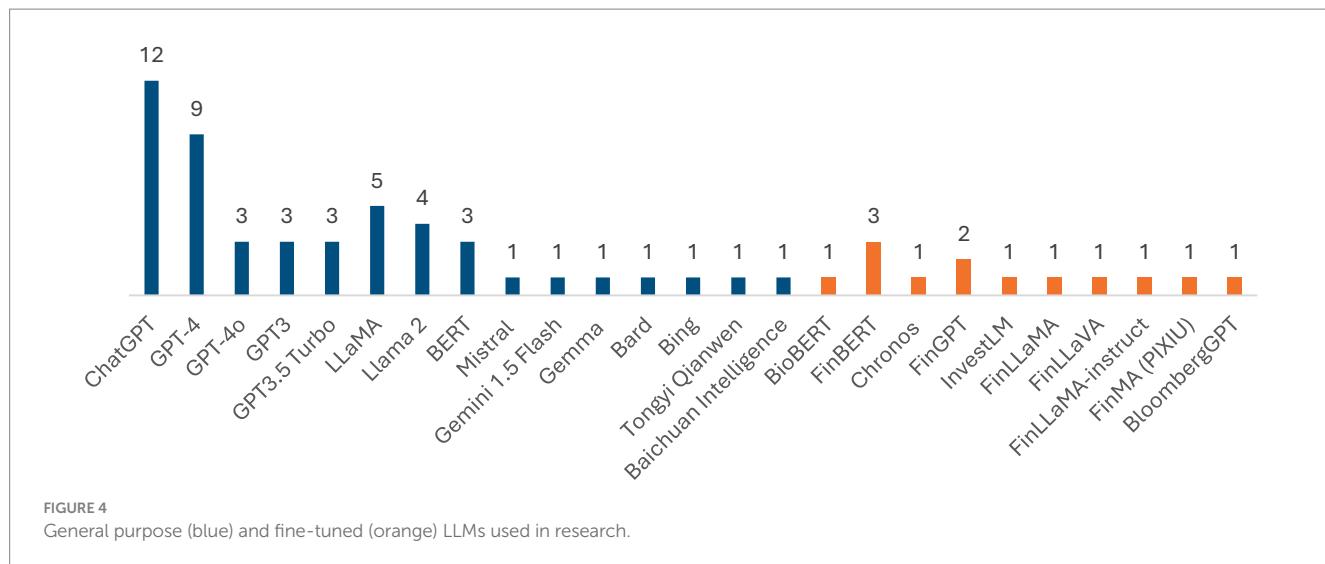


TABLE 1 Overview of financial LLMs for equity investing.

Model name	Base model	Parameters	Specialization	Key features
BloombergGPT	BLOOM	~50B	Financial text (news, reports, press)	Trained on combined general + financial corpus; strong performance on news-based tasks
FinGPT	Various	Varies	Financial datasets	Fine-tuned with Low-Rank Adaptation (LoRA); dissemination-aware and context-enriched
InvestLM	LLaMA-65B	65B	Financial investment tasks	Fine-tuned with a curated dataset for investment-related applications; focuses on advanced RL integration
FinLLaMA	LLaMA-2	52B tokens	Financial sentiment classification, trading simulations	Pre-trained on financial corpus; instruction fine-tuned with 573 K financial instructions
FinLlama	LLaMA-2 (7B)	7B	Sentiment analysis for algorithmic trading	Fine-tuned for sentiment valence and strength classification; optimized using LoRA
FinLLaVA	FinLLaMA	N/A	Multimodal financial data (text, tables, charts)	Trained with 1.43 M image-text instructions to handle complex financial data types
FinMA (PIXIU)	LLaMA (7B & 70B)	7B, 70B	Financial sentiment analysis and NLP tasks	Fine-tuned on financial datasets
FinBERT	BERT	N/A	Financial NLP tasks	Pre-trained on general and financial corpora; uses multi-task learning
FLANG	Custom	N/A	Financial corpus	Domain-specific model

- Benchmarking and Standardization:** Studies such as InvestorBench and FinanceQA offer benchmarking frameworks to systematically evaluate LLM performance in finance, providing a foundation for more structured comparisons.

- Research Delivery in Constrained environments:** Access to computational resources—such as GPUs for model fine-tuning could be a limitation. Similarly, lack of transparency and documentation on closed source models is another limitation for

experimentation and Model interpretability. However, the breadth of research so far despite these limitations is notable.

5.2 Limitations of existing research studies

Despite these strengths, several limitations remain:

- **Limited Real-World Testing:** While many studies demonstrate promising results in controlled environments, few have tested LLM-driven investment strategies in real-world trading conditions. Challenges such as execution slippage, transaction costs, and market impact are often overlooked.
- **Data Quality and Bias:** LLMs are highly sensitive to the quality of the training data and the data used for inference. Many LLM-driven models depend on real-time market data and unstructured text sources, making them vulnerable to biased, manipulated, or misleading information, leading to potential overfitting and poor generalization. There is a potential risk for bad actors and attackers to manipulate data to distort investing outcomes and money flow. Out of the 84 research papers, only few research studies cover the data quality limitations and some factor bias considerations.
- **Interpretability and Explainability Challenges:** The black-box nature of many LLM-based models limits transparency, making it difficult for investors and regulators to understand, trust, and justify AI-driven investment decisions. This lack of interpretability hinders broader adoption and raises regulatory concerns in financial markets. GPT based models used in many research paper so far are closed-source and do not provide visibility on the model specifications, limiting explainability.
- **Context window Limitations:** All existing research studies focus on data sets with limited sizes and scope. None of the studies have elaborated or found a scalable solution for the high context window of inputs required for stock investing use cases- examples include PDF files with 100 + pages. Evaluating and
- **Challenges in Back-Testing and Validation:** The risk of data leakage in back-testing remains a critical concern. Most studies fail to rigorously test against out-of-sample data or consider survivorship bias, which can inflate performance metrics.—mention few studies that account for this
- **Regional Coverage:** Most research studies are focused on the US and China markets with one study covering data from the Japan market. Scalability of the findings and frameworks to data from markets in other regions remains an open question.
- **Coverage across Investment types:** Majority of the research conducted so far focusses on broader stock investing in the construct of long term/value investing. Research focus on day trading use cases or High frequency trading is unexplored.
- **Limited Exploration of Non-Equity Asset Classes:** Most research focuses on stock investing, with limited exploration of LLM applications in commodities, fixed income, or options markets, which require different risk assessment models.

By addressing these gaps, future research can refine LLM applications in financial investing, making them more accurate, scalable, and aligned with industry needs.

5.3 Research gaps and future directions

To overcome current limitations and further enhance the utility of LLMs in stock investing, future research should focus on:

- **Hybrid Modeling Approaches:** Integrating LLMs with traditional quantitative and AI models (e.g., econometric, factor models) and can leverage the strengths of both methodologies, could improve predictive performance and decision-making reliability.
- **Reasoning Models:** Most general purpose and generic LLMs used in the research so far are derived from GPT, BERT and Llama. There is limited to none reference to usage of reasoning models such as GPT o1, GPT o1mini, Deep Seek R1, others.
- **Efficiency Improvements- Solving for Computational Overhead and Latency:** The large size and complexity of LLMs often result in high computational costs and latency issues, making real-time trading applications challenging. Advancements in model optimization—such as distillation, quantization, and efficient fine-tuning techniques like LoRA—could reduce computational overhead and latency, making real-time applications more feasible. There is a huge potential for model architectural and algorithmic advances to meet unique needs of stock investing use cases.
- **Scalability of Multi-Agent AI Systems:** While multi-agent LLM frameworks have shown initial promise in controlled setups, their scalability, coordination mechanisms and reliability in high-stakes financial environments remain an open challenge.
- **Enhanced Explainability and Interpretability:** Developing new interpretability frameworks tailored to LLMs in financial/stock investing applications will be crucial for building stakeholder trust and ensuring regulatory compliance. Research into techniques that demystify LLM outputs is needed. Future work must address the ethical implications of automated decision-making
- **Ethical & Regulatory considerations:** The ethical and regulatory implications of deploying LLMs in equity markets require more comprehensive investigation to ensure responsible adoption and compliance with financial regulations. Current research, while acknowledging issues such as data biases and model interpretability, often overlooks the broader ethical concerns, including the potential for LLMs to amplify market manipulations through biased or misleading data inputs, such as orchestrated social media campaigns or falsified financial reports. Additionally, the lack of transparency in closed-source models (e.g., GPT-based systems) raises concerns about accountability, particularly when these models influence high-stakes investment decisions. Regulatory frameworks, such as those enforced by the SEC or ESMA, demand rigorous validation and explainability of AI-driven strategies, yet few studies address compliance with these standards or the ethical risks of over-reliance on automated systems. Future research should prioritize developing frameworks for ethical AI governance, including robust auditing mechanisms to detect and mitigate biases, transparent reporting protocols for LLM-driven decisions, and alignment with global financial regulations to foster trust and ensure equitable market participation.
- **Solving for Large Context Window:** To ensure true scalability of use in financial applications, the context window limitations

- of LLMs will need to be solved. Research on this topic will be vital for true scalability.
- **Maintenance and Domain Adaptation:** Domain-specific LLMs require continuous fine-tuning and updates to remain effective as market conditions evolve. This can be resource-intensive and may limit scalability.
 - **Addressing Bias and Data Manipulation Risks:** Future research should develop more robust techniques to detect and mitigate biases in financial datasets, including adversarial attacks on AI-generated trading signals.
 - **Adaptive Learning and Market Feedback Integration:** Reinforcement learning from real market interactions should be further explored to enable self-improving models that adapt dynamically to changing market conditions
 - **Expansion to Broader Financial Instruments:** While most studies concentrate on equities, LLM applications should be extended to alternative asset classes, such as bonds, derivatives, and crypto markets, to assess their predictive power across different financial products.
 - **Cross-Regional and Cross-Market Evaluations:** LLM-based investment models should be tested across different geographic markets and economic conditions to assess their generalizability and robustness.

Author contributions

AJ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. VM: Formal analysis, Investigation,

Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. We used for literature review and re-writing some of the text concisely.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abdelsamie, M., and Wang, H. (2024). Comparative analysis of LLM-based market prediction and human expertise with sentiment analysis and machine learning integration. In Proceedings of the 2024 7th international conference on data science and information technology (DSIT) (pp. 1–6). IEEE.
- Aparicio, V., Gordon, D., Huayamares, S. G., and Luo, Y. (2024). BioFinBERT: finetuning large language models (llms) to analyze sentiment of press releases and financial text around inflection points of biotech stocks. ArXiv. Available online at: <https://arxiv.org/abs/2401.11011>
- Bhat, Rithesh H. Stock price trend prediction using emotion analysis of financial headlines with distilled LLM model. PETRA '24: Proceedings of the 17th international conference on pervasive technologies related to assistive environments. pp. 67–73.
- Bi, S., Deng, T., and Xiao, J. (2024). The role of AI in financial forecasting: ChatGPT's potential and challenges. ArXiv. Available online at: <https://arxiv.org/abs/2411.13562>
- Bond, S. A., Klok, H., and Zhu, M. (2023). Large language models and financial market sentiment. Available online at: <https://ssrn.com/abstract=4584928>
- Chen, S., Green, T. C., Gulen, H., and Zhou, D. (2024). What does ChatGPT make of historical stock returns? Extrapolation and miscalibration in LLM stock return forecasts. Available online at: <https://ssrn.com/abstract=4941906>
- Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected returns and large language models. Available online at: <https://ssrn.com/abstract=4416687>
- Chen, S., Peng, L., and Zhou, D. (2024). Wisdom or whims? Decoding investor trading strategies with large language models. Available online at: <https://ssrn.com/abstract=4867401>
- Cheng, Y., and Tang, K. (2023). GPT's idea of stock factors. Available online at: <https://ssrn.com/abstract=4560216>
- Chiang, H., Hynes, L., and Sandberg, D. (2025). Questioning the answers: LLMs enter the boardroom. S&P global market intelligence quantamental report. Available online at: <https://ssrn.com/abstract=5109196>
- Das, S., Zera, R. E., Lyngkhoi, M., Saha, S., and Maurya, A. (2024). Wealth guide: a sophisticated language model solution for financial trading decisions. In Proceedings of the eighth financial technology and natural language processing and the 1st agent AI for scenario planning, pp. 133–140, Jeju, South Korea.
- Deng, X., Bashlochkina, V., Han, F., Baumgartner, S., and Bendersky, M. (2022). What do LLMs know about financial markets? A case study on Reddit market sentiment analysis. WWW '23 companion: companion proceedings of the ACM web conference 2023, pp. 107–110
- Deng, Y., He, X., Hu, J., and Yiu, S. (2024). Enhancing few-shot stock trend predictions with large language models. ArXiv. Available online at: <https://arxiv.org/abs/2407.09003>
- Di, Z., Chen, J., Yang, Y., Ding, L., and Xiang, Y. (2024). LLM-driven knowledge enhancement for securities index prediction. In proceedings of the first international OpenKG workshop: large knowledge-enhanced models. Jeju Island, South Korea: CEUR Workshop Proceedings, 3818, 71–82.
- Ding, Y., Jia, S., Ma, T., Mao, B., Zhou, X., Li, L., et al. (2023). Integrating stock features and global information via large language models for enhanced stock return prediction. ArXiv. Available online at: <https://arxiv.org/abs/2310.05627>
- Ding, H., Li, Y., Wang, J., and Chen, H. (2024). Large language model agent in financial trading: a survey. ArXiv. Available online at: <https://arxiv.org/abs/2408.06361>
- Dolphin, R., Dursun, J., Chow, J., Blankenship, J., Adams, K., and Pike, Q. (2024). Extracting structured insights from financial news: An augmented LLM driven approach [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2407.15788>
- Fatemi, S., and Hu, Y. (2024). FinVision: a multi-agent framework for stock market prediction. ICAIF '24: Proceedings of the 5th ACM international conference on AI in finance, pp. 582–590.
- Fatouros, G., Metaxas, K., Soldatos, J., and Karathanassis, M. (2025). MarketSenseAI 2.0: enhancing stock analysis through LLM agents. ArXiv. Available online at: <https://arxiv.org/abs/2502.00415>

- Fatouros, G., Metaxas, K., Soldatos, J., and Kyriazis, D. (2024). Can large language models beat wall street? Unveiling the potential of AI in stock selection. ArXiv. Available online at: <https://doi.org/10.1007/s00521-024-10613-4>
- Fieberg, C., Hornuf, L., Streich, D., and Meiler, M. (2024). Using large language models for financial advice. Available online at: <https://ssrn.com/abstract=4850039>
- Glasserman, P., and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. ArXiv. Available online at: <https://arxiv.org/abs/2309.17322>
- Gu, J., Ye, J., Wang, G., and Yin, W. (2024). Adaptive and explainable margin trading via large language models on portfolio management. ICAIF '24: Proceedings of the 5th ACM international conference on AI in finance, pp. 248–256.
- Guo, T., and Hauptmann, E. (2024). Fine-tuning large language models for stock return prediction using newsflow. In Proceedings of the 2024 conference on empirical methods in natural language processing: industry track, pp. 1028–1045, Miami, Florida, US. Association for Computational Linguistics.
- Guo, J., and Shum, H. (2024). Large investment model. ArXiv. Available online at: <https://arxiv.org/abs/2408.10255>
- Huang, Z., Zhang, Z., Hua, C., Liao, B., and Li, S. (2024). Leveraging enhanced egret swarm optimization algorithm and artificial intelligence-driven prompt strategies for portfolio selection. *Sci. Rep.* 14:26681. doi: 10.1038/s41598-024-77925-2
- Kim, H., and Oh, H. Llm analyst: what stocks do you recommend today. Available at SSRN. Available online at: <https://ssrn.com/abstract=4899957>
- Kirtac, K., and Germano, G. (2024). Sentiment trading with large language models. *Financ. Res. Lett.* 62:105227. doi: 10.1016/j.frl.2024.105227
- Ko, H., and Lee, J. (2024). Can ChatGPT improve investment decisions? From a portfolio management perspective. *Financ. Res. Lett.* 64:105433. doi: 10.1016/j.frl.2024.105433
- Koa, K. J., Ma, Y., Ng, R., and Chua, T. (2024). Learning to generate explainable stock predictions using self-reflective large language models. In Proceedings of the ACM web conference 2024 (WWW '24) (pp. 4304–4315).
- Koa, K. J., Ma, Y., Ng, R., Zheng, H., and Chua, T. (2024). Temporal relational reasoning of large language models for detecting stock portfolio crashes. Arxiv. Available online at: <https://arxiv.org/abs/2410.17266>
- Kong, Y., Nie, Y., Dong, X., Mulvey, J., Poor, V., Wen, W., et al. (2024). Large language models for financial and investment management: applications and benchmarks. *J. Portfolio Manag. Quant. Tools* 51, 162–210. doi: 10.3905/jpm.2024.1.645
- Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., et al. (2024). Large language models for financial and investment management: models, opportunities, and challenges. *J. Portfolio Manag.* 51, 211–231. doi: 10.3905/jpm.2024.1.646
- Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T. G., and Mandic, D. (2024). FinLlama: financial sentiment classification for algorithmic trading applications. ICAIF '24: Proceedings of the 5th ACM international conference on AI in finance, pp. 134–141.
- Kou, Z., Yu, H., Peng, J., and Chen, L. (2024). Automate strategy finding with LLM in quant investment [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2409.06289>
- Krause, D. (2023). Large language models and generative AI in finance: an analysis of ChatGPT, Bard, and Bing AI. Available online at: <https://ssrn.com/abstract=4511540>
- Lee, H., Choi, Y., and Kwon, Y. Quantifying qualitative insights: leveraging LLMs to market predict. Available online at: <https://ssrn.com/abstract=5093626>
- Lefort, B., Benhamou, E., Ohana, J. -J., Saltiel, D., Guez, B., and Jacquiot, T. (2024). Sentiment analysis of bloomberg markets wrap using ChatGPT: application to the NASDAQ. Available online at: <https://ssrn.com/abstract=4780150>
- Li, H., Cao, Y., Yu, Y., Javaji, S. R., Deng, Z., He, Y., et al. (2024). INVESTORBENCH: a benchmark for financial decision-making tasks with LLM-based agent. ArXiv. Available online at: <https://arxiv.org/abs/2412.18174>
- Li, L., Chang, T., and Wang, H. (2023). Multimodal Gen-AI for fundamental investment research. ArXiv. Available online at: <https://arxiv.org/abs/2401.06164>
- Li, X., Li, Z., Shi, C., Xu, Y., Du, Q., Tan, M., et al. (2024). AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024) (pp. 773–783). ELRA and ICCL.
- Li, Y., Luo, B., Wang, Q., Chen, N., Liu, X., and He, B. (2024). CryptoTrade: A reflective LLM-based agent to guide zero-shot cryptocurrency trading. In Proceedings of EMNLP 2024 (pp. 1094–1106). Association for Computational Linguistics.
- Li, Y., Yu, Y., Li, H., Chen, Z., and Khashanah, K. (2023). TradingGPT: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2309.03736>
- Liang, Y., Liu, Y., Zhang, B., Wang, C. D., and Yang, H. (2024). FinGPT: enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched LLMs. ArXiv. Available online at: <https://arxiv.org/abs/2412.10823>
- Liu, C., Arulappan, A., Naha, R., Mahanti, A., Kamruzzaman, J., and Ra, I. H. (2024). Large language models and sentiment analysis in financial markets: a review, datasets, and case study. *IEEE Access* 12, 134041–134061. doi: 10.1109/ACCESS.2024.3445413
- Liu, X., Wang, G., Yang, H., and Zha, D. (2023). FinGPT: democratizing internet-scale data for financial large language models. ArXiv. Available online at: <https://arxiv.org/abs/2307.10485>
- Lopez-Lira, A., and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models [Working paper]. Available online at: <https://doi.org/10.2139/ssrn.4412788>
- Lu, F. (2025). Robo-advising meets large language models: educating investors on alpha and beta of mutual funds and stocks. Available online at: <https://ssrn.com/abstract=5083305>
- Lu, F., Huang, L., and Li, S. (2023). ChatGPT, generative AI, and investment advisory. Available online at: <https://ssrn.com/abstract=4519182>
- Mateega, S., Georgescu, C., and Tang, D. (2025). FinanceQA: a benchmark for evaluating financial analysis capabilities of large language models. ArXiv. Available online at: <https://arxiv.org/abs/2501.18062>
- Ni, H., Meng, S., Chen, X., Zhao, Z., Chen, A., Li, P., et al. (2024). Harnessing earnings reports for stock predictions: a QLoRA-enhanced LLM approach. ArXiv. Available online at: <https://doi.org/10.1109/DOCS363458.2024.10704454>
- Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., et al. (2024). A survey of large language models for financial applications: progress, prospects and challenges. ArXiv. Available online at: <https://arxiv.org/abs/2406.11903>
- Nishida, S., and Utsuro, T. (2025). Generating financial news articles from factors of stock price rise / decline by LLMs. Proceedings of the joint workshop of the 9th financial technology and natural language processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), pp. 184–195, Abu Dhabi, UAE, Association for Computational Linguistics.
- Papasotiriou, K., Sood, S., Reynolds, S., and Balch, T. (2024). “AI in investment analysis: LLMs for equity stock ratings” in ICAIF '24: Proceedings of the 5th ACM international conference on AI in finance (ACM), 419–427.
- Park, T. (2024). Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework. ArXiv. Available online at: <https://arxiv.org/abs/2403.19735>
- Perlin, M., Fogesatto, C., Muller, F. M., and Righi, M. Can AI beat a naive portfolio? An experiment with anonymized data. Available online at: <https://ssrn.com/abstract=4954881>
- Pop, A., Spörer, J., and Handschuh, S. (2024). The structure of financial equity research reports -- identification of the most frequently asked questions in financial analyst reports to automate equity research using Llama 3 and GPT-4. ArXiv. Available online at: <https://arxiv.org/abs/2407.18327>
- Romanko, O., Narayan, A., and Kwon, R. H. (2023). ChatGPT-based investment portfolio selection. *Oper. Res. Forum* 4:91. doi: 10.1007/s43069-023-00277-6
- Saqur, R., and Rudzicz, F. (2024). Stock price trend prediction using emotion analysis of financial headlines with distilled LLM model. Open Review. Available online at: https://openreview.net/forum?id=y3W1TVujii&referrer=%5Bthe%20profile%20of%20Raeid%20Saqr%5D%2Fprofile%3Fid%3D~Raeid_Saqr1
- Shi, J., and Hollifield, B. (2024). Predictive power of LLMs in financial markets. ArXiv. Available online at: <https://arxiv.org/abs/2411.16569>
- Swamy, M., Shukla, A., and Purtilo, J. LLM-based stock market trend prediction. Open Review. Available online at: <https://openreview.net/forum?id=ICwdNpmu2d>
- Tandon, R. (2024). Prediction of stock market trends based on large language models. *ITEGAM J. Eng. Technol. Ind. Appl. (ITEGAM-JETIA)* 11, a615–a622. Available at: <https://www.jetir.org/view?paper=JETIR2409071>
- Tong, H., Li, J., Wu, N., Gong, M., Zhang, D., and Zhang, Q. (2024). Ploutos: towards interpretable stock movement prediction with financial large language model. ArXiv. Available online at: <https://arxiv.org/abs/2403.00782>
- Valeyre, S., and Aboura, S. (2024). LLMs for time series: an application for single stocks and statistical arbitrage. ArXiv. Available online at: <https://arxiv.org/abs/2412.09394>
- Vidal, J. (2024). Efficacy of AI and other large language models in predicting stock prices. Available online at: <https://ssrn.com/abstract=4947135>
- Voigt, F., von Luck, K., and Stelldinger, P. (2024). Assessment of the applicability of large language models for quantitative stock price prediction. PETRA '24: Proceedings of the 17th international conference on pervasive technologies related to assistive environments, pp. 293–302.
- Wang, Q., Gao, Y., Tang, Z., Luo, B., and He, B. (2024). Enhancing LLM trading performance with fact-subjectivity aware reasoning [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2410.12464>
- Wang, S., Ji, T., Wang, L., Sun, Y., Liu, S., Kumar, A., et al. (2024). StockTime: a time series specialized large language model architecture for stock price prediction. ArXiv. Available online at: <https://arxiv.org/abs/2409.08281>
- Wang, H., Pan, Z., Zhang, H., Liu, M., Lin, Y., and Zhao, H. V. (2024). InvestAlign: Align LLMs with investor decision-making under herd behavior. In Adaptive foundation models: evolving AI for personalized and efficient learning, NeurIPS 2024.
- Wang, S., Yuan, H., Zhou, L., Ni, L. M., Shum, H., and Guo, J. (2023). Alpha-GPT: Human-AI interactive alpha mining for quantitative investment [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2308.00016>

- Wu, R. (2024). Portfolio performance based on LLM news scores and related economical analysis [Working paper]. Available online at: <https://ssrn.com/abstract=4709617>
- Xiao, Y., Sun, E., Luo, D., and Wang, W. (2024). TradingAgents: multi-agents LLM financial trading framework. ArXiv. Available online at: <https://arxiv.org/abs/2412.20138>
- Xing, F. (2024). Designing heterogeneous LLM Agents for financial sentiment analysis. *ACM Trans. Manag. Inf. Syst.* 16, 1–24. doi: 10.1145/3688399
- Yang, H., Zhang, B., Wang, N., Guo, C., Zhang, X., Lin, L., et al. (2024). FinRobot: an open-source AI agent platform for financial applications using large language models. ArXiv. Available online at: <https://arxiv.org/abs/2405.14767>
- Yang, Y., Zhang, Y., Wu, M., Zhang, K., Zhang, Y., Yu, H., et al. (2025). TwinMarket: A scalable behavioral and social simulation for financial markets [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2502.01506>
- Yu, B. (2023). Benchmarking large language model volatility [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2311.15180>
- Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., et al. (2023). FinMem: a performance-enhanced LLM trading agent with layered memory and character design [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2311.13743>
- Yu, Y., Yao, Z., Li, H., Deng, Z., Cao, Y., Chen, Z., et al. (2024). FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024). [Conference proceedings]. Available online at: https://proceedings.neurips.cc/paper_files/paper/2024/hash/f7ae4fe91d96f50abc2211f09b6a7e49-Abstract-Conference.html
- Yue, T., and Au, D. (2023). GPTQuant's conversational AI: simplifying investment research for all. Available online at: <https://ssrn.com/abstract=4380516>
- Zhang, C., Liu, X., Zhang, Z., Jin, M., Li, L., Wang, Z., et al. (2024). When AI Meets Finance (StockAgent): large language model-based stock trading in simulated real-world environments. ArXiv. Available online at: <https://arxiv.org/abs/2407.18957>
- Zhao, Z. (2024). Next-generation intelligent portfolio management [Institutional report]: DSpace@MIT. Cambridge, MA: Massachusetts Institute of Technology. Available at: <https://dspace.mit.edu/handle/1721.1/156635>
- Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., et al. (2024). Revolutionizing finance with LLMs: an overview of applications and insights [ArXiv preprint]. ArXiv. Available online at: <https://arxiv.org/abs/2401.11641>
- Zhao, Z., and Welsch, R. E. (2024). Aligning LLMs with human instructions and stock market feedback in financial sentiment analysis. ArXiv. Available online at: <https://arxiv.org/abs/2410.14926>
- Zhou, T., Wang, P., Wu, Y., and Yang, H. (2024). FinRobot: AI agent for equity research and valuation with large language models. ArXiv. Available online at: <https://arxiv.org/abs/2411.08804>

The Relevance of Web Traffic for Stock Prices of Internet Firms

Shivaram Rajgopal
University of Washington
Box 353200
Seattle, WA 98195
(206) 543-7525
Fax: (206) 683-9392
E-mail: rajgopal@u.washington.edu

Suresh Kotha
University of Washington
Box 353200
Seattle, WA 98195
(206) 543-4466
Fax: (206) 683-9392
E-mail: skotha@u.washington.edu

Mohan Venkatachalam
Graduate School of Business
Stanford University
Stanford, CA 94305
Tel: (650) 725-9461, Fax: (650) 725-0468
E-mail: vmohan@gsb.stanford.edu

October 2000

Rajgopal and Kotha appreciate funding from the University of Washington. Mohan Venkatachalam appreciates funding from the Stanford University. We are grateful to Tim Miller of Net Media Resources (webmergers.com) for generously providing us with the data on mergers of Internet companies. We thank an anonymous reviewer, Terry Shevlin, Dave Burgstahler, Nola Jean Bamberg for editorial assistance, and the workshop participants at the University of British Columbia, Oregon and Washington (UBCOW) in January 2000 and the European Finance Association (EFA) meetings at London in August 2000 for their comments. We also acknowledge Sunil Kumar, Rick Lambert, Anu Ramanathan of Amazon.com and Violina Rindova for many helpful discussions and suggestions.

The Relevance of Web Traffic for Stock Prices of Internet Firms

Abstract

This study shows that web traffic is an important non-financial indicator of the market values of Business to Consumer (B2C) Internet firms. We add three important insights to the literature on the value-relevance of traffic. We show that traffic is a summary measure of the strategies that firms use to attract visitors to their websites. The value-relevance of traffic disappears once the exogenous determinants of traffic (e.g., setting up an alliance with America Online, creating affiliate referral programs, generating media visibility, incurring marketing expenditure and constraints imposed by cash availability) are accounted for in the value-relevance model. We also demonstrate that traffic contains no predictive information about future revenues once past revenues are accounted for. The value-relevance of traffic does not stem merely from its role as a predictor of future sales. Finally, we show that the stock market appears to use traffic as a measure of the web businesses' ability to create network effects. Network effects occur when the value of a website to a visitor may depend on how many others visit that site. Consistent with Metcalfe's law of network economics, we find that market values of web businesses increase non-linearly with traffic.

The Relevance of Web Traffic for Stock Prices of Internet Firms

1. Introduction

Many see the Internet as a revolutionary technology that will alter the way business, commerce, medicine, science, communications, the law, politics, and government are conducted (Gates, 2000; Christensen and Overdorf, 2000). Andrew Grove, the chairman of Intel Corporation, predicts that the Internet will become so pervasive that in the future every business will be an Internet business or no business at all (Grove, 1996).

Online retail sales are forecasted to reach \$184 billion by 2004 as compared to \$700 million in 1996 (Modahl, 2000). A precursor to generating such retail sales is attracting traffic to the firm's website. Web traffic is a non-financial measure actively followed by the investment community to value firms in the Business to Consumer (B2C) segment (*The Houston Chronicle*, November 22, 1999). One of the important reasons for traffic's popularity as a value-driver among the investment community is the continuing prevalence of somewhat anomalous relations between market values and key financial measures. An Internet venture like Amazon.com has achieved a higher valuation than the entire traditional book retailing and publishing industries combined, even though it has yet to turn a profit (Evans and Wurster, 1999). Despite a steep decline in the stock prices of web businesses this year (Demers and Lev, 2000), a large number of B2C firms continue to trade at high prices relative to their operating performance. As of July 31, 2000, Yahoo had a Price to Earnings (P/E) ratio of 386, eBay of 568, and Amazon.com traded at a multiple to revenue of 4.66 with a market capitalization 400 times its book value. Such apparently anomalous associations between market values and key accounting measures raise questions about the role of non-financial information such as website traffic in explaining the variation in the market values of B2C Internet firms.

Website traffic as a non-financial metric has three appealing characteristics. First, web traffic provides information about the extent of consumer interest in the web business and is central to revenue generation and growth of B2C Internet firms. Second, traffic numbers are readily obtainable from third-party survey firms such as PC Data Online, Nielsens and Media Metrix. Third, traffic is cross-sectionally comparable across several B2C business models such as portals (e.g., Yahoo), e-tailers (e.g., Amazon.com), content and community sites (e.g., iVillage) and sites providing financial services (e.g., E*trade).

We examine the value-relevance of web traffic using a sample of 92 firms covering portals, content and community sites, financial services sites and e-tailers over the five quarters beginning with the first quarter of 1999. We find that the number of unique monthly visitors to a site -- our measure of web traffic drawn from PC Data Online -- is positively associated with stock prices and adds significant incremental explanatory power (24 percentage points) to a regression of just financial statement information against Internet share prices.

After establishing that traffic is highly value relevant, we ask three research questions: (1) Is traffic value relevant in its own right, or is it merely a proxy for the strategies firms use to generate traffic? (2) Is traffic valued by the market because it predicts future sales? (3) Is traffic valued because it measures the potential customer relationships that network effects can create?

In our first set of empirical analyses, we model the factors and constraints affecting website traffic. In conducting such analysis, we are motivated by concerns about the problematic issue of endogeneity while interpreting the value-relevance of non-financial indicators (Ittner and Larcker, 2000). Ittner and Larcker (2000) observe that if all organizations in the sample are optimizing with regard to their choice of traffic, one would observe no association between

traffic and organizational performance measures such as market value of equity, once the exogenous determinants of the choice of traffic are controlled for in the value-relevance model.

We posit that traffic levels are determined by several determinants such as an alliance with America Online (AOL), presence of an affiliate-marketing programs, the magnitude of marketing expenditure, the extent of media visibility attained by the firm, and the extent of cash available. Considering web traffic as an endogenous variable, we find that traffic is not value relevant in its own right, but is a proxy for the aforementioned strategies used by firms to generate traffic.

Our next objective is to explore why the stock market values traffic. One plausible explanation is that web traffic provides information about a firm's future sales. Consistent with Trueman et al. (2000b), we find that web traffic levels predict one- and two-quarter ahead sales. However, traffic *per se* does not explain future sales incremental to current sales of the firm. In other words, traffic has no incremental information about future revenues once the predictive information in past sales is controlled for. Moreover, both traffic and sales are incrementally value-relevant over each other. Hence, the market does not appear to value traffic merely because it predicts future sales.

Finally, we explore the possibility that web traffic is positively priced by the market because of potential future benefits from network effects generated by traffic. The value of a website to a visitor may depend on how many others visit that site. Once the number of visitors and hence the size of the virtual community created by the firm grows, more and more users find the firm's website attractive because of their ability to interact with other members of the community and their ability to share and contribute to member generated content. Moreover, accumulation of data about visitors' preferences makes it possible for vendors and advertisers to

tailor products and services to visitors, thus making the site even more attractive to future visitors. This, in turn, increases the potential for future long-run profitability.

We investigate whether the stock market values traffic based on the number of potential relationships site visitors can create among themselves. Metcalfe's law of network economics predicts that if there are n people in the network, the value of the network is proportional to the number of other users, i.e., $n \times (n-1) = n^2 - n$ (Shapiro and Varian, 1999). If Metcalfe's law is descriptive of the data, we would expect the market values of web businesses to increase in the squared transformation of the number of unique visitors to the firm's website. Consistent with Metcalfe's law, we find that the market values of web businesses are positively associated with the squared transformation of the number of unique visitors to the firm's website. We also examine traffic-based acquisitions for the period September 1999 to June 2000 to provide corroborative evidence on network effects. Our results indicate that not only do target firms' traffic numbers account for 95 percent of the cross sectional variation in acquisition prices, but these acquisition prices are also a positive function of the squared transformation of the number of unique visitors acquired. Thus, evidence from the market for corporate control is also consistent with traffic's value-relevance stemming from its ability to measure potential relationships among site visitors.

Our study extends the growing body of literature on the valuation of Internet firms in two ways. First, concurrent work (e.g., Trueman et al., 2000) on the value-relevance of traffic and other non-financial measures such as customer satisfaction (Ittner and Larcker, 1998) or population coverage in the telecommunication industry (Amir and Lev, 1996) implicitly assume that managers are non-optimizers. These studies are silent about why value-maximizing managers would not increase non-financial measures, such as traffic, infinitely in an attempt to

increase firm value (Lambert, 1998; Nagar, 2000). In contrast, we treat traffic as an endogenous choice variable that optimizing managers select based on other exogenous factors. We show that once the exogenous factors affecting traffic are controlled for, there is no relation between market values and traffic. This finding is consistent with the idea that managers set traffic levels optimally given the exogenous factors they face. Second, we contribute to the understanding of why traffic is value-relevant. In particular, we document evidence consistent with potential benefits from positive network externalities as one important explanation for the value-relevance of web traffic.

The remainder of the paper is organized as follows. Section 2 discusses the data and descriptive statistics. Section 3 explores the value-relevance of traffic. Section 4 examines alternative explanations for why web traffic is value-relevant, and section 5 presents some concluding remarks.

2. Data and Descriptive Statistics

2.1 Traffic data

We rely on web traffic data compiled by PC Data Online – an independent firm that measures Internet audiences. PC Data Online defines its Internet audience as individuals who access the World Wide Web or proprietary online areas such as America Online during the past 30 days using personal computers with Windows 95/98/NT as their operating system. PC data generates its data from a random panel of 100,000 participants who have installed the company's tracking software on their personal computers at home or at work. This software collects and stores a participant's web activities on his/her computer. Once the user has been online for 15 minutes, which may be split across one or more sessions, this data is encrypted and sent, in real time, via the Internet to PC Data Online.

Of the various metrics reported by PC Data Online, we focus on unique monthly visitors in our study.¹ PC Data Online defines unique visitors as the number of web-active individuals who visited a particular site(s) belonging to a web property (company) within a given time period. Each visitor is represented only once as a unique user. The data on unique monthly visitors for each month is usually posted within a week to fifteen days after the end of the month on PC Data Online's website. Traffic statistics compiled by PC Data Online are freely available to the public on PC Data's website.²

2.2 Sample and descriptive statistics

Our sample consists of a list of 92 publicly traded pure Internet firms (see Panel A of Table 1). We begin with a list of 120 firms from four categories of firms on the Internet Stock List at www.internet.com as of July 1, 2000: (i) Content and Community sites; (ii) E-tailers; (iii) Financial Services sites and, (iv) Portals. We focus on only the aforementioned categories because the business model for firms in these categories involves generating revenue by exploiting traffic attracted to their websites. The Internet Stock List compiled by Internet.com has been used to collect a sample of Internet firms in a number of previous studies (Trueman et al., 2000a, b; Hand, 2000a, b; Demers and Lev, 2000). To this list we add four firms (Excite, Geocities, Onsale, and Xoom.com) that have been acquired or merged before July 1, 2000. From the initial list of 124 firms, we exclude 18 firms for which traffic data was not available on PC Data Online for any quarter in our sample period. Fourteen more firms are dropped because we cannot find financial statements for any quarter during the sample period on the SEC's EDGAR database. Panel B of Table 1 presents a frequency distribution of firms sorted by industry type. As shown, e-tailers (38 of 92) and content and community sites (36 of 92) dominate the sample.

¹ We also consider the percentage of unique monthly visitors to total web population as an alternative measure and find similar results.

We hand-collect all financial data from 10-Qs and 10-Ks filed by firms available on the EDGAR database in SEC's website www.sec.gov. Information about unique monthly visitors for our sample firms comes from PC Data for the period February 1999 to March 2000. In particular, we use the quarterly average of unique monthly visitors (UNIVIS) for our empirical analyses. Thus, the quarterly average for the quarters ended March 31, 1999, June 30, 1999, September 30, 1999, December 31, 1999 and March 31, 2000 are lined with accounting data from 10-Qs for those quarters.^{3,4}

Because PC Data issues a press release for a particular month's traffic within thirty days of the end of that month, we measure the market value of the firm's equity thirty days after the 10-Q quarter-end. Stock prices are obtained from www.finance.yahoo.com. Of the possible 460 firm quarters (92 firms over 5 quarters), we are left with 303 firm-quarters for our empirical analyses. This is because all firms in our sample were not publicly traded throughout the sample period. Also, note that the number of observations reported in the empirical analyses that follow may not equal 303. This is because statistical outliers, defined as firm-quarter observations with absolute values of R-student measures greater than 3, are deleted when estimating the regressions.

Panel C of Table 1 presents descriptive data on UNIVIS and a number of other independent variables used in the analyses. For descriptive reasons, we also provide data on the quarterly average of a firm's REACH, defined by PC Data Online as the percentage of unique

² The site has recently started restricting free access.

³ Because PC data started reporting traffic numbers from February 1999, we assume that the average unique monthly visitors for the quarter ended March 1999 is the same as the average unique monthly visitors for February and March 1999.

⁴ A small number of firms do not follow the calendar year for reporting purposes. For these firms we align the observations with the calendar quarters depending on the fiscal quarter ending dates. If the fiscal quarter ends within one month of a calendar quarter we include the observations in that calendar quarter. Observations for a firm with fiscal quarter ending in April or February will belong to the first calendar quarter. However, when calculating

monthly visitors to a firm's site scaled by the total web population. It is interesting to note that the mean firm attracts 7% of the Internet population in a quarter or 4.1 million unique visitors on average in a quarter. At least three-fourths of the observations in the sample report negative earnings because the third quartile cut-off of the earnings distribution is negative. For the median firm, quarterly losses (\$10.68 million) actually exceed quarterly sales (\$10 million). However, the median firm still has a market to book ratio of about 5. Next, we explore whether traffic explains some of the variation in market values, after controlling for financial information.

3. The value-relevance of traffic

3.1 Levels specification

In the absence of strong priors about how traffic should be incorporated into a model that relates market values to accounting and non-accounting information, we follow Amir and Lev (1996) and introduce UNIVIS, our proxy for traffic, as a linear additive value driver in regression (1) below:

$$MVE_{jt} = \beta_0 + \beta_1 BVE_{jt} + \beta_2 E_{jt} + \beta_3 \text{Log}(TA)_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_{5i} \text{IND}_{jxit} + \beta_{6k} \text{QTR}_{jkt} + \varepsilon_{jt} \quad (1)$$

where MVE is market value of equity thirty days after the fiscal quarter end, E is earnings before extraordinary items, BVE is book value of common equity, Log(TA) is the natural logarithm of total assets, UNIVIS is average monthly unique visitors during the quarter, IND is an industry dummy that reflects the firm's membership in each of the four ($i=1,2,\dots,4$) industries studied (content and community sites, portals, e-tailers and financial services), QTR is a quarter dummy that identifies one of the five ($k=1,2,\dots,5$) quarters studied (i.e., 4 quarters in 1999 and the first quarter of 2000). Finally, j and t are firm and quarter subscripts respectively.

the average monthly unique visitors we use the average of visitors in the fiscal quarter and determine market value of equity using the stock price thirty days after the fiscal quarter end.

Because MVE is not scaled by a size deflator, we add the logarithm of total assets (Log(TA)) as an independent variable (Barth and Kallapur (1996)) and report White (1980) adjusted t-statistics to account for heteroscedasticity. Addition of Log(TA) as a scale control also helps us assess whether BVE is value-relevant in its own right as opposed to serving as a scale control. Industry dummies and quarter dummies are introduced to control for unaccounted omitted variables that may be correlated with industry membership or time.

Results of estimating equation (1) are presented in panel A of Table 2. We conduct the regression analyses in stages to document the incremental value-relevance of UNIVIS. In the first stage we consider only the financial variables, earnings and book value, along with the scale control, Log(TA). Consistent with Hand (2000) and Trueman et al. (2000), we find that the coefficient on earnings is negative and statistically significant (coefficient = -12.18, t-statistic = -1.91). The coefficient on book value is positive and statistically significant (coefficient = 4.49, t-statistic = 2.53). Earnings and book value, along with the scale control, explain 36.61% of the cross-sectional variation in market values of firms.

We then estimate a model that includes both financial variables and the traffic measure, UNIVIS. As shown in panel A of Table 2, the coefficient on UNIVIS is positive and significantly associated with stock prices (coefficient = 700.49, t-statistic = 4.54).⁵ This suggests that market participants appear to attach a value of \$700 per unique monthly visitor. We also find that UNIVIS provides significant incremental explanatory power (about 24%) for stock prices beyond that provided by financial measures. Notwithstanding the inclusion of UNIVIS, the signs and the significance of coefficients on book value and earnings remain unchanged. Thus, while traffic explains a significant portion of variation in stock prices, we cannot dismiss the

relevance of financial information. We explore the robustness of the value-relevance result below.

3.2 Robustness checks

We conduct two sets of checks to assess robustness of the above results. First, we address econometric concerns such as correlation in error terms (sub-section 3.2.1) and the choice of functional form used to assess value-relevance (sub-section 3.2.2). Second, we examine the economic interpretation of the coefficient on UNIVIS after accounting for the effects of the stock market crash in B2C stocks in April 2000 (sub-section 3.2.3). We explain these checks in greater detail in the following paragraphs.

3.2.1 Serial and cross correlation in error terms

The results in panel A of Table 2 are based on a pooled cross-sectional OLS (ordinary least squares) model. Although we introduced industry and quarter dummies into the specification, the error terms in the model are likely to suffer from serial and cross-correlation. To address this issue we re-estimate regression (1) every quarter with industry dummies. In untabulated results, we find that UNIVIS is statistically significant in all the 5 quarters examined (mean coefficient = 470.13; mean t-statistic = 2.77). We also estimate equation (1) using the Generalized Least Squares (GLS) approach. Unlike OLS, the GLS model does not set the covariance among the error terms to zero. In untabulated results, we find that the coefficient on UNIVIS is strongly positive and statistically significant (coefficient = 600.21, $p < 0.01$). Thus, our value-relevance result is robust after controlling for serial or cross-correlation in errors.

3.2.2 Returns Analysis

⁵ When Yahoo is excluded from the data set, the pricing multiple on a unique monthly visitor drops to \$127 but it is statistically significant at $p < 0.01$. We re-estimate all regressions reported in the paper after excluding Yahoo and find that the qualitative nature of the inferences remains unchanged.

To assess the robustness of the results reported in panel A of Table 2 to a returns specification, we estimate the following returns regression:

$$\text{Ret}_{jt} = \beta_0 + \beta_1 E_{jt} + \beta_2 \Delta E_{jt} + \beta_3 \Delta \text{UNIVIS}_{jt} + \beta_{4i} \text{IND}_{jxit} + \beta_{5k} \text{QTR}_{jkt} + v_{jt} \quad (2)$$

where Ret is the abnormal return measured as holding period return over a three-month period ending 30 days after fiscal quarter end adjusted for return on NASDAQ index, E is the earnings before extraordinary items, ΔE is the change in earnings before extraordinary items, ΔUNIVIS is the change in UNIVIS. All the independent variables are scaled by market value of equity determined 30 days after the previous fiscal quarter end. In equation (2) j and t are firm and quarter subscripts respectively, while IND and QTR are industry and quarter dummies respectively.

Following Easton and Harris (1991), earnings levels and earnings changes are introduced to control for accounting information. Panel B of Table 2 reports the results of estimating regression equation (2). While earnings levels are weakly significant, earnings changes are not statistically associated with abnormal returns. More important, we find that the coefficient on change in UNIVIS is positive and statistically significant (coefficient = 15.88, t-statistic = 1.91). Thus, the value-relevance of UNIVIS is robust to the changes specification.

3.2.3 Impact of the April 2000 stock market crash in B2C stocks

Demers and Lev (2000) examine the impact of April 2000 sell-off in B2C stocks on the pricing of various factors such as research and development expenditure, cash burn rate, strategic alliances and traffic. For descriptive reasons, we examine the effect of the April 2000 stock market crash on the implied pricing of web traffic for our sample firms. As noted, the market values corresponding to the accounting and traffic numbers are measured 30 days after the fiscal quarter end. Thus, for firms whose fiscal year ends on March 31, 2000 or later, market value of

equity will incorporate the April sell-off. We interact each regressor in equation (1) with a dummy variable that is set to 1 if the observation corresponds to the first quarter of 2000 and zero otherwise. We find that the market appears to have marked down the price of web traffic by about \$411 per unique monthly visitor. Nonetheless, the average weight placed by the market on a unique visitor after the market crash is still positive and statistically significant.

3.3 Traffic as a choice variable

The above analyses document that traffic exhibits systematic and robust value-relevance. However, an implicit assumption behind the valuation equation (1) is that traffic is not a choice variable for firms (Ittner and Larcker, 1998; Lambert, 1998; Nagar, 2000). The result that greater traffic implies greater market value begs the question as to why managers do not increase traffic even further to garner greater market values for their firms. Surely, there must be costs or constraints associated with increasing traffic. If such constraints and other exogenous determinants of traffic are controlled for in the value-relevance model, the traffic measure would cease to be value-relevant. That is, traffic may not be value-relevant in its own right. It might merely serve as a proxy for the underlying drivers of traffic. As Ittner and Larcker (2000) point out in their recent survey of managerial accounting research:

“One particularly difficult endogeneity problem arises when the researcher wants to assess whether some managerial accounting choice [**traffic, in our case**] is associated with improved performance. As discussed in Demsetz and Lehn (1985), if all organizations in the sample are optimizing with regard to the accounting system choice [**traffic, in our case**], there should no be association between organizational performance and the observed (endogenous) choice, once the exogenous determinants of the choice are controlled in the structural model.”
(bold type added)

To address this endogeneity problem, we adopt a two-stage approach. In the first stage, we model traffic as a linear function of five exogenous, but not necessarily mutually exclusive, determinants: (i) an alliance with AOL (AOL), (ii) the presence of an affiliate program (AFF),

(iii) the extent of media visibility that the firm attracts (VIS), (iv) the extent of marketing expenditure incurred (M&A), and (v) the availability of cash balances (CASH).

$$\text{UNIVIS} = f(\text{AOL}, \text{AFF}, \text{VIS}, \text{M&A}, \text{CASH}) \quad (3)$$

where f is a linear function operator. The determinants of traffic are discussed in greater detail in section 3.3.1. In the second stage, we assess whether UNIVIS ceases to be value-relevant once the exogenous determinants of traffic are introduced into the value-relevance model. Hence, we test whether β_4 in equation (4) below is statistically indistinguishable from zero:

$$\begin{aligned} \text{MVE}_{jt} = & \beta_0 + \beta_1 \text{BVE}_{jt} + \beta_2 \text{E}_{jt} + \beta_3 \text{Log(TA)}_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_5 \text{AOL}_{jt} + \beta_6 \text{AFF}_{jt} \\ & + \beta_7 \text{VIS}_{jt} + \beta_8 \text{M&A}_{jt} + \beta_9 \text{CASH}_{jt} + \beta_{10i} \text{IND}_{jxit} + \beta_{11k} \text{QTR}_{jkt} + \varepsilon_{jt} \end{aligned} \quad (4)$$

3.3.1 Modeling the Determinants of Traffic

In this section we discuss in detail the various determinants of web traffic.

Alliance with AOL (AOL): AOL, or America Online, is the world's largest Internet Service Provider (ISP). A significant amount of traffic is channeled to the Internet through its ISP service. AOL's user base not only includes the paying subscribers of AOL's ISP service (23.2 million as of June 30, 2000) but also users of AOL's other portals and services such as MapQuest.com, AOL Moviefone, Netcenter (more than 28 million registered users), ICQ.COM (with more than 20 million active registered users) and Digital City. One way for firms to promote themselves online is, therefore, to enter into an advertising alliance with AOL to maximize their website's exposure. Such alliances usually feature premier placement on AOL's welcome page or sponsorship of particular online areas or web pages for designated time periods. The alliances may also involve placing the firm's graphic links on the search results page in the AOL directory and the category pages in the AOL directory. Access to AOL's subscribers is so important to firms that they advertise the AOL keyword by which their sites can

be found on AOL's network. As compensation for such advertisement, AOL receives cash payments and the opportunity for revenue sharing on sales made through its network. AOL may also take a portion of its compensation in the form of equity in the advertising firm. However, not all firms may choose to enter into an alliance with AOL because such an alliance is potentially expensive.

To identify firms that had a co-marketing alliance with AOL, we scan the press releases made by both AOL and our sample firms since 1997. If an Internet firm has an advertising alliance with AOL, we coded a variable AOL as one for every quarter during which the alliance is active. Otherwise, we set the variable AOL to zero.

Affiliate Programs (AFF): An affiliate program is a referral service from other websites on the Internet to the firm's website. When traffic is channeled from an associate web site to a firm's web site, the associate site earns referral fees for sales generated at the firm's site (Kotha 1998). Setting up affiliate programs is an efficient way to expand a firm's presence on the web and create a community of retailers working for the firm. Commenting on Amazon.com's affiliate program, the *Economist* (1997, p.10) points out

Amazon.com knows that it will probably never be the best site for rock climbing information or quantum physics discussions, but that the sites specializing in such subjects would be great places to buy books. A link to Amazon is easy, and potentially lucrative, way for such specialist sites to do that at one remove: a click on the link takes a viewer to Amazon's relevant page.

Thus, setting up an affiliate program leverages the capabilities of the Internet without incurring any additional overhead, unlike physical stores that require a large outlay of financial capital. The goal of the affiliate program is to gain greater name and brand recognition on the Internet where over 1.6 million stores currently operate (Hoffman and Novak, 2000). In such a noisy and fragmented environment, capturing the consumer's attention is critical to attracting

traffic. Furthermore, given that analysts estimate that less than 2% or 3% of the people who see an advertisement on the web actually click through to see more, promoting the firm's presence via affiliate programs greatly increases the probability of drawing traffic (Kotha 1998).

Information about a firm's affiliate programs is collected by scanning the firm's press releases. If a firm announced an affiliate program, we code the variable AFF as 1 for every quarter after the program initiation date. Otherwise, the variable AFF was set to zero.

Media visibility (VIS): The amount of attention the media dedicates to an Internet firm may be critical to generating customer traffic to the firm's website. In the off-line world, consumer traffic depends on geographical location. However, web consumers move easily and instantaneously across the Internet, guided primarily by their awareness of firms' websites, not geographical proximity. Hence, increasing awareness through greater media exposure improves the probability of attracting new visitors to a firm's website. The greater the number of articles written about a firm, the more information online visitors have to draw on in forming impressions about a firm. Because media exposure is generally beyond the direct control of the firm, the information provided by the media also tends to have higher source credibility than a firm's own marketing efforts (Wartick, 1992). Thus, the amount of media exposure is likely to increase the extent of consumer interest in a firm's site.

We measure media visibility (VIS) as the total number of articles published about the Internet firm in the "Major Newspapers" database of the *Lexis/Nexis* electronic database for quarterly periods for each firm. We select this database because it includes daily newspapers that reflect the focus of the current media and general public attention.

Marketing expenditures (M&A): Marketing and advertising expenditures could generate traffic to a firm's website by creating awareness of and acceptance for its products or services.

Marketing expenditure also enables a firm to differentiate itself from its competition (Porter, 1980). We use quarterly marketing and advertising expenditures reported in firms' 10-Qs as the measure (M&A) in our empirical analysis. Because firms do not fully disclose marketing expenditures on specific strategies such as raising media visibility or entering into an AOL alliance, we expect to see correlation between M&A and other determinants of traffic creation.

Cash constraints (CASH): The above discussion suggests that firms can increase traffic by adopting several strategies. However, financial constraints may prevent firms from devoting infinite resources just to chase web traffic. We proxy for such financial constraints by the cash holdings (CASH), measured as short term investments and cash equivalents, at the end of the quarter reported in the firm's 10-Q. The greater the CASH, the larger the traffic levels that the firm can achieve. Alternatively, the level of cash availability might constrain firms from attaining traffic levels higher than the one actually achieved by the firm.

Using the above hypothesized determinants we model UNIVIS as follows:

$$\text{UNIVIS}_{jt} = \delta_0 + \delta_1 \text{AOL}_{jt} + \delta_2 \text{AFF}_{jt} + \delta_3 \text{VIS}_{jt} + \delta_4 \text{M\&A}_{jt} + \delta_5 \text{CASH}_{jt} + \delta_6 \text{Log(TA)}_{jt} + \delta_7 \text{i IND}_{jxit} + \delta_8 \text{k QTR}_{jkt} + \eta_{jt} \quad (5)$$

As before, Log(TA) is added to serve as a scale control, and we control for heteroskedasticity using White's (1980) correction. The industry and quarter dummies are introduced to account for uncontrolled omitted variables that vary with industry membership and time.

Panel C of Table 1 provides descriptive statistics about the exogenous factors that determine traffic. There is significant dispersion in the extent of media visibility that firms are able to muster. The average firm in the sample is mentioned in the major newspapers 42.21 times in a quarter, whereas the inter-quartile gap ranges from 2 to 25 mentions per quarter. The average firm spends \$15.54 million a quarter on marketing and advertising – a substantial sum when compared to negative earnings of \$19.48 million for the average firm. The average firm

has \$147.57 million in cash relative to \$445.22 million in total assets. The relatively high cash levels probably reflect proceeds from initial public offerings awaiting deployment into operating or investing activities. We also note (not tabled) that 24% of the firm quarters have AOL alliances while 39% have affiliate programs.

For descriptive purposes, we report the Pearson and Spearman correlations between UNIVIS and the hypothesized exogenous determinants of UNIVIS. Panel A of Table 3 shows that UNIVIS is significantly correlated (under both Spearman and Pearson correlations) with all the hypothesized determinants of traffic, with the exception of the affiliate programs variable. For the affiliate programs variable (AFF) only the Spearman correlation is significantly positive.

The multivariate regressions reported in panel B show that AFF and VIS are positive, as hypothesized, and statistically significant at conventional levels. In particular, the effect of media visibility on UNIVIS is significantly positive (coefficient 0.04, t-statistic = 8.87). The coefficients on AOL and CASH are not statistically significant in the multivariate model. One interpretation of this result is that the strategies used to create traffic are not mutually exclusive. Consistent with this interpretation, the correlation between select determinants of traffic is quite high. For example, the Pearson correlation between M&A and AOL is 0.32 while the Pearson correlation between M&A and CASH is 0.54. Contrary to expectations, the coefficient on M&A is negative and significant (coefficient = -0.03, t-statistic = -3.37). However, the univariate correlation between M&A and UNIVIS reported in panel A of Table 3 is positive and significant, as expected. The adjusted R-square of the multivariate model is 78%, suggesting that the hypothesized exogenous determinants explain a substantial portion of the cross-sectional variation in UNIVIS. To confirm further that endogeneity of traffic is a potential problem in

making inferences about the value-relevance of traffic, we conduct the Hausman test and find that the Hausman t-statistic is 19.43 and significant at $p < 0.01$.

3.3.2 Value-relevance of traffic after controlling for the determinants of web traffic

Next, we assess whether the value-relevance of traffic disappears when the exogenous determinants of traffic are introduced in the value-relevance model. In particular, if UNIVIS is merely a proxy for the exogenous determinants that create traffic, we would expect β_4 , the coefficient on UNIVIS in regression equation (4), to be statistically indistinguishable from zero.

Table 4 reports the results of estimating equation (4). As expected, β_4 is not statistically significant (coefficient = -7.40, t-statistic = -0.17). Of the exogenous determinants of traffic, media visibility (coefficient = 35.43, t-statistic = 9.84) and cash balances (coefficient = 2.20, t-statistic = 2.37) are strongly associated with market value. In sum, the value-irrelevance of traffic in the presence of the exogenous determinants of traffic suggests that traffic is a summary measure of the strategies used by a firm to generate traffic.⁶

4. Economic motivations for the value-relevance of web traffic

4.1 Web traffic as a leading indicator of future revenues

In this section we examine the economic reasons that web traffic is value-relevant. Firms attract traffic to their websites primarily to convert web surfers to customers. Even if surfers do not purchase goods and services during one visit, the firm can build relationships with the website visitors that can be converted to future sales. A good example of such relationship building is “the Eyes” program offered by Amazon.com (Kotha, 1998). The program is a personal notification service in which customers can register their interests in a particular topic

⁶ This inference is insensitive to two robustness checks. First, we consider the possibility that higher market values attract more traffic. This induces simultaneity in the market value specification. Therefore, we estimate equations (5) and (1) simultaneously after including the market value of equity as an additional variable in equation (1). Second, we estimate (4) and (5) using a Generalized Least Squares (GLS) specification.

or author on the website. Once customers register, they are notified by e-mail each time a book by their favorite author, topic, or interest is published (Kotha, 1998). Such notifications are likely to result in future sales for Amazon.com.

The above discussion suggests that traffic should be associated with future revenues of the firm. To assess whether this relation holds in the data, we conduct the following regression:

$$SALES_{jt+n} = \gamma_0 + \gamma_1 UNIVIS_{jt} + \gamma_2 Log(TA)_{jt} + \gamma_{3i} IND_{jit} + \gamma_{4k} QTR_{jkt} + \varphi_{jt} \quad (6)$$

where n=1, 2 and SALES = sales revenues. All other variables are as defined previously.

The dependent variables are one-quarter and two-quarter ahead sales.⁷ As before, Log(TA) serves as a scale control and industry and quarter dummies are introduced to account for industry or time related correlation in the error terms. The results of estimating (6) are presented in panel A of Table 5.

It is interesting to note that UNIVIS is strongly related to SALES for up to two quarters ahead. After controlling for size, one additional visitor when compared to the cross-sectional mean appears to be associated with \$2.78 in sales one quarter ahead and \$2.91 in sales two quarters ahead. The adjusted R-squares are 42.71% for the two-quarter ahead model and 43.04% for the one-quarter ahead model. These results are consistent with Trueman et al.'s (2000b) findings that traffic explains cross-sectional variation in future sales. However, note that Trueman et al. (2000b) do not control for time-series trends in revenue. To control explicitly for such time-series correlation in sales we introduce past sales into equation (6) (see equation (7) below).⁸ That is, we examine whether traffic is incrementally informative in predicting future sales once we control for past sales.

⁷ We are unable to examine the effect of traffic on longer time periods because we run out of observations to conduct a meaningful statistical analysis.

⁸ A more appropriate control variable would be the sales of the same quarter one year ago so that seasonal trends are controlled for. However, we do not have enough time series observations to control for seasonality in sales.

$$SALES_{jt+n} = \gamma_0 + \gamma_1 UNIVIS_{jt} + \gamma_2 Log(TA)_{jt} + \gamma_3 IND_{jkt} + \gamma_4 QTR_{jkt} + \gamma_5 SALES_{jt} + \phi_{jt} \quad (7)$$

where n=1,2.

Results of estimating equation (7) are presented in panel B of Table 5. Results indicate that traffic is not incrementally informative about future sales one or two quarters ahead, once past sales are controlled for in the model. The explanatory power (R^2) of model (7) is significantly higher than that of model (6), ranging between 94% and 95%. Thus, past sales completely swamps the information content of web traffic for future sales.

The dominance of past sales over web traffic in predicting future sales has potentially interesting implications for the value-relevance of sales and traffic. If web traffic is value-relevant because it merely captures information in future sales, then web traffic's value-relevance should disappear once current sales is controlled for in the model. To assess the value-relevance of traffic in the presence of sales in the model, we estimate equation (8):

$$\begin{aligned} MVE_{jt} = & \beta_0 + \beta_1 BVE_{jt} + \beta_2 E_{jt} + \beta_3 Log(TA)_{jt} + \beta_4 UNIVIS_{jt} + \beta_5 SALES_{jt} \\ & + \beta_6 IND_{jkt} + \beta_7 QTR_{jkt} + \varepsilon_{jt} \end{aligned} \quad (8)$$

Results of the above regression reported in Table 6 show that both SALES and UNIVIS are both positive and statistically significant. This indicates that traffic contains value relevant information above and beyond sales revenues, a result that is open to at least three alternative interpretations. First, it is plausible that traffic captures valuable information about sales beyond two quarters. Because we do not have access to a long time series of future sales observations, we cannot fully rule out this explanation. Second, market participants may value web traffic for strategic uses of information that a firm obtains from traffic to its websites, to develop sustainable competitive advantages. A web firm can learn valuable insights about customer behavior by tracking web visitors' click stream patterns. Tracking visitor behavior on a firm's website can provide important knowledge about the nature and needs of visitors even if such

visitors fail to buy goods and services. For example, Amazon.com continually analyzes search lists to identify products that visitors cannot find in its online stores (*Economist Intelligence Unit*, 2000, p.35). Such data on failed searches enable Amazon to learn about visitors' preferences and decide accordingly which new product lines to enter. Thus, traffic may provide information about firms' growth options. An attendant benefit to accumulating data on failed searches is that such data helps e-tailing firms and their supply chain partners to maintain optimal inventory both in terms of product mix and quantity stocked. A firm's ability to access and harness knowledge about customer behavior is likely to increase with the extent of traffic. This is because the statistical reliability of data about consumer preferences is likely to increase with the extent of visitor traffic to the website. Yet another explanation for the value-relevance of traffic, over and above sales revenues, is that the market views the firm's traffic as a measure of the network effects that users can generate. We explore this third explanation in greater detail below.

4.2. Does traffic capture network effects?

Network effects arise when the value of connecting to a network depends on the number of other people already connected to it (Shapiro and Varian, 1999; p.174). Once the number of visitors, and the size of the virtual community created by the firm grows, more and more users find the firm's website attractive because of their ability to interact with other members of the community and their ability to share and contribute to member generated content (e.g., book reviews generated by readers at Amazon.com). For instance, the ability to interact with more community members can be very valuable to an auction site like Ebay. Ebay's auction site is more popular than any other auction site (including free auction sites such as Yahoo Auctions) because of the huge virtual community that Ebay has created. A marginal buyer or seller has

strong incentives to transact on Ebay because this increases the probability of finding members who would take the other side of the trade.

A bigger member base creates opportunities for advertisers and vendors to market a range of products and services to those members. Accumulating data about member profiles and transaction profiles makes it possible to attract even more vendors and advertisers to tailor the products and services to the members, thus making it attractive for members to join the firm's virtual community (Hagel and Armstrong, 1997; Kotha, 1998). This increases the potential for revenue streams from advertisements and subscription-based revenues for content and community companies and portals. For E-tailers and financial services firms, the size of the virtual community increases the potential for selling goods and services to a wider audience with the added advantage of customers selling to one another via product reviews. Thus, chasing traffic in the earlier time periods may be value maximizing down the road because the leaders with larger traffic (or user base) can dominate their product space as positive feedback effects take hold (Shapiro and Varian, 1999; Hagel and Armstrong, 1997).

Network effects can be empirically detected by evaluating whether the value of the network increases non-linearly with the number of users in the network. Part of the motivation behind this empirical test is Metcalfe's law, named after Bob Metcalfe, the inventor of the Ethernet. According to Metcalfe's law, if there are n people in the network, the value of the network is proportional to the number of other users, i.e., $n \times (n-1) = n^2 - n$ (Shapiro and Varian, 1999, p.184).⁹

⁹ The intuition behind Metcalfe's law in his own words is as follows: "When you connect computers together, the cost of doing so is n , but the value is n^2 , because each of the machines that you hook up gets to talk to all of the other machines on the network. When you graph that, you see that over time your costs go down while the value of the network goes up." (Red Herring Magazine, Nov 1994)

To assess whether traffic is valued by the market as a barometer of the firm's ability to generate network effects, we introduce the squared transformation of UNIVIS variable (UNIVIS^2) in the value-relevance model (equation 1). If network effects drive the pricing of traffic by the stock market, we would expect the coefficient on UNIVIS^2 , to be positive and statistically significant in the following regression model:

$$\begin{aligned} \text{MVE}_{jt} = & \beta_0 + \beta_1 \text{BVE}_{jt} + \beta_2 E_{jt} + \beta_3 \text{Log(TA)}_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_5 \text{UNIVIS}_{jt}^2 \\ & + \beta_6 \text{IND}_{jt} + \beta_7 \text{k QTR}_{jkt} + \varepsilon_{jt} \end{aligned} \quad (9)$$

Results reported in Table 7 show that β_4 is negative but statistically insignificant (coefficient = -100.30, t-statistic = -0.77). More important, β_5 , the coefficient on UNIVIS^2 is positive and strongly significant (coefficient = 30.24, t-statistic = 11.23). Thus, this result is consistent with market participants valuing traffic because of potential network effects generated by the firm.

Acquisitions

Consistent with the notion of acquiring traffic in an attempt to generate network effects, some acquirers pay significant sums for acquiring website traffic even if the target has no revenues. As a case in point, in October 1999, Excite@home, an Internet portal, agreed to acquire Bluemountainarts.com, an electronic greeting card website, for \$780 million in cash and stock. Although Bluemountainarts.com had virtually no revenues and profits, acquisition of Blue Mountain Arts instantly added 9.2 million more monthly visitors to Excite's network and increased Excite's reach of the Internet population from 24% to 34% (Excite.com press release, October 25, 1999).

To provide corroborative evidence on whether traffic measures a firm's ability to create network effects, we examine the association between traffic and an alternative value indicator, i.e., acquisition prices in web mergers and acquisitions. We obtain a sample of 89 acquisitions of

Internet companies for the period January 1999 to June 2000 for which traffic information was also available. Because many of the acquired firms are privately held we are unable to obtain financial data for them. We conduct two sets of regressions to ascertain the relation between acquisition prices and the number of visitors acquired. In the first regression, unique visitors measured as of the month prior to the acquisition is related to traffic as a linear explanatory variable. In the second regression, a squared transformation of unique visitors is also introduced as a regressor. If the market for corporate control also views traffic as a measure of network effects, we would expect the acquisition prices to increase non-linearly with the number of unique visitors acquired.

Descriptive statistics and the regression results are provided in Table 8. We find that acquiring firms pay a mean (median) price of \$494.15 (\$104.13) per unique monthly visitor. Panel B shows that target firms' traffic numbers account for virtually all the cross-sectional variation (more than 95%) in acquisition prices, consistent with the hypothesis that the market for corporate control values future growth potential from network effects. The strong explanatory power of the regression suggests that firm-level financial information (such as book value or earnings) that is not available to us may not capture value relevant information not accounted for by traffic. Furthermore, the acquisition prices increase non-linearly with traffic as evidenced by the positive and statistically significant coefficient on the squared transformation of the number of visitors acquired (coefficient = 3.16, t-statistic=3.13). Thus, evidence from acquisitions is also consistent with the conjecture that traffic captures potential benefits from network effects.

5. Concluding remarks

In this study we explore the role of a key non-financial measure, web traffic, in explaining the cross-sectional variation of B2C Internet firms. We find that unique monthly visitors to a firm's website explain a substantial portion of the cross-sectional variation in equity values of Internet firms. This result obtains after controlling for traditional financial measures such as earnings and equity book values. However, the value-relevance of traffic disappears once we control for the strategies used by firms to increase traffic. Hence, we interpret web traffic as a summary measure of the extent of the firm's involvement in various strategies designed to improve traffic to its website. This result also suggests that firms are in equilibrium when managers set the traffic levels of their firms in response to the exogenous determinants the face.

We find that traffic has no incremental predictive power for future sales after controlling for past sales. This suggests that traffic may be value-relevant for reasons other than providing information about future sales. We conjecture that the capital market participants value potential network effects and potential customer relationships that traffic brings even though such traffic does not necessarily result in current sales.

Consistent with Metcalfe's law that the value of the network increases with the square of the number of users in the network, we find that market value of our sample firms is a positive function of the squared number of unique monthly visitors. Furthermore, the acquisition prices paid by acquirers also increases at an increasing rate with the number of unique visitors acquired. Our tests are, of course, constrained by the availability of only 5 quarters of time series observations. More direct tests of the network effects hypothesis can be conducted when enough time-series observations of Internet firm financial performance become available.

Explaining the cross-sectional variation in the market valuations of Internet firms represents an interesting and continuing challenge. Researchers (e.g., Trueman et al., 2000a,b; Hand, 2000b; Demers and Lev, 2000 and Rajgopal, Venkatachalam and Kotha, 2000) have, thus far, restricted the search of non-financial value drivers to B2C firms because clear quantifiable measures of consumer interest such as traffic or online customer experiences are available for such firms. Exploring the non-financial measures that would explain stock prices of Internet firms in other sectors such as infrastructure services or B2B (Business to Business) commerce is a potentially intriguing but challenging avenue for future research.

References

- Arthur, B. 1996. Increasing returns and the new world of business. *Harvard Business Review* 74, no.4 (July-August 1996): 100-109.
- Amir, E. and B. Lev. 1996. Value-relevance of nonfinancial information: The wireless communications industry. *Journal of Accounting and Economics* 22, 3-30.
- Barth, M.E. and S.Kallapur. 1996. The effects of cross-sectional scale differences on regression results in empirical accounting research. *Contemporary Accounting Research* 13, 527-567.
- Christensen, C.M. and M. Overdorf. 2000. Meeting the challenge of disruptive change. *Harvard Business Review*, March-April, 67-76.
- Demers, E. and B. Lev. 2000. A rude awakening: Internet value drivers in 2000, Working paper, University of Rochester and New York University.
- Evans, P.; Wurster, T. 1999. Getting real about virtual commerce, *Harvard Business Review*, 77: 84-94.
- The Economist. 1997. A survey of electronic commerce. May 10, 1-18.
- The Economist Intelligence Unit. 2000. *E-business transformation*.
- Gates, W. 2000. *Business @ the Speed of Thought: Succeeding in the Digital Economy*. Warner Books.
- Grove, A. 1996. *Only the Paranoid Survive*. New York: Doubleday.
- Hagel, J. and A.G. Armstrong. 1997. *Net Gain: Expanding Markets through Virtual Communities*. Harvard Business School Press, MA.
- Hand, J. 2000a. Profits, losses and the non-linear pricing of Internet stocks. Working paper, University of North Carolina, Chapel Hill, NC.
- Hand, J. 2000b. The role of accounting fundamentals, web traffic, and supply and demand in the pricing of U.S. Internet stocks. Working paper, University of North Carolina, Chapel Hill, NC.
- Hoffman, D.L. and T.P. Novak. How to acquire customers on the web? , *Harvard Business Review*, May-June 2000.
- Houston Chronicle. 1999. Internet stocks are a bird of a totally different color / Worth can't be measured by conventional means (November 22).

- Ittner, C. and D. Larcker. 1998. Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction. *Journal of Accounting Research* 36 (Supplement) 1-36.
- Ittner, C. and D. Larcker. 2000. Assessing empirical research in managerial accounting: A value-based management perspective. *Journal of Accounting and Economics* (Forthcoming).
- Kotha, S. 1998. Competing on the Internet: How Amazon.com is Rewriting the Rules of Competition. *Advances in Strategic Management*, 15, pp 239-265.
- Lambert, R. 1998. Customer Satisfaction and Future Financial Performance. Discussion of Are NonFinancial Measures Leading Indicators of Financial Performance? An Analysis of Customer Satisfaction. *Journal of Accounting Research* 36 (Supplement) 37-46.
- Modahl, M. 2000. *Now or never: How Companies Must Change Today to Win the Battle for Internet Consumers*. NY: Harper Business.
- Nagar, V. 2000. An economic model of the information content of non-financial measures: Evidence from the retail banking industry. Working paper, University of Michigan, Ann Arbor, MI.
- Porter, M. E. 1980. *Competitive Strategy*. New York, Free Press.
- Red Herring. 1994. Networking with Bob Metcalfe and Craig Benson (November Issue).
- Rajgopal, S., M.Venkatachalam and S.Kotha. 2000. Does online customer experience affect the performance of e-commerce firms? Working paper, University of Washington and Stanford University.
- Shapiro, C. & H. Varian. 1999. *Information Rules*. Boston, MA: Harvard Business School Press.
- Trueman, B. F.Wong and X. Zhang. 2000a. The eyeballs have it: Searching for the value in Internet stocks. Working paper. University of California, Berkeley. CA.
- Trueman, B., M.H. Wong, and X-J. Zhang, 2000b, Back to basics: Forecasting the revenues of Internet firms, Working paper, University of California, Berkeley.
- Wartick, S. L. 1992. The relationship between intense media exposure and change in corporate reputation. *Business & Society*, 31: 33-49.
- White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817-838.

Table 1
Descriptive Statistics

Panel A: Sample of firms

	Name		Name		Name		Name
1	About Com Inc	25	E-Stamp Corp	49	Infoseek	73	Priceline Com Inc
2	Alloy Online	26	Earthweb Inc	50	Infospace Inc	74	Quepasa Com Inc
3	Amazon Com Inc	27	Ebay Inc	51	Insweb Corp	75	Quokka Sports Inc
4	Ameritrade Holding	28	Edgar Online Inc	52	Internet.Com Corp	76	Smarterkids Com Inc
5	Artistdirect Inc	29	Egghead Com	53	Iturf Inc	77	Snowball Com Inc
6	Ashford Com Inc	30	Emusic.Com Inc	54	Ivillage Inc	78	Sportsline Com Inc
7	Ask Jeeves Inc	31	Etoys Inc	55	Knot Inc	79	Starmedia Network
8	Audible Inc	32	Excite	56	Launch Media Inc	80	Student Advantage(Ipo)
9	Audiohighway.Com	33	Expedia Inc	57	Looksmart Ltd	81	Switchboard Inc
10	Barnesandnoble Com Inc	34	Fashionmall Com Inc	58	Lycos Inc	82	Talk City Inc
11	Beyond Com Corp	35	Fatbrain Com Inc	59	Mapquest	83	Theglobe Com Inc
12	Bigstar Entmt Inc	36	FTD Com Inc	60	Marketwatch.Com	84	Thestreet.Com Inc
13	Bluefly Inc	37	Garden Com Inc	61	Mortgage Com Inc	85	Ticketmaster Online Cty
14	Buy Com Inc	38	Geocities	62	MP3 Com Inc	86	Ubid
15	Careerbuilder Inc	39	GO2NET Inc	63	Musicmaker Com Inc	87	Value America Inc
16	Cdnow / N2k Inc	40	GoTo Com Inc	64	NBC Internet Inc	88	Vitaminshoppe Com Inc
17	Cmgi Inc	41	Healthcentral Com	65	Netbank	89	Witcapital
18	CNET Networks Inc	42	Healthgate Data Corp	66	Netradio Corp	90	Women Com Networks
19	Crosswalk Com Inc	43	Homegrocer Com Inc	67	Nextcard Inc	91	Xoom.Com
20	Cyberian Outpost Inc	44	Homestore Com Inc	68	Onsale	92	Yahoo Inc
21	Drkoop Com Inc	45	Hoovers Inc	69	Peapod Inc		
22	Drugstore Com Inc	46	Ilife Com Inc	70	Pets Com Inc		
23	E Trade Group Inc	47	Improvenet Inc	71	Planetrx Com Inc		
24	E-Loan Inc	48	Infonautics Corp	72	Preview Travel		

Table 1 (continued)

Panel B: Sample firms by industry type

Type	Industry	No. of firms
1.	Content providers	36
2.	E-tailers	38
3.	Financial Services	8
4.	Portals	10
	Total	92

Panel C: Descriptive statistics

Variable	Mean	Std.dev.	Median	1 st quartile	3 rd quartile	(N=303)
UNIVIS (million)	4.10	6.82	1.65	0.50	4.72	
REACH	0.07	0.12	0.03	0.01	0.08	
E (\$ million)	-19.48	67.04	-10.68	-21.93	-5.37	
BVE (\$ million)	206.14	343.84	93.90	40.84	225.16	
TA (\$ million)	445.22	1188.32	133.53	58.54	352.57	
SALES (\$ million)	34.54	75.10	10.00	3.95	30.12	
MVE (\$ million)	2906.56	8440.27	455.81	159.75	1715.26	
MVBV	16.92	62.46	4.70	2.42	10.24	
VIS	42.21	110.26	10.00	2.00	25.00	
M&A	15.54	22.75	8.79	4.30	18.19	
CASH	147.57	260.89	67.55	30.22	143.04	

Notes:

Variables are defined as follows: UNIVIS = the average monthly unique visitors during a quarter, REACH = the average proportion of unique visitors to total web population during a quarter, E = income before extraordinary items, BVE = book value of equity, TA = total assets, SALES = sales revenues, MVE = market value of equity, MVBV = market to book ratio, VIS = media visibility measured as the number of articles in leading newspapers and magazines, M&A = marketing and advertisement expenditures, CASH = cash and cash equivalents.

Table 2

Summary statistics for the regression of market values and returns on financial measures and web traffic

Panel A: Levels specification

$$MVE_{jt} = \beta_0 + \beta_1 BVE_{jt} + \beta_2 E_{jt} + \beta_3 \text{Log}(TA)_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_{5i} \text{IND}_{jxit} + \beta_{6k} \text{QTR}_{jkt} + \varepsilon_{jt} \quad (1)$$

Variable	Pred. Sign	(N=296)		(N=295)	
		Coeff. Estimate	t-stat	Coeff. Estimate	t-stat
Intercept	?	-6463.46	-3.97*	849.19	0.53
BVE	+	4.49	2.53*	5.71	4.76*
E	+	-12.18	-1.91**	-10.28	-2.72*
Log(TA)	+	1012.92	3.16*	-347.04	-1.08
UNIVIS	+			700.49	4.54*
Adj. R²		36.61%		61.03%	

Panel B: Changes specification

$$\text{Ret}_{jt} = \beta_0 + \beta_1 E_{jt} + \beta_2 \Delta E_{jt} + \beta_3 \Delta \text{UNIVIS}_{jt} + \beta_{4i} \text{IND}_{jxit} + \beta_{5k} \text{QTR}_{jkt} + v_{jt} \quad (2)$$

Variable	Pred. Sign	(N=207)	
		Coeff. Estimate	t-stat
Intercept	?	-0.55	-6.26*
E	+	0.86	1.30***
ΔE	+	-0.19	-0.28
ΔUNIVIS	+	15.88	1.91**
Adj. R²		35.56%	

Notes:

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
2. Coefficients on quarter dummies and industry dummies have not been reported for expositional convenience.
3. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
4. Ret = abnormal stock return determined by subtracting return on NASDAQ index from the holding period return over the quarter, ΔE = change in income before extraordinary items, ΔUNIVIS = change in average monthly unique visitors during a quarter. Independent variables in equation (2) (except the industry and quarter dummies) are scaled by the market value of equity at the beginning of the return formation period. See the notes to Table 1 for other variable definitions.

Table 3
Correlation statistics and regression results of determinants of web traffic

Panel A: Pearson and Spearman correlation matrix

Variables	UNIVIS	AOL	AFF	VIS	M&A	CASH	Log(TA)
UNIVIS	1.00	0.18*	0.25*	0.51*	0.52*	0.53*	0.57*
AOL	0.10**	1.00	0.07	0.23*	0.23*	0.16*	0.23*
AFF	-0.00	0.07	1.00	0.12**	0.06	0.02	0.05
VIS	0.71*	0.20*	-0.00	1.00	0.58*	0.52*	0.62*
M&A	0.38*	0.32*	0.02	0.59*	1.00	0.54*	0.65*
CASH	0.45*	0.14**	-0.01	0.58*	0.54*	1.00	0.88*
Log(TA)	0.48*	0.24*	-0.00	0.46*	0.65*	0.64*	1.00

Panel B: Determinants of traffic

$$\text{UNIVIS}_{jt} = \delta_0 + \delta_1 \text{AOL}_{jt} + \delta_2 \text{AFF}_{jt} + \delta_3 \text{VIS}_{jt} + \delta_4 \text{M&A}_{jt} + \delta_5 \text{CASH}_{jt} \\ + \delta_6 \text{Log(TA)}_{jt} + \delta_7 \text{IND}_{jkt} + \delta_8 \text{QTR}_{jkt} + \eta_{jt} \quad (5)$$

(N=291)			
Variable	Pred. Sign	Coeff. Estimate	t-stat
Intercept	?	-4.75	-4.22*
AOL	+	-0.35	-0.98
AFF	+	0.74	2.54*
VIS	+	0.04	8.87*
M&A	+	-0.03	-3.37*
CASH	+	0.00	0.32
Log(TA)	+	0.97	5.32*
Adj. R ²		78.11%	

Notes:

1. In Panel A, Pearson correlation statistics are presented below the diagonal and Spearman correlation statistics are presented above the diagonal.
2. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
3. Coefficients on quarter dummies (QTR) and industry dummies (IND) have not been reported for expositional convenience.
4. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
5. See notes to Table 1 for variable definitions.

Table 4

Regression results of estimating the market value equation after accounting for factors that determine web traffic

$$MVE_{jt} = \beta_0 + \beta_1 BVE_{jt} + \beta_2 E_{jt} + \beta_3 \text{Log(TA)}_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_5 \text{AOL}_{jt} + \beta_6 \text{AFF}_{jt} + \beta_7 \text{VIS}_{jt} + \beta_8 \text{M\&A}_{jt} + \beta_9 \text{CASH}_{jt} + \beta_{10i} \text{IND}_{jxit} + \beta_{11k} \text{QTR}_{jkt} + \varepsilon_{jt} \quad (4)$$

Variable	Pred. Sign	Coeff. Estimate	t-stat
Intercept	?	-164.48	-4.87*
BVE	+	2.56	3.37*
E	+	-2.77	-1.47***
UNIVIS	?	-7.40	-0.17
AOL	+	244.31	0.75
AFF	+	-189.94	-0.96
VIS	+	35.43	9.84*
M&A	+	-4.22	-0.49
CASH	+	2.20	2.37*
Log(TA)	+	5.49	0.00
<hr/>		Adj. R ²	
		81.81%	

Notes:

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
2. Coefficients on quarter dummies (QTR) and industry dummies (IND) have not been reported for expositional convenience.
- 3 Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
4. See notes to Table 1 for variable definitions.

Table 5

Regression results examining the relation between web traffic and future revenues

Panel A: Regression of one-quarter and two-quarter ahead sales on traffic

$$SALES_{jt+n} = \gamma_0 + \gamma_1 UNIVIS_{jt} + \gamma_2 \text{Log(TA)}_{jt} + \gamma_3 i IND_{jkt} + \gamma_4 k QTR_{jkt} + \phi_{jt} \quad (6)$$

Variable	Pred. Sign	Dependent Variable			
		SALES _{jt+1} (N=207)		SALES _{jt+2} (N=128)	
		Coeff. Estimate	t-stat	Coeff. Estimate	t-stat
Intercept	?	-63.77	-2.88*	-77.34	-2.42**
UNIVIS	+	2.78	3.81*	2.91	2.77*
Log(TA)	+	18.59	4.14*	22.07	3.90*
Adj. R ²		43.04%		42.71%	

Panel B: Regression of one-quarter and two-quarter ahead sales on traffic after controlling for current sales information

$$SALES_{jt+n} = \gamma_0 + \gamma_1 UNIVIS_{jt} + \gamma_2 \text{Log(TA)}_{jt} + \gamma_3 i IND_{jkt} + \gamma_4 k QTR_{jkt} + \gamma_5 SALES_{jt} + \phi_{jt} \quad (7)$$

Variable	Pred. Sign	Dependent Variable			
		SALES _{jt+1} (N=205)		SALES _{jt+2} (N=127)	
		Coeff. Estimate	t-stat	Coeff. Estimate	t-stat
Intercept	?	-8.37	-1.15	-4.56	-0.37
UNIVIS	+	0.15	0.70	-0.28	0.81
Log(TA)	+	3.27	3.20*	2.00	1.34***
SALES	+	1.07	40.50*	1.60	36.95*
Adj. R ²		94.03%		94.91%	

Notes

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
2. Coefficients on quarter dummies (QTR) and industry dummies (IND) have not been reported for expositional convenience.
3. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
4. See notes to Table 1 for variable definitions.

Table 6

**Regression results examining the relation between web traffic and market value of equity
after controlling for sales**

$$\text{MVE}_{jt} = \beta_0 + \beta_1 \text{BVE}_{jt} + \beta_2 \text{E}_{jt} + \beta_3 \text{Log(TA)}_{jt} + \beta_4 \text{UNIVIS}_{jt} + \beta_5 \text{SALES}_{jt} + \beta_6 \text{IND}_{jt} + \beta_7 \text{QTR}_{jkt} + \varepsilon_{jt} \quad (8)$$

(N= 296)			
Variable	Pred. Sign	Coeff. Estimate	t-stat
Intercept	?	888.14	0.56
BVE	+	4.95	4.08*
E	+	-0.94	-0.28
Log(TA)	+	-559.59	-1.72***
UNIVIS	+	599.39	3.89*
SALES	+	25.79	4.07*
Adj. R ²		65.30%	

Notes:

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
2. Coefficients on quarter dummies (QTR) and industry dummies (IND) have not been reported for expositional convenience.
3. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
4. See notes to Table 1 for variable definitions.

Table 7

Summary statistics for the regression of market values on financial measures, web traffic measure (unique visitors) and the squared term of unique visitors

$$MVE_{jt} = \beta_0 + \beta_1 BVE_{jt} + \beta_2 E_{jt} + \beta_3 \text{Log(TA)}_{jt} + \beta_4 UNIVIS_{jt} + \beta_5 UNIVIS^2_{jt} + \beta_{6i} IND_{jxit} + \beta_{7k} QTR_{jkt} + \varepsilon_{jt} \quad (9)$$

(N= 295)			
Variable	Pred. Sign	Coeff. Estimate	t-stat
Intercept	?	-1863.85	-1.33
BVE	+	4.99	4.75*
E	+	-8.90	-2.17**
Log(TA)	+	294.11	1.52***
UNIVIS	?	-100.30	-0.77
UNIVIS ²	+	30.24	11.23*
Adj. R ²		81.61%	

Notes:

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction for heteroskedasticity.
2. Coefficients on quarter dummies (QTR) and industry dummies (IND) have not been reported for expositional convenience.
3. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
4. UNIVIS² refers to UNIVIS squared. See notes to Table 1 for other variable definitions.

Table 8
Results relating acquisition prices and web traffic

Panel A: Descriptive statistics

Variable	Mean	Std.dev.	Median	1 st quartile	3 rd quartile	(N=89)
ACQPR (\$ million)	365.75	1384.55	38.70	11.00	166.00	
UNIVIS _{acq} (millions)	1.88	4.92	0.39	0.17	1.11	
Price per unique monthly visitor	494.15	2183.78	104.13	38.12	236.84	

Panel B: Coefficient estimates from regressing acquisition prices on unique visitors

Variable	Pred. Sign	Coeff. Estimate	t-stat	Coeff. Estimate	t-stat	(N= 86)
Intercept	?	-47.69	-0.50	-42.52	-0.56	
UNIVIS _{acq}	?	260.91	20.79*	151.84	3.74*	
UNIVIS _{acq} ²	+			3.16	3.13*	
Adj. R ²		94.66%	96.58%			

Notes

1. ***, **, * represents significance at 10%, 5%, and 1% respectively. t-statistics are one-tailed where the sign is predicted, two-tailed otherwise. Reported t-statistics are adjusted for White's (1980) correction.
2. Regression results are presented after deleting outlier observations represented by the absolute value of R-student statistic greater than the three.
3. Variables are defined as follows: ACQPR = acquisition price, UNIVIS_{acq} = the monthly unique visitors of the acquired firm measured as of the month prior to the acquisition, UNIVIS_{acq}² = UNIVIS_{acq} squared.

**Attention, Acquisition Costs, or Insider Information? Using Big Data to Uncover the Sources of Local
Information Advantage**

Jung Koo Kang

Lorien Stice-Lawrence*

Forester Wong

University of Southern California

WOEFULLY PRELIMINARY AND RUEFULLY INCOMPLETE

PLEASE DO NOT CITE OR DISTRIBUTE

November 2018

Abstract

We use novel satellite data which tracks the number of cars in the parking lots of approximately 100,000 stores for 75 publicly listed U.S. retailers over the years 2009 to 2016 to disentangle the sources of local information advantage. We establish car counts as a timely measure of the performance of individual stores and show that local institutional investors appear to use this information to make profitable trades. Most importantly, we disentangle the extent to which this local information advantage is due to: increased attention (“attention channel”), information directly obtained from firm management (“insider information channel”), or decreased costs of information acquisition for local investors (“acquisition cost channel”). Our tests provide strong evidence in support of the acquisition cost channel, especially for geographically remote areas, weak evidence of the attention channel, and no evidence of the insider information channel. In an era when the common perception is that “the world is flat” and the costs to acquire information are decreasing, we find these results surprising and important to consider when making policy decisions about how or whether to close this gap.

*Corresponding author: Lorien Stice-Lawrence, Leventhal School of Accounting, University of Southern California, 3660 Trousdale Pkwy, ACC 308B, Los Angeles, CA 90089, sticelaw@usc.edu. We thank Orbital Insight for graciously providing the satellite data used in this paper.

I. Introduction

Investors have a consistent preference to invest in firms that are geographically close to them. In an international setting, this is referred to as “home bias” where investors hold a disproportionate number of assets originating in their home country (French and Poterba, 1991), while within the U.S. this has been referred to as “local bias,” where investors hold a disproportionate number of assets in firms located within their home state (Coval and Moskowitz, 1999). The question that naturally emerges is *why* investors exhibit this preference. While some papers have documented behavioral explanations driving some of these results such as familiarity (Huberman, 2001; Pool et al., 2012), patriotism (Morse and Shive, 2011), and favoritism (Cornaggia et al., 2018), a few others suggest that part of this behavior is driven by an information advantage on the part of local investors (Bernile et al., 2015; Coval and Moskowitz, 2001). However, there has been little to no evidence disentangling the potential sources of this local information advantage. Local investors could benefit from an information advantage because of differing attention and investment preferences, differential information sharing by management, or decreased barriers to information acquisition because of geographic proximity. Identifying which of these three channels are at play is critical because they each have very different policy implications for whether or how this gap should be addressed. We present evidence confirming the existence of a local information advantage for institutional investors and demonstrate that the largest driver of this advantage is greater costs of information acquisition for non-local investors, with attention also playing a smaller role.

In our study we define an investor as “local” to a given retail store location if the investor is located in the same state as that store. Because a retail firm may have stores in many states, this means that a given investor may be local to some of a retailer’s stores and non-local to others.

Conceptually, this captures the fact that operations of a firm are dispersed (Bernile et al., 2015) and investors in different locations have access to different sets of local information about a firm. Empirically, this definition allows us to hold an investor-firm relationship constant and examine how differences in location affect the extent to which an investor reacts to the information about individual stores. This design is key because we can use it to disentangle the sources of an investor's local information. Specifically, we test whether local information advantage is due to any of the following three channels: attention, insider information, or acquisition costs.

Local investors may gain an information advantage through the “attention channel” if being located near one store motivates the investor to pay more attention to the firm overall by expending more effort and costs to collect and analyze information about all of the firm’s locations. Alternatively, local investors may access information through the “insider information channel” if they have access to proprietary information about the overall firm via direct access to management. Lastly, geographic proximity to local stores may allow investors to observe or interpret information about local operations at a lower cost than non-local investors, which we refer to as the “acquisition cost channel.”

In this study we use satellite data on the number of cars in retail store parking lots as a timely measure of the performance of individual stores and validate it by showing that it predicts firm-level quarterly sales, income, and announcement returns. We confirm the existence of a local information advantage (Coval and Moskowitz, 2001; Bernile et al., 2015) by showing that the trades of local investors are more responsive to the performance of individual stores than the trades of non-local investors and that trades made using this local information are profitable. Most importantly, we are able to shed light on the channels through which firms obtain this information advantage. We find little evidence that the trades of investors respond more to overall firm

performance when the investor is near at least one store, ruling out attention to the firm as a whole as the main driver of local information advantage. Additionally, we find no evidence that investors located near the firm headquarters, where access to top-level management is especially likely to lead to insider information, have an advantage relative to investors local to other store locations, ruling out the insider information channel as the main driver of local information advantage. It appears that acquisition costs play the largest role in driving local information advantage, and this interpretation is further reinforced by the fact that the differential response of local and non-local investors is larger when stores are geographically remote.

We also conduct a variety of supporting analyses. For example, we show that being local has the largest effect on the trades of small investors. Further, although the satellite data used in this study became available to investors partway through our sample period, local investors still respond more to local information relative to non-local investors even at the end of our sample. This suggests that although acquisition costs for non-local investors may be decreasing, they are by no means negligible and remain a significant barrier to non-local investors.

Our results are important for several reasons. First, while a few papers have provided evidence consistent with an information advantage on the part of local investors (Baik et al., 2010; Bernile et al., 2015; Coval and Moskowitz, 2001), ours is the first to directly measure one type of local information and tie it to investor trades in order to rule out alternative explanations for this apparent advantage. Additionally, while the prevailing notion is that differences in information sets across local and non-local investors are disappearing (Bernile et al., 2018), we demonstrate that this difference still persists, even at the end of our sample period and even after our satellite data become available to investors who wish to purchase it. Our results indicate that this is due to differential information acquisition costs of local and non-local investors and not because of

different preferences or attention across investor groups or differential access to management. In other words, information frictions across geographic regions are still important drivers of information asymmetry even in this high-speed technological era. Further, our study adds to the emerging field which uses big data to examine fundamental questions (Frankel et al., 2016), as well as research examining the impact of big data on firms and investors (Zhu, 2018).

Most importantly, however, our paper is able to distinguish which of several different channels appear to contribute most to local information advantage. This finding is critical because the policy implications of these three channels are very different. If it is the case that local investors have better information about a firm because they choose to allocate more of their resources towards obtaining this information, then there is no information “problem” to fix. While the SEC has expressed a desire to level the playing field in terms of which information is provided to different investors (SEC, 2000), it is not within its purview to tell investors which firms they should consume information about. Alternatively, if top management appears to be sharing material information about the firm with a subset of investors, then this directly violates Reg FD which aims to provide equal access to firm-provided information to all investors. Evidence to support this channel would indicate that stricter enforcement of this regulation is necessary. If, however, local information advantage is driven by costs of information acquisition that have nothing to do with firm actions and rest entirely on investor location, then the regulatory problem is more nuanced.

While a key role of the SEC is ensuring all investors have equal access to basic firm information, the SEC has clarified that information differences among investors arising from “skill”, “diligence”, or “hard work” (SEC, 2000) are a normal and desirable aspect of securities markets. To the extent that differential acquisition costs of local and non-local investors enable

certain investors access to information that is not a result of their own effort, the SEC might believe there is a role for increased disclosure or greater restriction of what is currently viewed as non-material information on the part of firms in order to reduce this gap. This is especially likely given that this gap appears to disproportionately affect smaller investors and the SEC has previously instituted rules aimed at leveling the playing field between large and small investors (SEC, 2009). On the other hand, the SEC may view information asymmetry which arises as a function of geographic location (as opposed to intentional firm actions) as an inescapable part of the capital markets. The SEC may deem intervention unnecessary, especially to the extent that this disparity could only be eliminated through costly disclosure. The SEC may also rely on ongoing decreases in the overall costs of information acquisition and sharing made possible by advances in technology to further narrow this gap. Indeed, investors have already begun to address this information disparity by acquiring the very data that we use in this study, although our later tests indicate that the availability of the data itself does not close this gap. Thus while the policy implications of each channel for local information advantage may differ, we see our evidence as the necessary first step to begin the discussion.

Our study is subject to several important caveats. First, our interpretation of our findings may be incorrect. While it appears that the investors that we study are reacting to local information, there could be other factors that drive both local car count changes and changes in investor holdings. We believe our store-level design, which allows us to vary whether an investor is local to individual stores within a given investor-firm-quarter, mitigates this concern, but we cannot directly observe each investor's information set. Second, our results may have limited generalizability because we examine only institutional investors of retail firms, and we can only capture one aspect of local information using our car counts measure. It could be that the amount

and source of local information advantage are vastly different for non-retail firms and for different types of investors, or that other types of local information flow through different channels. We do find comfort that while our sample of firms is small, our set of investors represents a substantial portion (approximately 80%) of the 13F population. Additionally, it seems unlikely that other types of performance-related local information would flow through a substantially different mechanism for firms in different industries. While we acknowledge these caveats, we believe that our study is still able to provide important insights with respect to local information advantage.

II. Related Literature and Hypotheses

The literature on home bias and its domestic equivalent, local bias, has long puzzled over the propensity of investors to invest disproportionately in geographically proximate assets. While a variety of contributing factors have been proposed and investigated, this regularity has yet to be fully explained (Cooper et al., 2013). While many factors have been explored, including institutional barriers, trading costs, and behavioral biases, we focus on one prominent factor: information asymmetry between local and non-local investors.

A number of studies have documented evidence consistent with local investors having access to additional information (Bae et al., 2008; Baik et al., 2010; Bernile et al., 2015; Chi and Shanthikumar, 2017; Coval and Moskowitz, 2001); however, these studies have largely been agnostic as to the source of this information. In considering the various proposed mechanisms by which local investors could acquire private information, we categorize them into three main channels.

First, investors local to a firm may be motivated to direct more attention to that firm by expending more effort and costs to collect and analyze information about the firm overall. This

could be because of greater familiarity or salience of the firm to these local investors. As a result of this greater attention, local investors would be able to build an information advantage relative to non-local investors, but the additional information available to local investors via this “attention channel” would be available to any investor, local or non-local, who exerted the same effort and attention.

Second, investors local to a firm may have direct links to management who have access to proprietary information about the firm as a whole. This could be in the form of social ties arising from shared local social networks, or it could arise from greater direct access to management by local investors because their physical proximity allows them to interact with management more frequently than non-local investors. Management may share their proprietary information with these local investors thereby providing them with an information advantage relative to non-local investors without this direct access. We call this the “insider information channel.”

Lastly, investors who are local to a firm may face lower costs to acquire information about the firm from sources other than management (the “acquisition cost channel”). For example, their geographic proximity may allow them to directly observe business operations of the firm such as construction and shipping operations, or their knowledge of local economic conditions, laws, and institutions could help them predict and interpret information about local performance, or a broad network of non-management-level social ties could expose them to a “mosaic” of information about the general busyness of local operations, hiring activities, etc. Such information could potentially be obtained by a non-local investor, but only through more arduous means, for example careful examination of public reports, purchase of business intelligence, direct travel to the site, etc.

No studies in the prior literature have been able to fully distinguish among these three channels. Dyer (2018) documents that local institutional investors access filings on EDGAR more quickly, potentially indicating higher attention of local investors, but he also argues that his results could indicate that local investors had prior knowledge of the events leading to disclosure. Bernile et al. (2015 and 2018) show that local information advantage decreases around Reg FD, potentially indicating that there had previously been local insider information, but their simple pre/post tests cannot be directly tied to Reg FD (as opposed to other time-varying factors). Their results also do not tell us about the current sources of local information advantage prevalent today, almost two decades after Reg FD. Ellis et al. (2017) find that mutual funds invest more in firms when travel to firm headquarters becomes easier and these investments are more profitable, potentially indicating that acquisition costs contribute to local information advantage. However, their results may also stem from the ability of institutional investors to monitor the firm, which may directly contribute to performance, or they could capture the benefit of communicating directly with firm management via the insider information channel. Lastly, Engelberg and Parsons (2011) find that investors are more likely to trade a firm's stock if it was covered by a local newspaper, consistent either with increased attention or the ability of local newspapers to unearth and report value relevant information (Miller and Shanthikumar, 2010). Further, Van Nieuwerburgh and Veldkamp (2009) predict investors with an initial local information advantage will subsequently invest more resources and attention in collecting additional information about local investments, but empirical evidence cannot document this. In sum, the extent to which each of our three identified channels contributes to local information advantage remains an open question.

In order to disentangle the sources of local information advantage, we first must establish that a local information advantage exists. As described in detail in Section 3, we use the number

of cars in a retail store parking lot to capture that store's performance. To the extent that local investors actually have an information advantage about local stores relative to non-local investors, we would expect that they would be more likely to respond to this store-level information than their non-local counterparts, and trades based on this information would be profitable. Our first hypothesis is thus:

H1a: Changes in store-level car counts have a more positive relation with changes in the stock holdings of *local* investors than with changes in the stock holdings of *non-local* investors.

H1b: Changes in stock holdings associated with changes in local car counts have a positive relation with future stock returns.

We next examine the extent to which the attention channel drives local information advantage. Because investors may be local to some stores within a given firm and non-local to other stores, we can compare how investors incorporate information about local and non-local stores of the same firm into their trades. After controlling for changes in local car counts, this allows us to disentangle whether being local to at least one store increases the extent to which an investor incorporates information about *all* of the firm's stores into their trades, which would support the attention channel. Our attention hypothesis is:

H2: Changes in the stock holdings of investors who are local to at least one store will have a more positive relation with changes in the firm's overall car counts than the stock holdings of investors not local to any of the firm's stores.

Evidence supporting H2 would be consistent with attention to the overall firm as a channel which contributes to local information advantage. Further, if we found that local car counts no

longer had an incremental relation with changes in stock holdings, evidence supporting H2 would indicate attention was the *sole* driver of local information advantage. Because we define investors as being local or non-local at the store-level, and each firm in our sample has an average of 1,290 stores, it is very unlikely management of retail store locations would have material non-public information about stores located in other states. In other words, H2 is not a test of the insider information channel.

In order to distinguish the contribution of insider information to local information, we differentiate between investors which are inside the headquarter state versus those that are not. Because top level management has the greatest access to material, nonpublic information about overall firm performance and these individuals are generally located in the headquarter state, the insider information channel should be particularly pronounced for local investors in the headquarter state. Investors located in this state should be more responsive to overall firm performance than other local investors if they are privy to insider information from management. Stated formally:

H3: Changes in the stock holdings of investors located in a firm's headquarter state are more positively associated with changes in the firm's overall car counts than changes in the holdings of investors local to the firm's stores in other states.

While for the most part we identify the acquisition cost channel through the process of elimination, in Sections 4 and 5 we also discuss several cross-sectional tests which allow us to validate our interpretations of tests of the above hypotheses. First, however, we describe our data and our main results.

III. Empirical Setting and Variable Definitions

In this paper, we use data from five main sources to measure local information advantage from 2009 to 2016. First, we use accounting data from Compustat, price data from CRSP, and analyst forecast information from I/B/E/S. Next we use satellite data from Orbital Insight. Lastly, we use 13F data on institutional owner holdings from Thomson Reuters. We provide more detailed descriptions below.

3.1 Satellite Data

We use a unique dataset containing store-level car count data to disentangle the sources of local information advantage. The dataset is obtained from Orbital Insight,¹ who developed these car count data using satellite images from a wide array of providers including DigitalGlobe, Airbus, Landsat (a division of NASA) and Planet Labs.

Orbital Insight uses a proprietary algorithm to extract car count information from satellite images. First, the parking lots for stores of each retailer are carefully mapped and the algorithm extracts the number of cars in each parking lot from the images. Next, a number of additional steps are taken to ensure that the car count data are accurate and comparable. When a particular parking lot is shared between multiple retailers, the level of contamination is determined, and the car count measure is adjusted to reflect only those cars likely belonging to patrons of a particular store. Because different satellite images may be taken at different times of day and on different days of the week, Orbital Insight adjusts their car count data based on a host of additional information (e.g., surrounding traffic, the traffic at the same store historically, whether the particular day in question is a significant holiday, etc.) to make sure that these timing differences are correctly adjusted for. Finally, Orbital Insight carefully tracks changes to retailer parking lots and updates

¹ Orbital Insight was founded in 2013. In their latest round of funding, they raised \$20 million dollars with backing from notable investors including the CIA and Google Venture.

their parking lot mapping accordingly. At the time of writing, Orbital Insight's clients include more than 60 asset management firms, several U.S. government agencies, and two global non-profit organizations.

In Table 1 we provide descriptive information about the stores included in our satellite data from Orbital Insight. Our dataset spans from March 2009 to December 2016, although they were only available for purchase beginning in 2015. After merging with key accounting variables from Compustat² and 13F investor holdings from Thomson Reuters, our sample covers 96,742 unique stores distributed across all states in the United States, with the exception of Hawaii and the addition of Puerto Rico and Washington D.C. These 96,742 stores are owned by 75 unique publicly listed retailers and have 877,325 quarterly observations in our data (an average of 9 quarters per store). These 75 retailers range from very large to very small, with the largest having 10,706 stores in our data and the smallest having 51 stores. However, even the smallest retailers in our sample have a relatively wide geographic spread, with the least geographically dispersed retailer operating in 10 states, and the most dispersed retailer operating in 49 states and the District of Columbia. As seen in Panel D, the retailers in our sample are larger than other firms (other retailers) covered by the Compustat database with average assets of \$2.5 billion compared to average assets of \$1.3 billion (\$1.3 billion) for all other firms (retailers) available on Compustat over the same time period.³

² In order to be included in our sample, retailers must have the following quarterly variables available from Compustat: sales, net income, selling, general, and administrative expense, stockholders' equity, and total assets. Additionally, firms must have earnings forecasts and analyst following from I/B/E/S and stock price from CRSP.

³ Other retailers in Compustat are identified as all other firms within the same SIC codes as the firms in our sample. They fall within 2-digit SIC codes 25, 28, 51-59, and 75.

3.2 Proxy for local information

We use the car count information as an empirical proxy for the performance of individual stores and indirectly for the information that local investors are able to obtain. While we do not think it is likely that investors directly observe the number of cars in individual parking lots, we believe that car counts will be correlated with foot traffic in stores, which is positively correlated with retail sales (Perdikaki et al., 2012). As a result, we expect parking lot car counts to be associated with the performance of retail stores. As described in the hypothesis section, we predict a variety of mechanisms by which investors could obtain information about the performance of local stores.

In order to create a timely measure of *new* store-level information, we focus on changes in store-level car counts. We specifically use the change in car counts for that store compared to the previous quarter in order to capture the most timely information.⁴ In addition, we weight each store's change in car counts by the likely contribution of that store to overall firm performance by giving a higher weight to larger stores. In particular, our car count measure is calculated as:

$$Car_Chg_W_{r,s,q} = Weights_{r,s,q} * \left[\frac{Avg(CarCount_{r,s,q}) - Avg(CarCount_{r,s,q-1})}{Avg(CarCount_{r,s,q-1})} * 100 \right] \quad (1)$$

Where:

$$Weights_{r,s,q} = \frac{Car_Count_{r,s,q}}{\sum_{i=1}^{n_{rq}} Car_Count_{r,i,q}} \quad (2)$$

Where r refers to one of our 75 retailers, s identifies an individual store, and q is the fiscal quarter. $n_{r,q}$ is the number of stores associated with a retailer in a given quarter. Equation (1)

⁴ Results are similar when we instead use the change in car counts relative to the same quarter in the prior year.

compares the average car counts for an individual store in a given quarter to the average car counts of the same store in the previous quarter. An average is used because a store may have multiple observations in the satellite data within a given quarter.⁵ $Weights_{r,s,q}$ gives a greater weight to individual stores with larger car counts relative to other stores for the same retailer in the current quarter because we believe the largest store locations will have the greatest effect on overall firm performance.⁶ To simplify interpretation of our regression analyses, we normalize our car count measure $Car_Chg_W_{r,s,q}$ to $Car_Chg_N_{r,s,q}$ with mean zero and unit variance by doing the following transformation:

$$Car_Chg_N_{r,s,q} = \frac{Car_Chg_W_{r,s,q} - \text{Mean}(Car_Chg_W_{r,s,q})}{\text{Std}(Car_Chg_W_{r,s,q})} \quad (3)$$

In Table 2, we show that our proxy for local information can be used to predict the underlying retailer's performance. In particular, we find that Car_Chg_N is positively correlated with a battery of contemporaneous performance indicators including quarterly changes in sales (column 1), changes in selling, general, and administrative expense (column 2), changes in net income (column 3), earning surprises (column 4), and the earnings announcement return (column 5). The coefficient of 0.555 for our normalized car count measure in column 1 indicates that a one standard deviation increase in the weighted car count for an individual store is associated with an increase in overall quarterly sales of 0.55% relative to the prior quarter. In other words, an individual who had access to this real-time information on car counts during the quarter would be able to use it to predict quarterly performance before it was announced by the company. Consequently, we feel comfortable that our car count measure captures one aspect of local performance that can provide valuable information to investors. We believe this measure provides

⁵ On average, car counts are observed 2.83 times for a store in a given quarter.

⁶ Results are similar when we weight based on prior quarter car counts, or using weights generated by Orbital Insight using a proprietary formula.

a key contribution of the paper because it allows us to directly observe one aspect of local firm performance that is available in real time. In contrast, prior studies such as Bernile et al. (2015) must rely on indirect measures of local performance (for example, based on a portfolio of nearby firms) using reported financial data which would only be available to investors after the quarter end.

3.3 Investor Holding Data

Ideally, we would like to have daily trading data directly tied to the information investors have acquired about each store for every investor in a given retailer. However, shares are traded at the retailer level and investors are not required to disclose their trades or holdings unless 1) they are a beneficiary owner of 5% or more of a firm's equity, or 2) they manage at least \$100 million of equity. In the first instance, investors are required to file a Form 13D or 13G within 10 days of reaching the 5% threshold. In the second case, hedge funds are required to file a Form 13F within 45 days after calendar quarter ends March 31, June 30, September 30, and December 31.

We use Form 13F data from Thomson Reuters to determine investor holdings at the retailer level. Form 13F contains a breakdown of the number and value of shares held by institutional investors for each company in which they own stock. Therefore, we are able to determine whether an investor's holding in a particular stock increased or decreased over the reported quarter. In particular, investor i 's change in holdings in firm r over quarter q is calculated as follows:

$$Chg_Holdings_{i,r,q} = \frac{(Shares_{i,r,q} - Shares_{i,r,q-1}) \times Price_{i,r,q}}{\sum_{f=1}^n (Shares_{i,f,q} * Price_{i,f,q})} * 100 \quad (4)$$

Where f stands for all the firms that investor i invested in and the other subscripts are as previously defined. In Table 3 we provide descriptive information about the institutional investors

that invested in at least one of our sample retailers within any quarters of our sample period.⁷ Our sample includes 2,975 unique managers and each manager holds 199 unique firms (21 unique retailers) on average, or an average of 14,000 of our sample stores. On average these investors hold \$3.44 billion (\$165 million) in total shares (shares in sample retailers). In terms of dollar holdings (number of investors), our sample covers 97% (79%) of the institutional investors in the 13F universe. Thus while the *firms* in our sample represent only a small fraction of publicly-listed firms, they are sufficiently prominent that our set of *investors* captures the vast majority of institutional investors in the U.S.

3.4 Identification of Local Investors and Descriptive Statistics

A key benefit of our store-level dataset is that we are able to observe the performance of individual stores and classify investors as local or non-local at the store-level. Specifically, an investor with stock holdings for a given retailer is considered local to a particular store managed by that retailer if the investor is located in the same state as that store. Therefore, it is possible that an investor may be local to one store, but non-local to another store managed by the same retailer. As shown in Table 4, 4% of our investor-store pairs have a value of *Local* equal to 1, meaning that 4% of the investor-store-quarter observations in our sample correspond to investor-store pairs which are located in the same state.

Table 4 provides descriptive statistics for the other variables we use in our analyses. We point out that although all of these variables are used in our remaining tests conducted at the investor-store-quarter level, they are measured at multiple levels. For example, we include some firm-level control variables which vary only at the retailer-quarter level, such as change in overall

⁷ 13F investors in our sample are required to have non-missing price data from CRSP for investments in their portfolio and location data from the 13F filing. We exclude investors where the state of the mailing and business address differ. Firm-quarter observations are linked with the first investor holdings data reported after the firm's fiscal quarter end.

firm sales (*Sales_Chg* with a mean of 2.84) and returns (*AbRet_3m* with a mean of -2.51). We also have investor-firm-level data, specifically the change in a given investor's holdings in a particular retail firm over a specific quarter (*Chg_Holdings* with a mean of 0.04). We further have store-level data which track variables for individual stores within the retailers in our sample, specifically our store-level car counts (*Car_Chg_W* with a mean of 1.75). Lastly, we have investor-store-level data, specifically our *Local* variable, which is measured for investor-store pairs.⁸

Because our main analyses are conducted at the investor-store level, some variables have repeated values within investor-store pairs. For example, all firm-level variables are identical for all stores in that firm-quarter. We cluster our standard errors by firm-quarter in order to account for shared variation within firms and investor-firm pairs. We provide a graphical representation of our data structure for two example firms and investors in Figure 1.

IV. Empirical Results

4.1 Establishing Presence of Local Information Advantage

A key feature in our dataset is that it allows us to observe the car count at each particular store, and hence we have an empirical proxy for economic fundamentals of individual stores. This is important because empirically we find large variation in the car counts of stores managed by the same retailer that are located in different states, so investors who observe performance of different stores can have a very different idea of overall performance. For each investor-store pair we classify the pair as either local or non-local, as described above. Our treatment observations are local investor-store pairs. Our control observations include all remaining non-local investor-store pairs. Importantly, these non-local control pairs include: 1) different stores managed by the same

⁸ Because each store is linked with one retail firm, this can also be described as investor-firm-store-level data.

retailer and owned by the same investor, in order to control for the overall relationship the investor has with the retailer in that quarter; or 2) the same store as a local investor-store pair but owned by a non-local investor, to control for the overall information environment about a given store; or 3) stores from other retailers paired with the same investor in order to control for aspects of investor skill; or 4) other stores from the same retailer held by other investors in order to control for the overall firm information environment.⁹ All of these types of non-local investor-store pairs can be seen in Figure 1 Panel B. Including all of these investor-store pairs in the control group helps to remove investor-, retailer-, and investor-retailer-level variation from our *Local* variable.

By examining how an investor's holdings react to store-level information, we can isolate the sources of local information advantage. If an investor's holdings are more responsive to the car counts of local stores relative to the car counts of non-local stores, then this would be consistent with H1a. In particular we estimate the following investor-store level regression:

$$Chg_Holdings_{i,r,q} = \alpha + \beta_1 Car_Chg_N_{s,r,q} + \beta_2 Local_{s,i,r,q} + \beta_3 Local_{s,i,r,q} \times \\ Car_Chg_N_{s,r,q} + CONTROLS \quad (5)$$

In Equation 5, the dependent variable $Chg_Holdings_{i,r,q}$ is investor i 's change in holdings in retailer r in quarter q . Because we examine *changes* in holdings of an investor in a given firm, we are controlling for the past relationship of the investor-firm pair using first differences, including any past information the investor may have had about the firm which led the investor to its level of holdings as of the beginning of the quarter. As mentioned above, $Car_Chg_N_{r,s,q}$ is the change in normalized and weighted car counts for store s belonging to retailer r in quarter q . This changes variable allows us to focus on information that arose during the current quarter, controlling for past performance and information effects that would have led an investor to make their original

⁹ These categories represent a subset of all non-local investor-store pairs and are not all mutually exclusive.

investment amount in that retail firm or, potentially, locate near that particular store. *Local* is an indicator variable which takes on the value of 1 if store s is within the same state as investor i in quarter q . Additionally we control for investor size and current performance (*Fund_Size* and *Fund_Return*), firm characteristics such as size and performance, and quarter and store fixed effects.

As shown in Table 5 Panel A, consistent with H1a the coefficient on the interaction term of *Local* \times *Car_Chg_N* is positive and significant, indicating that investors' holdings are more responsive to changes in the car counts of local stores than to those of non-local stores. This is consistent with local investors having greater access to information about the performance of local stores relative to non-local investors and adjusting their holdings accordingly.

Next, in order to confirm that we are documenting an actual local information *advantage* (as opposed to investors myopically overweighting local information signals when making their trades), we test whether these trades in response to local car counts are information-driven by linking them with future returns. In particular, we estimate whether the portion of an investor's change in holdings which is driven by local car count changes is positively associated with future returns. In Table 5 Panel B we show that an investors' predicted change in holdings from Equation (5) (*Predicted_Trades_Local*) is positively associated with the future abnormal return of the underlying retailer in the 3-month period following the reported investor holdings (*CAR_3m_future*), even after controlling for the predicted change in holdings *not* related to local car counts. We estimate the change in holdings not associated with local car counts as the predicted value from the following equation (*Predicted_Trades*), which is simply Equation (5) with the *Local* \times *Car_Chg_N* interaction omitted:

$$Chg_Holdings_{i,r,q} = \alpha + \beta_1 Car_Chg_N_{s,r,q} + \beta_2 Local_{s,i,r,q} + CONTROLS$$

(6)

Because *Predicted_Trades_Local* is positively associated with future returns, we are able to confirm that not only are the holdings of investors particularly sensitive to information about the performance of local stores, but that the portion of changes in holdings attributable to this local information is profitable, confirming H1b and the existence of a local information advantage. While we do not see these results as the main contribution of the paper, we do point out that the existence of a local information advantage is not universally supported in prior studies (Grinblatt and Keloharju, 2000; Seasholes and Zhu, 2010) and our results provide further proof in support of this phenomenon.

4.2 Attention Channel

In this subsection, we provide empirical evidence on whether local advantage can be explained by heightened attention. To examine this hypothesis, we identify investor-retailer pairs where the investor is local to at least one of the stores managed by the retailer in question. That is, if an investor is local to any of the stores managed by retailer r , he is “proximate” to retailer r (*Proximate*=1). We posit that if local information advantage is driven by heightened attention to local firms (H2), then proximate investors should be more responsive to the retailer’s fundamentals as a whole relative to non-proximate investors. In particular, we estimate the following equation:

$$Chg_Holdings_{i,r,q} = \alpha + \beta_1 Car_Chg_N_{s,r,q} + \beta_2 Car_Chg_Agg_N_{r,q} + \beta_3 Local_{s,i,r,q} + \beta_4 Proximate_{i,r,q} + \beta_5 Local_{s,i,r,q} \times Car_Chg_N_{s,r,q} + \beta_7 Proximate_{i,r,q} \times Car_Chg_Agg_N_{r,q} + CONTROLS$$

(6)

In Equation (6), *Proximate* is an indicator variable which takes on the value of 1 if investor i is local to *any* of retailer r 's stores. The variable $\text{Car_Chg_Agg_N}_{r,q}$ is the retailer r 's normalized average car count for the quarter across all stores.

As shown in Table 6 column 1, the coefficient on the interaction term of *Proximate* \times Car_Chg_Agg_N is positive and marginally significant. This indicates that when an investor is local to any of retailer r 's stores, they are also more responsive to the overall activity of retailer r . This is the case even after controlling for the investor's reaction to local stores and is consistent with the heightened attention hypothesis, H2. However, the coefficient on *Local* \times Car_Chg_N is still positive and significant, compared to the marginally significant coefficient on *Proximate* \times Car_Chg_Agg_N , indicating that attention is not the main driver of local information advantage.¹⁰

4.3 Insider Information Channel

Next, we examine whether part of the local advantage we document can be explained by information leakage from insiders. We posit that if local information advantage is driven by leakage from insiders, then investors who are located within close proximity to the retailer's headquarters are more likely to obtain additional information about overall firm performance because top-level managers have information about the overall performance of the retailer and are mostly stationed at the firm's headquarters. In particular, we estimate the following regression:

$$\begin{aligned} \text{Chg_Holdings}_{i,r,q} = & \alpha + \beta_1 \text{Car_Chg_N}_{s,r,q} + \beta_2 \text{Car_Chg_N_Agg}_{r,q} + \beta_3 \text{Local}_{s,i,r,q} + \\ & \beta_4 \text{Proximate}_{i,r,q} + \beta_5 \text{HQ}_{i,r,q} + \beta_6 \text{Local}_{s,i,r,q} \times \text{Car_Chg_N}_{s,r,q} + \beta_7 \text{Proximate}_{i,r,q} \times \\ & \text{Car_Chg_N_Agg}_{r,q} + \beta_8 \text{HQ}_{i,r,q} \times \text{Car_Chg_N_Agg}_{r,q} + \text{CONTROLS} \quad (7) \end{aligned}$$

¹⁰ We also estimated Equation 6 examining the interaction of *Proximate* with Car_Chg_N instead of Car_Chg_Agg_N (the performance of individual non-local stores instead of overall aggregate performance) and continued to find weak evidence of attention, with the interaction marginally significant in a 1-tailed test. Again, this supports our interpretation that the attention channel is not the main driver of local information advantage.

In Equation (7), $HQ_{i,r,q}$ is an indicator variable which takes on the value of 1 if investor i is located in the same state as retailer r 's headquarters. As shown in Table 6 column 2, we find no evidence that investors who are close to a retailer's headquarters react more to the overall activity of the retailer, whereas *Local x Car_Chg_N* and *Proximate x Car_Chg_Agg_N* continue to be positive and significant. We find no evidence in support of H3, the insider information channel.¹¹

4.4 Geographically Remote Stores

The results of Table 6 indicate that while the attention channel plays a reduced role in local information advantage and insider information plays little to no role in this post-Reg FD period, acquisition costs appear to be the main driver of local information advantage. That is, differential costs to acquire local information give local investors an advantage relative to non-local investors. One such cost is the cost to travel to a store to directly observe operations. If acquisition costs are really the main source of local information advantage (and travel costs are one of these acquisition costs), we should see that the local information effect we document in Table 5 varies with average travel costs to reach a given store. Geographically “remote” stores, located in states far from financial hubs, are likely to have the highest travel costs on average. In Table 7 we identify stores in our sample that are the most (least) geographically remote by identifying those which are in the top (bottom) quintile in terms of distance from the closest financial hub. We define financial hubs as New York, Chicago, San Francisco, and Los Angeles. Investors tend to be centered at these

¹¹ The coefficient on $HQ \times Car_Chg_Agg_N$ continues to be insignificant even when $Proximate \times Car_Chg_Agg_N$ is omitted. It is possible that insider information manifests by investors with insider information having better information on how individual store-level car counts map into overall firm performance. In that case, changes in investor holdings in response to local car counts would be more profitable for investors located near the headquarters (with greater probability of insider information). We re-estimate Table 5 Panel B, the profitability of local information-driven trades, for headquarter-state and non-headquarter-state investors separately. We find no significant difference in the profitability of local information-driven trades for these two subsamples. Overall, we do not document any evidence of a local information advantage due to insider information.

hubs, so to the extent that a store is far from the closest hub, it is far from the vast majority of investors.

The results in Table 7 show that the relation between *Local* \times *Car_Chg_N* continues to be significantly positive for remote stores, while the relation is no longer significant for firms that are closest to financial hubs (although the difference in coefficients is not significant). We see this as preliminary evidence further supporting the acquisition costs channel.

V. Additional Analyses

We perform a variety of additional analyses in order to support our interpretation of our main results. First, in all of the tests in this paper we use our car counts as a measure of the performance of local stores and indirectly as a measure of local information. To rule out the possibility that local investors are simply purchasing the data used in our study (which was commercially available at the end of our sample period), in untabulated analyses we restrict our sample to include only observations before the inception date for each store (when the data became available for purchase), and our results remain qualitatively similar. Similarly, to rule out the possibility that local information advantage, as captured using car counts, disappears after our satellite data became available, we run our main test in Table 5 Panel A in the last 2 years of our sample when the data were available for purchase. We still find a significant coefficient on *Local* \times *Car_Chg_N*.

Further, much of the prior research has studied local bias and local information advantage by identifying investors who are local to the firm's headquarters. A benefit of our study is that we are able to define investors as both local and non-local relative to individual stores within the same retailer, maximizing our ability to control for the relationship of a given investor-retailer pair.

However, it could be the case that our local advantage results are driven entirely by investor-store pairs located within the retailer headquarter state. This would disprove our assumption that local information can be spread geographically throughout the retailer and would decrease the benefits of our *Local* measure. In untabulated analyses, we test whether the relation between *Car_Chg_N* \times *Local* and *Chg_Holdings* is significant only for investor-store pairs located within the headquarter state and find this relation is still significant in non-headquarter states.

Additionally, the SEC has previously expressed a preference to intervene in cases where small investors are at a disadvantage relative to large investors (SEC, 2009). We therefore examine whether the existence of the local information advantage that we document is more pronounced for small investors relative to large investors. In Table 8 we report the results of Table 5 Panel A for two samples of investors; those investors with total assets under management above the sample median (column 1) and those with total assets under management below the sample median (column 2). We find that the relation between *Local* \times *Car_Chg_N* and *Chg_Holdings* is significant only for small investors (the difference in coefficients is significant at the 1% level), consistent with small investors being disproportionately affected by local information. This is intuitive, because small investors have fewer resources to acquire information about non-local stores, for example through travel or purchase of costly data. However, we must caveat this result by acknowledging that large investors are those that are most likely to have multiple offices, whereas 13F filings provide only the address of the main investor office, meaning that large investors may have branch offices local to stores in our sample which we do not identify. Although this works against our finding significant results in our main sample, it disproportionately affects large investors in this subsample analysis, meaning that the difference in coefficients could be due to

this measurement problem. Because of this, we view this subsample analysis as suggestive but do not draw strong conclusions from it.

Lastly, in Table 5 we provided evidence supporting the existence of a local information advantage by showing that investors react more to local car counts and that these trades are profitable. Prior research has found some evidence that local investors sometimes trade for non-information-driven reasons and can make suboptimal trades (Engelberg and Parsons, 2011; Hochberg and Rauh, 2013; Hornuf and Schmidt, 2017; Seasholes and Zhu, 2010). Our results on the presence of a local information advantage do not contradict these prior findings because local investors can trade for a variety of reasons, both information-driven and non-information driven. However, in untabulated analyses we tie our results with this prior literature by examining whether investors use their local information to make suboptimal decisions, specifically whether they overweight their local information when making trades. We document that store-level car counts are less predictive of overall sales and net income when the standard deviation of car count changes across stores within a given firm is high. Further, when car counts vary widely across stores, the trades of local investors are *less* profitable than the trades of non-local investors. It appears that local investors make less profitable investment decisions when local car counts are less predictive of overall performance, suggesting they overweight their local information. This is consistent with prior studies showing suboptimal behavior on the part of local investors. Further, this type of behavior is unlikely to occur if local investors' information is obtained directly from management (who have information on overall firm performance) or when the investors are paying more attention to the firm overall. This further supports the information acquisition cost channel as the main source of local information advantage.

VI. Conclusion

In this paper we use data from satellite images on the number of cars in retail store parking lots to disentangle the sources of local information. We establish store-level car counts as a measure of one type of local information (specifically, information on the performance of local stores) by linking store-level car counts during the quarter with reported firm-level quarterly sales, unexpected earnings, and earnings announcement returns. Further, we demonstrate that institutional investors adjust their stock holdings more in response to the performance of local stores than to that of non-local stores and that these trades are profitable, confirming that investors benefit from a local information advantage. In addition, because our car count data allow us to measure one specific type of local information, we are able to disentangle the channels by which local investors obtain this information. Specifically, we distinguish between a local information advantage arising from: greater attention to the firm as a whole as a result of being local to some stores (the “attention channel”); information obtained directly from firm management (“insider information channel”); and lower costs of information acquisition for local investors relative to non-local investors (“acquisition cost channel”). Overall our results provide strong evidence to support the acquisition cost channel, especially for firms with stores in geographically remote locations, weak evidence to support the attention channel, and no evidence of the insider information channel.

Prior research such as Bernile et al. (2015) has provided evidence that one driver of the well-documented local bias phenomenon (whereby investors hold a disproportionate amount of assets in firms from their home state) is that local investors have an information advantage relative to their non-local counterparts, as opposed to non-information based reasons such as familiarity. However, we are the first to be able to distinguish between the actual sources of this information.

We think our results are important from a policy perspective. While some sources of private material information are illegal and require regulatory intervention, others are more benign. Because our results do not provide evidence of the insider information channel in this post-Reg FD era, we do not believe that the local information we capture appears to be transferred via illegal communications between managers and investors. Rather, the main source of this information advantage appears to be the increased cost to non-local investors of acquiring this information. This finding is surprising in an age where the prevailing notion is that the “world is flat” and barriers to information acquisition and dissemination due to geographic factors are virtually zero. Access to “big data,” such as the data used in this study, is often perceived as a cure-all for these types of frictions. However, our results suggest that there are still significant barriers to acquiring timely information about different geographic areas, and the SEC’s stated goal to reduce differences in information sets across investors is still relevant, even among institutional investors.

References

- Bae, K. H., Stulz, R. M., and Tan, H. (2008). Do local analysts know more? A cross-country study of the performance of local analysts and foreign analysts. *Journal of Financial Economics*, 88(3), 581-606.
- Baik, B., Kang, J. K., and Kim, J. M. (2010). Local institutional investors, information asymmetries, and equity returns. *Journal of Financial Economics*, 97(1), 81-106.
- Bernile, G., Kumar, A., and Sulaeman, J. (2015). Home away from home: Geography of information and local investors. *The Review of Financial Studies*, 28(7), 2009-2049.
- Bernile, G., Kumar, A., Sulaeman, J., and Wang, Q. E. (2018). Has local informational advantage disappeared? Working Paper.
- Chi, S. S., and Shanthikumar, D. M. (2016). Local bias in Google search and the market response around earnings announcements. *The Accounting Review*, 92(4), 115-143.
- Cooper, I., Sercu, P., and Vanpée, R. (2013). The equity home bias puzzle: A survey. *Foundations and Trends in Finance*, 7(4), 289-416.
- Cornaggia, J., Cornaggia, K. J., and Israelsen, R. (2018). Where the Heart Is: Information Production and the Home Bias. Working Paper.
- Coval, J. D., and Moskowitz, T. J. (1999). Home bias at home: Local equity preference in domestic portfolios. *The Journal of Finance*, 54(6), 2045-2073.
- Coval, J. D., and Moskowitz, T. J. (2001). The geography of investment: Informed trading and asset prices. *Journal of Political Economy*, 109(4), 811-841.
- Dyer, T. A. (2018). Does Public Information Acquisiton Level the Playing Field or Widen the Gap? An Analysis of Local and Non-local Investors. Working Paper.
- Ellis, J., Madureira, L., and Underwood, S. (2017). The Causal Effects of Proximity on Investment: Evidence from Flight Introductions. Working Paper.
- Engelberg, J. E., and Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67-97.
- Frankel, R., Jennings, J., and Lee, J. (2016). Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, 62(2-3), 209-227.
- French, K. R., and Poterba, J. M. (1991). Investor Diversification and International Equity Markets. *The American Economic Review*, 81(2), 222.

- Grinblatt, M., and Keloharju, M. (2000). The investment behavior and performance of various investor types: a study of Finland's unique data set. *Journal of Financial Economics*, 55(1), 43-67.
- Hochberg, Y. V., and Rauh, J. D. (2013). Local overweighting and underperformance: Evidence from limited partner private equity investments. *The Review of Financial Studies*, 26(2), 403-451.
- Hornuf, L., and Schmitt, M. (2017). Does a Local Bias Exist in Equity Crowdfunding? Working Paper.
- Huberman, G. (2001). Familiarity breeds investment. *The Review of Financial Studies*, 14(3), 659-680.
- Morse, A., and Shive, S. (2011). Patriotism in your portfolio. *Journal of Financial Markets*, 14(2), 411-440.
- Miller, G., and Shanthikumar, D. (2010). Geographic location, media coverage and investor reactions. Working Paper.
- Perdikaki, O., Kesavan, S., and Swaminathan, J. M. (2012). Effect of traffic on sales and conversion rates of retail stores. *Manufacturing and Service Operations Management*, 14(1), 145-162.
- Pool, V. K., Stoffman, N., and Yonker, S. E. (2012). No place like home: Familiarity in mutual fund manager portfolio choice. *The Review of Financial Studies*, 25(8), 2563-2599.
- Seasholes, M. S., and Zhu, N. (2010). Individual investors and local bias. *The Journal of Finance*, 65(5), 1987-2010.
- Securities and Exchange Commission (SEC), 2000. Final Rule: Selective Disclosure and Insider Trading. Available at: https://www.sec.gov/rules/final/33-7881.htm#P12_1307. SEC Offices, Washington D.C.
- Securities and Exchange Commission. (2009). Interactive data to improve financial reporting. Retrieved July 5, 2012 from the SEC's website: <http://www.sec.gov/rules/final/2009/33-9002.pdf>.
- Van Nieuwerburgh, S., and Veldkamp, L. (2009). Information immobility and the home bias puzzle. *The Journal of Finance*, 64(3), 1187-1215.
- Zhu, C. (2018). Big Data as a Governance Mechanism. Working Paper.

APPENDIX A

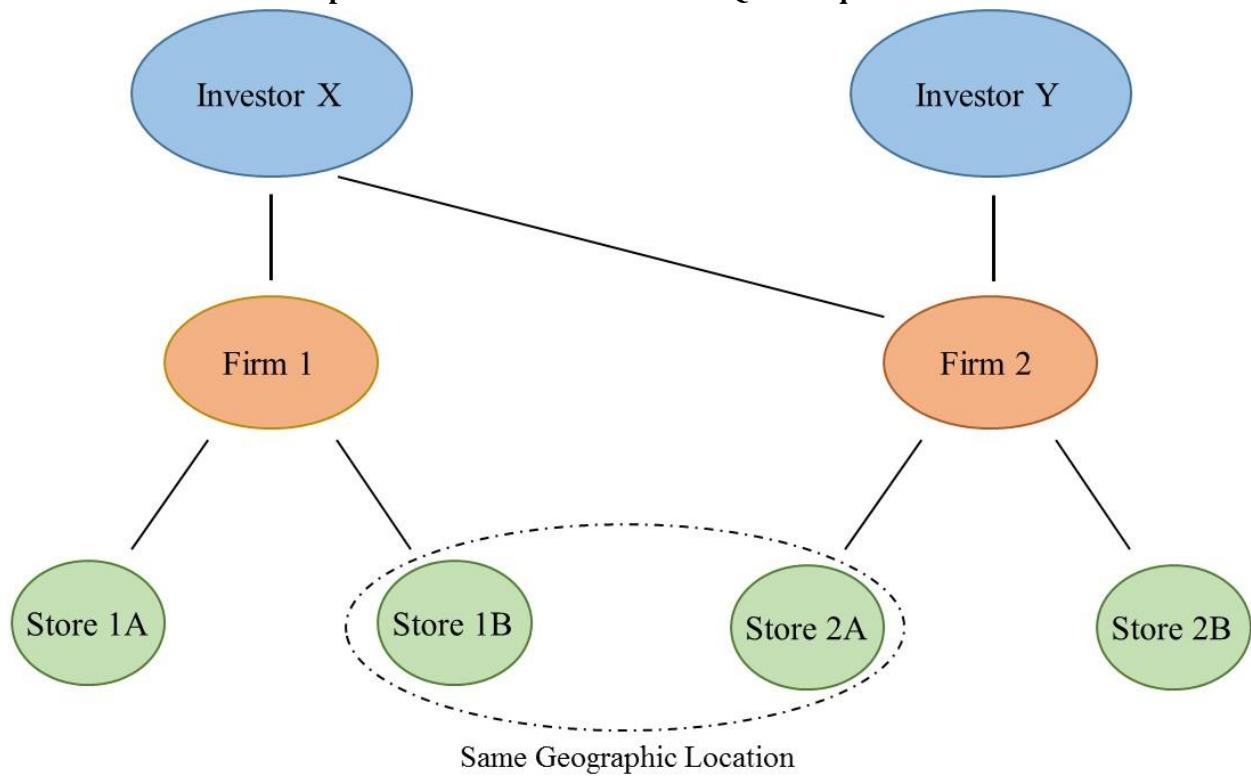
Variable Definitions

Variable	Definition	Source
<i>Analyst_Num</i>	= Total number of analysts following a firm	IBES
<i>AbRet_3m</i>	= Three month period cumulative abnormal return (CAR) = based on Fama-French four factor model prior to the <i>firm's</i> earnings announcement date (%)	CRSP, Eventus
<i>AbRet_3m_Lag</i>	= Lagged three month periods cumulative abnormal return (CAR) based on Fama-French four factor model prior to the <i>firm's</i> earnings announcement date (%)	CRSP, Eventus
<i>CAR_2d</i>	= Two day cumulative abnormal return (CAR) based on = Fama-French four factor model after and including the firm's earnings announcement date (%)	CRSP, Eventus
<i>CAR_3m_Future</i>	= Three month period cumulative abnormal return (CAR) = based on Fama-French four factor model starting after the <i>investor's</i> reported quarter end date (%)	CRSP, Eventus, 13F
<i>Car_Chg</i>	= The ratio of quarterly changes in average store level car counts to the average car counts at the prior quarter-end (%)	Orbital Insight
<i>Car_Chg_Agg</i>	= Quarterly average <i>Car_Chg</i> across stores for a firm	Orbital Insight
<i>Car_Chg_Agg_N</i>	= Normalized <i>Car_Chg_Agg</i> having zero mean and unit variance	Orbital Insight
<i>Car_Chg_W</i>	= <i>Car_Chg</i> times the ratio of the average store level car counts to the firm level average car counts (%)	Orbital Insight
<i>Car_Chg_N</i>	= Normalized <i>Car_Chg_W</i> having zero mean and unit variance	Orbital Insight
<i>Chg_Holdings</i>	$\frac{(Shares_{i,r,q} - Shares_{i,r,q-1}) \times Price_{i,r,q}}{\sum_{f=1}^n (Shares_{i,f,q} * Price_{i,f,q})} * 100$ = Where i refers to the investor, r refers to one of our 75 retailers, q is the fiscal quarter, f stands for all the firms that investor i invested in, <i>Shares</i> is the number of shares investor i holds in firm r and <i>Price</i> is the stock price for a firm r	13F, CRSP

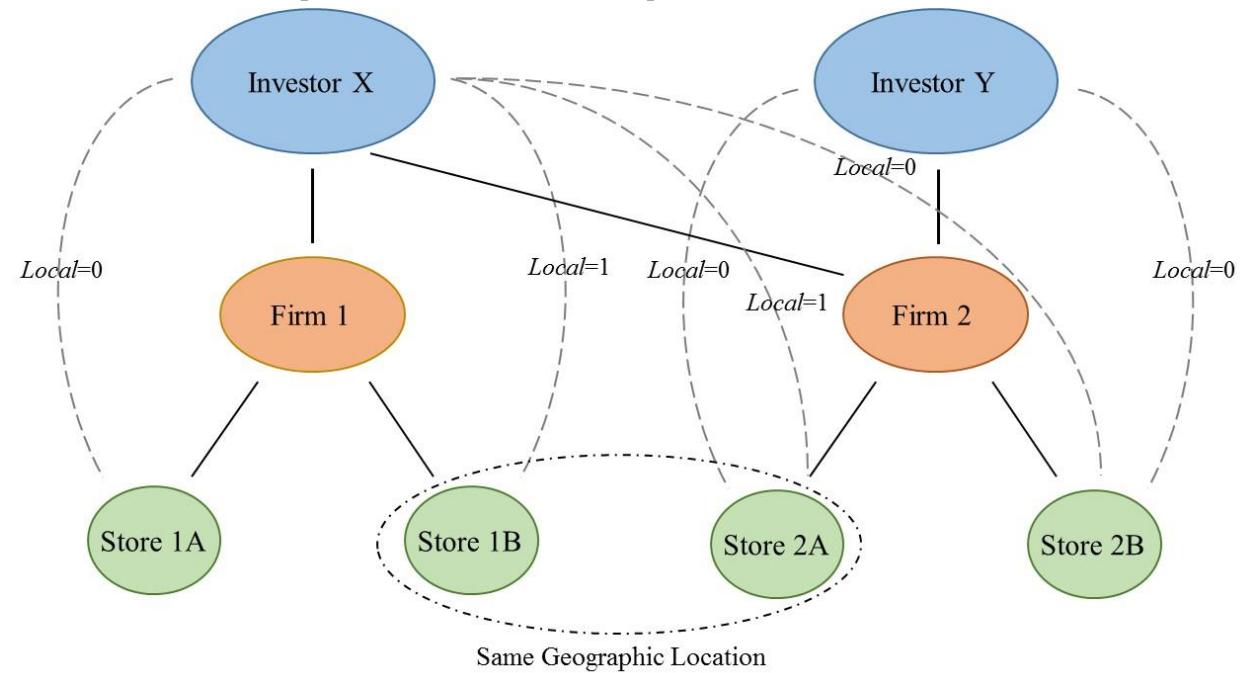
<i>Forecast_Vol</i>	= The standard deviation of the EPS forecasts scaled by the stock price at the prior quarter-end. (%)	IBES, CRSP
<i>Fund_Return</i>	= Quarterly weighted average return for the investment amounts of an investor	13F
<i>Fund_Size</i>	= The natural logarithm of the total investment amounts of an investor	13F
<i>HQ</i>	= An indicator equal to 1 if an investor's headquarter is same as a firm's headquarter, 0 otherwise.	13F, Compustat
<i>Leverage</i>	= The ratio of total liabilities (ltq) to the total equity (atq-ltq) of a firm	Compustat
<i>Leverage_Lag</i>	= The ratio of total liabilities (ltq) to the total equity (atq-ltq) of a firm in the prior quarter-end	Compustat
<i>Local</i>	= An indicator equal to 1 if an investor's headquarter state is same as the store's state, 0 otherwise.	13F, Orbital Insight
<i>Log_Assets</i>	= The natural logarithm of the total assets (atq)	Compustat
<i>Log_Mktcap</i>	= The natural logarithm of total market value (mkvaltq)	Compustat
<i>Log_Mktcap_Lag</i>	= The natural logarithm of total market value (mkvaltq) in the prior quarter-end	Compustat
<i>MTB</i>	= The ratio of market value (mkvaltq) to total equity (teqq)	Compustat
<i>NI_Chg</i>	= Quarterly changes in net income (niq) scaled by the net income (niq) in the prior quarter (%)	Compustat
<i>Proximate</i>	= An indicator equal to 1 if an investor's headquarter state overlaps with at least one state of a firm's stores, 0 otherwise.	13F, Orbital Insight
<i>Ret_Vol</i>	= Standard deviation of daily stock price return during the one year before the earnings announcement date scaled by the stock price at the prior quarter-end	IBES, CRSP
<i>ROA</i>	= The ratio of net income (niq) to the total assets (atq)	Compustat
<i>Sales_Chg</i>	= Quarterly changes in sales (saleq) scaled by the sales in the prior quarter (%)	Compustat
<i>SUE</i>	= The ratio of earnings per share minus the median value of earnings forecast to the stock price at the prior quarter-end (%)	IBES, Compustat CRSP

Figure 1. Data Structure

Panel A: Illustrative Example for 2 Investors and 2 Firms in Quarter q



Panel B: Illustrative Example--Local Variable in Quarter q



Panel C: Example Data for Quarter q

Investor	Firm	Store	Chg_Holdings	Sales_Chg	Fund_Size	Local	
			(Investor-Firm Level Variable)	(Firm Level Variable)	(Investor Level Variable)	(Investor-Store Level Variable)	
X	1	1A	Chg_Holdings _{X,1,q}	Sales _{1,q}	Fund_Size _{X,q}	Local _{1A,X,1,q} = 1	Firm-Quarter Clustering
X	1	1B	Chg_Holdings _{X,1,q}	Sales _{1,q}	Fund_Size _{X,q}	Local _{1B,X,1,q} = 0	
X	2	2A	Chg_Holdings _{X,2,q}	Sales _{2,q}	Fund_Size _{X,q}	Local _{2A,X,2,q} = 0	
X	2	2B	Chg_Holdings _{X,2,q}	Sales _{2,q}	Fund_Size _{X,q}	Local _{2B,X,2,q} = 0	
Y	2	2A	Chg_Holdings _{Y,2,q}	Sales _{2,q}	Fund_Size _{Y,q}	Local _{2A,Y,2,q} = 0	
Y	2	2B	Chg_Holdings _{Y,2,q}	Sales _{2,q}	Fund_Size _{Y,q}	Local _{2B,Y,2,q} = 1	

Table 1. Satellite Image Data

Panel A: Sample Selection

	Store-Quarter Observations	Unique Firms	Unique Stores
(1) Quarterly average car counts from Orbital Insight	2,000,483	80	106,776
(2) Eliminate obs missing accounting data from Compustat, analyst data from IBES, and price data from CRSP	1,167,654	78	104,692
(3) After merging with 13F data	877,325	75	96,742

Panel B: Store-Quarter Observations Per Geographic Location

State	Total	Unique Firms	Unique Stores	State	Total	Unique Firms	Unique Stores
AK	1,175	23	89	NC	31,608	70	3,896
AL	11,046	64	2,007	ND	1,657	43	191
AR	6,947	62	1,016	NE	5,952	55	629
AZ	37,087	66	2,164	NH	3,184	48	479
CA	131,597	62	8,400	NJ	19,642	60	2,096
CO	28,865	67	1,934	NM	8,319	55	718
CT	8,066	59	1,064	NV	16,407	64	892
DC	741	11	35	NY	27,644	64	3,515
DE	1,902	52	295	OH	22,470	69	4,385
FL	77,650	68	6,830	OK	12,664	65	1,259
GA	27,644	70	3,903	OR	9,742	56	954
IA	6,439	60	1,073	PA	24,660	67	4,049
ID	5,227	55	448	PR	16	2	2
IL	24,613	71	3,676	RI	3,184	50	298
IN	15,725	70	2,490	SC	17,777	65	2,036
KS	8,498	66	1,129	SD	1,623	47	234
KY	8,136	67	1,709	TN	18,484	69	2,605
LA	15,173	65	1,812	TX	82,865	71	8,669
MA	18,130	58	1,670	UT	10,535	59	845
MD	12,975	65	1,607	VA	25,886	68	3,037
ME	2,381	48	415	VT	660	36	138
MI	15,653	65	3,187	WA	14,060	55	1,677
MN	12,294	61	1,176	WI	8,108	62	1,304
MO	17,658	68	2,024	WV	1,543	48	610
MS	6,261	56	1,137	WY	947	36	183
MT	1,474	44	231				

Panel C: Store-Quarter Observations Over Time

Year	Total	Unique Firms	Unique Stores	% of Pre-Ingestion Periods
2009	32,276	50	13,391	100.0
2010	96,459	62	27,298	100.0
2011	103,852	65	31,051	100.0
2012	112,812	65	31,104	100.0
2013	113,548	69	31,812	100.0
2014	102,552	70	29,652	100.0
2015	132,574	72	36,956	79.6
2016	183,252	72	53,690	11.6
Total	877,325	75	96,742	78.4

Panel D: Sample Retailers vs. Other Firms

Characteristics (Mean)	Sample Retailers	Non-Sample Retailers	All Other Compustat
Assets (\$millions)	2,535.1	1,336.8	1,303.7
Market Cap (\$millions)	3,010.9	1,277	1,025.9
MTB	4.2	2.5	2.1
Liabilities/Assets	0.6	0.6	0.6
Revenue (\$millions)	4,041.1	1,344.6	731.3
NI (\$millions)	114.1	57.3	39.6
Loss	0.1	0.3	0.5

Table 2. Ability of Satellite Car Count Data to Predict Firm Performance

	<i>Sales_Chg</i>	<i>SG&A_Chg</i>	<i>NI_Chg</i>	<i>SUE</i>	<i>CAR_2d</i>
	(1)	(2)	(3)	(5)	(5)
<i>Car_Chg_N</i>	0.555*** (11.07)	0.239*** (3.78)	1.507** (2.24)	0.010* (1.90)	0.065*** (3.17)
<i>Log_Assets</i>	5.997* (1.700)	4.739 (1.180)	(27.880) (-1.16)	-0.356** (-2.23)	0.618 (0.430)
<i>Log_Mktcap</i>	-4.608** (-2.47)	-5.435* (-1.93)	10.904 (0.430)	(0.069) (-0.24)	-4.181*** (-5.32)
<i>AbRet_3m</i>	(0.042) (-1.21)	0.029 (0.650)	(0.457) (-0.94)	(0.001) (-0.27)	(0.018) (-1.13)
<i>AbRet_3m_Lag</i>	0.053* (1.830)	0.078 (1.540)	(0.839) (-1.15)	(0.008) (-1.57)	0.014 (1.020)
<i>MTB</i>	0.003 (0.250)	0.008 (0.510)	0.410 (0.970)	0.000 (-0.04)	0.003 (0.350)
<i>Ret_Vol</i>	-7.835** (-1.99)	-8.434* (-1.66)	6.000 (0.210)	(0.042) (-0.07)	(1.605) (-1.42)
<i>Analyst_Num</i>	0.122 (0.580)	(0.027) (-0.07)	0.370 (0.190)	0.005 (0.300)	(0.076) (-1.01)
<i>Forecast_Vol</i>	2.987 (1.530)	2.313 (1.520)	10.540 (0.930)	(0.023) (-0.10)	0.973 (1.510)
<i>Store_Num</i>	-10.113** (-2.11)	(11.273) (-1.38)	(10.805) (-0.26)	0.780** (2.080)	(0.757) (-0.50)
<i>Fixed Effects</i>	Quarter, Store	Quarter, Store	Quarter, Store	Quarter, Store	Quarter, Store
<i>Clusters</i>	Firm-Quarter	Firm-Quarter	Firm-Quarter	Firm-Quarter	Firm-Quarter
<i>Observations</i>	869,216	869,216	869,216	869,216	869,216
<i>Adj. R2</i>	0.078	0.016	0.051	0.146	0.062

Table 3. 13F Investor Holdings Data

Panel A: 13F Data Sample Selection

	Investor-firm-quarter observations	Unique Investors	Unique Firms
(1) Full 13F dataset (March 2009-December 2016)	24,325,256	4,779	18,074
(2) Eliminate obs missing location data	16,807,569	3,826	17,091
(3) Eliminate obs missing price data from CRSP	14,386,941	3,770	5,551
(4) After merging with satellite data	471,073	2,975	75

Panel B: Merged 13F and Satellite Data

	Investor-Store-Quarter Observations	Unique Investors	Unique Firms	Unique Stores
(1) Combined investor-store-quarter data	301,118,232	2,975	75	96,742

Panel C: Sample Investors vs. Others

	13F investors not in our sample with 13F investors available CRSP in our sample and location data	%
Total # of Managers	2,975	795
Total Assets Under Management (avg, \$billions)	10,244	334
Average # Unique Firms Held	199	21

Table 4. Descriptive Statistics

	N	Mean	Q1	Median	Q3	SD
<i>Firm-Quarter Level</i>						
<i>Sales_Chg</i>	2,022	2.84	-6.44	1.02	10.79	18.18
<i>SG&A_Chg</i>	2,022	2.18	-3.95	1.34	7.33	14.86
<i>NI_Chg</i>	2,022	-12.38	-53.12	-7.80	39.74	234.29
<i>SUE</i>	2,022	0.05	-0.03	0.03	0.13	0.87
<i>CAR_2d</i>	2,022	0.11	-4.55	0.14	4.89	7.90
<i>Log_Assets</i>	2,022	7.84	6.47	7.63	9.01	1.62
<i>Log_Mktcap</i>	2,022	8.01	6.57	7.88	9.35	1.76
<i>AbRet_3m</i>	2,022	-2.51	-13.36	-2.27	8.85	18.18
<i>AbRet_3m_Lag</i>	2,022	-1.17	-13.03	-1.35	10.26	19.30
<i>MTB</i>	2,022	4.17	1.64	2.78	4.43	16.74
<i>Ret_Vol</i>	2,022	0.17	0.03	0.06	0.17	0.37
<i>Analyst_Num</i>	2,022	17.68	8.00	19.00	26.00	9.18
<i>Forecast_Vol</i>	2,022	0.25	0.04	0.08	0.18	0.59
<i>Store_Num</i>	2,022	5.95	5.02	5.86	6.78	1.09
<i>Car_Chg_Agg</i>	2,022	0.64	-3.82	0.06	4.89	6.86
<i>Car_Chg_Agg_N</i>	2,022	0.00	-0.65	-0.08	0.62	1.00
<i>Investor-Quarter Level</i>						
<i>Fund_Size</i>	54,982	13.10	11.85	12.79	14.21	1.85
<i>Fund_Return</i>	54,982	0.03	0.00	0.03	0.07	0.07
<i>Store-Quarter Level</i>						
<i>Car_Chg_W</i>	877,325	1.75	-0.92	0.00	1.67	9.08
<i>Car_Chg_N</i>	877,325	0.00	-0.29	-0.19	-0.01	1.00
<i>Investor-Firm-Quarter Level</i>						
<i>Chg_Holdings</i>	471,073	0.04	-0.01	0	0.03	0.36
<i>CAR_3m_future</i>	471,073	-1.21	-7.92	-1.14	5.56	12.51
<i>HQ</i>	471,073	0.04	0.00	0.00	0.00	0.20
<i>Proximate</i>	471,073	0.86	1.00	1.00	1.00	0.34
<i>Investor-Store-Quarter Level</i>						
<i>Local</i>	301,118,232	0.04	0	0	0	0.19

Table 5. Local Information Advantage**Panel A: Changes in Investment Holdings in Response to Local Information**

	Chg_Holdings	
	(1)	(2)
<i>Car_Chg_N</i>	-0.0001876*	-0.00055445**
	(-1.76)	(-2.56)
<i>Local</i>	0.0004254	0.00026623
	(0.40)	(0.20)
<i>Local*Car_Chg_N</i>	0.0004898**	0.00049227**
	(2.56)	(2.24)
<i>Fund_Size</i>	-0.0135104***	-0.01335107***
	(-18.17)	(-16.57)
<i>Fund_Return</i>	-0.328942***	-0.33027887***
	(-3.54)	(-3.50)
<i>Log_Mktcap_Lag</i>	-0.0014739	-0.00830745***
	(-0.26)	(-7.52)
<i>Leverage_Lag</i>	-0.00000088	0.00
	(-1.36)	(1.10)
<i>ROA_Lag</i>	0.00	0.00303586*
	(1.58)	(1.66)
<i>MTB</i>	0.00	0.00000111***
	(0.39)	(4.56)
<i>Fixed Effects</i>	Quarter, Store	Quarter
<i>Clusters</i>	Firm-Quarter	Firm-Quarter
<i>Observations</i>	301,118,219	301,118,232
<i>Adj. R2</i>	0.013	0.010

Panel B: Profitability of Trades Based on Local Information

	CAR_3m_future
	(1)
<i>Predicted_Trades_Local</i>	1.833*
	(1.830)
<i>Predicted_Trades</i>	1.886*
	(1.655)
<i>Fixed Effects</i>	Quarter, Store
<i>Clusters</i>	Firm-Quarter
<i>Observations</i>	301,118,219
<i>Adj. R2</i>	0.383

Table 6. Sources of Local Information

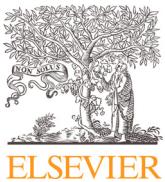
	<i>Chg_Holdings</i>	
	(1)	(2)
<i>Car_Chg_N</i>	-0.00021611** (-2.27)	-0.000219** (-2.30)
<i>Car_Chg_N_Agg</i>	-0.00062074** (-2.53)	-0.0006349** (-2.59)
<i>Local</i>	0.00195456 (0.72)	0.0003069 (0.30)
<i>Proximate</i>	0.00195456 (0.72)	0.0023537 (0.87)
<i>Local*Car_Chg_N</i>	0.00058700*** (2.68)	0.0006535*** (2.98)
<i>Proximate*Car_Chg_Agg_N</i>	0.00040540* (1.72)	0.0004415* (1.86)
<i>HQ</i>		-0.0095393* (1.90)
<i>HQ*Car_Chg_Agg_N</i>		-0.0002865 (0.68)
<i>Fund_Size</i>	-0.01347142*** (-18.23)	-0.013511*** (-18.46)
<i>Fund_Return</i>	-0.32985999*** (-3.55)	-0.3297177*** (-3.54)
<i>Mktcap_Lag</i>	-0.00139333 (-0.25)	-0.0013983 (-0.25)
<i>Leverage_Lag</i>	-0.000000089 (-1.38)	-0.000000089 (-1.38)
<i>ROA_Lag</i>	0.00187353 (1.58)	0.0018729 (1.58)
<i>MTB</i>	0.0000001 (0.40)	0.000000096 (0.40)
<i>Fixed Effects</i>	Quarter, Store	Quarter, Store
<i>Clusters</i>	Firm-Quarter	Firm-Quarter
<i>Observations</i>	301,118,219	301,118,219
<i>Adj. R2</i>	0.013	0.013

Table 7. Local Advantage at Geographically Remote Stores

	Chg_Holdings	
	Remote=1	Remote=0
	(1)	(2)
<i>Car_Chg_N</i>	-0.0000972 (-0.68)	-0.0003392* (-1.72)
<i>Local</i>	-0.0037971 (-1.25)	0.0035005 *** (1.76)
<i>Local*Car_Chg_N</i>	0.0010025** (1.96)	0.0006371 (1.64)
<i>Fund_Size</i>	-0.013718 *** (-16.96)	-0.0130493*** (-17.25)
<i>Fund_Return</i>	-0.3238069 *** (-3.21)	-0.3664097*** (-3.87)
<i>Mktcap_Lag</i>	-0.0034128 (0.51)	0.0007465 (0.15)
<i>Leverage_Lag</i>	-0.0000013 (-1.22)	-0.0000026 (-1.35)
<i>ROA_Lag</i>	0.0024116* (1.59)	0.0018428 (1.54)
<i>MTB</i>	-0.0000393 (-0.44)	-0.0001706 (-1.12)
<i>Fixed Effects</i>	Quarter, Store	Quarter, Store
<i>Clusters</i>	Firm-Quarter	Firm-Quarter
<i>Observations</i>	70,956,238	76,462,494
<i>Adj. R2</i>	0.013	0.013

Table 8. Local Advantage of Large and Small Investors

	Chg_Holdings	
	Large_Investor=1 (1)	Large_Investor=0 (3)
<i>Car_Chg_N</i>	-0.00005813 (-0.78)	-0.00058906* (-1.94)
<i>Local</i>	0.00277477** (2.25)	-0.00165214 (-1.10)
<i>Local*Car_Chg_N</i>	-0.00001156 (-0.05)	0.00112156*** (2.97)
<i>Fund_Size</i>	-0.00963240*** (-14.43)	-0.01646320*** (-9.79)
<i>Fund_Return</i>	-0.07518222 (-1.20)	-0.49395602*** (-3.86)
<i>Mktcap_Lag</i>	0.00431059 (1.51)	-0.0066925 (-0.74)
<i>Leverage_Lag</i>	-0.00000028 (-0.35)	-0.000000552* (-1.83)
<i>ROA_Lag</i>	0.00070203 (0.99)	0.00349317* (1.8800000)
<i>MTB</i>	-0.00005119 (-0.82)	-0.00025264 (-1.12)
<i>Fixed Effects</i>	Quarter, Store	Quarter, Store
<i>Clusters</i>	Firm-Quarter	Firm-Quarter
<i>Observations</i>	135,318,630	165,799,336
<i>Adj. R2</i>	0.007	0.012



Sentiment analysis on social media for stock movement prediction

Thien Hai Nguyen ^{a,*}, Kyoaki Shirai ^a, Julien Velcin ^b

^a School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

^b University of Lyon (ERIC, Lyon 2), 5 Avenue Pierre Mendes-France, 69676 Bron Cedex, France



ARTICLE INFO

Keywords:
 Sentiment analysis
 Opinion mining
 Classification
 Prediction
 Stock
 Social media
 Message board

ABSTRACT

The goal of this research is to build a model to predict stock price movement using the sentiment from social media. Unlike previous approaches where the overall moods or sentiments are considered, the sentiments of the specific topics of the company are incorporated into the stock prediction model. Topics and related sentiments are automatically extracted from the texts in a message board by using our proposed method as well as existing topic models. In addition, this paper shows an evaluation of the effectiveness of the sentiment analysis in the stock prediction task via a large scale experiment. Comparing the accuracy average over 18 stocks in one year transaction, our method achieved 2.07% better performance than the model using historical prices only. Furthermore, when comparing the methods only for the stocks that are difficult to predict, our method achieved 9.83% better accuracy than historical price method, and 3.03% better than human sentiment method.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Stock price forecasting is very important in the planning of business activity. However, building an accurate stock prediction model is still a challenging problem. In addition to historical prices, the current stock market is affected by the mood of society. The overall social mood with respect to a given company might be one of the important variables which affect the stock price of that company. Nowadays, the emergence of online social networks makes available large amounts of mood data. Therefore, incorporating information from social media with the historical prices can improve the predictive ability of models.

The goal of our research is to develop a model to predict the stock price movement (whether the price will be up or down) using information from social media (Message Board). In our proposed method, a model that predicts the stock value at t using features derived from information at $t - 1$ and $t - 2$, where t stands for a transaction date, will be trained by supervised machine learning. Apart from the mood information, the stock prices are affected by many factors such as microeconomic and macroeconomic factors. However, this research only focuses on how the mood information from social media can be used to predict the stock price. We will mainly aim at extracting the mood information by sentiment analysis on social

media data. Then, these sentiments will be integrated into a model to predict stocks. To achieve this goal, discovering the topics and sentiments in a large amount of social media is very important to get the opinions of investors. However, sentiment analysis on social media is difficult. The text is usually short, contains many misspellings, uncommon grammar constructions and so on. In addition, the literature shows conflicting results in sentiment analysis for stock market prediction. Some researchers report that sentiments from social media have no predictive capabilities (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001), while other researchers have reported either weak or strong predictive capabilities (Bollen, Mao, & Zeng, 2011). Therefore, how to use opinions in social media for stock price predictions is still an open problem.

One contribution of this paper is that we propose a novel feature 'topic-sentiment' to improve the performance of stock market prediction. It is important to recognize what topics are discussed in social media and how people feel about these topics. The 'topic-sentiment' feature, which represents the sentiments of the specific topics of the company (product, service, dividend and so on), are used for prediction of stock price movement. This feature is obtained in two ways: by using the existing topic model called the joint sentiment/topic model (JST) and by our own proposed method. The extracted topics and sentiments in the former method are hidden (latent), whereas not hidden in the latter. To the best of our knowledge, this is the first research trying to extract topics and sentiments simultaneously and utilize them for stock market prediction. Another contribution is a large scale evaluation. The effectiveness of the sentiments in social media in stock market prediction is still uncertain because a

* Corresponding author. Tel.: +81 80 2956 5927.

E-mail addresses: nhthien8x@gmail.com, [nhthien@jaist](mailto:nhthien@jaist.ac.jp).[c.jp](http://jaist.ac.jp) (T.H. Nguyen), kshirai@jaist.ac.jp (K. Shirai), julien.velcin@univ-lyon2.fr (J. Velcin).

relatively small data was used for evaluation in the previous work. This paper investigates whether the sentiments in the social media are really useful on the test data containing many stocks and transaction dates.

The rest of the paper is organized as follows. **Section 2** introduces some previous approaches on sentiment analysis for stock prediction. **Section 3** describes our dataset. **Section 4** describes our proposed method. We also propose a novel feature for stock prediction based on the topics and the sentiments associated with them. **Section 5** assesses the results of the experiments. Finally, **Section 6** concludes our contribution.

2. Related work

Stock market prediction is one of the most attracted topics in academic as well as real life business. Many researches have tried to address the question whether the stock market can be predicted. Some of the researches were based on the random walk theory and the Efficient Market Hypothesis (EMH). According to the EMH (Fama, 1991; Fama, Fisher, Jensen, & Roll, 1969), the current stock market fully reflects all available information. Hence, price changes are merely due to new information or news. Because news in nature happens randomly and is unknowable in the present, stock prices should follow a random walk pattern and the best bet for the next price is current price. Therefore, they are not predictable with more than about 50% accuracy (Walczak, 2001). On the other hand, various researches specify that the stock market prices do not follow a random walk, and can be predicted at some degree (Bollen et al., 2011; Qian & Rasheed, 2007; Vu, Chang, Ha, & Collier, 2012). Degrees of directional accuracy at 56% hit rate in the predictions are often reported as satisfying results for stock predictions (Schumaker & Chen, 2009b; Si, Mukherjee, Liu, Li, Li, & Deng, 2013; Tsibouris & Zeidenberg, 1995).

Besides the efficient market hypothesis and the random walk theories, there are two distinct trading philosophies for stock market prediction: fundamental analysis and technical analysis. The fundamental analysis studies the company's financial conditions, operations, macroeconomic indicators to predict stock price. On the other hand, the technical analysis depends on historical and time-series prices. Price moves in trends, and history tends to repeat itself. Some researches have tried to use only historical prices to predict the stock price (Cervelló-Royo, Guijarro, & Michniuk, 2015; Patel, Shah, Thakkar, & Kotecha, 2015a, 2015b; Ticknor, 2013; Zuo & Kita, 2012a, 2012b). To discover the pattern in the data, they used Bayesian network (Zuo & Kita, 2012a, 2012b), time-series method such as Auto Regressive model, Moving Average model (Patel et al., 2015a, 2015b), Auto Regressive Moving Average model (Zuo & Kita, 2012a) and so on.

While these previous methods did not consider the sentiments on the social media, in this paper our work aims at incorporating them to improve the performance of the stock market prediction.

Most of the research tried to predict only one stock (Bollen et al., 2011; Qian & Rasheed, 2007; Si et al., 2013) and the number of instances (transaction dates) in a test set is very low such as 14 or 15 instances (Bollen et al., 2011; Vu et al., 2012). With only a few instances in the test set, the conclusion might be insufficient. To the best of our knowledge, there is no research showing a good prediction result on a data consisting of many stocks in a long time period. Our research tried to solve this issue by predicting 18 stocks over a period of one year.

2.1. Use of opinions from text for stock market prediction

Sentiment analysis has been found to play a significant role in many applications such as product reviews and restaurant reviews (Liu & Zhang, 2012; Pang & Lee, 2008). There are some researches

trying to apply sentiment analysis on an information source to improve the stock prediction model (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). There are two main sources from which authors have incorporated information aggregated from textual content into financial models. In the past, the main source was the news (Schumaker & Chen, 2009a, 2009b), and in recent years, social media sources. Then, these sentiments are integrated into prediction models. A simple approach is combining the textual content with the historical prices through the linear regression model.

Most of the previous work primarily used the bag-of-words as text representation that are incorporated into the prediction model. Schumaker and Chen (2009b) tried to use different textual representations such as bag-of-words, noun phrases and named entities for financial news. Then this information was integrated with linear regression and support vector machine regression as predictive models. They applied their models to estimate a discrete stock price 20 min. after a news article was released. The results show 0.04261 mean square error, 57.1% directional accuracy, and 2.06% return in a simulated trading engine. However, the textual representations are just the words or named entity tags, not exploiting so much about the mood information.

Antweiler and Frank (2004) used naive Bayes to classify the messages from message boards into three classes: buy, hold and sell. The number of relevant messages in these three classes was aggregated into a single measure of bullishness. They investigated three aggregation functions as a number of alternatives to bullishness. They were integrated into the regression model. However, they concluded that their model does not successfully predict stock returns.

Zhang, Fuehres, and Gloor (2011) measured collective hope and fear on each day and analyzed the correlation between these indices and the stock market indicators. They used the mood words to tag each tweet as fear, worry, hope and so on. They concluded that the emotional tweet percentage significantly negatively correlated with Dow Jones, NASDAQ and S&P 500, but had significant positive correlation to VIX. However, they did not use their model to predict the stock price values.

Two mood tracking tools, OpinionFinder and Google Profile of Mood States, were used to analyze the text content of daily Twitter (Bollen et al., 2011). The former measures positive and negative mood. The latter measures mood in terms of six dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). They used the Self Organizing Fuzzy Neural Network model to predict DJIA values. The results show 86.7% direction accuracy (up or down), Mean Absolute Percentage Error 1.79%. However, their test period is very short (from December 1 to December 19, 2008). Even though, they achieved high accuracy, there are only 15 transaction dates in their test set. With such a short period, it might not be sufficient to conclude the effectiveness of their method.

Xie, Passonneau, Wu, and Creamer (2013) proposed a novel tree representation based on semantic frame parsers. They indicated that this representation performed significantly better than bag-of-words. By using stock prices from Yahoo Finance, they annotated all the news with labels in a transaction date as going up or down categories. However, the weakness of this assumption is that all the news in one day will have the same category. In addition, this becomes a document classification problem, not stock prediction.

Rechenthin, Street, and Srinivasan (2013) incorporated Yahoo Finance Message Board into the stock movement prediction. They tried to use various classification models to predict stock. They used the explicit sentiments and predicted sentiments obtained by a classification model with the bag-of-words and meta-features.

A keyword-based algorithm was proposed to identify the sentiment of tweets as positive, neutral and negative for stock prediction (Vu et al., 2012). Their model achieved around 75% accuracy. However, their test period is very short, from 8th to 26th in September, 2012 which contains only 14 transaction dates.

[Si et al. \(2013\)](#) developed a non-parametric topic model for Twitter messages to predict the stock market. They proposed a continuous Dirichlet Process Mixture (cDPM) model to learn the daily topic set. Then, a sentiment time series was built based on these topics. The advantage of this method is that the model estimates the number of topics inherent in the data itself. However, the time period of their dataset is quite short, only three months.

A series of the previous work discussed in this subsection tried to extract the overall opinion or sentiment of the document. However, the opinions are often expressed toward the topics or aspects. For the prediction of stock prices, it is important to know on which topics of the company people have positive or negative opinions. In our proposed method, the sentiments of the topics or aspects are extracted, then they are incorporated into the stock prediction models.

Next subsection will discuss some related work for the identification of aspect-oriented sentiments.

2.2. Aspect based sentiment analysis

There are some researches trying to extract both topic and sentiment for other domains such as online product review, restaurant review and movie review dataset ([Dermouche, Kouas, Velcin, & Loudcher, 2015](#)). [Jo and Oh \(2011\)](#) proposed ASUM model for extracting both aspect and sentiment for online product review dataset. The model assumes that all words within a sentence are generated from one topic.

The joint sentiment/topic model (JST) was proposed to detect sentiment and topic simultaneously for movie review dataset ([Lin & He, 2009](#)). This model assumes that each word is generated from a joint topic and sentiment distribution. Because this model can extract topic and sentiment simultaneously, we will use it to extract topic-sentiment features.

[Lakkaraju, Bhattacharyya, Bhattacharya, and Merugu \(2011\)](#) proposed the FACTS, CFACTS, FACTS-R, and CFACTS-R model to perform sentiment analysis on a product review data. These models assume a word fallen into three categories: facet word, sentiment word or other category (background word, stop word, function word, etc.). Based on its category, a word is generated from corresponding facet, sentiment or background distribution. In addition, they introduced a window, which is a contiguous sequence of words. All facet words within a window are assumed to be derived from the same facet topic, and all sentiment words from the same sentiment topic.

[Zhao, Jiang, Yan, and Li \(2010\)](#) proposed the MaxEnt-LDA hybrid model to jointly discover both aspects and aspect-specific opinion words on a restaurant review dataset. Besides the general opinion words, they considered the aspect-specific opinion words. Therefore, a word is fallen into five categories: background, specific aspect, general aspect, specific opinion and general opinion. Based on these categories, words are generated from corresponding distributions.

The above methods tried to extract the hidden (latent) topic-sentiment associations. In our proposed method, both hidden and non-hidden topic-sentiment are considered by using JST topic model and proposed algorithms that will be discussed in [Sections 4.5](#) and [4.6](#), respectively.

3. Dataset

We used two datasets for our stock prediction model. The first one is the historical price dataset, and the second one is the mood information dataset.

3.1. Historical prices

Historical prices are extracted from Yahoo Finance for the 18 stocks. The list of stock quotes and company names is shown in the [Table 1](#). For each transaction date, there are open, high, low, close

Table 1
Quotes and company names.

Stocks	Company names
AAPL	Apple Inc.
AMZN	Amazon.com Inc.
BA	The Boeing Company
BAC	Bank of America Corporation
CSCO	Cisco Systems Inc.
DELL	Dell Inc.
EBAY	eBay Inc.
ETFC	E Trade Financial Corporation
GOOG	Google Inc.
IBM	International Business Machines Corporation
INTC	Intel Corporation
KO	The Coca-Cola Company
MSFT	Microsoft Corporation
NVDA	NVIDIA Corporation
ORCL	Oracle Corporation
T	AT&T Inc.
XOM	Exxon Mobil Corporation
YHOO	Yahoo! Inc.

and adjusted close prices. The adjusted close prices are the close prices which are adjusted for dividends and splits. The adjusted close price is often used for stock market prediction as in other researches ([Rechenthin et al., 2013](#)). Therefore, we chose it as the stock price value for each transaction date.

3.2. Message board dataset

To get the mood information of the stocks, we collected the 18 message boards of the 18 stocks from Yahoo Finance Message Board for a period of one year (from July 23, 2012 to July 19, 2013)¹. On the message boards, users usually discuss company news, prediction about stock going up or down, facts, comments (usually negative) about specific company executives or company events. In 15.6% messages in this dataset, when users posted messages on these message boards, they annotated each message as one of the following sentiment tags: Strong Buy, Buy, Hold, Sell and Strong Sell. There are two kinds of messages. The first one is the messages created by starting a new topic. The other is reply messages to existing messages. Most of users' posts are reply messages. They form a complicated communication network. In our research, however, we treated all messages independent from each other.

[Fig. 1](#) shows an example message from AAPL Message Board. In this message, on July 6, 2012 a username "keepshorting" posted the message "Looks like the competition is heating up. \$199 tablet, what is next? \$999 laptops and then \$499 laptops? the margins are impossible to keep up. impossible folks." to reply to another message of another user. In addition, this user selected the sentiment for this stock as "Strong Sell".

The stock market is not opened at the weekend and holiday. To assign the messages to the transaction dates, the messages which were posted from 4 pm of the previous transaction date to 4 pm of the current transaction date will belong to the current transaction. We choose 4 pm because that is the time of closing transaction. There are 249 transaction dates from the one year period of our dataset. [Table 2](#) summarizes the statistics of our dataset for each transaction date about the min, median, mean, max of the number of messages and the mean of the number of the existing sentiments annotated by users.

Some previous works used Twitter as the mood information source for sentiment analysis related to a particular stock. There are

¹ The AAPL message board has the highest number of messages. Because of the limitation on the number of web pages, we can only collect for a period of seven months for this stock.

Fig. 1. A message from AAPL message board.

Table 2
Statistics of our dataset for each transaction date.

Stocks	The number of messages				Mean of the number of human sentiments
	Min	Median	Mean	Max	
AAPL	0	1093	1678	11220	350
AMZN	24	154	192	1963	28
BA	46	173	203	1053	16
BAC	94	282	343	1366	49
CSCO	69	247	274	972	10
DELL	0	18	42	587	10
EBAY	1	17	29	267	3
ETFC	2	42	56	315	12
GOOG	10	69	93	1305	16
IBM	3	14	20	195	3
INTC	37	177	200	958	29
KO	0	6	8	89	2
MSFT	27	139	172	815	53
NVDA	10	65	80	410	11
ORCL	5	67	79	372	6
T	10	52	59	251	8
XOM	10	37	44	202	4
YHOO	22	121	141	860	27

some reasons why in our research Twitter is not chosen as a mood source. The first one is the information in Twitter seems to be messier than that in the Message Board. In the Twitter, users discuss about many things. Even though tweets can be filtered by some rules such as using hashtag (#AAPL, \$AAPL and so on) to find relevant tweets, the lack of consistency among posters in hashtag use might be problematic. Furthermore, the high level of noise makes finding a post related to a specific stock difficult. The second reason is the way to collect tweets. There are two ways to collect tweets from the Twitter. The first one is from the Twittter Searching API. This only allows searching tweets from one week in the past for free. The other way is using Twitter Streaming API. It allows collecting the real time tweets rather than search from the history. However, to collect tweets in one year period, it takes one year. Those make difficult to gather a large amount of data from the Twitter. Finally, there are no explicit sentiments annotated by posters in the Twitter. There is no way to compare between human sentiment and automatic sentiment extraction.

However, as in other mood information sources, the messages on the Message Board are also messy. The text is usually short, contains many misspellings, uncommon grammar constructions and so on. Moreover, the false and unrelated information also exists.

4. Methods for stock movement prediction

The Support Vector Machine (SVM) has long been recognized as being able to efficiently handle high dimensional data and has been shown to perform well on classification (Joachims, 1998; Nguyen & Shirai, 2013). Therefore, we chose the SVM with the linear kernel as the prediction model. To assess the effectiveness of sentiment analysis on the message boards, six sets of features are designed. The first

Table 3
Features of the prediction model.

Method	Features
Price only	$price_{t-1}$, $price_{t-2}$
Human sentiment	$price_{t-1}$, $price_{t-2}$, $Hsent_{i,t}$, $Hsent_{i,t-1}$
Sentiment classification	$price_{t-1}$, $price_{t-2}$, $Csent_{i,t}$, $Csent_{i,t-1}$
LDA-based method	$price_{t-1}$, $price_{t-2}$, $lda_{i,t}$, $lda_{i,t-1}$
JST-based method	$price_{t-1}$, $price_{t-2}$, $jst_{i,j,t}$, $jst_{i,j,t-1}$
Aspect-based sentiment	$price_{t-1}$, $price_{t-2}$, $Asent_{i,t}$, $Asent_{i,t-1}$, $I_{i,t}$, $I_{i,t-1}$

one used only the historical prices. The other methods incorporated the mood information into the prediction model. All the feature values were scaled into $[-1, 1]$. Table 3 summarizes our features used in the prediction model to predict the price movement at the transaction date t . The details of each feature will be explained in the next subsections.

4.1. Price only

In this method, only historical prices are used to predict the stock movement. The purpose of this method is to investigate whether there are patterns in the history of the stock or not. In addition, this model was used as a baseline to evaluate whether integration of the sentiments is effective by comparing with other sentiment models. Features used for the training of SVM are $price_{t-1}$ and $price_{t-2}$ which are the price movements (up, down) at the transaction dates $t - 1$ and $t - 2$, respectively.

4.2. Human sentiment

In addition to historical prices, this model integrated the sentiments annotated by human into the prediction model. As discussed in Section 3.2, in 15.6% of the MessageBoard dataset, the users explicitly select a sentiment label with their posts. These sentiment labels are "strong buy", "buy", "hold", "sell" and "strong sell". Instead of using all the messages, we tried to use only the messages with annotated sentiments by the users, and discard the other messages. From these messages, we used only the explicit sentiment and remove other information such as message content. The purpose of this method is that how mood annotated by human can be used to predict the stock. Because the sentiments are annotated by human, this feature is one of the strongest features for stock prediction.

For each transaction date t , the percentage of each class (Strong Buy, Buy, Hold, Sell, and Strong Sell) was calculated. The percentage of a class is the number of messages having sentiments as that class label divided by the number of messages in the current transaction date t . Then, we integrated them into the prediction model.

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $Hsent_{i,t}$ and $Hsent_{i,t-1}$. $Hsent_{i,t}$ and $Hsent_{i,t-1}$ are the percentages of the number of messages belonging to the sentiment class i ($i \in \{\text{Strong Buy}, \text{Buy}, \text{Hold}, \text{Sell}, \text{Strong Sell}\}$).

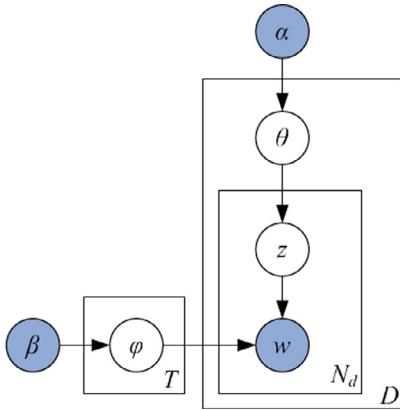


Fig. 2. Graphical model representation of LDA.

Buy, Hold, Sell, and Strong Sell) at the transaction dates t and $t - 1$, respectively.

4.3. Sentiment classification

To utilize the remaining 84.4% of the messages without the explicit sentiments, we tried to build a model to extract the sentiments for those messages. A classification model was trained from the messages with annotated sentiments on the training dataset. Then it was used to classify the remaining messages into five classes (Strong Buy, Buy, Hold, Sell, and Strong Sell).

We removed the stop words from the messages. Then, all the words are lemmatized by the Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014). The feature representation is the bag-of-words from the title and content of the message. The feature weighting is TF-IDF. We chose SVM with the linear kernel as the classification model.

As in the human sentiment feature, we also calculated the percentage of the number of messages of each class for each transaction date. Features used for the training of SVM are $price_{t-1}$, $price_{t-2}$, $Csent_{i,t}$ and $Csent_{i,t-1}$. $Csent_{i,t}$ and $Csent_{i,t-1}$ are similar to $Hsent_{i,t}$ and $Hsent_{i,t-1}$, but both messages with human annotated sentiment and automatically classified sentiments are used to calculate the percentages of the number of messages belonging to the sentiment class i .

4.4. LDA-based method

In this model, we consider each message as a mixture of hidden topics. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Therefore, we choose the LDA as a simple topic model to discover these hidden topics². Fig. 2 shows the graphical model representation of LDA. Notations in Fig. 2 are shown in Table 4.

We removed the stop words from messages. Then, all the words are lemmatized by the Stanford CoreNLP. We train the LDA on the training set, and infer the topics for unseen messages on the test set. Topics are inferred by the Gibbs Sampling with 1000 iterations. We chose 50 as the number of topics. After that, the probability of each topic for each message is calculated. Next, for each transaction date t , the probability of each topic is defined as the average of the probabilities of that topic in the messages belonging to that transaction date. Then we integrated these probabilities into the prediction model.

Table 4
Notations in LDA.

Notation	Definition
α, β	Hyperparameters
φ	The distribution over words
T	The number of topics
θ	The message specific topic distribution
z	A topic
w	A word in the message d
N_d	The number of words in the message d
D	The number of messages

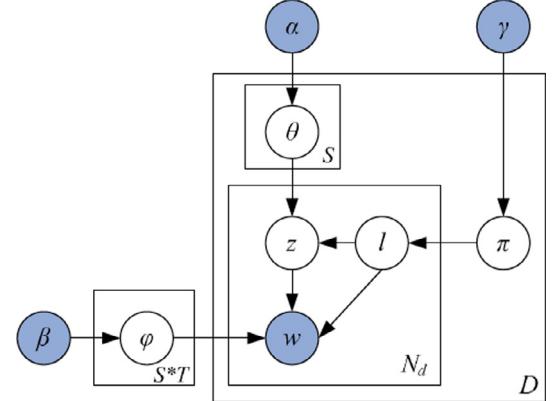


Fig. 3. Graphical model representation of JST.

Table 5
Notations in JST.

Notation	Definition
α, β, γ	Hyperparameters
φ	The distribution over words
T	The number of topics
S	The number of sentiments
θ	The message and sentiment specific topic distribution
z	A topic
w	A word in the message d
l	A sentiment label
π	The message specific sentiment distribution
N_d	The number of words in the message d
D	The number of messages

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $lدا_{i,t}$ and $lدا_{i,t-1}$. $lدا_{i,t}$ and $lدا_{i,t-1}$ are the probabilities of the topic i ($i \in \{1, \dots, 50\}$) for the transaction dates t and $t - 1$.

4.5. JST-based method

The opinion is often expressed on a topic or aspect. When people post the message on the social media to express their opinion for a given stock, they tend to talk their opinions for a certain topic such as profit and dividend. Based on pairs of topic-sentiment, they would think that the future price of that stock goes up or down. From that intuition, we propose a new feature topic-sentiment for the stock prediction model. To extract pairs of topic-sentiment, we tried to use two kinds of models. The first one is a latent topic based model, the JST model (Lin & He, 2009). The second one is Aspect-based Sentiment model which will be discussed in the next Section 4.6.

We consider each message as a mixture of hidden topics and sentiments. The JST model was used to extract topics and sentiments simultaneously. Fig. 3 shows the graphical model representation of JST. Notations in Fig. 3 are shown in Table 5. In LDA model, there is only one document specific topic distribution for each document. In contrast, each document in JST is associated with S sentiment labels. Each

² We used the LDA implementation from the Mallet library.

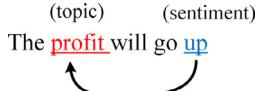


Fig. 4. An example sentence with topic and its sentiment.

of sentiment labels is associated with a document specific topic distribution with the same number of topics. A word in the document is drawn from distribution over the words defined by the topic and sentiment label.

After removal of stop words and lemmatization, the JST model is trained from the training set, and topics on the test set are inferred by the Gibbs Sampling with 1000 iterations. We chose 50 as the number of topics and 3 as the number of sentiments. Next, the joint probability of each pair of topic and sentiment is calculated for each message. After that, for each transaction date t , the joint probability of each topic-sentiment pair is defined as the average of the joint probabilities of that in the messages belonging to that transaction date. Then we integrated these probabilities into the prediction model.

Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $jst_{i,j,t}$ and $jst_{i,j,t-1}$. $jst_{i,j,t}$ and $jst_{i,j,t-1}$ are the joint probabilities of the sentiment i ($i \in \{1, 2, 3\}$) and topic j ($j \in \{1, \dots, 50\}$) for the transaction dates t and $t - 1$.

4.6. Aspect-based sentiment

Instead of considering the mixtures of hidden topics and sentiments as in the previous model, in this model the mixtures are not

hidden. Each message is represented as a list of topics and their corresponding sentiment values. In our proposed method, the topic is the consecutive nouns in the sentence. For example, the message "The profit will go up." contains the topic "profit" and a positive sentiment "up" for that topic as in Fig. 4.

We propose a new model to calculate the sentiment values of the topics in a sentence. For each message, the sentences are split. Then, we used the Stanford CoreNLP for POS tagging and lemmatization of each word in each sentence. First, we extracted the topics in the training dataset by using the algorithm shown in Fig. 5. We extract the consecutive nouns as the topics in the sentence. To eliminate rare topics, topics occurring less than 10 times are removed from the list of the topics. Next, based on the topic list, we extracted their sentiment values in each sentence by using the algorithm shown in Fig. 6. For each sentence, opinion words are identified based on the list of opinions from SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010). SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns each word three sentiment scores: positivity, objectivity and negativity. We combined scores of positivity and negativity into a single opinion value. The closer between the topic phrase and the opinion word, the higher affection of that opinion on the topic phrase. Therefore, the sentiment value of a topic phrase in a sentence is the summation of overall opinion values divided by their distance to that topic.

For each message, the sentiment value of each topic is defined as the average of the sentiment scores of that topic in the sentences. Finally, for each transaction date t , the sentiment value for each topic is defined as the average of the sentiment

Input: Training dataset
Output: List of topics of this dataset
1 Extract consecutive nouns in each sentence as a topic ;
2 Remove topics that appear less than 10 times in the training dataset ;

Fig. 5. Algorithm for extracting topics from dataset.

Input: A sentence
Output: List of pairs $(topic, sentimentValue)$ for this sentence
1 Extract topics in the sentence (Based on the list of topics extracted from the algorithm shown in Figure 5) ;
2 Extract opinion words in the sentence by using SentiWordNet ;
3 for each topic t_i in the sentence do
4 for each opinion o_j in the sentence do
5 Calculate $distance(t_i, o_j)$ = position distance between topic t_i and opinion word o_j ;
6 Get pos_score , neg_score of opinion o_j from SentiWordNet ;
7 Calculate $opinionValues(o_j) = \frac{pos_score - neg_score}{pos_score + neg_score}$;
8 $sentimentValue_{t_i} = sentimentValue_{t_i} + \frac{opinionValues(o_j)}{distance(t_i, o_j)}$;
9 Add $(t_i, sentimentValue_{t_i})$ to the list of pairs $(topic, sentimentValue)$
10 end
11 end

Fig. 6. Algorithm for extracting topics and their sentiment values.

Table 6
Results of accuracies of 18 stocks.

Stocks	Baseline models				Our models	
	Price only	Human sentiment	Sentiment classification	LDA-based method	JST-based method	Aspect-based sentiment
AAPL	0.3951	0.5679	0.4938	0.5802	0.5802	0.5432
AMZN	0.4605	0.4868	0.4605	0.5132	0.5921	0.7105
BA	0.6316	0.6053	0.5132	0.5526	0.6316	0.5921
BAC	0.5658	0.5921	0.5658	0.5526	0.5658	0.4474
CSCO	0.5526	0.4474	0.5263	0.4737	0.5132	0.4605
DELL	0.5395	0.5921	0.4737	0.5132	0.4342	0.6447
EBAY	0.5921	0.4605	0.4605	0.5658	0.4079	0.5789
ETFC	0.5789	0.5921	0.5789	0.4868	0.4342	0.5526
GOOG	0.5000	0.5658	0.5789	0.5658	0.5395	0.5263
IBM	0.4868	0.4737	0.4868	0.5395	0.4474	0.5526
INTC	0.4474	0.4605	0.4342	0.5000	0.4868	0.5263
KO	0.4079	0.4868	0.5132	0.5658	0.5132	0.4474
MSFT	0.5789	0.6579	0.5921	0.5526	0.5526	0.5263
NVDA	0.6053	0.5789	0.6184	0.3947	0.5000	0.5395
ORCL	0.4868	0.5263	0.5263	0.5921	0.5000	0.5395
T	0.5526	0.4737	0.4868	0.5000	0.5658	0.5132
XOM	0.4868	0.6447	0.4868	0.4342	0.5658	0.5395
YHOO	0.5526	0.5526	0.5395	0.5263	0.4474	0.5526
AVERAGE	0.5234	0.5425	0.5187	0.5227	0.5154	0.5441

values of that topic in the messages belonging to that transaction date.

In addition to the sentiment values of the topics, the importance of the topics for each transaction date were also considered. Intuitively, some topics have more impact on the prediction than others. If a topic was discussed in many messages, it might be an important topic in the given transaction date. The importance of a topic i in a transaction date t was calculated as in Eq. (1). It is defined as the fraction between the number of messages containing the current topic i in the transaction date t and the number of messages in that transaction date.

$$I_{i,t} = \frac{N_{i,t}}{N_t} \quad (1)$$

where:

$I_{i,t}$: the importance of topic i in the transaction date t .

$N_{i,t}$: the number of messages containing the topic i in the transaction date t .

N_t : the number of messages in the transaction date t .

The sentiment scores of the topics at the transaction date level and their important were used in the prediction model. Features used for training SVM are $price_{t-1}$, $price_{t-2}$, $Asent_{i,t}$, $Asent_{i,t-1}$, $I_{i,t}$ and $I_{i,t-1}$. $Asent_{i,t}$ and $Asent_{i,t-1}$ are the sentiment values of the topic i at the transaction dates t and $t - 1$. While, $I_{i,t}$, $I_{i,t-1}$ are the importance of the topic i at the transaction dates t and $t - 1$.

5. Evaluation

5.1. Experiment setup

We divided the time series into two parts: the period from July 23, 2012 to March 28, 2013 for training containing 171 transaction dates, and April 01, 2013 to July 19, 2013 for testing containing 78 transaction dates³. We assigned each transaction date a label (up, down) by comparing its price with the previous transaction date's price. The performance is evaluated using the Accuracy metric. Accuracy is the proportion of true results in the test set.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad (2)$$

³ For AAPL: the period July 06–October 01, 2012 for training containing 61 transaction dates, and November 12, 2012–March 13, 2013 for test containing 83 transaction dates.

where:

tp : the number of samples correctly categorized for positive samples.

tn : the number of samples correctly rejected for the negative samples.

fp : the number of samples incorrectly categorized for the positive samples.

fn : the number of samples incorrectly rejected for the negative samples.

5.2. Experiment results

The results of accuracy measure are shown in Table 6. In addition to the result of each stock, we also calculated the average of 18 stocks for each model for easy comparison. Using Aspect-based sentiment feature achieved the best result with 54.41% average accuracy for 18 stocks. As discussed in Section 2, degrees of accuracy of 56% hit rate are often reported as satisfying results for stock prediction. In addition, the number of instances (transaction dates) in test set of most of other researches is small, and the number of stock is usually only one. In contrast, the advantage of this work is that we used the training and test data on a long period (one year) containing many instances, and for many stocks (18 stocks). For some stocks, the accuracies are quite high, such as 71.05% for AMZN stock, 64.47% for DELL stock and so on.

To assess the effectiveness of integrating mood information, we compare our Aspect-based sentiment method with the Price only method. The results show that the model using mood information outperforms 2.07% on the average accuracy than the model without mood. Furthermore, comparing the Human sentiment with Price only method, it indicated that the prediction accuracy is improved 1.91% by using the sentiments annotated by human. Therefore, we can conclude that integration of the sentiments of both overall documents and specific topics from social media could help to improve the stock market prediction.

To assess the effectiveness of automatic sentiment analysis and human sentiment, we compare our Aspect-based sentiment method with the Human sentiment method. The results show that our automatically extracted sentiment is slightly higher than using the sentiment annotated by human. Therefore, our method is comparable to the human sentiment method. Note that the advantage of our method

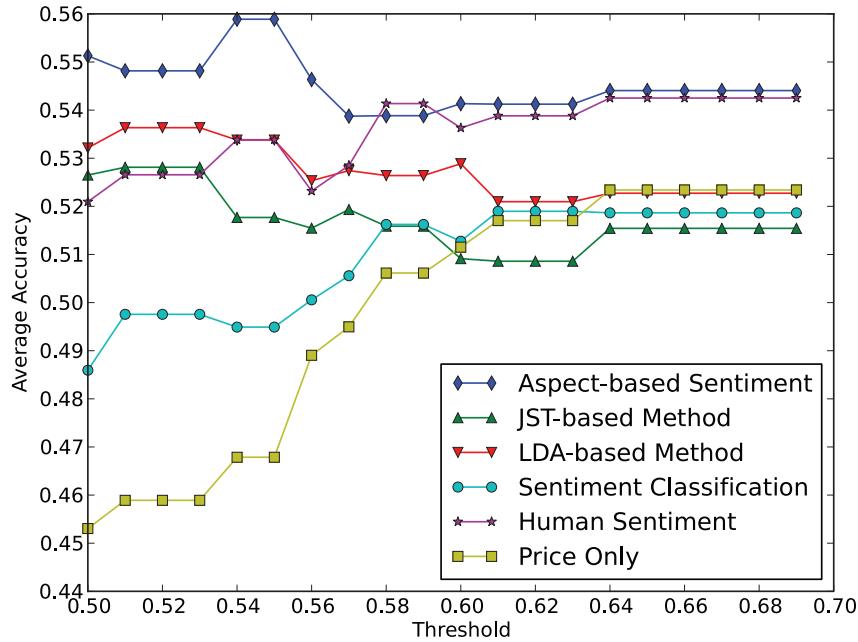


Fig. 7. Comparison of the models for different threshold α .

is that it can be applicable for other social media without human annotated sentiment such as the Twitter.

The Aspect-based sentiment method outperformed over 2.54%, 2.14% and 2.87% on average accuracy compared to Sentiment classification, LDA-based method and JST-based method, respectively. The LDA-based and JST-based method seem to be not successful in this experiment. The limitation of these methods is that we have to specify the number of hidden topics in LDA and the number of hidden topics and sentiments in JST. In our experiment, we specify 50 topics for all of the stock. This assumption would be not appropriate in general. For the individual stock, the number of discussed topics depends on the content of the messages. Therefore, the appropriate number of hidden topics may be varied for different stocks. However, there is no way to determine the number of topics in the model of LDA and JST. One of the solutions is a grid search trying different number of topics and finding the best value. However, since the running time of the Gibbs Sampling depends on the size of the dataset, it takes very long time to run it repeatedly on a big dataset of 18 stocks for a long period. Therefore, a grid search cannot be tested in our experiment.

To assess the effectiveness of topic-sentiment feature, we compare our Aspect-based sentiment method with Sentiment classification method (using only sentiment information) and LDA-based method (using only topic information). The results indicate that using topic-sentiment feature is better than using only sentiments by 2.54% accuracy, and using only topics by 2.14% accuracy. Therefore, understanding on which topics the sentiments are expressed is useful for stock market prediction. In other words, the topic-sentiment feature is better than only topic or sentiment feature.

Although the sentiment information is effective for the stock prediction on average, in the comparison on the individual stocks, the model with sentiment analysis is worse than the price only model for several stocks. There are many possible reasons for it. As discussed in Section 2, the stock market is influenced by many factors. Some proposed that they are not predictable with more than 50% accuracy. One reason is that the sentiment might not be a factor which causes the stock price moving. Another reason is that even though sentiment might be one of the factors which affect price moving, the extracted sentiments from the Message Boards do not reflect the price because of the messy, fault comment or fault prediction of human when they post the messages.

A simple assumption about the effectiveness of sentiment feature is that the sentiment analysis may not provide any additional information if the stock movement can be predicted well by the historical price only. If the accuracy of the price only model is high, there are trends and historical repetition in the stock. In such cases, only historical prices might be enough to predict, and integration of the sentiment may not improve the accuracy much. On the other hand, if the accuracy of the price only model is low, the stock seems to have no pattern in its history. For such stocks, the use of sentiment may be effective for the prediction.

To investigate the above assumption, we compare the models from another point of view. First, we define a threshold α . If the accuracy of the stock in the Price only method ($A_{PriceOnly}$) is higher than α , this stock is discarded from the evaluation. In other words, we compared the average accuracy for the stocks where $A_{PriceOnly} < \alpha$. Fig. 7 shows the average accuracies against various thresholds. It is found that the difference between the models with and without sentiment information becomes greater when α is set smaller. At the threshold 50%, using our Aspect-based sentiment model improved the accuracy over 9.83% compared to Price only, over 3.03% compared to Human sentiment method. In addition, in most of the thresholds, our method achieved the best accuracy compared with other methods.

6. Conclusion & future work

Stock price prediction is a challenging task because the stock prices are affected by many factors. This paper presents the novel method to integrate the sentiments in social media for the prediction of stock price movement. The contribution of this study can be summarized as follows. First, while the overall sentiments in the documents are considered in the previous research, this research proposed a method using the sentiment of the topic for stock market prediction. Second, we proposed two methods to capture these topic-sentiment associations. One is JST-based method that relies on the existing topic model, the other is Aspect-based sentiment method where the topics and sentiments are identified by the proposed method. Finally, this is the first research to show the effectiveness of incorporation of the sentiment analysis by investigation on a large scale test data. From a practical point of view, although the average accuracy is only 54.41%, the proposed method can predict the stock

price movement with more than 60% accuracy for a few stocks, and performs much better than other methods for the stocks that are difficult to predict with only past prices.

A limitation of this research is that we specified the number of topics and sentiment beforehand for the LDA and JST-based method. To overcome this weakness, a non-parametric topic model that can infer the number of topics and sentiments automatically is useful to extract the topic and sentiment simultaneously for the stock prediction. This will be done in our future work.

The current model only predicts if the stock price is up or down. However, people may want to forecast drastic movement of the stock market. In that sense, the proposed model is insufficient. However, our model can be extended to predict the degree of the change by setting more fine grained classes such as 'great up', 'little up', 'little down', 'great down' and so on.

One of the weaknesses of our method is that only the historical prices and sentiments derived from social media are considered. In future, we will try to find and integrate more factors which can affect the stock prices to develop a more accurate stock prediction model. For example, co-variance between stocks, macroeconomic indicators and the financial conditions of the company, which can be guessed from the income statement, balance sheet and cash flow, are important factors to be considered in the stock prediction model.

References

- Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), 1259–1294.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation: vol. 10* (pp. 2200–2204).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications*, 42(14), 5963–5975.
- Dermouche, M., Kouas, L., Velcin, J., & Loudecher, S. (2015). A joint model for topic-sentiment modeling from text. In *ACM/SIGAPP symposium on applied computing (sac)* (pp. 819–824).
- Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, 46(5), 1575–1617.
- Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, 10(1), 1–21.
- Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 815–824). ACM.
- Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. Springer.
- Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the eleventh SIAM international conference on data mining* (pp. 498–509). SIAM /Omnipress.
- Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 375–384). ACM.
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).
- Nassiroussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: a systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
- Nguyen, T. H., & Shirai, K. (2013). Text classification of technical papers based on text segmentation. In *Natural language processing and information systems - 18th international conference on applications of natural language to information systems* (pp. 278–284).
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
- Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015a). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), 259–268.
- Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015b). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4), 2162–2172.
- Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1), 25–33.
- Rechenthin, M., Street, W. N., & Srinivasan, P. (2013). Stock chatter: using stock sentiment to predict price direction. *Algorithmic Finance*, 2(3), 169–196.
- Schumaker, R. P., & Chen, H. (2009a). A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5), 571–583.
- Schumaker, R. P., & Chen, H. (2009b). Textual analysis of stock market prediction using breaking financial news: the azfin text system. *ACM Transactions on Information Systems*, 27(2), 12:1–12:19.
- Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st annual meeting of the association for computational linguistics, volume 2: short papers* (pp. 24–29). The Association for Computer Linguistics.
- Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, 40(14), 5501–5506.
- Tsibouris, G., & Zeidenberg, M. (1995). Testing the efficient markets hypothesis with gradient descent algorithms. In *Neural networks in the capital markets* (pp. 127–136). Wiley: Chichester.
- Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3), 41–51.
- Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in Twitter. In *24th international conference on computational linguistics* (pp. 23–38).
- Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of Management Information Systems*, 17(4), 203–222.
- Xie, B., Passonneau, R. J., Wu, L., & Creamer, G. (2013). Semantic frames to predict stock price movement. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 873–883).
- Zhang, X., Fuehres, H., & Gloo, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26(0), 55–62.
- Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 56–65). Association for Computational Linguistics.
- Zuo, Y., & Kita, E. (2012a). Stock price forecast using Bayesian network. *Expert Systems with Applications: An International Journal*, 39(8), 6729–6737.
- Zuo, Y., & Kita, E. (2012b). Up/down analysis of stock index by using Bayesian network. *Engineering Management Research*, 1(2), 46–52.



Quantifying Trading Behavior in Financial Markets Using Google Trends

SUBJECT AREAS:

STATISTICAL PHYSICS,
THERMODYNAMICS AND
NONLINEAR DYNAMICS

APPLIED PHYSICS

COMPUTATIONAL SCIENCE

INFORMATION THEORY AND
COMPUTATION

Received
25 February 2013

Accepted
3 April 2013

Published
25 April 2013

Correspondence and
requests for materials
should be addressed to
T.P. (Tobias.Preis@
wbs.ac.uk)

* These authors
contributed equally to
this work.

Tobias Preis^{1*}, Helen Susannah Moat^{2,3*} & H. Eugene Stanley^{2*}

¹Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK, ²Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA, ³Department of Civil, Environmental and Geomatic Engineering, UCL, Gower Street, London, WC1E 6BT, UK.

Crises in financial markets affect humans worldwide. Detailed market data on trading decisions reflect some of the complex human behavior that has led to these crises. We suggest that massive new data sources resulting from human interaction with the Internet may offer a new perspective on the behavior of market participants in periods of large market movements. By analyzing changes in *Google* query volumes for search terms related to finance, we find patterns that may be interpreted as “early warning signs” of stock market moves. Our results illustrate the potential that combining extensive behavioral data sets offers for a better understanding of collective human behavior.

The increasing volumes of ‘big data’ reflecting various aspects of our everyday activities represent a vital new opportunity for scientists to address fundamental questions about the complex world we inhabit^{1–7}. Financial markets are a prime target for such quantitative investigations^{8,9}. Movements in the markets exert immense impacts on personal fortunes and geopolitical events, generating considerable scientific attention to this subject^{10–19}. For example, a range of recent studies have focused on modeling financial markets^{20–25} and on performing network analyses^{26–29}.

At their core, financial trading data sets reflect the myriad of decisions taken by market participants. According to Herbert Simon, actors begin their decision making processes by attempting to gather information³⁰. In today’s world, information gathering often consists of searching online sources. Recently, the search engine *Google* has begun to provide access to aggregated information on the volume of queries for different search terms and how these volumes change over time, via the publicly available service *Google Trends*. In the present study, we investigate the intriguing possibility of analyzing search query data from *Google Trends* to provide new insights into the information gathering process that precedes the trading decisions recorded in the stock market data.

A recent investigation has shown that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country³¹. Further studies exploiting the temporal dimension of *Google Trends* data have demonstrated that changes in query volumes for selected search terms mirror changes in current numbers of influenza cases³² and current volumes of stock market transactions³³. This demonstration of a link between stock market transaction volume and search volume has also been replicated using *Yahoo!* data³⁴. Choi and Varian³⁵ have shown that data from *Google Trends* can be linked to current values of various economic indicators, including automobile sales, unemployment claims, travel destination planning and consumer confidence. A very recent study has shown that Internet users from countries with a higher per capita GDP are more likely to search for information about years in the future than years in the past³⁶.

Here, we suggest that within the time period we investigate, *Google Trends* data did not only reflect the current state of the stock markets³³ but may have also been able to anticipate certain future trends. Our findings are consistent with the intriguing proposal that notable drops in the financial market are preceded by periods of investor concern. In such periods, investors may search for more information about the market, before eventually deciding to buy or sell. Our results suggest that, following this logic, during the period 2004 to 2011 *Google Trends* search query volumes for certain terms could have been used in the construction of profitable trading strategies.

Results

We analyze the performance of a set of 98 search terms. We included terms related to the concept of stock markets, with some terms suggested by the *Google Sets* service, a tool which identifies semantically related keywords. The set of terms used was therefore not arbitrarily chosen, as we intentionally introduced some financial bias. We explain our strategy based on changes in search volume with reference to the term *debt*, a

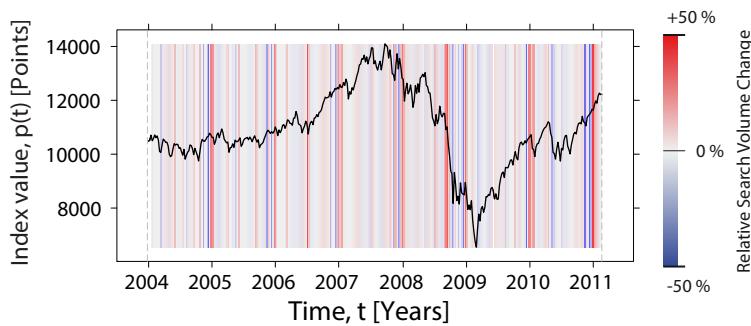


Figure 1 | Search volume data and stock market moves. Time series of closing prices $p(t)$ of the *Dow Jones Industrial Average* (DJIA) on the first day of trading in each week t covering the period from 5 January 2004 until 22 February 2011. The color code corresponds to the relative search volume changes for the search term *debt*, with $\Delta t = 3$ weeks. Search volume data are restricted to requests of users localized in the United States of America.

keyword with an obvious semantic connection to the most recent financial crisis, and overall the term which performed best in our analyses.

To uncover the relationship between the volume of search queries for a specific term and the overall direction of trader decisions, we analyze closing prices $p(t)$ of the *Dow Jones Industrial Average* (DJIA) on the first trading day of week t . We use *Google Trends* to determine how many searches $n(t-1)$ have been carried out for a specific search term such as *debt* in week $t-1$, where *Google* defines weeks as ending on a Sunday, relative to the total number of searches carried out on *Google* during that time. We find that search volume data change slightly over time due to *Google*'s extraction procedure. For each search term, we therefore average over three realizations of its search volume time series, based on three independent data requests in consecutive weeks. The variability of *Google Trends* data across different dates of access is irrelevant for our results, and it can be shown that the data are consistent with reported real world events (see Fig. S1 in the *Supplementary Information*).

To quantify changes in information gathering behavior, we use the relative change in search volume: $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ with $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \dots + n(t-\Delta t))/\Delta t$, where t is measured in units of weeks. In Fig. 1, we depict relative search volume changes for the term *debt*, and their relationship to DJIA closing prices.

To investigate whether changes in information gathering behavior as captured by *Google Trends* data were related to later changes in stock price in the period between 2004–2011, we implement a hypothetical investment strategy for a portfolio using search volume data, called ‘*Google Trends* strategy’ in the following. Profit can only be made in a trading strategy if at least some future changes in the stock price are correctly anticipated, in particular around large market

movements. We implement this strategy by selling the DJIA at the closing price $p(t)$ on the first trading day of week t , if $\Delta n(t-1, \Delta t) > 0$, and buying the DJIA at price $p(t+1)$ at the end of the first trading day of the following week. Note that mechanisms exist which make it possible to sell assets in financial markets without first owning them. If instead $\Delta n(t-1, \Delta t) < 0$, then we buy the DJIA at the closing price $p(t)$ on the first trading day of week t and sell the DJIA at price $p(t+1)$ at the end of the first trading day of the coming week. At the beginning of trading, we set the value of all portfolios to an arbitrary value of 1. If we take a ‘short position’—selling at the closing price $p(t)$ and buying back at price $p(t+1)$ —then the cumulative return R changes by $\log(p(t)) - \log(p(t+1))$. If we take a ‘long position’—buying at the closing price $p(t)$ and selling at price $p(t+1)$ —then the cumulative return R changes by $\log(p(t+1)) - \log(p(t))$. In this way, buy and sell actions have symmetric impacts on the cumulative return R of a strategy’s portfolio. In using this approach to analyze the relationship between *Google* search volume and stock market movements, we neglect transaction fees, since the maximum number of transactions per year when using our strategy is only 104, allowing a closing and an opening transaction per week. We of course do not dispute that such transaction fees would impact profit in a real world implementation.

In Fig. 2, the performance of the *Google Trends* strategy based on the search term *debt* is depicted by a blue line, whereas dashed lines indicate the standard deviation of the cumulative return from a strategy in which we buy and sell the market index in an uncorrelated, random manner (‘random investment strategy’). The standard deviation is derived from simulations of 10,000 independent realizations of the random investment strategy. Fig. 2 shows that the use of the *Google Trends* strategy, based on the search term *debt* and $\Delta t = 3$ weeks, would have increased the value of a portfolio by 326%. The

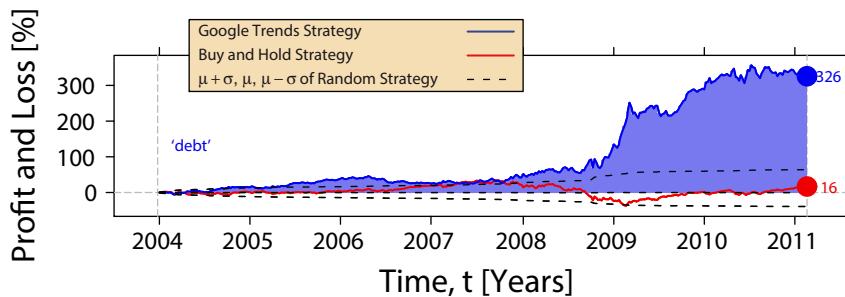


Figure 2 | Cumulative performance of an investment strategy based on *Google Trends* data. Profit and loss for an investment strategy based on the volume of the search term *debt*, the best performing keyword in our analysis, with $\Delta t = 3$ weeks, plotted as a function of time (blue line). This is compared to the “buy and hold” strategy (red line) and the standard deviation of 10,000 simulations using a purely random investment strategy (dashed lines). The *Google Trends* strategy using the search volume of the term *debt* would have yielded a profit of 326%.

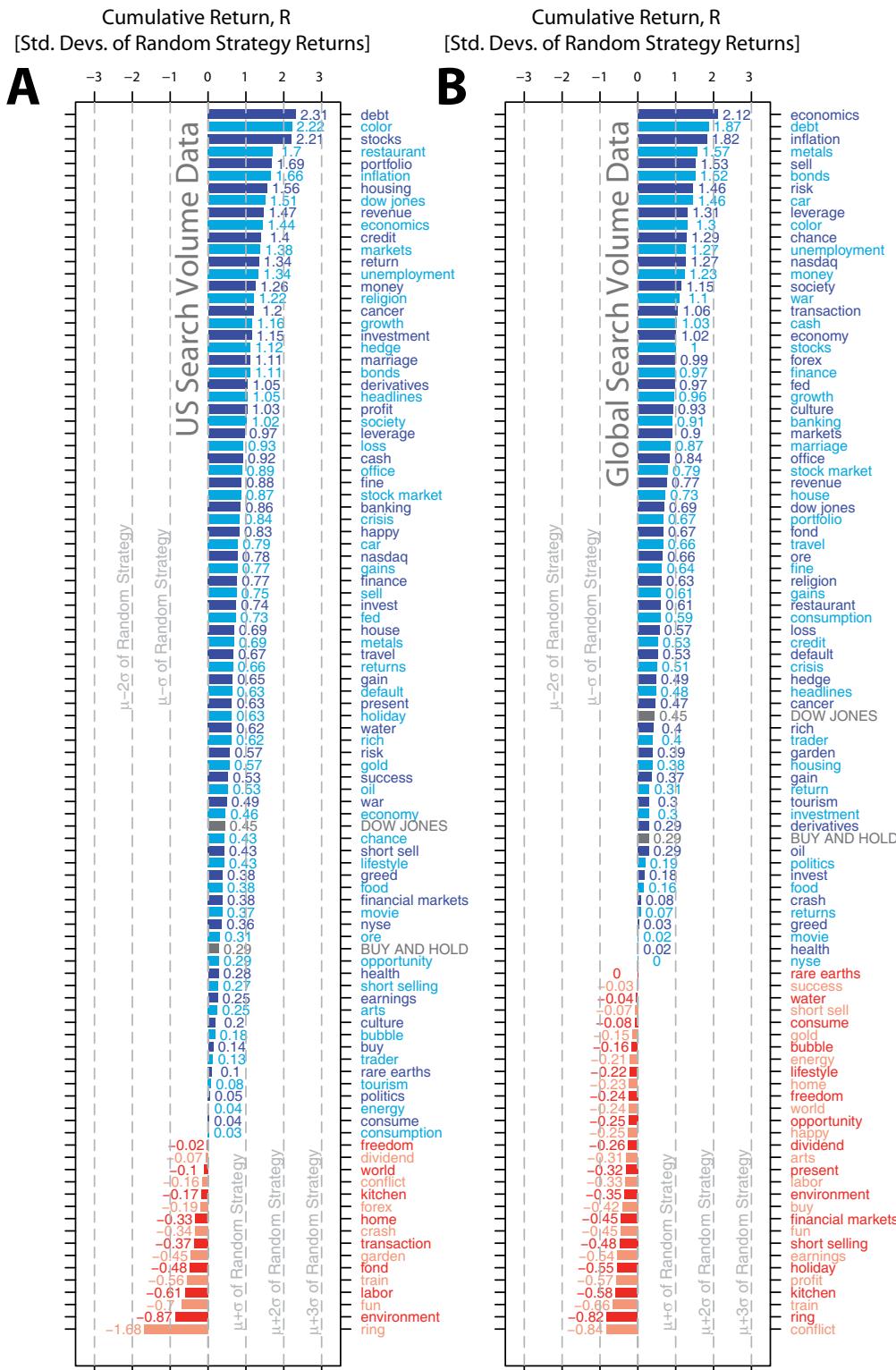


Figure 3 | Performances of investment strategies based on search volume data. (A) Cumulative returns of 98 investment strategies based on search volumes restricted to search requests of users located in the United States for different search terms, displayed for the entire time period of our study from 5 January 2004 until 22 February 2011—the time period for which *Google Trends* provides data. We use two shades of blue for positive returns and two shades of red for negative returns to improve the readability of the search terms. The cumulative performance for the “buy and hold strategy” is also shown, as is a “Dow Jones strategy”, which uses weekly closing prices of the *Dow Jones Industrial Average* (DJIA) rather than *Google Trends* data (see gray bars). Figures provided next to the bars indicate the returns of a strategy, R , in standard deviations from the mean return of uncorrelated random investment strategies, $\langle R \rangle_{\text{RandomStrategy}} = 0$. Dashed lines correspond to -3 , -2 , -1 , 0 , $+1$, $+2$, and $+3$ standard deviations of random strategies. We find that returns from the *Google Trends* strategies tested are significantly higher overall than returns from the random strategies ($\langle R \rangle_{\text{US}} = 0.60$; $t = 8.65$, $df = 97$, $p < 0.001$, one sample t-test). (B) A parallel analysis shows that extending the range of the search volume analysis to global users reduces the overall return achieved by *Google Trends* trading strategies on the U.S. market ($\langle R \rangle_{\text{US}} = 0.60$, $\langle R \rangle_{\text{Global}} = 0.43$; $t = 2.69$, $df = 97$, $p < 0.01$, two-sided paired t-test). However, returns are still significantly higher than the mean return of random investment strategies ($\langle R \rangle_{\text{Global}} = 0.43$; $t = 6.40$, $df = 97$, $p < 0.001$, one sample t-test).

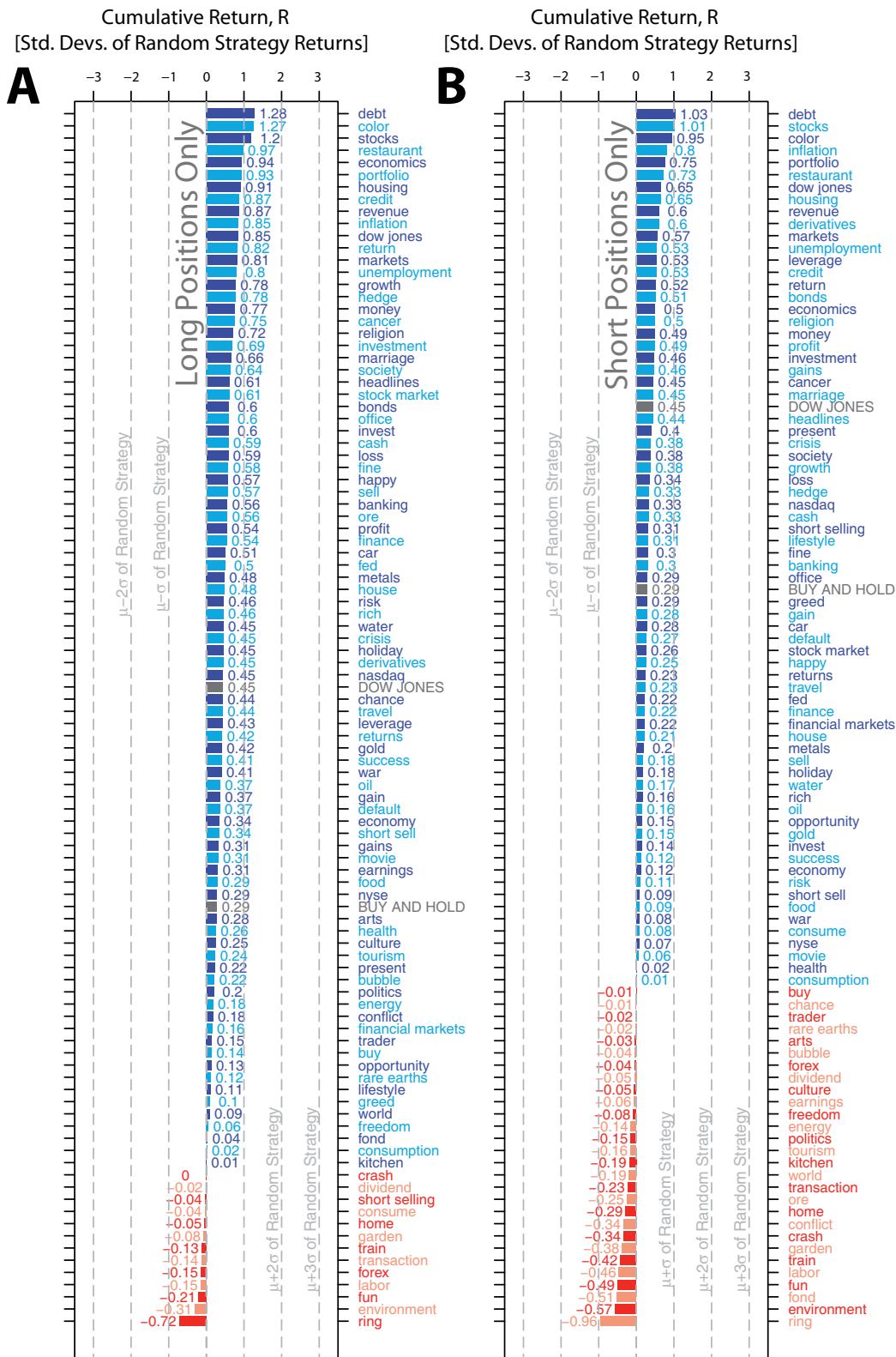


Figure 4 | Analysis using strategies in which we take long or short positions only, using U.S. search volume data. (A) We implement *Google Trends* strategies in which we take long positions following a decrease in search volume, and never take short positions. We find that returns from these long position *Google Trends* strategies are significantly higher overall than returns from the random investment strategies ($\langle R \rangle_{USLong} = 0.41$; $t = 11.42$, $df = 97$, $p < 0.001$, one sample t-test). Again, we find a positive correlation between our indicator of financial relevance and returns from these strategies (Kendall's tau = 0.242, $z = 3.53$, $N = 98$, $p < 0.001$). (B) We also implement *Google Trends* strategies in which we take short positions following an increase in search volume, and never take long positions. In line with our results from the long position *Google Trends* strategies, we find that returns from the short position *Google Trends* strategies are significantly higher overall than returns from the random investment strategies ($\langle R \rangle_{USShort} = 0.19$; $t = 5.28$, $df = 97$, $p < 0.001$, one sample t-test), and that there is a positive correlation between our indicator of financial relevance and short position *Google Trends* returns (Kendall's tau = 0.275, $z = 4.01$, $N = 98$, $p < 0.001$).



performance of *Google Trends* strategies based on all other search terms that we analyze is depicted in Figures S3-S100 in the *Supplementary Information*.

We rank the full list of the 98 investigated search terms by their trading performance when using search data for U.S. users only (Fig. 3A) and when using globally generated search volume (Fig. 3B). In order to ensure the robustness of our results, the overall performance of a strategy based on a given search term is determined as the mean value over the six returns obtained for $\Delta t = 1 \dots 6$ weeks. Returns of the strategies are calculated as the logarithm of relative portfolio changes, following the usual definition of returns. The distribution of final portfolio values resulting from the random investment strategies is close to log-normal. Cumulative returns from the random investment strategy, derived from the logarithm of these portfolio values, therefore follow a normal distribution, with a mean value of $\langle R \rangle_{RandomStrategy} = 0$. Here we report R , the cumulative returns of a strategy, in standard deviations of the cumulative returns of these uncorrelated random investment strategies.

We find that returns from the *Google Trends* strategies we tested are significantly higher overall than returns from the random strategies ($\langle R \rangle_{US} = 0.60; t = 8.65, df = 97, p < 0.001$, one sample t-test).

We compare the performance of these search terms with two benchmark strategies. The ‘buy and hold’ strategy is implemented by buying the index in the beginning and selling it at the end of the hold period. This strategy yields 16% profit, equal to the overall increase in value of the DJIA in the time period from January 2004 until February 2011. We further implement a ‘Dow Jones strategy’ by using changes in $p(t)$ in place of changes in search volume data as the basis of buy and sell decisions. We find that this strategy also yields only 33% profit with $\Delta t = 3$ weeks, or when determined as the mean value over the six returns obtained for $\Delta t = 1 \dots 6$ weeks, 0.45 standard deviations of cumulative returns of uncorrelated random investment strategies (Figs. 3A and 3B; see also Fig. S101 in the *Supplementary Information*).

Our results show that performance of the *Google Trends* strategy differs with the search term chosen. We investigate whether these differences in performance can be partially explained using an indicator of the extent to which different terms are of financial relevance—a concept we quantify by calculating the frequency of each search term in the online edition of the *Financial Times* from August 2004 to June 2011, normalized by the number of *Google* hits for each search term (see Fig. S2 in the *Supplementary Information*). We find that the return associated with a given search term is correlated with this indicator of financial relevance (Kendall’s tau = 0.275, $z = 4.01, N = 98, p < 0.001$) using Kendall’s tau rank correlation coefficient³⁷.

It is widely recognized that investors prefer to trade on their domestic market, suggesting that search data for U.S. users only, as used in analyses so far, should better capture the information gathering behavior of U.S. stock market participants than data for *Google* users worldwide. Indeed, we find that strategies based on global search volume data are less successful than strategies based on U.S. search volume data in anticipating movements of the U.S. market ($\langle R \rangle_{US} = 0.60, \langle R \rangle_{Global} = 0.43; t = 2.69, df = 97, p < 0.01$, two-sided paired t-test).

Our empirical results so far are consistent with a two part hypothesis: namely that key increases in the price of the DJIA were preceded by a decrease in search volume for certain financially related terms, and conversely, that key decreases in the price of the DJIA were preceded by an increase in search volume for certain financially related terms. However, our trading strategy can be decomposed into two strategy components: one in which a decrease in search volume prompts us to buy (or take a long position) and one in which an increase in search volume prompts us to sell (or take a short position).

In order to verify that both strategy components play a significant role in our results, such that we have evidence for both parts of this

hypothesis, we implement and test one strategy in which we take long positions following a decrease in search volume but never take short positions (Fig. 4A), and another strategy in which we take short positions following an increase in search volume but never take long positions (Fig. 4B). We find that returns from both *Google Trends* strategy components are significantly higher overall than returns from a random investment strategy (long position strategies: $\langle R \rangle_{USLong} = 0.41; t = 11.42, df = 97, p < 0.001$, one sample t-test; short position strategies: $\langle R \rangle_{USSHort} = 0.19; t = 5.28, df = 97, p < 0.001$, one sample t-test).

Discussion

In summary, our results are consistent with the suggestion that during the period we investigate, *Google Trends* data did not only reflect aspects of the current state of the economy, but may have also provided some insight into future trends in the behavior of economic actors. Using historic data from the period between January 2004 and February 2011, we detect increases in *Google* search volumes for keywords relating to financial markets before stock market falls. Our results suggest that these warning signs in search volume data could have been exploited in the construction of profitable trading strategies.

We offer one possible interpretation of our results within the context of Herbert Simon’s model of decision making²⁸. We suggest that *Google Trends* data and stock market data may reflect two subsequent stages in the decision making process of investors. Trends to sell on the financial market at lower prices may be preceded by periods of concern. During such periods of concern, people may tend to gather more information about the state of the market. It is conceivable that such behavior may have historically been reflected by increased *Google Trends* search volumes for terms of higher financial relevance.

We find that strategies based on search volume data for U.S. users are more successful for the U.S. market than strategies using global search volume data. Given the assumption that the population of U.S. Internet users contains a higher proportion of traders on the U.S. markets than the worldwide population of Internet users contains, this finding is in line with the intriguing suggestion that these data sets may provide insights into different stages of decision making within the same population.

In this work, we provide a quantification of the relationship between changes in search volume and changes in stock market prices. Future work will be needed to provide a thorough explanation of the underlying psychological mechanisms which lead people to search for terms like *debt* before selling stocks at a lower price. It is clear that many opportunities also remain to extend our analyses to further financial data sets.

The results of our investigation suggest that combining large behavioral data sets such as financial trading data with data on search query volumes may open up new insights into different stages of large-scale collective decision making. We conclude that these results further illustrate the exciting possibilities offered by new big data sets to advance our understanding of complex collective behavior in our society.

Methods

How related are search terms to the topic of finance? We quantify financial relevance by calculating the frequency of each search term in the online edition of the *Financial Times* (<http://www.ft.com>) from August 2004 to June 2011, normalized by the number of *Google* hits (<http://www.google.com>) for each search term. Details are given in the *Supplementary Information*.

Data retrieval. We retrieved search volume data by accessing the *Google Trends* website (<http://www.google.com/trends>) on 10 April 2011, 17 April 2011, and 24 April 2011. The data on the number of hits for search terms in the online edition of the *Financial Times* was retrieved on 7 June 2011. The numbers of *Google* hits for these terms were obtained on 8 June 2011.

1. Axtell, R. L. Zipf distribution of US firm sizes. *Science* **293**, 1818–1820 (2001).
2. King, G. Ensuring the Data-Rich Future of the Social Sciences. *Science* **331**, 719–721 (2011).



3. Vespignani, A. Predicting the Behavior of Techno-Social Systems. *Science* **325**, 425–428 (2009).
4. Lazer, D. *et al.* Computational Social Science. *Science* **323**, 721–723. (2009).
5. Perc, M. Evolution of the most common English words and phrases over the centuries. *J. R. Soc. Interface* **9**, 3323–3328 (2012).
6. Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E. & Perc, M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports* **2**, 943 (2012).
7. Christakis, N. A. & Fowler, J. H. *Connected: The surprising power of our social networks and how they shape our lives* (Little, Brown and Company, 2009).
8. Fehr, E. Behavioural science - The economics of impatience. *Nature* **415**, 269–272 (2002).
9. Shleifer, A. *Inefficient Markets: An Introduction to Behavioral Finance* (Oxford University Press, Oxford, 2000).
10. Lillo, F., Farmer, J. D. & Mantegna, R. N. Econophysics - Master curve for price-impact function. *Nature* **421**, 129–130 (2003).
11. Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, H. E. A theory of power-law distributions in financial market fluctuations. *Nature* **423**, 267–270 (2003).
12. Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D. & Ben-Jacob, E. Quantifying the Behavior of Stock Correlations Under Market Stress. *Scientific Reports* **2**, 752 (2012).
13. Preis, T., Schneider, J. J. & Stanley, H. E. Switching processes in financial markets. *PNAS* **108**, 7674–7678 (2011).
14. Preis, T. Econophysics - complex correlations and trend switchings in financial time series. *European Physical Journal Special Topics* **194**, 5–86 (2011).
15. Bunde, A., Schellnhuber, H. J. & Kropff, J., eds. *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes* (Springer, Berlin, 2002).
16. Vandewalle, N. & Ausloos, M. Coherent and random sequences in financial fluctuations. *Physica A* **246**, 454–459 (1997).
17. Podobnik, B., Horvatic, D., Petersen, A. M. & Stanley, H. E. Cross-correlations between volume change and price change. *PNAS* **106**, 22079–22084 (2009).
18. Sornette, D., Woodard, R. & Zhou, W. X. The 2006–2008 oil bubble: Evidence of speculation, and prediction. *Physica A* **388**, 1571–1576. (2009).
19. Watanabe, K., Takayasu, H. & Takayasu, M. A mathematical definition of the financial bubbles and crashes. *Physica A* **383**, 120–124 (2007).
20. Bouchaud, J. P., Matacz, A. & Potters, M. Leverage effect in financial markets: the retarded volatility model. *Physical Review Letters* **87**, 228701 (2001).
21. Hommes, C. H. Modeling the stylized facts in finance through simple nonlinear adaptive systems. *PNAS* **99**, 7221–7228 (2002).
22. Haldane, A. G. & May, R. M. Systemic risk in banking ecosystems. *Nature* **469**, 351–355 (2011).
23. Lux, T. & Marchesi, M. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **397**, 498–500 (1999).
24. Krugman, P. *The Self-Organizing Economy* (Blackwell, Cambridge, Massachusetts, 1996).
25. Sornette, D. & von der Becke, S. Complexity clouds finance-risk models. *Nature* **471**, 166 (2011).
26. Schweitzer, F. *et al.* Economic Networks: The New Challenges. *Science* **325**, 422–425 (2009).
27. Garlaschelli, D., Caldarelli, G. & Pietronero, L. Universal scaling relations in food webs. *Nature* **423**, 165–168 (2003).
28. Onnela, J. P., Arbesman, S., Gonzalez, M. C., Barabasi, A. L. & Christakis, N. A. Geographic Constraints on Social Network Groups. *PLoS One* **6**, e16939 (2011).
29. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).
30. Simon, H. A. A behavioral model of rational choice. *Quarterly Journal of Economics* **69**, 99–118 (1955).
31. Mondria, J., Wu, T. & Zhang, Y. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics* **82**, 85–95 (2010).
32. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
33. Preis, T., Reith, D. & Stanley, H. E. Complex dynamics of our economic life on different scales: insights from search engine query data. *Phil. Trans. R. Soc. A* **368**, 5707–5719 (2010).
34. Bordino, I. *et al.* Web Search Queries Can Predict Stock Market Volumes. *PLoS One* **7**, e40014 (2012).
35. Choi, H. & Varian, H. Predicting the Present with Google Trends. *The Economic Record* **88**, 2–9 (2012).
36. Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. Quantifying the Advantage of Looking Forward. *Scientific Reports* **2**, 350 (2012).
37. Kendall, M. A New Measure of Rank Correlation. *Biometrika* **30**, 81–89 (1938).

Acknowledgements

We thank Didier Sornette, Dirk Helbing, and Steven R. Bishop for comments. This work was partially supported by the German Research Foundation Grant PR 1305/1-1 (to T.P.). This work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00285 and by the National Science Foundation (NSF), the Office of Naval Research (ONR), and the Defense Threat Reduction Agency (DTRA). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Author contributions

T.P., H.S.M. and H.E.S. developed the design of the study, performed analyses, discussed the results, and contributed to the text of the manuscript.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports/>

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

How to cite this article: Preis, T., Moat, H.S. & Stanley, H.E. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci. Rep.* **3**, 1684; DOI:10.1038/srep01684 (2013).



Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns

Hong Kee Sul

*Wharton Research Data Service, The Wharton School, University of Pennsylvania
Philadelphia, PA 19104, e-mail: hongkee@wharton.upenn.edu*

Alan R. Dennis

*Operations and Decision Technologies Department, Kelley School of Business
Indiana University, Bloomington, IN 47405, e-mail: ardennis@indiana.edu*

Lingyao (Ivy) Yuan[†]

*Department of Supply Chain and Information Systems, College of Business, Iowa State
University, Ames, IA 50011, e-mail: lyuan@iastate.edu*

ABSTRACT

Decision making is often based on the rational assessment of information, but recent research shows that emotional sentiment also plays an important role, especially for investment decision making. Emotional sentiment about a firm's stock that spreads rapidly through social media is more likely to be incorporated quickly into stock prices (e.g., on the same trading day it was expressed), while sentiment that spreads slowly takes longer to be incorporated into stock prices and thus is more likely to predict stock prices on future days. We analyzed the cumulative sentiment (positive and negative) in 2.5 million Twitter postings about individual S&P 500 firms and compared this to the stock returns of those firms. Our results show that the sentiment in tweets about a specific firm from users with less than 171 followers (the median in our sample) had a significant impact on the stock's returns on the next trading day, the next 10 days, and the next 20 days. Interestingly, sentiment in tweets from users with fewer than 171 followers that were *not* retweeted had the greatest impact on future stock returns. A trading strategy based on these findings produced meaningful economic gains on the order of an 11–15% annual return. [Submitted: December 4, 2014. Revised: April 1, 2016. Accepted: April 12, 2016.]

Subject Areas: *Twitter, Emotion, Sentiment, Stock returns, and S&P500.*

INTRODUCTION

Almost 75% of adult Internet users use social media, and this percentage is increasing (Pew-Research, 2014). Twitter is one of the most popular social media

[†]Corresponding author.

platforms in the world. Not only has the number of people using social media increased dramatically, so too has the amount of use. In 2015, there were about 300 million Twitter users worldwide, who sent an average of 500 million tweets per day ("About Twitter, Inc.," 2015). Users have integrated social media into many aspects of their daily life (Ellison, 2007), including investment decision making (Oh & Sheng, 2011). Numerous professional and amateur investors and analysts use Twitter to post news articles, and opinions, often providing information and comments more frequently than the professional news media (Sprenger, Tumasjan, Sandner, & Welpe, 2014).

Stock returns, or the profits from trading stocks, are influenced by many factors. Along with fundamental factors and transaction costs, investor sentiment also plays an important role in influencing stock return (Baker & Wurgler, 2007). Market sentiment can be expressed in many ways. The development of social media provides a new meaningful channel for users to share information and their personal feelings. As such, it also serves as a convenient method to capture market sentiment.

Prior research has studied whether the emotional content of tweets can be used to predict stock returns. Bollen, Mao, and Zeng (2011b) assessed the emotional state (calm, alert, sure, vital, kind, and happy) in 10 million tweets that were not related to the stock market. They found that the amount of one state, "calm," was significantly positively correlated with changes in the Dow Jones Industrial average (DJIA) several days later; in other words, when there was a great deal of "calm" in tweets on a given day, the DJIA tended to rise over the following days. Oh and Sheng (2011) examined 200,000 tweets from StockTwits that focused on specific stocks and classified each tweet as "bullish," "bearish," or "neutral" to create a "bullishness" index for each stock. They found the 5-day rolling average of the bullishness index was useful in predicting stock price movements. Sprenger et al. (2014) also used machine learning to create a different bullishness index that they too found to be predictive of stock returns several days later. Smailović, Grčar, Lavrač, and Žnidaršić (2014) used machine learning to examine sentiment (i.e., positive emotion) in tweets and found it to be predictive of stock returns several days later. Risius, Akolk, and Beck (2015) examined emotional states (happiness, affection, satisfaction, fear, anger, depression, contempt) and positive and negative sentiment, and found negative sentiment and "depression" to predict stock returns on the following day.

These findings are promising in suggesting that the emotional state and sentiment in tweets can be used to predict stock returns, but there are still many unanswered questions. Although empirical research has shown that certain emotional states and sentiments in tweets can predict stock price movements, there is a lack of theory to explain why they influence stock returns days later. We argue that the Gradual Information Diffusion model (Hong & Stein, 1999) is useful in understanding how tweets are linked to future stock returns. Under this theoretical perspective, information (in our case, sentiment) influences stock prices as it spreads through the investing public. Sentiment that spreads quickly has an immediate influence on prices, while sentiment that spreads slowly has a slower effect. Sentiment that spreads slowly opens the door for a trading strategy that capitalizes on the stock returns from slowly rising or falling prices.

We analyzed almost 2 years' worth of data collected from Twitter and linked it to the average daily stock returns of firms in the S&P 500. Our results show that the sentiment in tweets about specific firms was significantly related to stock returns on subsequent days. Tweets from individuals with fewer followers had a stronger impact on future returns than tweets from those with many followers, because their tweets took longer to spread. Likewise, tweets that were *not* retweeted took longer to spread and were linked to greater future stock returns.

PRIOR RESEARCH AND THEORY

Information and Stock Return Prediction

Whether stock returns can be predicted has long been a debate. Based on the Efficient Market Hypothesis (EMH), early research argued that stock returns are random and cannot be predicted (Eppen & Fama, 1969; Dockery & Kavussanos, 1996). Research shows that new information, especially news, is a major factor influencing stock returns and quickly leads to stock price changes (Malkiel & Fama, 1970; Hong, Lim, & Stein, 2000; Qian & Rasheed, 2007). Under EMH, positive or negative news (e.g., merger, terrorist attack) is quickly factored into a stock price within minutes, so there is little opportunity to profit, unless of course, one has insider knowledge of an event before it occurs. Mass media outlets play an important role in disseminating information to a broad audience, especially individual investors (Fang & Peress, 2009). This suggests that information contained in social media such as Twitter, which also reaches a broad audience, may be linked to stock returns (Bollen et al., 2011a; Oh & Sheng, 2011; Smailović et al., 2014; Sprenger et al., 2014; Risiis et al., 2015).

EMH assumes that information travels quickly and that investors are rational and capable of understanding the full implications of the information they receive (Hong & Stein, 1999). An alternative view is the Gradual Information Flow (GIF) model of Hong and Stein (1999), although the term GIF did not emerge until later (Hong & Stein, 2007). GIF argues that the modern world differs in two important ways from that assumed by EMH. First, some information is private, known only to some investors, and this information diffuses more slowly than public information. Second, investors have cognitive limitations and biases that limit their ability to fully process all implications of the information they receive. In general, investors are either news followers who use fundamental firm information to make investment decisions or momentum traders who use past changes in stock prices to make investment decisions. Both act under bounded rationality, and because they focus primarily on the information relevant to their investing style, they overlook other types of information they receive, and thus prices do not respond to new information as quickly as EMH would predict.

GIF predicts that the speed of information diffusion through the investing public influences how quickly stock prices change in response to new information. Under reasonably efficient markets, information diffuses rapidly among the investing public and is quickly incorporated into stock prices (Hong & Stein, 1999). Conversely, if information diffuses more slowly, it will take longer for that information to be fully incorporated into their prices, and thus there may be opportunities

to profit from information before it is fully incorporated into prices (Hong et al., 2000). Under GID, information should spread rapidly for stocks covered by the mass media but more slowly for stocks not covered by the media. Research shows that stocks not covered by the mass media earn significantly higher future returns than stocks that are covered, after controlling for risk characteristics (Merton, 1987; Fang & Peress, 2009), suggesting that the speed of information diffusion is important in understanding how information may be used to predict future stock prices and thus the returns that can be made by investing.

Information Diffusion in Social Media

Twitter is a social media platform in which users post short text messages of up to 140 characters, called tweets. Anyone can open a Twitter account and begin sending tweets. Users can subscribe to or “follow” other users, and the followers are notified immediately when a user tweets. Many Twitter users have few followers, while commentators, journalists, and celebrities have thousands or more. The median number of followers has gradually increased over time and was about 100 in 2014 (Liu, Kilman-Silver, & Mislove, 2014). Most users follow more people than they have followers; the median number of users followed has gradually increased over time and was around 140 in 2014 (Liu et al., 2014).

During the past several years, Twitter has drawn interests of researchers from multiple disciplines. Current research on Twitter includes several streams. One stream is its impact on information diffusion and supporting communication/collaboration (Honey & Herring, 2009) in many different contexts. Using Twitter during a talk show decreased the psychological distance between the host and his/her audience (Larsson, 2013). In the context of education, Twitter is a potential learning tool in classrooms (Dhir, Buragga, & Boreqqah, 2013). Twitter has become an important tool to spread information during natural disasters and social crises (Sakaki, Okazaki, & Matsuo, 2010; Oh, Agrawal, & Rao, 2013).

Another research stream using Twitter is designing and developing network analysis techniques and algorithms. The abundant data exchanged on Twitter every minute provide researchers, especially those in computer science, the opportunity to observe the social network change. Other related techniques, such as text mining and data mining techniques, also became more refined by studying Twitter data. A third stream is using Twitter to predict individual behavior. Using opinion mining tools and sentiment analysis techniques, researchers are able to predict election results (Tumasjan, Sprenger, Sandner, & Welpe, 2010), hospital-associated mortality (Daley et al., 1988), and heart disease in middle-aged and older persons (Gordon, Castelli, Hjortland, Kannel, & Dawber, 1977).

Due to its popularity, the investment community has adopted Twitter. This community uses the convention of tagging stock-related tweets with a dollar sign (\$) followed by the firm’s stock ticker symbol. For example, an individual tweeting about PepsiCo would include \$PEP in the tweet. A sample tweet from our data: “\$PEP has been strong all day. And who doesn’t love those Frito-Lay snacks? Be honest.”

Any Twitter user can send a tweet and include a stock ticker with a dollar sign to indicate that he or she thinks the tweet contains financial information.

Depending upon how many followers that user has, that information may reach a few users, many users, or even tens of thousands of users. Other users can “retweet” the information to their followers so that the information in the original tweet will spread throughout a broad audience of Twitter users—and to non-Twitter users if some users choose to spread the information using other media.

In his seminal work on networks, Barabási (2002) shows just how interconnected we are. Twitter is a directed graph network in that connections are directional. I receive information from users I follow, but they do not receive information from me unless they follow me back. The speed at which information spreads through such a network depends upon how many followers a user has (Barabási, 2002). The general formula for the number of hops it takes to reach any other node in a network is $d = \log N / \log k$, where N is the total number of nodes in the network, and k is the average number of connections per node (i.e., followers) (Barabási, 2002). The number of active traders is on the order of 10 million depending upon how one defines active (trade-IQ, 2011). It is difficult to estimate the average number of followers in this community; links in networks typically follow a power law distribution—not a normal distribution (Barabási, 2002)—and our data were no different, so we use the median of 171. Using these data, we see that it takes about 3.1 hops for information from one node to reach any other node. Of course, Twitter is not the only mechanism through which information is spread. Individual investors can talk with or e-mail other investors. Most people know 200–5,000 people by name (Barabási, 2002), which suggests that we are three to four hops away from anyone else on the planet (Barabási, 2002).

The speed of information diffusion influences whether the information is quickly incorporated in stock prices or takes longer—perhaps days—to be fully disseminated and incorporated into stock prices (Hong & Stein, 1999). Empirical studies of networks show that the number of connections is not randomly distributed (Barabási, 2002). In every network, there are hubs, individuals who have substantially more connections than the average (Barabási, 2002). These hubs are often opinion leaders who facilitate the rapid diffusion of information through the network (Barabási, 2002). Some professional analysts routinely tweet information and thus have a large number of followers; Jim Cramer of CNBC’s *Mad Money*, for example, has over 650,000 followers. These individuals are the hubs in the diffusion of investment information, reaching a significant proportion of investors in one hop.

If a Twitter user is a hub (i.e., has many followers), the information he or she tweets will spread more quickly than if the same information is tweeted by a user who has few followers (Barabási, 2002). Tweets from a hub not only reach more people in a single hop, but also tend to be more influential (Barabási, 2002). If a Twitter user has many followers, any information he or she tweets will be quickly disseminated, and stock prices should quickly change to incorporate that information that same day, and there should be little or no effect on stock returns on future days. For example, during the day on July 7, 2011, Twitter account howardlindzon (which has about 200,000 followers) tweeted, “Looks like Howard will be adding more \$aapl on a good close. He’s predictable.” In this tweet of 14 words, two were positive, making it a positive tweet on Apple. Apple had closed the previous day at \$351.76, and rose to close at \$357.20 that day.

Conversely, if a Twitter user has few followers, information should be slower to disseminate because it will take more hops to reach a critical mass of investors and because it will be less influential than tweets from a hub (Barabási, 2002). Therefore, there is more likely to be a relationship between that information and stock returns on future days because it will take longer for that information to reach many investors and be incorporated into stock prices. For example, during the day on August 8, 2011, several tweets came from multiple twitter accounts with fewer than 100 followers. User ibshakey tweeted, "Still love \$AAPL. Continue to love gold. The US in general, not so much." User cronked tweeted, "After that last quarter, how can you not buy \$AAPL here? There are some bright spots out there. Not all is lost.,," and drewmethey tweeted, "\$AAPL I have to buy Apple here. It's just too cheap!!!! Can't resist." Each tweet contained two positive words, making all three a positive tweet on Apple, Inc. From the previous close price of \$373.62, the stock price fell to \$353.21 at the end of the day. However, on the next day and over the next 10 days, the stock price gradually rose, closing a month later, on September 8, at \$384.14.

Sentiment and Contagion

Much of the investment information shared using traditional media and social media is facts and opinions, but individual behavior is not only the outcome of rational decision making. Emotions triggered by these facts and opinions can also influence decisions (Bechara & Damasio, 2005). Twitter provides a good environment to foster the sharing of emotion (Bollen, Pepe, & Mao, 2011b) because the length of each tweet is restricted to 140 characters. The limitation on length of tweets encourages users to be brief and get to the point (Oh & Sheng, 2011). Thus, a short message can provide a focused and more intense trigger for the receiver.

Individual moods, emotions, and other affects are influenced by both internal factors and external factors. Internal factors include personality, individual competency, and so on. External factors include experiences, and information the individual receives. Different affects have different impacts on individuals (Frijda, 1994). Affects can be broad and vague or acute and specific. Affects may have a long-term influence; their effects can also be short term.

Emotion, as one type of affect, has the characteristics of having a clear trigger and a short but more intense effect (Frijda, 1994). Emotion is a subjective feeling related, triggered by a stimulus such as an event, an object, or information in one's environment. Once the stimulus conditions, the stimulus itself, or the supporting cognition, perceptions, or other triggers are no longer active, the emotion will disappear. Emotion can be highly contagious (Schoenewolf, 1990; Hatfield, Cacioppo, & Rapson, 1993).

There are many ways to conceptualize the way emotion is expressed, but two dominant approaches have emerged (Russell, 1980, 2003; Calvo & Kim, 2013). The classic approach, used by Bollen et al. (2011a) and Risius et al. (2015), is to consider specific emotional states, such as joy, anger, sadness, etc. The other approach, used by Smailović et al. (2014), is the dimensional model in which emotional affect is conceptualized as having two dimensions: valence (positive or negative) and arousal (high or low) (Osgood, Suci, & Tannenbaum, 1957; Russell,

1980, 2003; Cacioppo, Petty, Losch, & Kim, 1986); some authors also include a third dimension of dominance (Bradley & Lang, 1994). Neither approach is more or less “correct” (Calvo & Kim, 2013), and it is straightforward to map emotional states onto the dimensional model (Russell, 1980, 2003; Bradley & Lang, 1994). For example, the “calm” emotional state studied by Bollen et al. would be considered neutral valence and low arousal (Russell, 2003).

Both models are commonly used, although Calvo and Kim (2013) conclude that researchers in natural language processing are more likely to use the emotional states model, while researchers in psychology are more likely to use the dimensional model. When natural language processing researchers use the dimensional model, they commonly focus on only the valence dimension, which they term “sentiment” (e.g., Wiebe, Wilson, & Cardie, 2005; Abbasi & Chen, 2008; Oh & Sheng, 2011; Smailović et al., 2014). In this study, we focus on sentiment, as has been common in financial research (e.g., Tetlock, 2007).

Sentiment affects decision making (Bakamitsos, 2006). According to Construal Level Theory (CLT), positive and negative sentiment may have different effects (Liberman & Trope, 1998; Bar-Anan, Liberman, & Trope, 2006; Fujita, Trope, Liberman, & Levin-Sagi, 2006). Positive sentiment increases abstract construal, that is, the adoption of abstract, future goals, while negative sentiment triggers a focus on immediate and proximal concerns and reduces the adoption of abstract future goals (Liberman & Trope, 1998; Eyal, Liberman, Trope, & Walther, 2004; Bar-Anan et al., 2006; Fujita et al., 2006; Labroo & Patrick, 2009). Positive sentiment is more likely to induce individuals to make a decision than negative sentiment, which tends to slow the decision process (Qiu & Yeung, 2008). Positive sentiment also may induce an individual to act on a decision (Frijda, 1994). Positive sentiment can increase consumers’ impulse to buy in the context of electronic commerce (Parboteeah, Valacich, & Wells, 2009) but increase an individual’s resistance to temptation in other contexts (Fedorikhin & Patrick, 2010). Thus, positive sentiment and negative sentiment are more than the opposite ends of the same dimension; they can trigger different behaviors. For this reason, researchers often have measured them separately to capture their full effects (e.g., Tetlock, Saar-Tsechansky, & Macskassy, 2008; Risius et al., 2015).

Sentiment is contagious (Schoenewolf, 1990; Hatfield et al., 1993). Social contagion is “the tendency to automatically mimic and synchronize expressions, vocalizations, postures, and movements with those of another person’s and to converge emotionally” (Hatfield et al., 1993). Contagion happens implicitly and explicitly (Singer & Lamm, 2009). Sentiment is expressed through facial impressions, physical gestures, vocal tones, and written words. When individuals exchange messages via written text, photos, audio, or even video on social media, the message sender’s sentiment also is exchanged. Thus, tweets and the sentiment they contain have the potential to influence the receiver’s behavior (Risius et al., 2015).

Hypotheses

We argue that the sentiment contained in social media tweets will have a direct effect on stock returns in a manner similar to the effects that professional news

media have on stock returns. Positive sentiment should be associated with positive returns and negative sentiment should be associated with negative returns (Tetlock et al., 2008).

The effects of sentiment spread in the same manner in which information spreads through the network. Thus, the speed of diffusion is important. If a tweet about a specific firm is sent by a hub (a user who has many followers), the sentiment it contains will spread faster than the sentiment sent by a user who has few followers because more individuals will see it immediately, and it will be more influential (Barabási, 2002). Sentiment that is spread more quickly will be incorporated into prices faster, so that it will have an effect on returns sooner (Hong & Stein, 1999; Tetlock et al., 2008). Its effects are more likely to be seen on the same trading day on which it was tweeted. Thus, it will have little effect on stock returns on future days, because its effects are immediately incorporated into prices (Hong & Stein, 1999; Tetlock et al., 2008).

In contrast, sentiment contained in tweets from those with fewer followers will take longer to disseminate because fewer people will see them on the first hop, and it will be less influential (Barabási, 2002). Thus, the number of followers affects the speed of sentiment diffusion. This less visible sentiment from users with few followers is spread more slowly and will take longer to affect stock prices (Hong & Stein, 1999). Therefore, it will have a larger effect on stock returns on future trading days (Hong & Stein, 1999; Tetlock et al., 2008). Thus:

H1: The sentiment in tweets about a specific firm sent by individuals with few followers is directly related to stock returns on future trading days.

This sentiment diffusion process will also be affected by the extent to which tweets are “retweeted”—that is, whether an individual who receives a tweet resends it to his or her followers. A study of 37 billion public tweets found that the percentage of retweets has increased over time: about 5% in 2010, 10% in 2011, 20% in 2012, and 25% in 2013 (Liu et al., 2014).

Individuals retweet for a variety of reasons. The most common reasons are because they believe the tweet’s information would be of interest to their followers or to express support for the original tweeter (Macskassy & Michelson, 2011; Liu et al., 2014). In the investing context where tweets are deliberately tagged with the \$ and ticker symbol, we theorize that most tweets are retweeted because the sender believes they have potential information for other investors.

Retweets affect the diffusion process. Retweeting someone else’s tweet is a deliberate signal that the user believes the tweet would be of interest to his or her followers. Retweeting spreads the sentiment in the tweet faster than if the tweet was not retweeted and makes the tweet more influential because now two people advocate for its content, not one. Sentiment in tweets that are retweeted will be more quickly incorporated in stock price, so it will have *less* of an effect on returns on future trading days. Therefore, it is the combination of few followers and *not* being retweeted that leads to the greatest stock returns on future days because the sentiment in these tweets will take the longest to diffuse through the network. Therefore, we hypothesize:

H2: The sentiment in tweets about a specific firm sent by individuals with few followers that are not retweeted is directly related to stock returns on future trading days.

METHODOLOGY

Financial Data

To ensure sufficient reliability of Twitter data, we focused only on firms that are part of the S&P 500. Financial data, including the closing price of each stock in the S&P 500, were obtained from Compustat, Center for Research in Security Prices (CRSP), Institutional Brokers' Estimate System (IBES), and Kenneth French's Web site (Rai, Patnayakuni, & Seth, 2006). The sample period is from March 2011 to January 2013.

Twitter Data

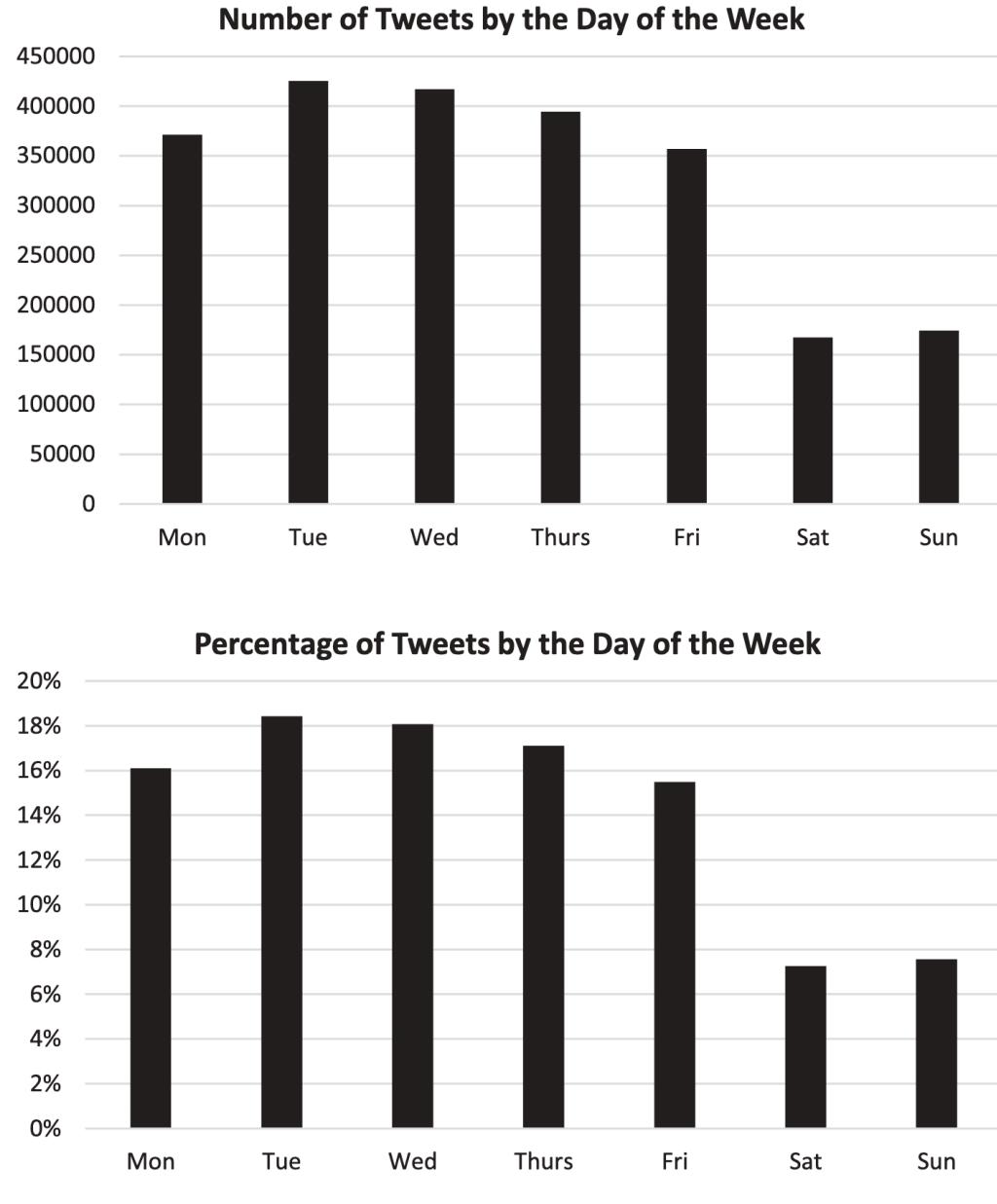
This study used data collected from Twitter. The focus of this article is on whether the sentiment in tweets about an individual firm can predict stock returns. Thus, it is important to match tweets to specific firms. The convention in Twitter is to precede the stock ticker symbol with a dollar sign (\$) to indicate that a tweet contains investment information about a firm. We collected all public tweets that contained the relevant \$ symbol with an S&P 500 stock ticker from Twitter using a developer account. We retrieved 3,475,428 tweets during the sample time period. Of all the tweets, 16.02% were retweets. We excluded all the tweets that contained more than one ticker symbol because we could not be sure if the information in the tweet pertained to one firm or all firms equally. For example, a tweet like "I also like long \$AAPL @347.40 ... and short \$RIMM @62.70" would be excluded from the analysis. This produced a final sample of 2,503,385 tweets. An inspection of 500 randomly selected tweets found no tweets from the firm itself. Figure 1 shows the distribution of the tweets by days of the week.

The Sentiment in Tweets

There are many approaches to sentiment analysis (Feldman, 2013). We used the word analysis strategy. Each word in a tweet was matched to a dictionary of terms to determine its sentiment. We used the Harvard-IV dictionary (Jorgenson & Vu, 2005), which is a commonly used source for word classification in the financial content analysis of popular press articles and Web news sites, used, for example, by Tetlock (2007), Tetlock et al. (2008), and Da, Engelberg, and Gao (2011). There are other dictionaries that could be used (e.g., the financial dictionary of Loughran and McDonald (2011)), but these dictionaries are designed for the analysis of legal and financial documents which contain formal English (e.g., 10K filings), not the slang version of English used in Twitter.

CLT argues that positive and negative sentiment may have different effects (Liberman & Trope, 1998; Bar-Anan et al., 2006; Fujita et al., 2006), so it is important to track both positive and negative sentiment because they may have different effects. Empirical research on stock returns has shown that sometimes

Figure 1: Number of tweets (top) and percentage distribution (bottom) by the day of the week.



positive sentiment has an impact (Smailović et al., 2014), while other times negative sentiment has an impact (Risius et al., 2015).

We counted all words in the tweets that had the “NEG” tag in the Harvard-IV dictionary as words that conveyed a negative sentiment. We counted “POS” tagged words as words conveying a positive sentiment. Although this approach has been widely used in prior research (Tetlock et al., 2008), it is an imperfect measure of sentiment, because it cannot detect subtle meanings in English, such as sarcasm or semantic word groups that combine positive and negative words (e.g., “not bad”:

Xie et al., 2015). Likewise, we did not include emoticons in our analysis, so this is another limitation (Xie et al., 2015). An analysis of 100,000 randomly selected tweets from our sample found 430 to contain emoticons (i.e., less than 1%).

We used three separate measures to better model sentiment, as has been done in prior research (Tetlock et al., 2008). If our measures do not accurately capture sentiment, then we are less likely to find a significant relationship, so this approach is a more conservative test of the relationship between sentiment and stock returns than human analysis of the tweets, which would be effectively impossible given our sample size of over 2.5 million tweets. Table 1 shows descriptive statistics about the tweets.

Because we are using daily stock returns as our dependent variable, we combined all tweets for each firm on a given day. Daily returns are defined as close-to-close daily returns, so we match day t return with firm level Twitter content on day t up to the market close time of 4 p.m. New York's time. Any tweet that was posted after 4 p.m. was treated as day $t + 1$. Following Tetlock et al. (2008), we used three variables to measure sentiment. Sentiment is measured as following, where P , N , and T are the daily aggregate number of positive, negative, and total words for each day for a given firm.

$$\text{Sentiment} = \begin{cases} \text{neg1} \equiv \frac{N}{T} \\ \text{pos1} \equiv \frac{P-N}{P+N} \\ \text{pos2} \equiv \log\left(\frac{1+P}{1+N}\right) \end{cases}. \quad (1)$$

Conceptually, neg is the ratio of the amount of negative sentiment to the total communication (positive, negative, and neither). Pos1 is a normalized ratio (on a -1 to +1 scale) of the overall positive or negative sentiment expressed (omitting words with no sentiment). Pos2 is an unstandardized ratio of positive to negative sentiment, but log adjusted to capture the potential for diminishing marginal effects. All three measures may produce similar results, but we included all three for greater insight. Descriptive statistics can be found in Table 2.

Analysis

To answer the question of whether social media have sentiment information that can predict future returns, we examine whether the speed of information dissemination (i) reflected by the number of followers and (ii) retweet history is associated with future returns.

To answer the first research question, we test the following equation:

$$CAR_{t,t+n}^i = \alpha + \beta_0 \text{sentiment}_{u_t}^i + \beta_1 \text{sentiment}_{o_t}^i + \gamma CV_t^i + \epsilon_t^i, \quad (2)$$

where $CAR_{t,t+n}^i$ is the cumulative abnormal return about firm i from day $t + 1$ to day $t + n$; $\text{sentiment}_{u_t}^i$ is the sentiment about firm i on day t expressed in tweets from users with a number of followers at or under a given threshold; $\text{sentiment}_{o_t}^i$ is the sentiment about firm i on day t expressed in tweets from users with a number of followers over a given threshold; and CV are five control variables, as described below.

Table 1: Descriptive statistics of tweets.

	March 2011	April 2011	May 2011	June 2011	July 2011	August 2011	September 2011	October 2011	November 2011	December 2011	January 2012	February 2012
Total stock tickers	497	497	498	497	498	498	498	498	498	498	498	497
Total number of tweets	126,926	216,303	197,978	160,119	189,655	244,763	205,893	215,702	199,471	171,531	190,144	189,079
Average number of tweets per firm	255.38	435.22	397.55	322.17	380.83	491.49	413.44	433.14	400.54	344.44	381.82	380.44
Average number of words per tweet for all firms	15.299	15.615	15.437	15.358	15.161	15.130	14.804	14.953	14.662	15.267	15.160	14.735
Average number of positive words per tweet for all firms	0.400	0.433	0.401	0.359	0.261	0.331	0.289	0.322	0.309	0.322	0.349	0.336
Average number of negative words per tweet for all firms	0.245	0.264	0.247	0.223	0.254	0.220	0.204	0.217	0.195	0.203	0.229	0.207
Average percentage of positive words for all firms	0.026	0.027	0.026	0.023	0.023	0.022	0.019	0.021	0.021	0.021	0.023	0.023
Average percentage of negative words for all firms	0.016	0.017	0.016	0.014	0.016	0.014	0.014	0.014	0.013	0.013	0.015	0.013
SD % positive words for all firms	0.012	0.013	0.011	0.012	0.012	0.012	0.013	0.012	0.012	0.011	0.011	0.011
SD % negative words for all firms	0.009	0.010	0.010	0.008	0.010	0.008	0.009	0.008	0.009	0.008	0.008	0.008

Table 2: Descriptive statistics of the firm/trading day data.

	N	Mean	Variance	SD	Min	25th Percentile	50th Percentile	75th Percentile	Max
Pos1 Sentiment i,t	119,727	0.254	0.474	0.688	-1.000	-0.111	0.333	1.000	1.000
Pos 2 Sentiment i,t	119,727	0.347	0.704	0.839	-4.500	-0.182	0.452	0.860	5.226
Neg Sentiment i,t	119,727	0.040	0.002	0.042	0.000	0.000	0.035	0.060	0.667
Surprise i,t	119,727	0.001	0.001	0.031	-2.080	0.000	0.000	0.000	3.950
Control 2 $_{i,t-30,t-2}$	112,938	-0.005	0.008	0.092	-1.010	-0.052	0.000	0.047	0.652
Control 1 $i,t-1$	112,858	0.000	0.000	0.019	-0.673	-0.008	0.000	0.008	0.532
Upgrade i,t	119,727	0.023	0.029	0.171	0.000	0.000	0.000	0.000	9
Downgrade i,t	119,727	0.021	0.023	0.151	0.000	0.000	0.000	0.000	5

Notes: Surprise i,t : Earnings surprise, relative to median analyst estimate.

Control 1 $i,t-30,t-2$: Past returns, cumulative abnormal return from the [-30, -2] trading window.

Control 2 $i,t-1$: The abnormal return on the prior trading day.

Upgrade i,t : The number of financial analyst upgrades for company i on day t .

Downgrade i,t : The number of financial analyst downgrades for company i on day t .

Table 3: Definitions of variables.

AR_t^i	The AR_t^i is the abnormal return of firm i on date t , adjusted using the size and book-to-market matched characteristic portfolio's return. $AR_t^i = R_t^i - Pfo_t^i,$ where $R_t^i = \ln(r_t^i + 1)$, and $Pfo_t^i = \frac{1}{n} \sum_{j=1}^n w_j(R_t^j)$. Note that r_t^i is the daily holding period return (ret) in the CRSP daily stock database (CRSP.DSF) including and w_j is the value weight of the j th firm in the portfolio of firms $j = 1 \sim n$, such that $\sum_{j=1}^n w_j = 1$. The size and book-to-market characteristic portfolio was formed using the 30th and 70th NYSE book-to-market percentiles and the median NYSE market equity.
$CAR_{t,t+n}^i$	The $CAR_{t,t+n}^i$ is future cumulative abnormal return of firm i on date t , the dependent variable in our regressions. It is the summation of the abnormal returns of the next n days starting day $t + 1$.
	$CAR_{t,t+n}^i = \sum_{j=1}^n (AR_{t+j}^i)$
$Control1_{i,t}$	The $Control1_{i,t}$ is the abnormal return of firm i for the date $t - 1$. $Control1_{i,t} = AR_{t-1}^i$
$Control2_{i,t}$	The $Control2_{i,t}$ is the cumulative abnormal return, or the summation of the abnormal return, of firm i for the previous 30 days excluding the abnormal return of the previous date.
	$Control2_{i,t} = \sum_{j=2}^{30} (AR_{t-j}^i)$
Earnings Surprise $_{i,t}$	An earnings surprise is calculated for each firm i on each earnings announcement date. The Earnings Surprise $_{i,t}$ is calculated as the difference between the actual EPS (actual) and the median EPS (medest) from the IBES summary statistics database (IBES.STATSUM).
$Upgrade_{i,t}$	An upgrade/downgrade is recorded as 1 if an analyst increased/decreased the IBES recommendation code (ireccd) from IBES recommendation detail database (IBES.RECDDET). The $Upgrade_{i,t}$, $Downgrade_{i,t}$ is the summation of the number of upgrade/downgrade for all analysts for firm i on the same date.
$Downgrade_{i,t}$	

Equation (2) examines future abnormal returns for days 1 to n after the tweets were made. Table 3 shows how we calculated $CAR_{t,t+n}^i$. We have chosen to use three time periods: next day returns (i.e., $n = 1$), next-day-to-10th-day returns (i.e., $n = 10$), and next-day-to-20th-day returns for a longer view (i.e., $n = 20$). These are trading days, so 10 days is approximately 2 weeks, and 20 days is approximately one month. These time periods are consistent with prior research (e.g., Tetlock et al., 2008; Fang & Peress, 2009; Chen, De, Hu, & Hwang, 2014).

Table 4a: Correlations among the sentiment variables used in H1.

	Pos1 Over	Pos2 Over	Neg Over	Pos1 Under	Pos2 Under
Pos2 Over	.905				
Neg Over	-.800	-.741			
Pos1 Under	.253	.253	-.193		
Pos2 Under	.242	.276	-.196	.905	
Neg Under	-.191	-.193	.199	-.811	-.766

Notes: All correlations are significant at $p < .001$. Cells in gray are correlations among variables used in the same regression analysis.

We used five control variables. Stock returns exhibit autocorrelation so we included two control variables to control for autocorrelation: control1 and control2 capture the abnormal returns on the day before (i.e., $t - 1$) and the cumulative return over the prior 30 days, respectively (Tetlock et al., 2008; Chen et al., 2014) (Table 3). The third control variable is earnings surprise, calculated as the actual earnings per share for a given firm announced on a given day minus the median analyst earnings per share prediction, where the median analyst prediction is the “Median Estimate” from IBES Summary. The last two control variables were the upgrades and downgrades on the company from professional stock analysts as control variables because upgrades and downgrades can influence stock returns (Chen et al., 2014). We counted the number of upgrades and downgrades on the specific firm’s stock on the same trading day as the tweets and included these numbers as controls. As is commonly done in financial research, we obtained analyst recommendations from the IBES, categorized each change in recommendation as either an upgrade or downgrade, and counted the number of each on each trading day (Chen et al., 2014) (Table 3).

To examine H1, the impact of the number of followers, we need to test whether β_0 is positive when sentiment is pos1 or pos2 and test whether β_0 is negative when sentiment is neg for the different trading periods. We split the tweets into two groups based on the number of followers of the tweeters, those with many followers and those with few followers. The question is, what is “many” and “few?” The median number of followers in our sample was 171, so we selected this as the break point for assigning tweets into groups of users with few followers and many followers. The sample size was 48,538 because we can analyze the data only when there are tweets from individuals both over and under the threshold on the same day for the same firm.

To examine H2, the combined impact of number of followers and retweets, we divided the tweets into four groups: many followers and retweeted; many followers and not retweeted; few followers and retweeted; and few followers and not retweeted. We used the same break point (171) as the threshold for assigning tweets into groups with few followers and many followers. The sample size was 8,245 because we can analyze data only when there are tweets in all four groups on the same day for the same firm.

Tables 4a and (b) shows the correlations among the sentiment variables. If the sentiment in tweets from those with few and many followers are highly correlated, multicollinearity could bias the results. The correlations indicate little risk due to

Table 4b: Correlations among the sentiment variables used in H2.

	1	2	3	4	5	6	7	8	9	10	11
1. Pos1 over not retweeted											
2. Pos1 over retweeted	.189										
3. Pos1 under not retweeted	.103	.067									
4. Pos1 under retweeted	.078	.209	.102								
5. Pos2 over not retweeted	.927	.235	.125	.117							
6. Pos2 over retweeted	.174	.910	.072	.265							
7. Pos2 under not retweeted	.096	.077	.915	.115	.091						
8. Pos2 under retweeted	.074	.232	.101	.063	.074	.141					
9. Neg over not retweeted	.793	.151	.085	.063	.142	.405	.021				
10. Neg over retweeted	.145	.436	.007	.074	.142	.405	.021	.087			
11. Neg under not retweeted	.069	.495	.003	.001	.050	.002	.002	.002	.066		
12. Neg under retweeted	.067	.073	.010	.411	.041	.041	.041	.383	.065	.304	.199

Note: All correlations shown in bold are significant at $p < .001$. Cells in gray are correlations among variables used in the same regression analysis.

multicollinearity, with the highest correlation between variables in the same model being less than .30. We included Variance Inflation Factors in all analyses and found that most were less than 1.2, and none exceeded 2.0, indicating that it is highly unlikely that our data suffer from multicollinearity.

One of the issues with large data sets is that the traditional approach of using p values can be misleading because the large sample size means that any relationship is likely to be significant (Lin, Lucas, & Shmueli, 2013). Lin et al. (2013) offer several strategies for the analysis of large data sets. We adopt three of their recommendations, plus a fourth traditionally used in the analysis of large sample stock return data. First, we present confidence intervals for the size of effects. Second, we conduct a series of robustness checks using alternate models to see the extent to which our results are dependent on the specific models we use. Third, we examine the predictive ability of the models by comparing them to a controls-only model using symmetric mean absolute percent error (SMAPE), the absolute value of the difference between the predicted and actual divided by the mean of the absolute value of the predicted and the absolute value of the actual (Armstrong, 1985; Makridakis, 1993; Tofallis, 2015). Finally, one of the strongest tests of the practical significance of models used to predict stock returns is a trading strategy analysis—a test of whether an investor who builds a trading strategy using the results would experience a profit after accounting for trading costs (Tetlock et al., 2008).

RESULTS

Impact of Followers

We begin with H1, which argues that future abnormal returns would be directly related to the sentiment in tweets from users with few followers. Table 5 shows that the beta coefficients are significant and in the hypothesized direction for all three measures (pos1, pos2, and neg) for all three time periods (next day, next-to-10th-day, and next-to-20th-day), except for next day returns for neg, which is in the hypothesized direction but not significant.

The adjusted R^2 for these analyses are in a similar range to those in other studies of stock returns, such as (Chordia, Roll, & Subrahmanyam, 2002; Tetlock, 2007; Tetlock et al., 2008; Bollen et al., 2011a; Chen et al., 2014). SMAPE values for the models with only the five control variables for next day returns, next-to-10th-day, and next-to-20th-day are 1.9231, 1.8685, and 1.8210, respectively. The SMAPE values for all nine models in Table 5 are below the SMAPE values for the controls-only models, indicating they are a better fit. Based on these results (8 of 9 hypothesis tests significant, R^2 equivalent to R^2 in prior research, and lower SMAPE), we conclude that H1 is supported.

Combined Impact of Followers and Retweets

H2 argues that future abnormal returns would be directly related to the sentiment in tweets from users with few followers that are not retweeted. Table 6 shows that the beta coefficients on the sentiment in tweets from those with few followers that were not retweeted are significant and in the hypothesized direction for all

Table 5: Regression results of emotional sentiment on abnormal returns by number of followers.

(a) Pos1	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.140	1.020*	1.103
Sentiment_Under i,t	0.388**	1.512***	2.582***
Control 1 $i,t-1$	0.004	0.017	-0.011
Control 2 $i,t-30,t-2$	0.001	0.013***	0.022***
Surprise i,t	0.010***	0.004	0.001
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.003
Intercept	0.000***	-0.003***	-0.006***
SMAPE	1.9208	1.8662	1.8200
Adj. R^2	0.001	0.001	0.001

(b) Pos2	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.105	0.618	0.133
Sentiment_Under i,t	0.349**	1.405***	2.674***
Control 1 $i,t-1$	0.004	0.017	-0.010
Control 2 $i,t-30,t-2$	0.001	0.012***	0.022***
Surprise i,t	0.010***	0.004	0.001
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.003
Intercept	0.000***	-0.004***	-0.007***
SMAPE	1.9190	1.8648	1.8187
Adj. R^2	0.002	0.001	0.002

(c) Neg	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	2.313	-13.140	-6.829
Sentiment_Under i,t	-3.265	-22.735***	-52.175***
Control 1 $i,t-1$	0.004	0.018	-0.010
Control 2 $i,t-30,t-2$	0.001*	0.013***	0.022***
Surprise i,t	0.010***	0.004	0.001
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.003
Intercept	0.000***	-0.001**	-0.003***
SMAPE	1.8200	1.8187	1.8201
Adj. R^2	0.001	0.001	0.002

Notes: The coefficients are multiplied by 1,000.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

three measures (pos1, pos2, and neg) for all three time periods (next day, next-to-10th-day, and next-to-20th-day), except for next-to-10th day returns for neg which is in the hypothesized direction but not significant. Table 7 presents confidence intervals for the betas.

The adjusted R^2 for these analyses are equivalent to or substantially higher (by an order of magnitude—i.e., 1,000%) than adjusted R^2 in prior studies (Chordia

Table 6: Regression results of emotional sentiment and retweeting on abnormal returns by number of followers.

(a) Pos1	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	-0.531	2.043	1.871
Sentiment_OverRe i,t	0.162	-0.407	-1.000
Sentiment_UnderNo i,t	0.917*	3.598**	5.850**
Sentiment_UnderRe i,t	0.468	-0.037	-0.122
Control 1 $i,t-1$	0.020*	0.006*	0.036
Control 2 $i,t-30,t-2$	0.003	0.043***	0.082***
Surprise i,t	0.008**	0.000	0.007
Upgrade i,t	0.002**	-0.002	-0.005
Downgrade i,t	-0.001	-0.005*	0.004
Intercept	-0.001***	-0.005***	-0.009***
SMAPE	0.2219	0.2148	0.2172
Adj. R ²	0.002	0.008	0.013

(b) Pos2	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	-0.126	0.778	-0.116
Sentiment_OverRe i,t	-0.327	-0.412	-1.214
Sentiment_UnderNo i,t	0.831**	3.353***	4.649***
Sentiment_UnderRe i,t	0.522	0.221	0.901
Control 1 $i,t-1$	0.020*	0.064	0.036
Control 2 $i,t-30,t-2$	0.004	0.043***	0.008***
Surprise i,t	0.008*	0.000	0.007
Upgrade i,t	0.002	-0.002	-0.005
Downgrade i,t	-0.001	0.005	-0.004
Intercept	-0.001***	-0.005***	-0.009***
SMAPE	0.2217	0.2146	0.2172
Adj. R ²	0.003	0.009	0.013

(c) Neg	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	3.008	-57.433	-79.755
Sentiment_OverRe i,t	8.017	10.249	25.681
Sentiment_UnderNo i,t	-14.274*	-40.149	-67.144*
Sentiment_UnderRe i,t	-7.962	3.638	2.038
Control 1 $i,t-1$	0.020*	0.064*	0.035
Control 2 $i,t-30,t-2$	0.003	0.004***	0.082***
Surprise i,t	0.009*	0.000	0.007
Upgrade i,t	0.002*	-0.002	-0.005
Downgrade i,t	-0.001	0.005*	0.004
Intercept	0.000	0.000	-0.003
SMAPE	0.2219	0.2148	0.2172
Adj. R ²	0.002	0.008	0.012

Notes: The coefficients are multiplied by 1,000.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table 7: Confidence intervals for beta for sentiment in non-retweeted tweets from individuals with 171 or fewer followers.

		Next Day	Next-to-10th-Day	Next-to-20th-Day
(a) Pos1 Sentiment_UnderNo _{i,t}		0.053 to 1.781	0.946 to 6.250	2.048 to 9.652
(b) Pos2 Sentiment_UnderNo _{i,t}		0.213 to 1.449	1.459 to 5.247	1.932 to 7.366
(c) Neg Sentiment_UnderNo _{i,t}		Next Day -0.005 to -28.543	Next-to-10th-Day 3.635 to -83.933	Next-to-20th-Day -4.361 to -129.927

et al., 2002; Tetlock, 2007; Tetlock et al., 2008; Bollen et al., 2011a; Chen et al., 2014). SMAPE values for the models with only the five control variables for next day returns, next-to-10th-day, and next-to-20th-day are 1.8476, 1.7763, and 1.7042, respectively. The SMAPE values for all nine models in Table 6 are below the SMAPE values for their matching controls-only models, indicating they are a better fit. Based on these results (eight of nine hypothesis tests significant, R^2 equivalent to or higher than R^2 in prior research, and lower SMAPE), we conclude that H2 is supported.

These results support our arguments that the speed of sentiment diffusion affects future returns. When sentiment spreads the slowest (i.e., tweets sent by those with fewer than the median number of followers [171] that are not retweeted), it affects stock returns on future trading days.

Robustness Checks

We conducted several robustness checks. We ran separate analyses for H1 and H2 treating missing values as zero emotion (which produced sample sizes of 83,891) and found the same pattern of results.

We conducted a separate analysis for H1 using a different split in the number of followers. We used 1,000 followers as the threshold between many and few. A user with 1,000 followers falls in the top 4% of all Twitter users (“About Twitter, Inc.,” 2014), so they are what Barabási (2002) would call hubs. The Twitter users with over 1,000 followers were typically well-known media or analysts, such as CNN, ABC, WSJ, CNBC, Fox News, *Fortune Magazine*, and Jim Cramer, who is a writer, TV show host, and co-founder of *TheStreet.com*. This produced the analysis of “few” and “many” followers, with a 96–4% split. The split using 1,000 followers followed the same pattern as with the median split (Table A1 in the Appendix). We conclude that this hypothesis is robust to the choice of threshold for the number of followers. For this hypothesis, as long as one is not a hub (i.e., the top 4% of all Twitter users), it takes days before sentiment in your tweets spreads.

We conducted a similar analysis for H2 using the 1,000 follower threshold (96–4% split). The sample size here was 9,014. The pattern here was different; only three of the nine hypothesis tests were significant (Table A2 in the Appendix). We conclude that this hypothesis is not robust to the choice of threshold for what “few” followers means. If we consider retweeting behavior, then the threshold

Table 8: Annualized returns from a trading strategy using sentiment in non-retweeted tweets from individuals with 171 or fewer followers.

	Holding Period		
	1 Day	10 Days	20 Days
Without trading costs	11.44%	17.91%	12.59%
With trading costs	-28.57%	15.65%	11.41%

number to identify “few” followers must be such that we do not consider only hubs and nonhubs in the network.

We examined the effects of the sentiment contained in tweets only from those with many followers, omitting the sentiment in tweets from those with few followers. Four of the nine hypothesis tests were significant (Table A3 in the Appendix), but the SMAPE values are not improved. We conclude that the sentiment in tweets from those with many followers is not consistently related to future returns.

If the sentiment in tweets does affect the market, we are likely to see an effect on the same day by users with many followers. For example, if Jim Cramer tweets positively or negatively about a specific stock then its price should move quickly. Table A4 in the Appendix shows the same day effects (i.e., $t = 0$); there is a significant same day effect for the sentiment in tweets from users with many followers both with and without considering the sentiment in tweets from those with few followers. All three betas on the sentiment from those with many followers are greater than the corresponding betas for those with few followers, with significance of $p < .001$. The SMAPE values are lower than those of the corresponding controls-only models.

Effectiveness of a Trading Strategy

One important question is whether these results can be used to build a profitable trading strategy. We theorized, and the empirical results show that the tweets from users with few followers that are *not* retweeted lead to the greatest abnormal returns. We followed the approach of Tetlock et al. (2008) and constructed two equally weighted portfolios, one long, one short. At the close of each trading day, we analyzed the sentiment in that day’s tweets about specific firms using pos2 (we choose pos2 because it takes into account both positive and negative sentiment). We purchase firms in the top 10% and short sell the firms in the bottom 10%. Not all firms receive tweets each day, so the number of firms varies from day to day. We used three different holding periods (1 day, 10 days, and 20 days), and at the end of the holding period, we close out our long and short positions. Because we are simultaneously taking long and short positions, there is no need to consider market return as a control; any rise or fall in the market as a whole is controlled for by the simultaneous long and short positions.

Table 8 presents the annualized returns of the trading strategy with and without trading costs for the three different holding periods. Following Tetlock

et al. (2008), we assume round-trip trading costs of 10 basis points (i.e., the total cost to buy and sell). The 1-day holding period produces positive returns, but because the strategy executes trades every day, the returns become negative after including trading costs. The trading strategies using 10- and 20-day holding periods produce significant positive returns, both before and after trading costs. These returns compare favorably with those in Tetlock et al. (2008), who found trading strategies using sentiment in news stories produced returns of 23.17% before trading costs and -2.71% after trading costs (i.e., a loss). In other words, the results in Table 8 show that a trading strategy with a 10- or 20-day holding period that balances long and short positions results in meaningful positive returns.

DISCUSSION

We theorized that the sentiment in tweets is related to stock returns of individual stocks based on how fast that sentiment spreads through the market. The sentiment in tweets sent by users with few followers, which diffuses more slowly than sentiment in tweets sent by those with many followers, is significant in predicting the firm's stock returns one trading day, 10 trading days, and 20 trading days after the tweets were posted. The sentiment in tweets from those with few followers that were not retweeted had the strongest effect on returns on future days, as this sentiment takes the longest to spread through the market. A trading strategy with a 10- or 20-day holding period built on these factors shows meaningful annual returns.

We argue that a social contagion process is at work. Tweets spread positive or negative sentiment about a stock through the market and can influence prices, and thus the returns from trading those stocks. Sentiment can spread quickly; for example, network hubs like Jim Cramer can send a tweet that has positive or negative sentiment about a stock and his 650,000 followers immediately see its sentiment. If these followers act, the stock price can respond very quickly. Thus, we conclude that Twitter users with many followers have a market impact similar to traditional news media; the impact of the sentiment in their tweets disseminates rapidly and is quickly incorporated into stock price. However, there are no significant stock returns on future trading days and thus it is difficult to profit from a trading strategy based on them. In contrast, if a user with few followers sends a tweet, few people will see it, and even if they quickly act on its sentiment, the small number of trades will have little immediate impact on the stock price. Over time, however, as the sentiment diffuses through the market, the sentiment will gradually affect the stock price. The sentiment in tweets from users with fewer followers had a stronger impact on stock returns the next trading day and over the next 10–20 days compared to tweets from users with many followers. Because the change is gradual, there is an opportunity to profit from this as a trading strategy. The diffusion is slower for tweets that are not retweeted, and thus they have the greatest impact on returns on future trading days and offer the greatest opportunities to profit from a trading strategy based on this.

Our results offer similar conclusions to other research based on the GIFT model of information diffusion. GIFT argues that markets are generally efficient at the macro level, but if we examine them at a micro level, we see that it is possible

to uncover situations in which markets are not perfectly efficient, because human behavior is not perfectly efficient at spreading information. Studies in line with GIF have suggested that it takes longer for information about stocks that are not routinely covered by the mass media to be absorbed into their price (Merton, 1987; Fang & Peress, 2009). Thus, it is possible to use information to predict future stock returns for firms with systematically slower information diffusion, even after controlling for risk characteristics (Merton, 1987; Fang & Peress, 2009). Thus, we believe the GIF model provides a good theoretical foundation for understanding why Twitter sentiment can be used to predict future stock returns in some cases but not in others.

There exist at least two possible theoretical explanations for how the speed of information diffusion influences stock returns. The first is that the sentiment of tweets “causes” changes in stock prices. Individuals post tweets when they believe they have useful comments about an individual stock. These comments may have facts as well as an underlying sentiment. Sentiment is highly contagious (Schoenewolf, 1990; Hatfield et al., 1993), and it influences how investors make buy/sell decisions as the sentiment spreads through the public. A cumulative positive sentiment triggers positive thoughts about the company and leads to a purchase decision, raising the stock price. A cumulative negative sentiment induces negative thoughts and thus leads to sell decisions, decreasing the stock price. The rate at which this sentiment spreads through the market is influenced by the number of followers of the sender and whether the tweet is retweeted, so that sentiment in tweets from those with few followers that are not retweeted takes a longer time to spread and thus takes longer to influence stock prices; this leads to significant stock returns on future trading days.

A second possible explanation is that tweets “reflect” the underlying information that influences individual stock returns. In this case, it is not the sentiment of the tweets themselves that influence stock returns, but rather the tweets reflect how investors feel about the stock and are a leading indicator of their buy/sell decisions. Investors planning to buy a stock have positive sentiment about the stock and communicate this sentiment in their tweets. Likewise, investors planning to sell a stock communicate negative sentiment in their tweets.

We believe that the first explanation, that the sentiment of tweets causes stock price changes, best explains our findings because the number of followers and whether the tweet was retweeted were significantly related to stock price changes. Two assumptions would have to hold for the explanation that tweets reflect information to be viable. First, investors would have to tweet their information days before acting on it, which is illogical; no rational investor would share information likely to affect prices before acting on it. Second, the underlying information would need to spread through the social network in the same manner as the tweets and retweets but via a different mechanism in order for us to find the relationships we did. This is a less likely explanation than the simpler explanation that it is the tweets themselves that influence behavior.

CLT argues that positive and negative sentiment may have different effects (Liberman & Trope, 1998; Bar-Anan et al., 2006; Fujita et al., 2006). Previous research has shown that both positive and negative sentiment in tweets can affect stock returns, but no study has found both to have effects (Smailović et al., 2014;

Risius et al., 2015). Interestingly, we found both positive and negative sentiment to directly affect returns.

Investors make investment decisions using a variety of information sources, with Twitter being just one of many possible sources. The economic magnitude of the relationships in our study is moderate to high (Tetlock et al., 2008). The trading analysis showed positive annual returns for 10- and 20-day holding periods after considering trading costs. The economic significance of these effects is meaningful.

Limitations

This study also has several limitations. We only studied firms in the S&P 500. We have no empirical data to argue that our results apply or do not apply to smaller firms or firms traded in other markets that are not covered by the S&P 500. We studied one specific time period in the life of the market, so it could be that the market conditions that led to our findings no longer apply. Likewise, we studied the same time period in the life of Twitter, and because Twitter behavior changes over time (Liu et al., 2014), it may be that Twitter users behave differently now, and the behaviors we observed no longer occur.

The fundamental theory underlying our research is the GIF model. Our results are driven by the speed of diffusion of sentiment, so one important theoretical limitation is if this sentiment is based on already widely diffused fundamental information (e.g., a rise in oil prices that could negatively influence transportation stocks) then this sentiment is likely to have little effect on stock returns, because investors have already acted and prices have already changed. We did not examine the extent to which the sentiment in the tweets we analyzed was based on already disseminated fundamental information, so this is an avenue for future research.

Another potential limitation is homophily, the possibility that individuals similar to each other tend to post similar Tweets (Aral, Muchnik, & Sundararajan, 2009; Shalizi & Thomas, 2011). Under this argument, the changes in stock prices are not due to social contagion but are because people similar to each other in the number of followers use similar trading strategies. This is possible, but we view this as less likely than social contagion because it could only be true if trading behaviors were related to the number of followers and the retweet history of tweets. This is possible, but requires additional, somewhat convoluted theorizing to link the number of Twitter followers and retweeting history to trading behaviors. While homophily is useful in understanding some tweeting behaviors, it is often not as powerful as other theoretical models (Macskassy & Michelson, 2011). So, using Occam's razor, we conclude that the social contagion of sentiment is a better explanation for our results.

Implications for Research

Despite these limitations, we believe that these results have implications for future research. Our work builds on recent research showing that the "calmness" or "depression" in tweets (Bollen et al., 2011a; Risius et al., 2015), their "bullishness" or "beariness" (Oh & Sheng, 2011; Sprenger et al., 2014), and their sentiment (Smailović et al., 2014; Risius et al., 2015) can be useful in predicting stock returns. We use social contagion based on the GIF model as underlying theory and show that factors which influence the speed of sentiment diffusion (number

of followers, retweeting) significantly affect the stock returns on future trading days. We offered two possible explanations for the theoretical mechanism that links the sentiment in tweets to future stock returns. We need more research to better understand the underlying theoretical mechanism that links sentiment to stock returns.

Our research shows that *who* sends the tweets is an important factor in explaining how the sentiment in tweets affects stock returns. Ironically, users with many followers (i.e., those with more than the median number of 171) have *no* significant influence on stock returns on the next trading day or subsequent days. The sentiment in their tweets is quickly incorporated into stock prices leading to no future returns. In contrast, sentiment expressed by Twitter users who have few followers—and thus diffuses slowly—has significant and meaningful impacts on stock returns on future trading days. We used a simple analysis that divided users into two groups, over and under the median number of followers in our data set. We believe that this calls for more research into who expresses the sentiment in tweets and how this can be used to explain stock price movements and better predict stock returns. The number of connections in a social network typically follows a power law distribution (Barabási, 2002), so an analysis that uses more than two categories to better capture this distribution may better model the speed of sentiment diffusion and provide additional insight.

Our research also shows that what happens to the tweet *after* it is sent has a significant impact on stock returns. Tweets that are retweeted have a faster impact on stock prices and thus do not predict stock returns on future trading days, whereas tweets that are not retweeted can predict future returns. We believe this calls for more research into retweeting behavior. For example, how do investors react to the tweets they receive that are and are not retweeted? The most common reason for retweeting is because the sender believes the tweet's information would be of interest to their followers (Macskassy & Michelson, 2011). Do retweeted tweets appear more important and thus get more attention, so they are more likely to influence behavior?

Prior studies examining how emotion is linked to stock returns have used different approaches to measuring it, including emotional states (e.g., calm, happiness, depression) and sentiment (e.g., positive emotional valence, "bullishness") (Calvo & Kim, 2013). We used the Harvard IV Psychological Dictionary to assess the positive and negative sentiment expressed in the tweets. There are many dictionaries designed to categorize words based on sentiment, such as Loughran and McDonald Financial Sentiment Dictionaries (Loughran & McDonald, 2011). We did not include emoticons in our analysis, which could be examined in future research.

We used the formulas of Tetlock et al. (2008) to build three different measures of sentiment (pos1, pos2, neg) that provided essentially the same conclusions (with some minor differences among them). There are many other formulas and machine learning techniques that can be used to develop sentiment metrics that are more sophisticated (e.g., Oh & Sheng, 2011; Smailović et al., 2014; Sprenger et al., 2014). One key challenge in sentiment analysis is understanding semantics in groups of words. A Twitter post may have both positive and negative terms, so if one considers the semantic rules of groups of words, the meaning may become

clearer (e.g., “this is not bad.”) (Xie et al., 2015). Additional research is needed that uses different, more sophisticated, analysis strategies to better understand if different approaches are better at predicting stock returns on future trading days.

In this study, we use Twitter as the social media platform. There are many other social media platforms that may also provide insights into future stock returns. We hope our work can spawn future research on this topic. What are the impacts of Facebook, LinkedIn, or other Web media?

Finally, we examined the impact of sentiment on stock returns at a daily level. Future research could use market microstructural data to examine how emotional state and sentiment impact markets in real time.

Implications for Practice

We believe that this study has two implications for practice. The first is providing guidance to investment decision making. Our results show that a trading strategy built on the analysis of the sentiment in tweets from users with few followers that are not retweeted produces significant positive returns after considering trading costs. Tweets are available publically and can be retrieved using Twitter development accounts, so this may be an investable trading strategy. Combining this with a focus on firms that have little coverage from the traditional media may also increase returns (Merton, 1987; Fang & Peress, 2009).

A second implication is that firms should carefully monitor how they use Twitter. Most firms manage formal financial information that could impact stock prices because there are numerous financial regulations in place. Because the sentiment of tweets is linked to future stock prices, firms need to monitor the sentiment in their tweets in addition to the “rational” information they contain.

CONCLUSION

We found that the sentiment in social media postings can predict stock returns on future trading days. Tweets from users with few followers (i.e., less than the median of 171 followers) that were not retweeted had an impact on future returns 10 and 20 days later, while those from users with many followers and those that were retweeted had no impact on future returns. The findings are consistent with our hypothesis that sentiment that is diffused slowly takes longer to be incorporated into prices, while sentiment that is diffused faster will be quickly incorporated into prices and thus will have little association with returns on future days.

REFERENCES

- Abbasi, A., & Chen, H. (2008). CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, 32(4), 811–837.
- About Twitter, Inc. (2015). Accessed March 14, 2016. Available at: <https://about.twitter.com/company>
- Aral, S., Muchnik, L., & Sundararajan, A., (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51), 21544–21549.

- Armstrong, J. S. (1985). *Long range forecasting: From crystal ball to computer* (2nd ed.). New York, NY: Wiley.
- Bakamitsos, G. A. (2006). A cue alone or a probe to think? The dual role of affect in product evaluations. *Journal of Consumer Research*, 33(3), 403–412.
- Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), 129–151.
- Bar-Anan, Y., Liberman, N., & Trope, Y. (2006). The association between psychological distance and construal level: Evidence from an implicit association test. *Journal of Experimental Psychology-General*, 135(4), 609–622.
- Barabási, A. L. (2002). *Linked*. Cambridge, MA: Persues Publishing.
- Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52(2), 336–372.
- Bollen, J., Mao, H., & Zeng, X. (2011a). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Bollen, J., Pepe, A., & Mao, H. (2011b). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.
- Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50(2), 260–268.
- Calvo, R. A., & Kim, S. M. (2013). Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3), 527–543.
- Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5), 1367–1403.
- Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics*, 65(1), 111–130.
- Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5), 1461–1499.
- Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N., & Walker, J. (1988). Predicting hospital-associated mortality for Medicare patients. *JAMA: The Journal of the American Medical Association*, 260(24), 3617–3624.
- Dhir, A., Buragga, K., & Boreqqah, A. A. (2013). Tweeters on campus: Twitter a learning tool in classroom? *Journal of Universal Computer Science*, 19(5), 672–691.
- Dockery, E., & Kavussanos, M. G. (1996). Testing the efficient market hypothesis using panel data, with application to the Athens stock market. *Applied Economics Letters*, 3(2), 121–123.

- Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- Eppen, G. D., & Fama, E. F. (1969). Cash balance and simple dynamic portfolio problems with proportional costs. *International Economic Review*, 10(2), 119–133.
- Eyal, T., Liberman, N., Trope, Y., & Walther, E. (2004). The pros and cons of temporally near and distant action. *Journal Personality and Social Psychology*, 86(6), 781–795.
- Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5), 2023–2052.
- Fedorikhin, A., & Patrick, V. M. (2010). Positive mood and resistance to temptation: The interfering influence of elevated arousal. *Journal of Consumer Research*, 37(4), 698–711.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82–89.
- Frijda, N. H. (1994). Varieties of affect: Emotions and episodes, moods, and sentiments. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions*. Oxford, UK: Oxford University, 197–202.
- Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, 90(3), 351–367.
- Gordon, T., Castelli, W. P., Hjortland, M. C., Kannel, W. B., & Dawber, T. R. (1977). Predicting coronary heart disease in middle-aged and older persons. *JAMA: The Journal of the American Medical Association*, 238(6), 497–499.
- Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, 2(3), 96–99.
- Honey, C., & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. *Proceedings of the 42nd Hawaii International Conference on System Science*, Waikoloa.
- Hong, H., Lim, T., & Stein, J. C. (2000). Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance*, 55(1), 265–295.
- Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6), 2143–2184.
- Hong, H., & Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2), 109–128.
- Jorgenson, D. W., & Vu, K. (2005). Information technology and the World economy. *The Scandinavian Journal of Economics*, 107(4), 631–650.
- Labroo, A. A., & Patrick, V. M. (2009). Psychological distancing: Why happiness helps you see the big picture. *Journal of Consumer Research*, 35(5), 800–809.
- Larsson, A. O. (2013). Tweeting the viewer—Use of Twitter in a talk show context. *Journal of Broadcasting & Electronic Media*, 57(2), 135–152.

- Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology*, 75(1), 5–18.
- Lin, M., Lucas, H. C. Jr., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4), 906–917.
- Liu, Y., Kilman-Silver, C., & Mislove, A. (2014). The tweets they are a-changin': Evolution of Twitter users and behavior. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, Ann Arbor, MI.
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.
- Macskassy, S. A., & Michelson, M. (2011). "Why do people retweet? Anti-homophily wins the day!" *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona.
- Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4), 527–529.
- Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42(3), 483–510.
- Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, 37(2), 407–426.
- Oh, O., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *Proceedings of the International Conference on Information Systems*, Shanghai.
- Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois.
- Parboteeah, D. V., Valacich, J. S., & Wells, J. D. (2009). The influence of website characteristics on a consumer's urge to buy impulsively. *Information Systems Research*, 20(1), 60–78.
- Pew-Research. (2014). Social media use over time, accessed November 11, 2014, available at <http://www.pewinternet.org/data-trend/social-media/social-media-use-all-users/>.
- Qian, B., & Rasheed, K., (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1), 25–33.
- Qiu, C., & Yeung, C. W. (2008). Mood and comparative judgment: Does mood influence everything and finally nothing? *Journal of Consumer Research*, 34(5), 657–669.
- Rai, A., Patnayakuni, R., & Seth, N. (2006). Firm performance impacts of digitally enabled supply chain integration capabilities. *MIS Quarterly*, 30(2), 225–246.
- Risiis, M., Akolk, F., & Beck, R. (2015). Differential emotions and the stock market—The case of company-specific trading. *European Conference on*

- Information Systems*, Completed Research Papers, Munster, Germany, Paper 147.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.
- Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. *Proceedings of the 19th International Conference on the World Wide Web*. Raleigh, NC.
- Schoenewolf, G. (1990). Emotional contagion: Behavioral induction in individuals and groups. *Modern Psychoanalysis*, 15(1), 49–61.
- Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2), 211–239.
- Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156(1), 81–96.
- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršić, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285(32), 181–203.
- Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5), 926–957.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168.
- Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467.
- Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation, *Journal of the Operational Research Society*, 66(8), 1352–1362.
- trade-IQ. (2011). The U.S. active trader market; Report Preview, accessed December 8, 2015, available at <http://tradeiq.blogspot.com/2012/02/us-active-trader-market-report-preview.html>.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M., (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC.
- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 165–210.
- Xie, Y., Chen, Z., Zhang, L. K., Cheng, Y., Honbo, D., & Agrawal, A. (2015).et al. MuSES: Multilingual sentiment elicitation system for social media data, *IEEE Intelligent Systems*, 29(4), 34–42.

Appendix

Table A1: Regression results of emotional sentiment on abnormal returns by number of followers using a breakpoint of 1,000 followers (a 96–4% split). The SMAPE values for the controls-only models are 1.9182, 1.8694, and 1.8116.

(a) Pos1	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.193	0.428	0.088
Sentiment_Under i,t	0.384**	1.248**	1.871**
Control 1 $i,t-1$	0.006	0.022	0.007
Control 2 $i,t-30,t-2$	0.002*	0.016***	0.027***
Surprise i,t	0.010***	0.006	0.005
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.002
Intercept	0.000***	-0.003***	-0.006***
SMAPE	1.9146	1.8594	1.8113
Adj. R^2	0.002	0.001	0.001

(b) Pos2	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.164	0.114	-0.625
Sentiment_Under i,t	0.301**	1.145***	1.988***
Control 1 $i,t-1$	0.006	0.022	0.007
Control 2 $i,t-30,t-2$	0.002*	0.016***	0.027***
Surprise i,t	0.010***	0.006	0.005
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002**	0.000	-0.002
Intercept	0.000***	-0.003***	-0.006***
SMAPE	1.9142	1.8589	1.8107
Adj. R^2	0.002	0.001	0.001

(c) Neg	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	1.885	-4.643	-1.104
Sentiment_Under i,t	-2.922	-20.715**	-14.126***
Control 1 $i,t-1$	0.006	0.023*	0.026***
Control 2 $i,t-30,t-2$	0.002*	0.016***	0.007
Surprise i,t	0.010***	0.006	0.005
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.002
Intercept	0.000*	-0.002***	-0.004***
SMAPE	1.9170	1.8597	1.8112
Adj. R^2	0.002	0.001	0.001

Notes: The coefficients are multiplied by 1,000.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table A2: Regression results of emotional sentiment and retweeting on abnormal returns by number of followers using a breakpoint of 1,000 followers (a 96–4% split). The SMAPE values for the controls-only models are 1.8577, 1.7660, and 1.7168, respectively.

(a) Pos1	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	-1.021	2.814	0.672
Sentiment_OverRe i,t	0.146	-0.481	0.309
Sentiment_UnderNo i,t	1.185*	1.240	3.737
Sentiment_UnderRe i,t	0.526	0.998	0.922
Control 1 $i,t-1$	-0.003	0.058*	0.037
Control 2 $_{i,t-30,t-2}$	0.008***	0.047***	0.082***
Surprise i,t	0.007*	0.001	-0.005
Upgrade i,t	0.002	-0.002	-0.005
Downgrade i,t	-0.002***	0.003	0.002
Intercept	-0.001**	-0.005***	-0.008***
SMAPE	0.3253	.3145	0.3158
Adj. R^2	0.004	0.009	0.012

(b) Pos2	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	-0.655	1.079	-0.745
Sentiment_OverRe i,t	-0.165	-1.127	-0.604
Sentiment_UnderNo i,t	0.896*	1.444	3.643*
Sentiment_UnderRe i,t	0.384	1.186	0.856
Control 1 $i,t-1$	-0.003	0.059*	0.038
Control 2 $_{i,t-30,t-2}$	0.008***	0.047***	0.083***
Surprise i,t	0.008*	0.001	-0.005
Upgrade i,t	0.002	-0.002	-0.005
Downgrade i,t	-0.002***	0.003	0.002
Intercept	-0.001**	-0.005***	-0.008
SMAPE	0.3252	0.3144	0.3156
Adj. R^2	0.004	0.009	0.012

(c) Neg	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_OverNo i,t	6.373	-59.201*	-39.345
Sentiment_OverRe i,t	4.147	32.940	29.771
Sentiment_UnderNo i,t	-13.255	-27.298	-65.048
Sentiment_UnderRe i,t	-8.617	-21.707	-22.808
Control 1 $i,t-1$	-0.003	0.059*	0.037
Control 2 $_{i,t-30,t-2}$	0.008***	0.046***	0.082***
Surprise i,t	0.008*	0.001	-0.005
Upgrade i,t	0.002	-0.002	-0.005
Downgrade i,t	-0.002***	0.003	0.002
Intercept	0.000	0.000	-0.002
SMAPE	0.3257	0.3145	0.3157
Adj. R^2	0.004	0.009	0.012

Notes: The coefficients are multiplied by 1,000.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table A3: Regression results of emotional sentiment on abnormal returns using only those with more than 171 followers. The SMAPE values for the controls-only models are .9678, .9399, and .9216.

(a) Pos1	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.056	0.832**	0.952*
Control 1 $i,t-1$	0.002	0.011	-0.014
Control 2 $i,t-30,t-2$	0.001	0.007***	0.013***
Surprise i,t	0.012***	0.012*	0.014
Upgrade i,t	0.002***	0.001	-0.001
Downgrade i,t	-0.002***	0.001	-0.004**
Intercept	0.000***	-0.003***	-0.006***
SMAPE	0.9677	0.9399	0.9216
Adj. R^2	0.001	0.000	0.001

(b) Pos2	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	-0.041	0.587*	0.395
Control 1 $i,t-1$	0.002	0.011	-0.010
Control 2 $i,t-30,t-2$	0.001	0.012***	0.013***
Surprise i,t	0.012***	0.004	0.014
Upgrade i,t	0.002***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.004***
Intercept	0.000***	-0.003***	-0.004***
SMAPE	0.9677	0.9399	0.9216
Adj. R^2	0.001	0.000	0.001

(c) Neg	Next Day	Next-to-10th-Day	Next-to-20th-Day
Sentiment_Over i,t	0.149	-12.808**	-9.950
Control 1 $i,t-1$	0.001	0.018	-0.013
Control 2 $i,t-30,t-2$	0.001	0.013***	0.013***
Surprise i,t	0.012***	0.004	0.014
Upgrade i,t	0.001***	0.000	-0.001
Downgrade i,t	-0.002***	0.000	-0.003
Intercept	0.000***	-0.001**	-0.003**
SMAPE	0.9678	0.9399	0.9216
Adj. R^2	0.001	0.001	0.001

Note: The coefficients are multiplied by 1,000.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table A4: Regression results of emotional sentiment on abnormal returns by number of followers for the same trading day. The SMAPE values for the controls-only model are 1.7935 and .9286, respectively.

(a) Pos1	Both Over and Under	Over Only
Sentiment_Over _{i,t}	1.635***	1.135***
Sentiment_Under _{i,t}	0.508**	
Control 1 _{i,t-1}	-0.003	-0.003
Control 2 _{i,t-30,t-2}	0.001	0.001
Surprise _{i,t}	0.027***	0.022***
Upgrade _{i,t}	0.011***	0.010***
Downgrade _{i,t}	-0.017***	-0.015***
Intercept	0.000***	0.000***
SMAPE	1.7409	0.9018
Adj. <i>R</i> ²	0.036	0.027

(b) Pos2	Both Over and Under	Over Only
Sentiment_Over _{i,t}	1.627***	1.318***
Sentiment_Under _{i,t}	0.506***	
Control 1 _{i,t-1}	-0.004	-0.004
Control 2 _{i,t-30,t-2}	0.001	0.001
Surprise _{i,t}	0.027***	0.022***
Upgrade _{i,t}	0.011***	0.009***
Downgrade _{i,t}	-0.017***	-0.015***
Intercept	0.000***	0.000***
SMAPE	1.7206	0.8910
Adj. <i>R</i> ²	0.038	0.028

(c) Neg	Both Over and Under	Over Only
Sentiment_Over _{i,t}	-23.415***	-16.503***
Sentiment_Under _{i,t}	-7.596**	
Control 1 _{i,t-1}	-0.002	-0.003
Control 2 _{i,t-30,t-2}	0.001	0.001
Surprise _{i,t}	0.027***	0.023***
Upgrade _{i,t}	0.011***	0.009***
Downgrade _{i,t}	-0.017***	-0.015***
Intercept	0.000***	0.000***
SMAPE	1.7495	0.9086
Adj. <i>R</i> ²	0.035	0.027

Notes: The coefficients are multiplied by 1,000.

p* ≤ .05, ** *p* ≤ .01, * *p* ≤ .001.

Hong Kee Sul is a research support director at Wharton Research Data Services, The Wharton School, University of Pennsylvania. In 2015, he graduated with a PhD in Finance from Indiana University, Kelley School of Business. Dr. Sul received a Master of Science degree from Seoul National University and a Bachelor of Science degree from Korean Advanced Institute of Science and Technology. His work has been presented in several conferences, including the 2016 American Economic Association Annual Meetings, 2014 Financial Management Association Annual Meeting Doctoral Consortium, and the 47th Annual Hawaii International Conference on System Sciences.

Alan Dennis is a professor of information systems and holds the John T. Chambers Chair of Internet Systems in the Kelley School of Business at Indiana University. He is a Fellow of the AIS and has written more than 150 research papers. His research focuses on three main themes: the use of computer technologies to support team creativity and decision making; IS for the subconscious; and the use of the Internet to improve education. He is the AIS Vice President of Conferences, the editor in Chief of *Foundations and Trends in Information Systems*, and co-editor in Chief of *AIS Transactions on Replication Research*. Prof. Dennis also has written four books, two on data communications and networking, and two on systems analysis and design. His most recent start-up company is NameInsights.com, which uses big data and analytics to help parents pick baby names.

Lingyao (Ivy) Yuan is an assistant professor of information systems of College of Business at Iowa State University. Her research interests include the impact of noncognition behavior and decision making, especially the impact of emotion, on computer-mediated communication, decision making, and collaboration. She has conducted research in the fields of electronic commerce and social media. She has been published in the *Group Decision and Negotiation* as well as several conferences, including *47th Annual Hawaii International Conference on System Sciences*, and *2013 INFORMS Annual Meeting*. She received a Master of Science in Information Technology from University of North Carolina Charlotte in 2011 and Bachelors of Management Information Systems from University of International Business and Economics in 2009.

NBER WORKING PAPER SERIES

USING SATELLITE IMAGERY TO UNDERSTAND AND
PROMOTE SUSTAINABLE DEVELOPMENT

Marshall Burke
Anne Driscoll
David Lobell
Stefano Ermon

Working Paper 27879
<http://www.nber.org/papers/w27879>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

We thank Jenny Xue, Brian Lin, and Zhongyi Tang for excellent research assistance, and thank USAID Bureau for Food Security, the Global Innovation Fund, Darpa World Modelers program, and the Stanford King Center on Global Development for funding. Data and code for replication of all results will be made public upon publication. M.B., D.L., and S.E. are co-founders of AtlasAI, a company that uses machine learning to measure economic outcomes in the developing world. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2020 by Marshall Burke, Anne Driscoll, David Lobell, and Stefano Ermon. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Using Satellite Imagery to Understand and Promote Sustainable Development
Marshall Burke, Anne Driscoll, David Lobell, and Stefano Ermon
NBER Working Paper No. 27879
October 2020
JEL No. C45,C55,O1

ABSTRACT

Accurate and comprehensive measurements of a range of sustainable development outcomes are fundamental inputs into both research and policy. We synthesize the growing literature that uses satellite imagery to understand these outcomes, with a focus on approaches that combine imagery with machine learning. We quantify the paucity of ground data on key human-related outcomes and the growing abundance and resolution (spatial, temporal, and spectral) of satellite imagery. We then review recent machine learning approaches to model-building in the context of scarce and noisy training data, highlighting how this noise often leads to incorrect assessment of models' predictive performance. We quantify recent model performance across multiple sustainable development domains, discuss research and policy applications, explore constraints to future progress, and highlight key research directions for the field.

Marshall Burke
Department of Earth System Science
Stanford University
Stanford, CA 94305
and NBER
mburke@stanford.edu

Anne Driscoll
Stanford University
616 Serra St
Stanford, CA 94305
anne.driscoll@stanford.edu

David Lobell
Stanford University
Department of Environmental
Earth System Science
Y2E2 Bldg - MC4205
473 Via Ortega, room 367
Stanford, CA 94305
dlobell@stanford.edu

Stefano Ermon
Stanford University
353 Serra Mall
Stanford, CA 94305
ermon@cs.stanford.edu

1 Introduction

Humans have long sought to image their habitat from above the ground. Socrates purportedly stated in 500 B.C.E. that “Man must rise above the earth – to the top of the atmosphere and beyond – for only thus will he fully understand the world in which he lives”.¹ His lofty goal was taken up in earnest after the advent of photography in the mid-nineteenth century C.E., with earth observation data collected by strapping cameras to balloons, kites, and pigeons. The first known image of earth from space was taken nearly a century later (1946) by American scientists using a captured Nazi rocket, revealing blurry expanses of the American Southwest.² This was followed decades later by the launch of the first civilian earth-observing satellite, Landsat I, in 1972, which ushered in the modern era of satellite-based remote sensing. As of early 2020, there are an estimated 713 active non-military earth observation satellites in orbit, 75% of which were launched in the last five years.³ These satellites are now capturing imagery of the earth in unprecedented temporal, spatial, and spectral frequency.

Here we review and synthesize a rapidly growing scientific literature that seeks to use this satellite imagery to measure and understand various human outcomes, including a range of outcomes directly linked to the Sustainable Development Goals. We pay particular attention to recent approaches that use methods from artificial intelligence to extract information from images, as these methods typically outperform earlier approaches, enabling new insights. Our focus is on settings and applications where humans themselves, or what they produce, are the outcome of interest, and where these outcomes are being predicted using satellite imagery. We quantify existing performance in these domains across a large set of studies, explore key constraints to future progress, and highlight a number of research directions that we believe are key if these approaches are going to be improved and adopted by practitioners.

We do not review and assess the large literature on using remote sensing for other earth observation tasks (e.g. environmental monitoring), or efforts that use other sources of non-traditional, unstructured data (e.g. data from social media or cell phones) to measure human-related outcomes. We discuss this work if these other unstructured data sources are used in combination with imagery for sustainability tasks. Our review complements existing sector specific reviews, including the use of remote sensing in agriculture,^{4,5} in economic applications,⁶ and in the detection of informal settlements,⁷ drawing common lessons across these and other domains.

Our review makes four main points. First, satellite-based performance in predicting key sustainable development outcomes is reasonably strong and appears to be improving. Indeed, analyses suggest that reported model performance likely understates true performance in many settings, given

the noisy data on which predictions are evaluated and the types of noise typically observed in sustainability applications. For multiple outcomes of interest, satellite-based estimates can now equal or exceed the accuracy of traditional approaches to outcome measurement.

Second, perhaps the largest constraint to model development is now training data rather than imagery. While imagery has become abundant, the scarcity and (in many settings) unreliability of quality ground data make both training and validation of satellite-based models difficult. Expanding the quantity and – in particular – the quality of labels will quickly accelerate progress in this field.

Third, despite the growing power of satellite-based approaches, we argue that in most settings, these approaches will amplify rather than replace existing ground-based data collection efforts. Many outcomes of interest will likely never be accurately estimated with satellites; for outcomes where satellites do have predictive power, high-quality local training data can nearly always improve model performance.

Finally, there remain few documented cases where satellites have been operationalized into public-sector decision-making processes in the sustainable development domains where we focus – with applications in population and agricultural measurements being the main exceptions. Limited adoption is likely driven by a number of forces, including the recency of the technology, the lack of accuracy (perceived or real) of the models, lack of model interpretability, and entrenched interests in maintaining the current data regime. We discuss how some of these constraints might be overcome.

2 The availability and reliability of data

2.1 Key data are scarce, and often scarcest in places where most needed

Household- or field-level surveys remain the main data collection tool for key development-related outcomes, including poverty, agricultural productivity, population, and many health outcomes. Methodologies for such data collection are well developed, and are implemented by national statistical agencies and other organizations in nearly all countries of the world. For livelihood surveys designed to generate regionally or nationally-representative estimates, sampling strategies typically follow two-stage designs, where survey “enumeration areas” (or “clusters”, often the size of a village or a neighborhood) are first sampled proportional to population, and then a given number of households or individuals are randomly sampled within each cluster. Typically survey sizes for surveys such as the Demographic and Health Surveys (DHS) or Living Standard Measurement Surveys (LSMS) are a few hundred to a few thousand clusters, and then 10-20 households per cluster, yielding total household sample sizes typically between 2000 and 20,000 for a given country.

Such surveys provide critical information – and often incredible detail – on a range of outcomes, and are the bedrock on which many sustainable development related outcomes have and will continue to be measured. But their implementation and use also faces a number of important challenges. First, nationally-representative surveys are expensive and time-consuming to conduct. Conducting a DHS or LSMS survey in one country for one year typically costs \$1.5-2 million USD,⁸ with the entire survey operation taking multiple years and involving the training and deployment of enumerators to often remote and insecure locations. Population censuses are substantially more expensive, costing tens to hundreds of millions of USD in a typical African country.⁹

An implication of this expense is that many countries conduct surveys infrequently, if at all. In half of African nations at least 6.5 years pass between nationally representative livelihood surveys, as shown in Figure 1a (compare to sub-annual frequency in most wealthy countries). Globally, the frequency of these economic household surveys is on average substantially lower in less wealthy countries (Fig 1b), meaning that data on livelihood outcomes are often lacking where they are arguably the most needed. Surveys are also much less common in less democratic societies (Fig 1c), which could at least partly reflect the desire and ability of some autocrats to limit awareness of poor economic progress.¹⁰ The frequency of agricultural and population censuses also varies widely around the world (Fig 1d,g). For instance, 25% (n = 53) of countries have gone more than 15 years since their last agricultural census, and 8% (n = 17) countries more than 15 years since their last population census. Restricting to just African nations, 34% of countries have gone more than 15 years since their last agricultural census. For both agricultural and population data, the relationship between survey recency, income, and level of democracy is less clear, perhaps reflecting the more important role of these data in developing economies.

A second challenge for many downstream applications are that surveys are typically only representative at the national or (sometimes) regional level, meaning they often cannot be used to generate accurate summary statistics at a state, county, or more local level. This represents a challenge for a range of research or policy applications that require individual or local-level information – for instance an anti-poverty program attempting to target an intervention (e.g cash transfer) to a particular group, or a research effort aimed at studying the impact of such an intervention.

Third, underlying household or cluster-level observations are not made publicly available in many surveys, including nearly all the surveys that contribute to official poverty statistics (such as those depicted in Fig 1a), and no geographic information is publicly provided on where in a country the data were collected. These factors further deepen the challenge of using such data to conduct local research or policy evaluation, or to train models to predict local outcomes using these data. Even when local-level anonymized georeferenced data are made public in some form, data are typically

released more than a year after survey completion, hampering real-time knowledge of livelihood conditions on the ground.

Finally, as explored below, ground data can have multiple sources of noise or bias, further limiting their reliability and utility in research and decision-making. This in turn has important implications for how satellite-based models trained on these data are validated and interpreted.

2.2 Existing ground data can be unreliable

Even where ground data are present, several key sources of error can limit their utility. First, most outcomes are not measured directly, but rather inferred from responses to surveys. This can introduce large amounts of both random and systematic measurement errors, for example in the case of self-reported household consumption^[11] or agricultural production^[12] surveys. For instance, in household consumption expenditure surveys, changes to the recall period or the list of items households are questioned about can lead to household expenditure estimates that are >25% too low relative to gold standard household diaries.^[11]

Lack of reliability also extends to agricultural contexts. In recent reviews of agricultural statistical systems, the World Bank noted that the “practice of ‘eye observations’ or ‘desk-based estimation’ is commonly used by agricultural officers”, leading to often-conflicting estimates of key agricultural outcomes by different government ministries, and to variation over time in published statistics that cannot easily be reconciled with events on the ground.^[12] Current practices are likely to have a bias toward overestimation, further weakening the quality of food security assessments.^{[12][13]}

An additional key source of noise comes from sampling variability. As noted, surveys are typically designed to be representative at very large scales (e.g. nationally), and this representativeness is typically obtained by taking small random samples of households or fields across many cluster locations. Because most agricultural and economic outcomes of interest often exhibit substantial variation even at very local levels (e.g. coefficients of variation > 1 at the village level), these small samples thus represent an unbiased but potentially very noisy measure of average outcomes in a given locality.

The combined effects of both measurement error and sampling variability can be appreciated when comparing two independent measures of the same outcome for the same administrative level. In Figure 2, average maize yields (in units of tons per ha of land) are compared at the first administrative level (e.g., province or state) as obtained from household surveys covered by the LSMS-ISA program versus by official government ministry estimates in three African countries. This comparison reveals both a systematic bias towards higher yields in official government data

than in household responses, and a relatively low correlation between the two measures, with the highest observed correlation equal to $r = 0.39$ for Ethiopia.

A third source of error, particularly relevant to researchers relying on access to data acquired by others, is noise purposefully introduced to protect the privacy of surveyed households. Adding jitter to village coordinates is common practice for most of the publicly released datasets based on household surveys, for instance with up to 2km of random jitter added in urban areas and 5km in rural areas. Below we explore the implications of these three sources of error for model development and evaluation.

2.3 Availability of satellite imagery changing rapidly

Information from satellite imagery has long offered a potential inroad into helping solve problems of data scarcity and unreliability in sustainability. Such information has been used in both agricultural and socioeconomic applications for decades.^{[14][15]} However, thanks to both public and private sector investment, recent years have seen a remarkable increase in the temporal, spatial, and spectral information available from satellites. These investments have largely undone the traditional trade-off between temporal and spatial resolution, and are helping to undo the trade-off between spectral and temporal/spatial resolution.

To quantify this increase and understand how it varies across developing and developed countries, we randomly sample 100 locations in Africa and 100 additional across the US and EU (sampling proportional to population), and query the availability of cloud-free imagery (defined as <30% cloud cover) at each location in 2010 and 2019 for all available optical sensors, using multiple online tools (see Supplemental Information for details on this process). We calculate region- and year-specific average revisit rates as the number of available cloud-free images across locations divided by the number of locations times the number of days. We calculate this separately for each sensor and also calculate an imagery-resolution “frontier”, defined as overall revisit rate across sensors at or below a given spatial resolution.

Results are shown in Figure 3. Many new public and private-sector entrants since 2010 (Fig 3a) have lessened the traditional temporal/spatial tradeoff in imagery, particularly at resolutions $\geq 3\text{m}$. Although the revisit rate of very high resolution ($< 1\text{m}$) sensors over Africa has seen only slight improvement over the last decade (Fig 3b), and very-high-resolution revisit rates remain lower in Africa as compared to the US/EU (Fig 3c), revisit rates for high resolution (1-5m) and moderate-to low-resolution sensors has increased dramatically. Images at this resolution are now captured multiple times per week rather than multiple times per year and equitable capture between Africa versus the US/EU.

Figure 3 provides additional detail and sample imagery for a number of sensors in African locations. Information on human activity is readily visible even in moderate-resolution sensors (5-30m), and indices constructed from moderate-resolution multispectral imagery provide an increasingly clear picture of a broad range of human activity at very local scale, including urban infrastructure development, agricultural activity, and moisture availability (Fig 3f). The increasingly high revisit rate of such imagery also provides key insight into development-relevant activities that change seasonally, such as the location and productivity of croplands (Fig 3g).

3 Modeling approaches using satellite imagery to predict sustainability outcomes

Researchers have taken many different modeling approaches in using this large amount of new imagery to measure and understand sustainable development. We use “model” to mean any function or set of functions mapping inputs (e.g., satellite images) to outputs (e.g., a wealth index or yield estimates for an area). Such models are often simple, such as linear regression models that relate satellite-derived vegetation indices to crop yields^[10] or that relate nighttime lights to economic outcomes.^[11] When there is substantial prior knowledge of the likely relationship between satellite-derived features and the outcome of interest, as in the case of many agricultural variables, such approaches can often work well. However, even in these settings, machine learning approaches that seek to more flexibly learn – rather than specify – the mapping of inputs to outputs can often improve predictive performance.

Machine learning approaches start by defining a suitable model family, i.e., a set of candidate functions used to represent the relationship between inputs and outputs. These could be decision trees, random forests, support vector machines, or fully-connected neural networks with a fixed structure and varying weights.^[12] When inputs and outputs have explicit spatial or temporal structure (e.g. images, or images over time) it is typically advantageous to use functions tailored to this structure. These include convolutional neural networks for images, recurrent neural networks for sequential data, and convolutional autoencoders when both inputs and outputs have spatial structure^[13] (e.g., segmentation of agricultural fields).

Training data for these models consists of a set of inputs with their corresponding ground-truth outputs, e.g., images of villages and their corresponding poverty levels, or a sequence of images of a field captured during the growing season and the corresponding crop yield. A model in the family is chosen by training, which typically involves the minimizing of a suitable loss function that describes the difference between predicted and observed values of the outcome. For regression, the loss could be squared loss or absolute value, and for classification a common choice is cross-

entropy. After training, the loss function is evaluated on held-out data – i.e. data not used to train the model. Evaluation on held-out data is important because training data are often limited and the model family complex (often with many orders of magnitude more model parameters than training observations), and overfitting is thus a major concern. Regularization techniques such as weight decay, dropout, and early stopping using a validation set are often employed in practice to mitigate overfitting.

Suitable preprocessing of inputs is also often important in achieving good performance. Common pre-processing steps include median compositing across images to mitigate the effect of occlusions due to clouds, imputing missing values, scaling to put all the inputs on the same scale, centering, whitening, and harmonic preprocessing for temporal data. While deep models can in principle learn these transformations, these are tailored to existing learning algorithms and initialization schemes and will generally make learning more stable. Tiling and rescaling is also often necessary to match the input requirements (e.g. pixel dimensions) of exiting neural architectures of neural networks.

Here we provide an overview of the range of modeling approaches that have been used to relate satellite images to sustainable development outcomes.

Shallow models based on hand-crafted features. In some domains, prior knowledge of the physics, chemistry, or biology of the relevant processes suggest that certain functions of the inputs are likely useful for prediction. This is the case for numerous vegetation indexes (VI), which are computed from raw imagery as simple ratios of reflectances at different wavelengths and are known to be related to vegetation health. Simple regression models such as linear regression or random forests can be used to make pixel-wise predictions directly from these hand-crafted features to the outputs of interest, e.g. predicting yield with VIs observed over time (see ref²⁰ for a recent review in the agricultural domain). When the input has spatial structure, simple aggregation strategies can be used to map pixel-wise features to image-wise features. These include simple statistics such as taking the mean, quantiles (min,median,max), or histograms of binned values as inputs to a regression or ML model. As an example, this strategy is very effective for predicting GDP with nightlights¹⁷ or aggregate crop yields at the county and state level from multispectral images.²¹ However, these simple aggregation strategies discard most of the spatial structure, which can be undesirable.

Models that use spatial structure in the imagery. In computer vision, spatial context can often greatly improve prediction accuracy for image prediction and analysis tasks. Machine learning models with filters designed to take into account spatial structure, such as convolutional neural networks (CNNs), often perform much better than hand-crafted features and aggregation strategies.

Models such as VGG,²² or deeper models with residual connections such as DenseNET or ResNet²³ are often employed. In this case, features are automatically learned from the data rather than hand-crafted. This is currently the leading approach in most computer vision applications, including in the satellite space when training data are plentiful. Use of this approach in sustainable development applications has proliferated in recent years, including in the measurement of population,²⁴⁻²⁶ economic livelihoods,²⁷⁻³⁰ infrastructure quality,³¹⁻³² land use,³³⁻³⁴ informal settlements,³⁵⁻³⁶ fishing activity,³⁷⁻³⁸ and many others.

Models that use spatial and temporal structure in the imagery. When available, multiple images of the same location over time can reduce ambiguity (e.g., due to partial cloud cover) and provide crucial information about changes occurring on the ground. Such a sequence of images is similar to a video, and architectures from video prediction in computer vision can be brought to bear for prediction and regression tasks. These include recurrent neural network variants such as long-short term memory networks (LSTMs),³⁹ convolutional LSTMs,⁴⁰ and 3-D CNNs, where images are fed in sequence into the model before it makes a prediction. These models have been successfully used for crop classification,⁴¹⁻⁴³ crop yield prediction,²¹⁻⁴⁴ predicting landslide susceptibility,⁴⁵ assessing building damage after disasters⁴⁶⁻⁴⁷ among many other tasks.

Models that use several modalities. When multiple data modalities are available, such as measurements from different satellites, it is often possible to combine all the inputs into a single deep learning model. Approaches include stacking the inputs as additional channels of a single network, or multi-branch architectures where data modalities are processed separately to extract features which are then concatenated before a final prediction layer. Examples of this approach include models that combine multiple sources of satellite information³⁰ or models that combine imagery with data from weather sensors,⁴⁸ cell phones,²⁹ Wikipedia,⁴⁹ social media,⁵⁰ street-level imagery⁵¹ or Open Street Map⁵² to predict development-related outcomes.

3.1 Model development with limited training data

An additional set of techniques have been developed to utilize the above modeling approaches in the context of limited training data – a common problem in sustainability applications. For instance, standard convolutional neural network architectures contain millions to tens of millions of trainable parameters,⁵³ whereas training data for specific sustainability tasks can often number in the hundreds. This limited amount of labeled data is often insufficient for “end-to-end” training of deep networks, i.e. training a model to directly predict the outcome of interest on the available labeled data by minimizing a suitable loss function. Multiple strategies have been deployed to address this problem.

Using synthetic data. A first approach is to generate and use synthetic data to train models. In some cases, domain knowledge about the relevant physical process exists in the form of validated simulators. These simulators can be used to provide synthetic training data, i.e., synthetic inputs of what the process would look like from space paired with simulated outputs. These synthetic pairs can be used to augment the training data. For example, crop model simulations have been used to augment field data collection for satellite-based yield mapping in smallholder systems, and have been shown to perform on par or better than approaches that calibrate directly to limited field data.^{[16][54]}

Transfer learning. A second approach, transfer learning, is a common strategy in deep learning. The idea is that a neural network can be pre-trained on a different but related task for which large amounts of labeled data are available (such as ImageNet in computer vision, or Functional Map of the World^[55] and WikiSatNet^[56] for satellite images). The neural network is then “fine-tuned” on the task of interest. For example, Jean et al.^[27] showed how transfer learning could be used to predict a very small (<500) number of observations of economic livelihoods in Africa from imagery. A neural network was first trained to predict nightlights (a plentiful proxy for economic development) from daytime imagery, thus learning to recognize features in the high-resolution daytime imagery related to economic activity. Features were then extracted for daytime images in locations where livelihoods data were available, and a simpler model (e.g. regularized regression such as ridge or lasso) used to predict livelihoods from these features. Another recent approach applied a trained object identifier to high resolution data to identify buildings, vehicles, and other objects, and then used these objects as features in a regularized regression to predict economic wellbeing in Uganda with high accuracy.^[57]

Transfer learning can also be done spatially, with models trained using data from one region where labels are often plentiful, and then “fine-tuned” on the target geography of interest where labels are sparse. To be successful, this approach requires relevant features to be similar between training and target geographies, but does not require the mapping of features to outcomes to be the same between regions (e.g. having productive crops near your house could signify wealth in one region but relative poverty in another). For example, a model trained to predict infrastructure quality in Africa could be finetuned to a specific country using only a small amount of labeled data.^[58] The main challenge with spatial transfer learning is that changes in the input data distribution from one region to another (e.g. the appearance of houses or crops) will decrease predictive performance.

Unsupervised or semi-supervised learning. A third approach uses unsupervised or semi-supervised learning, which take advantage of the fact that while labels are often scarce in sustainability applications, obtaining large amounts of unlabeled satellite imagery is relatively easy.

Utilizing large amounts of unlabeled data to pre-train neural networks and learn useful features has recently shown great progress in computer vision,^{[58][59]} narrowing the gap with fully supervised methods. Among others,^[60] Tile2Vec is an unsupervised pre-training technique tailored specifically to satellite images that performs well on a range of tasks, such as crop type classification and predicting economic wellbeing in Africa.^[61] Semi-supervised learning strategies attempt to improve model performance by additionally leveraging a small amount of labeled data. These are often based on the assumption that data is clustered, and decision boundaries should separate these clusters as much as possible. This idea has been extended to regression problems, with resulting performance improvements in predicting economic well-being from satellite imagery.^[62]

3.2 Model development and evaluation with noisy data

The performance of satellite-based models, particularly in settings beyond where they were trained, is perhaps the most common and important concern for researchers and policy makers interested in potential applications in sustainable development. Noisy training data can degrade model performance in two ways. First, it can diminish the ability of a model to learn features in imagery that are predictive of the outcome of interest. Second, and more subtly, the model might learn relevant features but perform poorly in predicting test data, precisely because the test data has noise. This latter outcome would lead researchers to underestimate the model’s true performance. As noisy datasets are increasingly employed for model development, researchers must contend with the dual challenges of not overfitting to noise and not underestimating the performance of a model with respect to reality. Both challenges are potentially important, with existing work mainly highlighting how noise in training data can degrade model performance.^[63] But in many sustainable development settings, we believe models can learn to separate signal from noise in training data, and that the more fundamental – and underappreciated – challenge is in accurately assessing model performance in light of noisy test data. We quantify this insight and discuss methods for addressing it.

Noisy training versus noisy test data Studies in the broader computer vision/deep learning domain have demonstrated how models trained on noisy but numerate labels can still perform well when evaluated on high-quality test data, even when high-quality labels are massively outnumbered by low-quality labels in training data.^{[64][66]} Under suitable assumptions on the noise, these empirical results can be explained from a theoretical point of view.^{[66][67]} In sustainable development settings, while noisy training can certainly still degrade model performance when the amount of training data is limited (ref^[68] provides one example in Indian smallholder wheat systems) or errors non-random (as in the poor-quality government data in Fig [2]), numerous recent studies highlight how such noise can be overcome so long as training data are reasonably numerous and errors are

largely random. For instance, in Uganda, a model trained to predict maize yields from relatively noise data performed twice as well when evaluated on high-quality test data as when evaluated on noisy data held-out from training.⁵⁴ In India, a satellite-based crop classification model trained on labels derived from millions of imperfectly geolocated smartphone photos was able to exceed the performance of benchmark satellite-based classifiers.⁶⁹ A global study showed how noisy object labels from Open Street Map could be used to train a model to make accurate predictions of the location of urban structures.⁷⁰

Using data and imagery from an earlier study of asset wealth across thousands of African villages,³⁰ we use simulation to explore the influence on model performance of three types of error common in publicly-available training data: (1) noise due random noise (“jitter”) purposely added to village geo-coordinates to protect respondent privacy, (2) sampling variability noise from the construction of village-level estimates from small numbers of respondent households, and (3) noise from households’ misreporting of asset ownership. We add a given type of noise to the observed wealth estimates, train a random forest model to predict these labels from nightlights imagery on 5 folds of the data, and evaluate performance on the remaining test data that has either been similarly degraded or unaltered; we use nightlights and random forest rather than a CNN and/or optical imagery to make these experiments tractable.

As shown in Fig 4a-c, when evaluated on noisy training data, model performance degrades as increasing amounts of each type of noise is added. However, when models trained on increasingly noisy data are evaluated on un-degraded test data, model performance remains highly stable, even for large amounts of training noise. This holds true for all three common types of training data noise we explore, again suggesting that ML models can be surprisingly robust to various types of training noise.

Accurately assessing model performance. Most existing work has focused on techniques to avoid overstating model performance, including strategies discussed above to avoid overfitting during training, and the typical practice of testing models on held-out data. Here we discuss two strategies for dealing with the opposite problem: understating model performance due to noise in test data.

A first approach is to ensure that a small amount of very high-quality ground data is available for model testing. Often this can require additional investment in data collection. In using these data, the typical practice of splitting a dataset from a single heritage into training, validation, and test sets is then replaced by a practice with two different measurement approaches for training and validation on the one hand, and testing on the other – with the high-quality data reserved for testing. Typically, the data volumes needed for testing are far fewer than for training, and thus the expenses

associated with obtaining “gold-standard” measures for testing are more likely tractable.

A second strategy, particularly useful if ground data are unavailable, is to identify a variable that previous work has identified as being associated with the outcome of interest, such as weather in the case of economic output, or fertilizer in the case of agricultural productivity. This strength of association between this variable and model predictions, as measured for instance by correlation, can then be compared to the association between the variable and the (noisy) training data for the model. Because these third variables (e.g. weather) are often readily available for most locations in the world, this approach should have broad applicability.

To illustrate both of these strategies, Figure 4d-f draws on a recent study of maize yields in Uganda.⁵⁴ The left panel shows the agreement between satellite-based yield estimates and the data on which the model was trained. In this case, the training data comprised 8mx8m crop cuts (i.e. harvests from small, randomly-selected portions of a field) from 125 different maize fields in the region. Although crop-cuts are low-error measurements of productivity for the portions of a field they sample, we consider this noisy training data because of high heterogeneity within fields and the potential spatial mismatch between the crop-cut location and the satellite pixels (which are 10x10m and not perfectly aligned with the crop cut). As judged by the training fit, the model has a relatively modest explanatory power ($r^2 = 0.25$, Fig 4d). Yet the model performance is much better when predictions are compared to the “gold-standard” measure of full plot harvests, which were available for a smaller number of randomly-selected fields (Fig 4e). Similarly, the correlation between satellite estimates and self-reported fertilizer or objective measures of soil quality were the same as the correlation between crop cut yields and these measures, suggesting the “signal” in the satellite measures was as strong as that from the ground measure (Fig 4f). A similar finding was obtained in Kenya when pitting satellite-estimated maize yields against self-reported yield data.¹⁶

Another example of both strategies is given in ref,³⁰ where estimates of wealth from satellites and from ground data are each compared against independent wealth measures from census data (considered high quality) and against a measure of annual temperature, which has been shown to correlate strongly to economic outcomes. Ground data and model predictions showed similar correlation against the independent wealth measure, and both uncovered similar non-linear relationships between temperature and wealth, suggesting that the satellite-based wealth measure was roughly as trustworthy as the original ground data.

4 Applications

Researchers are actively evaluating the usefulness of satellite imagery for a range of sustainable development applications, with more work thus far focused on whether satellites can be used

to make reliable measurements of key variables of interest and comparatively less devoted to using derived measures for downstream research tasks or policy decisions. Rather than try to provide a comprehensive survey of all applications of satellite-based remote sensing in sustainable development, we focus on four domains where recent work on satellite-based measurement has been particularly active and where comparable quantitative results exist across studies. Our goal is to provide rough performance benchmarks across these domains and, where possible, diagnose constraints to further improvement. In making these comparisons, we included all published or posted (e.g. on arxiv) studies where the test statistic of interest could be obtained for the outcome of interest in a developing-world geography.

We then review the more limited set of cases where these and other satellite-based measurements have been used for research or policy tasks. Our focus is again on domains directly involving human activity, and does not encompass progress in all realms of earth or environmental observation.

Smallholder agriculture. Roughly 2.5 billion individuals, and over half of the worlds poor, are estimated to live in “smallholder” households that primarily depend on farming small plots of land for their livelihoods.^[71] While remote sensing has been used in agricultural applications for decades, coarse sensor resolutions and a paucity of training data had until recently largely precluded its application in smallholder agriculture, where field sizes are often $<0.1\text{ha}$ (or roughly 1 30m Landsat pixel).

Here we assemble data from recent studies attempting to predict yield at the field scale in heterogeneous smallholder environments (ref^[72] provide a nice overview of yield prediction performance at more aggregate scales). Field-scale yield prediction is useful for a range of development applications, including the targeting and evaluation of agricultural interventions and the rapid monitoring of rural livelihoods. We found 11 published studies that reported comparable performance metrics for field-scale yield prediction on smallholder fields, spanning multiple continents and seven crops. All studies used relatively simple models to relate hand-crafted features (typically, vegetation indices constructed from ratios of reflectances in the visible and near-infrared wavelengths) to ground-measured yields, and nearly all evaluated models on training rather than held-out test data. While predictive performance differed widely across and within crops (Fig 5a), likely due to the enormous temporal and spatial heterogeneity present in smallholder agriculture, re-analysis of multiple studies for which replication data were available allowed insight into the determinants of model performance.

First, models trained and evaluated on more “objective” ground data – i.e. harvest data collected from crop cuts or full plot harvests – performed on average substantially better than models trained

on farmer self-reported data (Fig 5b). This finding again highlights the importance of ground-based measurement error in training and evaluating remote sensing models. Second, in settings where average field sizes were small, model performance was much higher on larger fields (Fig 5c). This difference is likely because for certain sources of error, e.g. error in field area measurement or in the georeferencing of field data, the same magnitude error is more consequential for smaller fields; a 10m georeferencing error is more consequential for a 10m-wide field as compared to a 100m-wide field.

Finally, because collecting high quality ground data is expensive and time consuming, we studied the extent to which additional training samples improve model performance. At very small sample sizes, additional training samples rapidly improved performance on held out test data (as measured by root mean squared error; Fig 5d), up to around 30-50 samples. Performance was largely stable beyond that, suggesting that – at least in the African settings represented here – adequate performance for yield prediction could be achieved with only a few dozen high quality training samples. See Table S1 for the full list of studies and estimates we included.

Population A second area in which satellite information has played an important role is helping generate local-level population estimates. Accurate knowledge of where people are is a critical input into an immense range of research and policy applications. Because population census are infrequent in many developing countries and fine-scale data from existing censuses are often not made public, generating fine-scale model-based estimates of settlement locations and population density has been an areas of substantial research focus for decades.

The traditional approach to generating local-level population estimates takes a “top-down” approach in which available admin-level census data is redistributed down to a finer-scale grid (1km or finer), using satellite-derived information and other covariates as input. Because population data are almost never available for training or validation at the target fine scale, one common approach uses the coarse-scale data from census to model the relationship between satellite features (e.g. nighttime lights imagery or satellite-derived estimates of land use), other ancillary data such as the location of transportation infrastructure, and census-based population estimates, and then applies the trained model to available fine-scale features.^{[73][74]} Another approach generates a binary population mask at fine scale using estimates of building or settlement locations derived from imagery, and then applies this mask to coarse-scale census data.^[75] Both approaches typically use machine learning at some step, e.g a random forest to predict coarse census data, or computer vision approaches to identify settlement locations. For either approach, predictions can only be readily evaluated at coarse scale; the fine scale gridded predictions cannot be easily validated. In the absence of clear evaluation opportunities, a consortium of data producers have built useful tools in which different

gridded estimates can be visually compared at local scale (<https://popgrid.org>).

As additional quantitative comparison, we study three commonly-used population rasters that used satellite data as at least one input in their production: WorldPop,⁷⁴ GHSL,⁷⁵ and LandScan.⁷³ We harmonize each to a consistent 1km grid and compare population estimates for grid cells with non-zero estimates across all three rasters. Estimates show modest agreement ($r=0.62-0.78$) when comparing across all global pixels (Figure 6), with lowest agreement between LandScan and the other rasters. Agreement was often substantially lower in the developing world. On the African continent, the average pairwise correlation between the 3 datasets across 47 African countries is $r = 0.45$, perhaps in part due to the relative paucity of census data on which to train models. Overall disagreements in African and globally could also result from differences in conceptualization of population used in dataset construction, with LandScan attempting to measure “ambient population” averaged over 24 hours and the other datasets attempting to measure population at individuals’ usual residences. Agreement improves when comparisons are made at increasingly aggregate levels, with correlations approaching $r = 1.0$ when estimates are aggregated to 100km pixels.

Multiple studies have sought to further validate estimates of one or more of these datasets in settings where fine-scale population data are available. Using very-fine-scale (100m) administrative population data from Sweden available over a 25-yr period (none of which used in the creation of any of the gridded datasets), researchers found cell-wise correlations between the admin data and GHSL, WorldPop, and LandScan of $r = 0.83$, $r = 0.82$, and $r = 0.7$, respectively, with predictive performance improving slightly in later years.⁷⁶ The authors caution that performance in Sweden (where model predictions were highly correlated, see Fig 6) might not reflect performance elsewhere, given the high quality of ancillary data available in Sweden. Other studies in China and Europe found similar or higher performance of individual gridded datasets evaluated at somewhat more aggregate scales, but (as in Fig 6) found that performance was not uniform and tended to degrade at finer spatial scales.^{77,78} Overall performance on this population prediction task appears roughly on par with performance predicting asset wealth described below.

Because standard approaches to generating these estimates is to disaggregate official census estimates, final estimates are unavoidably affected by any inaccuracies in the official census data – for instance due to the most recent census having occurred a decade or more prior. An alternative that does not face this problem is to train “bottom-up” models to directly predict local-level population estimates, and these approaches have shown promise in multiple settings.^{9,26,79} Such approaches are beginning to be incorporated into global gridded products (e.g. WorldPop) for countries where censuses are particularly out of date,⁸⁰ and have been shown to be a cost-effective way for generating reliable national-scale population estimates.⁹

Economic livelihoods

Predicting variation in local-level economic outcomes is another domain where the combination of machine learning and satellite imagery has seen recent application, again motivated by the paucity of existing data (Fig 1) and the broad range of applications for which such data could be useful. As in the agricultural setting, existing work spans diverse geographies and seeks to predict a range of outcomes, making quantitative comparison of different models or sensors difficult.

We focus on 12 studies that used imagery – either alone or in combination with other data – to predict asset wealth at local level in the developing world. Asset wealth is a commonly used measure of households' longer-run economic wellbeing, and is consistently measured in a number of georeferenced nationally-representative household surveys, making it appealing training data in this domain. Fig 6a shows 16 asset wealth estimates across these 11 studies. All studies applied convolutional neural networks to imagery to generate features used to predict wealth, and reported evaluation statistics on held-out test data.

While study intercomparison was challenging even for this group of studies that measured the same outcome due to the varied geographic settings (spanning Africa, Asia, and the Caribbean), the various spatial scales at which predictions were evaluated (from village level to district level), and some studies' inclusion of additional data input data not from satellites, results allowed some generalizations. First, information derived from satellites could always explain more than half, and often more than 75%, of the variation in the survey-measured asset wealth, with performance appearing to trend upward over time. For reasons described above, these estimates likely understate true model performance, as test data are almost always from publicly-available survey data with known sources of noise. Second, although small samples make generalization tenuous, studies that made predictions at more aggregate spatial scales, and studies that combined satellite information with data from other sources, tended to outperform village-level satellite-only models. These data fusion approaches have become increasingly common, with researchers demonstrating how combining imagery with data from cell phones,²⁹ Wikipedia,⁴⁹ social media,⁵⁰ or Open Street Map⁵² can improve predictions.

Table S2 describes results from additional studies that looked at other measures of economic livelihoods, including consumption expenditure and multi-dimensional poverty indices. Prediction performance for consumption expenditure (the measure on which official poverty estimates are based) is typically lower than that for asset wealth, a difference which has been in part attributed to relatively higher noise in the consumption data^{27,30} and the extreme paucity of public georeferenced public data on which to train models.

Informal settlements A final related area where there has been much recent work is in the detection of informal settlements (sometimes called “slums”). Urban populations are growing rapidly throughout much of the developing world, and about 30% of developing-country urban populations are estimated to live in slums – settled areas where inhabitants lack access to essential services, durable housing, and/or tenure security.^[81] Systematic data on the location and size of such settlements is lacking, making it difficult to monitor and target service delivery and to protect residents against eviction, among other challenges.^[7] Some governments, lacking reliable data on informal settlements, do not officially acknowledge their existence.^[81]

Because the spatial structure (e.g density, size and type of buildings) can differ substantially between informal settlements and surrounding regions, researchers have sought to use imagery to measure the location and size of these settlements (see ref^[7] for a recent review). We focus on 23 studies that used satellite imagery to segment or classify informal settlements in the developing world. These studies use a variety of methods, with some focused on creating rule bases for classification and others on directly using machine learning for classification. The fuzzy-logic rule bases are sometimes generated using machine learning (eg. decision trees) and sometimes are human generated from ontologies (formalized descriptions of expert knowledge from a certain perspective) of local informal settlements.

As with the other domains discussed, the literature spans diverse geographies where informal settlements can be very structurally dissimilar from each other, making study intercomparison difficult. However, in 17 studies that reported classification accuracy (evaluated against typically small numbers of ground observations), accuracy exceeded 80% in most studies and appeared to be improving over time (Fig 6g). Table S3 shows results from additional studies that reported alternate performance metrics.

4.1 Application in research

Here we highlight a number of settings in which measures derived from satellite-based remote sensing, including those discussed above, are being used for some downstream research task in the developing world.

The widest adoption of satellite-derived measures in research and policy has been in the realm of population estimates, with existing gridded population data being used in an impressive array of research applications. These include in public health, disaster response, economic development, climate change research, and others; see refs^{[9][80][82]} for excellent recent reviews.

Satellite imagery has also been widely used to better understand agricultural productivity, including

why some fields or some regions are more productive than others⁵ and whether particular management practices have been adopted.⁸³ Satellite estimates are also increasingly being used to identify fields most likely to respond to a particular input^{16,54} or new management practice.⁸⁴

Fisheries and animal production are additional food-related domains where satellite imagery is becoming increasingly used in research and policy. Recent work shows how multiple satellite sensors and deep learning can shed light on overall patterns of global fishing activity³⁷ as well as on specific activities like illegal fishing.^{38,85}

Researchers in economics also increasingly utilize satellite imagery – and particularly night-time lights imagery – for a variety of applications (see ref⁴ for a review). Nightlights have been used to assess the validity of official government statistics,^{17,86} to understand the growth and activity of urban versus rural areas,^{87,88} and to assess the role of local and federal institutions, transport costs, and other factors on economic development.⁸⁹⁻⁹² While the use of optical imagery beyond nightlights remains somewhat more limited, recent papers have shown how high-resolution optical imagery can be used to measure compliance with conservation programs⁹³ and to understand how ethnic favoritism shapes economic investment.⁹⁴ Recent work⁹⁵ also shows how to combine satellite-derived estimates with survey data to obtain tighter confidence intervals and improve regression analyses.

More recent work has shown how satellites can be useful in the experimental evaluation of interventions in both the agricultural and economic sphere. Jain et al⁸⁴ show how remote sensing estimates can be used to measure the effectiveness of a new agricultural technology on productivity and quantify who benefits most from the adoption of the technology. Huang⁹⁶ shows how a deep learning model trained to identify housing quality in high-resolution imagery can be used to estimate the livelihood impact of a randomized cash transfer program in Kenya, with estimates benchmarked against ground survey data. Jayachandran et al⁹³ show how high-resolution imagery can be used to measure compliance in an experimental evaluation of a payment-for-forest-protection program. While all of these studies focus on settings where changes induced by an intervention are readily apparent in imagery – an aspect that might not hold in other settings – they demonstrate the large potential for satellite imagery to contribute to the quantitative evaluation of many development interventions.

4.2 Use in decision making

While satellite-based measures are now being used in a variety of research applications, documented examples of their operational use in public-sector decision-making and policy in the developing world is much more limited. Systematic information on operational use in the private sector is

even more sparse, although use is likely widespread and growing; the same is true of military applications. Here we only consider public-sector non-military use.

As in research, the widest application of satellite-based measures in public-sector decision-making is in the population domain. For instance, the UN World Food Programme and US government both used gridded population estimates to inform needs assessments and target humanitarian response following natural disasters.⁸⁰ Gridded population data are also being used to inform sampling strategies for ground surveys.⁸⁰

In agriculture, remote-sensed vegetation indices and satellite-derived rainfall estimates are key inputs into short-term forecasting of food insecurity, which directly informs food aid and other humanitarian resource allocation.⁹⁷ Numerous systems that track agricultural growing conditions and crop output around the world also make ample use of remote sensing information, and output from these systems are used in a wide array of tasks, including in early warning alerts, foreign aid decisions, analysis of commercial trends, and in trade policy.⁹⁸ Data from remote detection of fishing activity is also being used by numerous governments and other organizations to manage fisheries and design protected areas.⁹⁹

Across other domains – e.g. economic livelihood measurement – documented use in decision-making appears limited or non-existent, although anecdotally there is rapidly growing interest in the policy community in exploring these measures.¹⁰⁰

We hypothesize on why adoption in these and other domains has been relatively limited. The simplest explanation is that the combination of satellite information and machine learning is still quite new in many domains, and decision-makers might not be familiar with these approaches or convinced they are “good enough”. Our view is that in many settings, including smallholder agricultural and livelihood measurement, the true accuracy of satellite-derived estimates can rival or exceed that of traditional survey-based measures. It remains the job of the research community to help make this clear, and the job of the user community to transparently define the counterfactual: if not satellite-based data, what alternative data would be used to make a decision, and what do we know about its reliability?

Even if satellite-based measures are accurate, they might not yet be operational. To our knowledge there exist no updated, global-scale estimates of smallholder crop productivity, economic well-being, or informal settlements that a decision-maker could immediately use (estimates are beginning to exist for individual countries). The research community is arguably not well positioned to generate and update such estimates over time, and partnerships with public-sector institutions or the private sector to scale and operationalize these estimates could be important in enabling their

sustained use.

Even when models are operational, decision-makers might be understandably hesitant to adopt a measure they cannot fully explain. Deep learning models tend to sacrifice interpretability for predictive performance, and researchers are often satisfied if a model is working well (as evaluated on held-out data) even if they cannot explain why. But understanding why a model makes the predictions it does can help build trust that predictions are accurate and fair. Well-publicized instances of algorithmic bias in other settings (e.g. predictive policing, sentencing, and hiring decisions^[101]), and concerns by civil rights groups that further deployment of algorithmic decision-making might worsen racial and socioeconomic inequalities,^{[102][103]} understandably amplify worries that predictions from these new approaches could be either inaccurate or unfair.

Existing guidelines for Fairness, Accountability, and Transparency in Machine Learning (“FAT ML”),^[104] if followed, could help navigate these issues. The guidelines aim to ensure that researchers are aware of potential discriminatory impacts of their algorithms and are able to investigate and provide redress should issues arise. While implementation of the guidelines certainly has its own challenges^[105] (e.g. defining “fairness”), we are not aware of any of the papers we review above – including our own – having fully engaged with these guidelines.

A final reason for limited adoption is that some actors might see benefit in not having certain outcomes be measured. Autocratic regimes already collect less data (recall Fig [1]), and certain countries have passed laws (since reversed) that make it a crime to publish independent estimates of key economic outcomes.^[106]

5 Conclusions and directions for future work

We draw four main conclusions from the above analysis, and lay out open challenges and directions for future work. First, satellite-based performance in predicting key sustainable development outcomes is reasonably strong and appears to be improving. Estimates are being used in a wide variety of research applications and, in some cases, are already actively informing decision-making. Indeed, analyses suggest that reported model performance likely understates true performance in many settings, given the noisy data on which predictions are evaluated, and that satellite-based estimates can equal or exceed the accuracy of traditional approaches to measuring key outcomes. For certain outcomes, satellite-based approaches can already add substantial information at broad scale and low cost compared to what can be collected on the ground. Numerous quantitative approaches now exist to assist researchers and practitioners in better understanding – and not underestimating – the performance of satellite-based approaches relative to traditional alternatives.

Second, perhaps the largest constraint to model development is now training data rather than imagery. While imagery has become abundant, the scarcity and (in many settings) unreliability of quality labels make both training and validation of satellite-based models difficult. Expanding the quantity and – in particular – the quality of labels will quickly accelerate progress in this field, and allow both researchers and practitioners to measure new outcomes and to accurately assess model performance.

Third, despite the growing power of satellite-based approaches, there are many domains where such approaches are likely to contribute little in the near term – for instance, in measuring female empowerment, educational outcomes, or conflict events. Even in settings where satellites are likely to be useful, satellite-based approaches will likely amplify rather than replace existing ground-based data collection efforts. High-quality local training data can nearly always improve model performance, and will remain essential for convincing both researchers and decision-makers that satellite-based approaches are working.

Finally, there remain limited documented cases where satellites have been operationalized into decision-making processes in the sustainable development domains where we focus – with satellite-informed population estimates being the main exception. Limited adoption is likely driven by a number of forces, including the recency of the technology, the lack of accuracy (perceived or real) of the models, lack of model interpretability, and entrenched interests in maintaining the current data regime.

Helping to overcome these constraints constitute key tasks for researchers and policymaker going forward. We suggest nine specific areas where we believe future work would be particularly useful:

1. *More accurate, more numerous training data.* Many applications of deep learning outside sustainable development have been advanced by the curation of reference datasets that are then made available to the community. These datasets lower the barriers to entry and make comparison of different approaches more straightforward, yet they are lacking for sustainable development outcomes. Particularly needed are datasets that track outcomes over time so that models can be optimized to detect changes. These datasets are a major public good and investment in their collection would greatly accelerate research progress. Collecting and publishing location data from existing and ongoing ground surveys (using appropriate privacy safeguards already widely in use) would also greatly benefit research efforts in this area.
2. *More evaluation in the context of specific use cases.* Most evaluation of satellite estimates have focused on agreement with a ground-based measure of a particular outcome. Fewer

studies have then gone the next step to evaluate the actual application of the outcome measure, such as to test the impact of a randomized control trial or target an intervention to a sub-population. These downstream tasks often provide a more tangible example of the utility to potential users, and can avoid the pitfalls of direct comparisons to noisy ground measures. A related task will be to define and utilize meaningful loss functions for the specific task at hand; for instance, a poverty targeting application might be more tolerant of small errors at the wealthy end of the distribution than the poorer end.

3. *Improved model interpretability and transparency.* Especially in cases where satellite-based prediction is being used to make decisions that directly impact people (e.g. targeting aid) it is especially important that predictions be explainable and that decisions based on those predictions be transparent. Applying FAT ML or similar guidelines to research output will be increasingly important as research gets operationalized.
4. *Creative data fusion.* Combining information from multiple different optical sensors of different temporal and spatial resolutions, combining different types of imagery (e.g. optical + radar), and/or combining satellite imagery with other relevant data (e.g. from cell phones), appear to be particularly promising approaches to improving model performance. As much of these additional data are collected by the private sector, sustained and enforceable data-sharing agreements between companies and researchers will be key.^[107]
5. *Scaling estimates.* Researchers typically have more incentive to innovate on methods than they do (e.g.) to apply validated methods across large geographies and update estimates as new data come in – the later being what is often required to make outputs useful to decision-makers. Partnerships between academic researchers and public- or private-sector organizations who have the skills and resources to do this scaling will be key to operationalizing many promising research advances in the satellite/ML domain.
6. *Measuring changes over time.* Much of the literature reviewed above makes predictions at a given point in time. However, many applications require measuring changes over time. While the relationship between inputs and outputs over time is reasonably stable in some domains (e.g. vegetation indices and yields in agriculture), this might not be true in other domains (e.g. economic development). Unfortunately, temporal evaluation at a local level is difficult because there exist few ground datasets that repeatedly and reliably measure the same locations over time. Curating these datasets and using them to develop and validate temporal predictions will be key for tracking the evolution of key sustainability outcomes.
7. *Using imagery to actively guide ground data collection.* As predictive performance of satellite-based models improve, their output could be used to optimally guide further data collection on the ground – for instance, to collect data in locations where model predictions are least certain. Research should explore to what extent such sampling strategies could

- improve outcome measurement compared to traditional sampling approaches.
8. *Understanding potential pitfalls in causal inference applications.* For instance, can poverty predictions from a satellite-based model be used to study the impact of new road construction on poverty, if there is a chance that the model looks for a road to decide whether a location is poor? How do we proceed if we're concerned that image-derived proxies for a dependent variable of interest are themselves the independent variable of interest?
 9. *Improved guidelines for privacy.* As predictions become increasingly granular and accurate, who has access to these data? How can precisely georeferenced ground data (which is increasingly collected) be used to train or validate models without undermining privacy? Guidelines for navigating these issues are increasingly critical as models improve.

References

- [1] Moore, G. K. What is a picture worth? a history of remote sensing. *Hydrological Sciences Bulletin* **24**, 477–485 (1979).
- [2] Waxman, O. B. Aerial photography's surprising role in history (2018). URL <https://time.com/longform/aerial-photography-drones-history/>.
- [3] Union of Concerned Scientists. The ucs satellite database (2020). URL http://www.ucsusa.org/satellite_database.
- [4] Mulla, D. J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering* **114**, 358–371 (2013).
- [5] Lobell, D. B. The use of satellite data for crop yield gap analysis. *Field Crops Research* **143**, 56–64 (2013).
- [6] Donaldson, D. & Storeygard, A. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* **30**, 171–98 (2016).
- [7] Kuffer, M., Pfeffer, K. & Sliuzas, R. Slums from space—15 years of slum mapping using remote sensing. *Remote Sensing* **8**, 455 (2016).
- [8] Network, S. D. S. *Data for development: A needs assessment for SDG monitoring and statistical capacity development* (Sustainable Development Solutions Network., 2015).
- [9] Wardrop, N. *et al.* Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences* **115**, 3529–3537 (2018).
- [10] Devarajan, S. Africa's statistical tragedy. *Review of Income and Wealth* **59**, S9–S15 (2013).
- [11] Beegle, K., De Weerdt, J., Friedman, J. & Gibson, J. Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of Development Economics* **1**, 3–18 (2012).
- [12] Carletto, C., Jolliffe, D. & Banerjee, R. From tragedy to renaissance: Improving agricultural data for better policies. *The Journal of Development Studies* **51**, 133–148 (2015).
- [13] Braimoh, A. *et al.* Capacity needs assessment for improving agricultural statistics in kenya. Tech. Rep., The World Bank (2018).

- [14] MacDonald, R. B. & Hall, F. G. Global crop forecasting. *Science* **208**, 670–679 (1980).
- [15] Elvidge, C. D. *et al.* Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing* **18**, 1373–1379 (1997).
- [16] Burke, M. & Lobell, D. B. Satellite-based assessment of yield variation and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences* **114**, 2189–2194 (2017).
- [17] Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *American economic review* **102**, 994–1028 (2012).
- [18] Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).
- [19] Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, vol. 1 (MIT press Cambridge, 2016).
- [20] Weiss, M., Jacob, F. & Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment* **236**, 111402 (2020).
- [21] You, J., Li, X., Low, M., Lobell, D. & Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, 4559–4566 (2017).
- [22] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [23] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
- [24] Tiecke, T. G. *et al.* Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839* (2017).
- [25] Zong, Z., Feng, J., Liu, K., Shi, H. & Li, Y. Deepdpm: Dynamic population mapping via deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 1294–1301 (2019).
- [26] Hu, W. *et al.* Mapping missing population in rural india: A deep learning approach with satellite imagery. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 353–359 (2019).
- [27] Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
- [28] Head, A., Manguin, M., Tran, N. & Blumenstock, J. E. Can human development be measured with satellite imagery? In *ICTD*, 8–1 (2017).
- [29] Steele, J. E. *et al.* Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* **14**, 20160690 (2017).
- [30] Yeh, C. *et al.* Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications* **11**, 1–11 (2020).
- [31] Cadamuro, G., Muhebwa, A. & Taneja, J. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699* (2018).
- [32] Oshri, B. *et al.* Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proc. 24th ACM SIGKDD Conference* (2018).

- [33] Albert, A., Kaur, J. & Gonzalez, M. C. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1357–1366 (2017).
- [34] Helber, P., Bischke, B., Dengel, A. & Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**, 2217–2226 (2019).
- [35] Mboga, N., Persello, C., Bergado, J. R. & Stein, A. Detection of informal settlements from vhr images using convolutional neural networks. *Remote sensing* **9**, 1106 (2017).
- [36] Persello, C. & Stein, A. Deep fully convolutional networks for the detection of informal settlements in vhr images. *IEEE geoscience and remote sensing letters* **14**, 2325–2329 (2017).
- [37] Kroodsma, D. A. *et al.* Tracking the global footprint of fisheries. *Science* **359**, 904–908 (2018).
- [38] Park, J. *et al.* Illuminating dark fishing fleets in north korea. *Science Advances* **6**, eabb1197 (2020).
- [39] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).
- [40] Xingjian, S. *et al.* Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810 (2015).
- [41] Ji, S., Zhang, C., Xu, A., Shi, Y. & Duan, Y. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing* **10**, 75 (2018).
- [42] Rußwurm, M. & Körner, M. Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **42**, 551 (2017).
- [43] M Rustowicz, R. *et al.* Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 75–82 (2019).
- [44] Sun, J., Di, L., Sun, Z., Shen, Y. & Lai, Z. County-level soybean yield prediction using deep cnn-lstm model. *Sensors* **19**, 4363 (2019).
- [45] Xiao, L., Zhang, Y. & Peng, G. Landslide susceptibility assessment using integrated deep learning algorithm along the china-nepal highway. *Sensors* **18**, 4436 (2018).
- [46] Xu, J. Z., Lu, W., Li, Z., Khaitan, P. & Zaytseva, V. Building damage detection in satellite imagery using convolutional neural networks. *arXiv preprint arXiv:1910.06444* (2019).
- [47] Ci, T., Liu, Z. & Wang, Y. Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression. *Remote Sensing* **11**, 2858 (2019).
- [48] Davenport, F. M. *et al.* Using out-of-sample yield forecast experiments to evaluate which earth observation products best indicate end of season maize yields. *Environmental Research Letters* **14**, 124095 (2019).
- [49] Sheehan, E. *et al.* Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2698–2706 (2019).
- [50] Fatehkia, M., Coles, B., Ofli, F. & Weber, I. The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 934–938 (2020).

- [51] Cao, R. *et al.* Integrating aerial and street view images for urban land use classification. *Remote Sensing* **10**, 1553 (2018).
- [52] Tingzon, I. *et al.* Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In *AI for Social Good ICML 2019 Workshop* (2019).
- [53] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
- [54] Lobell, D. B. *et al.* Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics* **102**, 202–219 (2020).
- [55] Christie, G., Fendley, N., Wilson, J. & Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180 (2018).
- [56] Uzkent, B. *et al.* Learning to interpret satellite images using wikipedia. *IJCAI* (2019).
- [57] Ayush, K., Uzkent, B., Burke, M., Lobell, D. & Ermon, S. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612* (2020).
- [58] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).
- [59] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).
- [60] Basu, S. *et al.* Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, 1–10 (2015).
- [61] Jean, N., Wang, S., Azzari, G., Lobell, D. & Ermon, S. Tile2vec: Unsupervised representation learning for remote sensing data. In *arXiv preprint arXiv:1805.02855* (2018).
- [62] Jean, N., Xie, S. M. & Ermon, S. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *NIPS* (2018).
- [63] Elmes, A. *et al.* Accounting for training data error in machine learning applied to earth observations. *Remote Sensing* **12**, 1034 (2020).
- [64] Krause, J. *et al.* The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, 301–320 (Springer, 2016).
- [65] Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694* (2017).
- [66] Natarajan, N., Dhillon, I. S., Ravikumar, P. K. & Tewari, A. Learning with noisy labels. In *Advances in neural information processing systems*, 1196–1204 (2013).
- [67] Charikar, M., Steinhardt, J. & Valiant, G. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 47–60 (2017).
- [68] Paliwal, A. & Jain, M. The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. *Frontiers in Sustainable Food Systems* **4**, 25 (2020).

- [69] Wang, S. *et al.* Mapping crop types in southeast india with smartphone crowdsourcing and deep learning. *working paper* (2020).
- [70] Kaiser, P. *et al.* Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing* **55**, 6054–6068 (2017).
- [71] Christen, R. P. & Anderson, J. Segmentation of smallholder households: Meeting the range of financial needs in agricultural families. Tech. Rep., The World Bank (2013).
- [72] Chivasa, W., Mutanga, O. & Biradar, C. Application of remote sensing in estimating maize grain yield in heterogeneous african agricultural landscapes: a review. *International Journal of Remote Sensing* **38**, 6816–6845 (2017).
- [73] Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C. & Worley, B. A. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing* **66**, 849–857 (2000).
- [74] Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one* **10**, e0107042 (2015).
- [75] Schiavina, M., Freire, S. & MacManus, K. Ghs population grid multitemporal (1975, 1990, 2000, 2015) r2019a. *Eur. Comm. JRC* (2019).
- [76] Bustos, M. F. A., Hall, O., Niedomysl, T. & Ernstson, U. A pixel level evaluation of five multitemporal global gridded population datasets: a case study in sweden, 1990–2015. *Population and Environment* (2020).
- [77] Calka, B. & Bielecka, E. Ghs-pop accuracy assessment: Poland and portugal case study. *Remote Sensing* **12**, 1105 (2020).
- [78] Bai, Z., Wang, J., Wang, M., Gao, M. & Sun, J. Accuracy assessment of multi-source gridded population distribution datasets in china. *Sustainability* **10**, 1363 (2018).
- [79] Engstrom, R., Newhouse, D. & Soundararajan, V. Estimating small-area population density in sri lanka using surveys and geo-spatial data. *PloS one* **15**, e0237063 (2020).
- [80] Thematic Research Network on Data and Statistics. Leaving no one off the map: A guide for gridded population data for sustainable development. Tech. Rep., UN Sustainable Development Solutions Network (2020).
- [81] Habitat, U. Habitat iii issue paper 22—informal settlements. *New York: UN Habitat* (2015).
- [82] Leyk, S. *et al.* The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data* **11** (2019).
- [83] Kubitza, C., Krishna, V. V., Schulthess, U. & Jain, M. Estimating adoption and impacts of agricultural management practices in developing countries using satellite data. a scoping review. *Agronomy for Sustainable Development* **40**, 1–21 (2020).
- [84] Jain, M. *et al.* The impact of agricultural interventions can be doubled by using satellite data. *Nature Sustainability* **2**, 931–934 (2019).
- [85] Belhabib, D. *et al.* Catching industrial fishing incursions into inshore waters of africa from space. *Fish and Fisheries* **21**, 379–392 (2020).

- [86] Pinkovskiy, M. & Sala-i Martin, X. Lights, camera... income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics* **131**, 579–631 (2016).
- [87] Henderson, J. V., Squires, T., Storeygard, A. & Weil, D. The global distribution of economic activity: nature, history, and the role of trade. *The Quarterly Journal of Economics* **133**, 357–406 (2018).
- [88] Harari, M. Cities in bad shape. *American Economic Review* **110**.
- [89] Michalopoulos, S. & Papaioannou, E. Pre-colonial ethnic institutions and contemporary african development. *Econometrica* **81**, 113–152 (2013).
- [90] Michalopoulos, S. & Papaioannou, E. National institutions and subnational development in africa. *The Quarterly journal of economics* **129**, 151–213 (2014).
- [91] Storeygard, A. Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. *The Review of economic studies* **83**, 1263–1295 (2016).
- [92] Pinkovskiy, M. L. Growth discontinuities at borders. *Journal of Economic Growth* **22**, 145–192 (2017).
- [93] Jayachandran, S. *et al.* Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* **357**, 267–273 (2017).
- [94] Marx, B., Stoker, T. M. & Suri, T. There is no free house: Ethnic patronage in a kenyan slum. *American Economic Journal: Applied Economics* **11**, 36–70 (2019).
- [95] Zhao, S., Yeh, C. & Ermon, S. A framework for sample efficient interval estimation with control variates. In *International Conference on Artificial Intelligence and Statistics*, 4583–4592 (2020).
- [96] Huang, L. Y. Measuring the impacts of poverty alleviation programs with satellite imagery and deep learning (2020). <http://luna-yue-huang.com/assets/pdf/jmp.pdf>.
- [97] Brown, M. E. *Famine early warning systems and remote sensing data* (Springer Science & Business Media, 2008).
- [98] Fritz, S. *et al.* A comparison of global agricultural monitoring systems and current gaps. *Agricultural systems* **168**, 258–272 (2019).
- [99] Watch, G. F. *Ocean sustainability through transparency, data-sharing and collaboration* (2020). URL <https://globalfishingwatch.org/wp-content/uploads/GFW-program-2020.pdf>
- [100] Blumenstock, J. Machine learning can help get covid-19 aid to those who need it most. *Nature* (2020).
- [101] Cossins, D. Discriminating algorithms: 5 times ai showed prejudice. *New Scientist* **12** (2018).
- [102] Milner, Y. Data for black lives. URL <https://d4bl.org/about.html>.
- [103] on Civil Human Rights, L. C. Civil rights principles for the era of big data. URL <https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/>.
- [104] Diakopoulos, N. *et al.* Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML* (2017).
- [105] Gajane, P. & Pechenizkiy, M. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

- [106] It is no longer a crime to publish statistics in tanzania. *The Citizen* (2019). URL <https://www.thecitizen.co.tz/news/It-is-no-longer-a-crime-to-publish-statistics-in-Tanzania-/1840340-5174870-wjjdxhz/index.html>
- [107] Lazer, D. M. *et al.* Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).
- [108] Solt, F. The standardized world income inequality database, version 8. *Cambridge: Harvard Dataverse* (2019).
- [109] World development indicators (2014). URL <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD>
- [110] Marshall, M. & Gurr, T. Polity5: Political regime characteristics and transitions, 1800-2018. *Center for Systemic Peace*. <http://www.systemicpeace.org/inscr/p5manualv2018.pdf> (2020).
- [111] Team, P. Planet application program interface: In space for life on earth (2017). URL <https://api.planet.com>
- [112] Team, L. Landinfo worldwide mapping llc. URL <http://search.landinfo.com/>
- [113] Gorelick, N. *et al.* Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* (2017). URL <https://doi.org/10.1016/j.rse.2017.06.031>
- [114] Jain, M. *et al.* Mapping smallholder wheat yields and sowing dates using micro-satellite data. *Remote sensing* **8**, 860 (2016).
- [115] Lambert, M.-J., Blaes, X., Traoré, P. S. & Defourny, P. Estimate yield at parcel level from s2 time serie in sub-saharan smallholder farming systems. In *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 1–7 (IEEE, 2017).
- [116] Guan, K. *et al.* Mapping paddy rice area and yields over thai binh province in viet nam from modis, landsat, and alos-2/palsar-2. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**, 2238–2252 (2018).
- [117] Karst, I. G. *et al.* Estimating yields of household fields in rural subsistence farming systems to study food security in burkina faso. *Remote Sensing* **12**, 1717 (2020).
- [118] Jin, Z., Azzari, G., Burke, M., Aston, S. & Lobell, D. B. Mapping smallholder yield heterogeneity at multiple scales in eastern africa. *Remote Sensing* **9**, 931 (2017).
- [119] Schulthess, U., Timsina, J., Herrera, J. & McDonald, A. Mapping field-scale yield gaps for maize: An example from bangladesh. *Field Crops Research* **143**, 151–156 (2013).
- [120] Zhao, Q. *et al.* Detecting spatial variability of paddy rice yield by combining the dndc model with high resolution satellite images. *Agricultural Systems* **152**, 47–57 (2017).
- [121] Lobell, D. B. *et al.* Sight for sorghums: Comparisons of satellite-and ground-based sorghum yield estimates in mali. *Remote Sensing* **12**, 100 (2020).
- [122] Kim, J. H., Xie, M., Jean, N. & Ermon, S. Incorporating spatial context and fine-grained detail from satellite imagery to predict poverty (2016).
- [123] Perez, A. *et al.* Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654* (2017).
- [124] Engstrom, R., Hersh, J. & Newhouse, D. Poverty from space: using high-resolution satellite imagery for estimating economic well-being (2017).

- [125] Pokhriyal, N. & Jacques, D. C. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences* **114**, E9783–E9792 (2017).
- [126] Njuguna, C. & McSharry, P. Constructing spatiotemporal poverty indices from big data. *Journal of Business Research* **70**, 318–327 (2017).
- [127] Perez, A. *et al.* Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. *arXiv preprint arXiv:1902.11110* (2019).
- [128] Smith, B. & Wills, S. Left in the dark? oil and rural poverty. *Journal of the Association of Environmental and Resource Economists* **5**, 865–904 (2018).
- [129] Irvine, J. M., Wood, R. J. & McBee, P. Viewing society from space: Image-based sociocultural prediction models (2017).
- [130] Li, G., Cai, Z., Liu, X., Liu, J. & Su, S. A comparison of machine learning approaches for identifying high-poverty counties: robust features of dmsp/ols night-time light imagery. *International journal of remote sensing* **40**, 5716–5736 (2019).
- [131] Xie, M., Jean, N., Burke, M., Lobell, D. & Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).
- [132] Watmough, G. R. *et al.* Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences* **116**, 1213–1218 (2019).
- [133] Watmough, G. R., Atkinson, P. M., Saikia, A. & Hutton, C. W. Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: An example from assam, india. *World Development* **78**, 188–203 (2016).
- [134] Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A. & Crooks, A. T. Detecting and mapping slums using open data: a case study in kenya. *International Journal of Digital Earth* **13**, 683–707 (2020).
- [135] Rhinane, H., Hilali, A., Berrada, A., Hakdaoui, M. *et al.* Detecting slums from spot data in casablanca morocco using an object based approach. *Journal of Geographic Information System* **3**, 217 (2011).
- [136] Hofmann, P., Strobl, J., Blaschke, T. & Kux, H. Detecting informal settlements from quickbird data in rio de janeiro using an object based approach. In *Object-based image analysis*, 531–553 (Springer, 2008).
- [137] Stoler, J. *et al.* Assessing the utility of satellite imagery with differing spatial resolutions for deriving proxy measures of slum presence in accra, ghana. *GIScience & Remote Sensing* **49**, 31–52 (2012).
- [138] Maiya, S. R. & Babu, S. C. Slum segmentation and change detection: A deep learning approach. *arXiv preprint arXiv:1811.07896* (2018).
- [139] Duque, J. C., Patino, J. E. & Betancourt, A. Exploring the potential of machine learning for automatic slum identification from vhr imagery. *Remote Sensing* **9**, 895 (2017).
- [140] Gram-Hansen, B. J. *et al.* Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368 (2019).
- [141] Engstrom, R. *et al.* Mapping slums using spatial features in accra, ghana. In *2015 Joint Urban Remote Sensing Event (JURSE)*, 1–4 (IEEE, 2015).

- [142] Williams, T. K.-A., Wei, T. & Zhu, X. Mapping urban slum settlements using very high-resolution imagery and land boundary data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 166–177 (2019).
- [143] Verma, D., Jana, A. & Ramamritham, K. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. *Habitat International* **88**, 101981 (2019).
- [144] Leonita, G., Kuffer, M., Sliuzas, R. & Persello, C. Machine learning-based slum mapping in support of slum upgrading programs: The case of bandung city, indonesia. *Remote sensing* **10**, 1522 (2018).
- [145] Graesser, J. *et al.* Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **5**, 1164–1176 (2012).
- [146] Fallatah, A., Jones, S. & Mitchell, D. Object-based random forest classification for informal settlements identification in the middle east: Jeddah a case study. *International Journal of Remote Sensing* **41**, 4421–4445 (2020).
- [147] Jochem, W. C., Bird, T. J. & Tatem, A. J. Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computers, environment and urban systems* **69**, 104–113 (2018).
- [148] Wurm, M., Taubenböck, H., Weigand, M. & Schmitt, A. Slum mapping in polarimetric sar data using spatial features. *Remote sensing of environment* **194**, 190–204 (2017).
- [149] Kuffer, M., Pfeffer, K., Sliuzas, R. & Baud, I. Extraction of slum areas from vhr imagery using glcm variance. *IEEE Journal of selected topics in applied earth observations and remote sensing* **9**, 1830–1840 (2016).
- [150] Kit, O. & Lüdeke, M. Automated detection of slum area change in hyderabad, india using multitemporal satellite imagery. *ISPRS journal of photogrammetry and remote sensing* **83**, 130–137 (2013).
- [151] Hofmann, P. & Bekkarnayeva, G. Object-based change detection of informal settlements. In *2017 Joint Urban Remote Sensing Event (JURSE)*, 1–4 (IEEE, 2017).
- [152] Kohli, D., Sliuzas, R. & Stein, A. Urban slum detection using texture and spatial metrics derived from satellite imagery. *Journal of spatial science* **61**, 405–426 (2016).
- [153] Kohli, D., Warwadekar, P., Kerle, N., Sliuzas, R. & Stein, A. Transferability of object-oriented image analysis methods for slum identification. *Remote Sensing* **5**, 4209–4228 (2013).

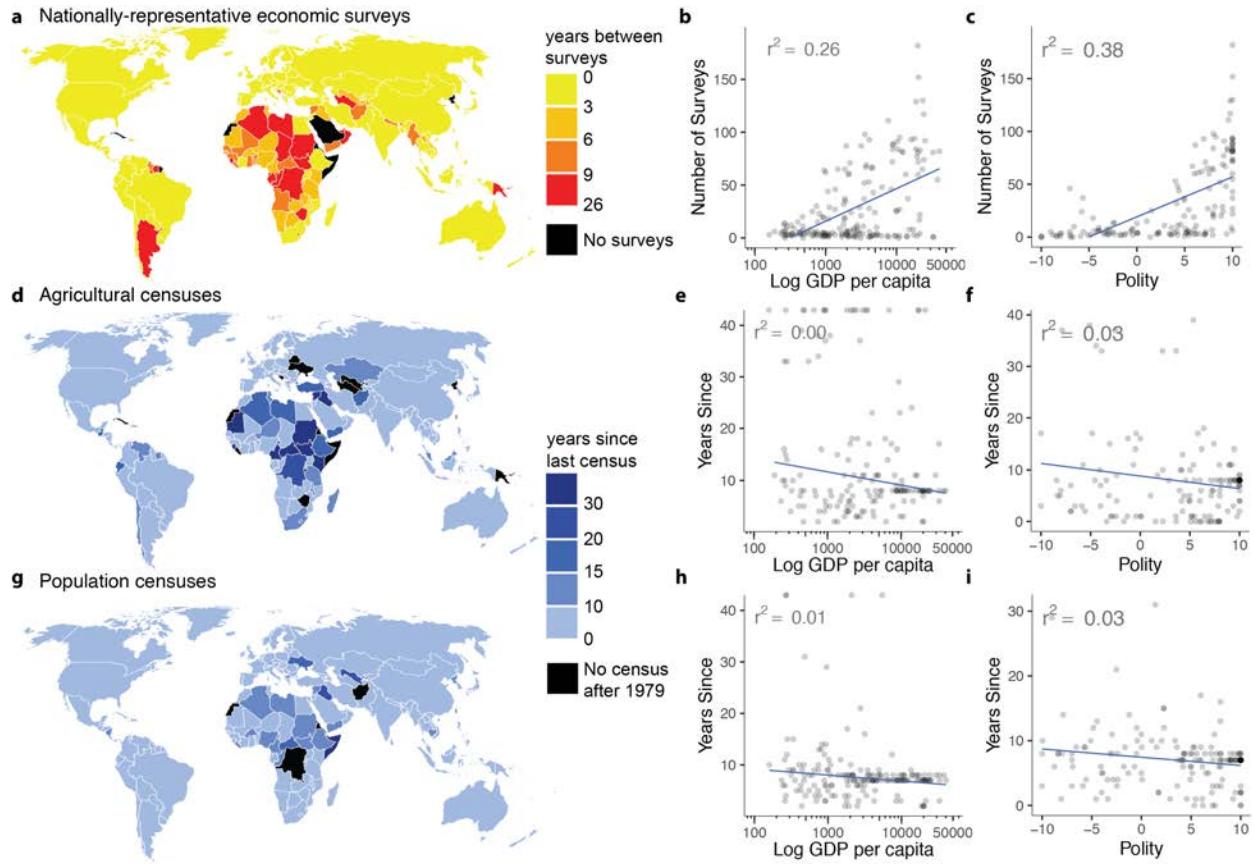


Figure 1: Nationally-representative economic, agricultural, and population data are collected infrequently in much of the world. **a** The average interval between nationally representative economic surveys (of Average or High quality) for the period of 1993 to 2018 from the UN World Income Inequality Database.^[108] **b** Relationship between GDP per capita^[109] and number of surveys in the study period. Nations with higher GDP per capita tend to have more surveys. **c** Relationship between the Polity Score of each country (+10 is fully democratic, -10 is fully autocratic)^[110] and the number of surveys in the study period. **d-f** Years since last agricultural census. **e-f** Relationship between GDP per capita, Polity score and years since last agricultural census. **g-i** As in d-f but for population censuses.

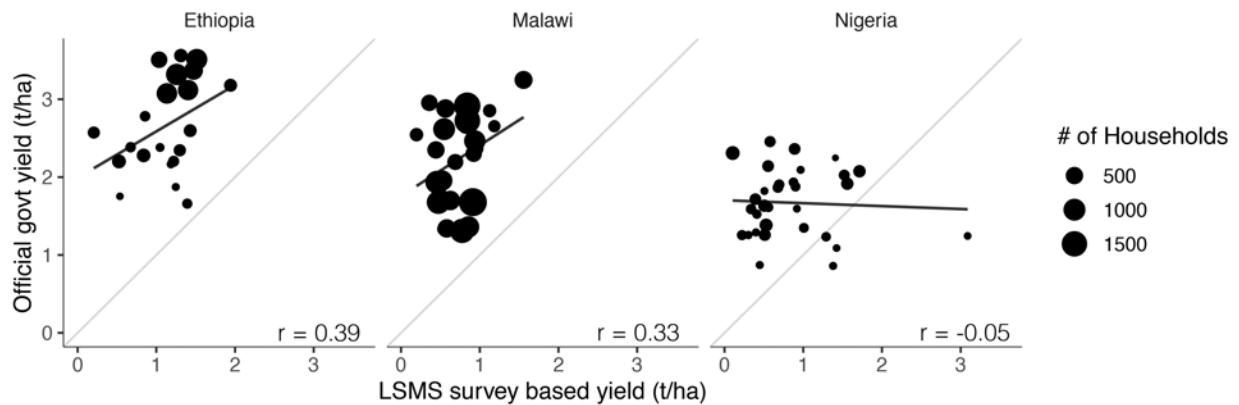


Figure 2: Government and household-survey based data on maize productivity are not well correlated at the district level. Using government data from eAtlas and household level yield data from LSMS-ISA surveys, maize yields are compared by averaging across all households in a given district. Data include 2011, 2013, and 2015 data in Ethiopia, 2013 data in Malawi, and 2010 and 2012 data in Nigeria. Comparison is restricted to district-years with at least 30 households. Grey line is 1:1 line, while black lines show linear fits within each country. Points are sized relative to the number of households contributing to each estimate in the LSMS data.

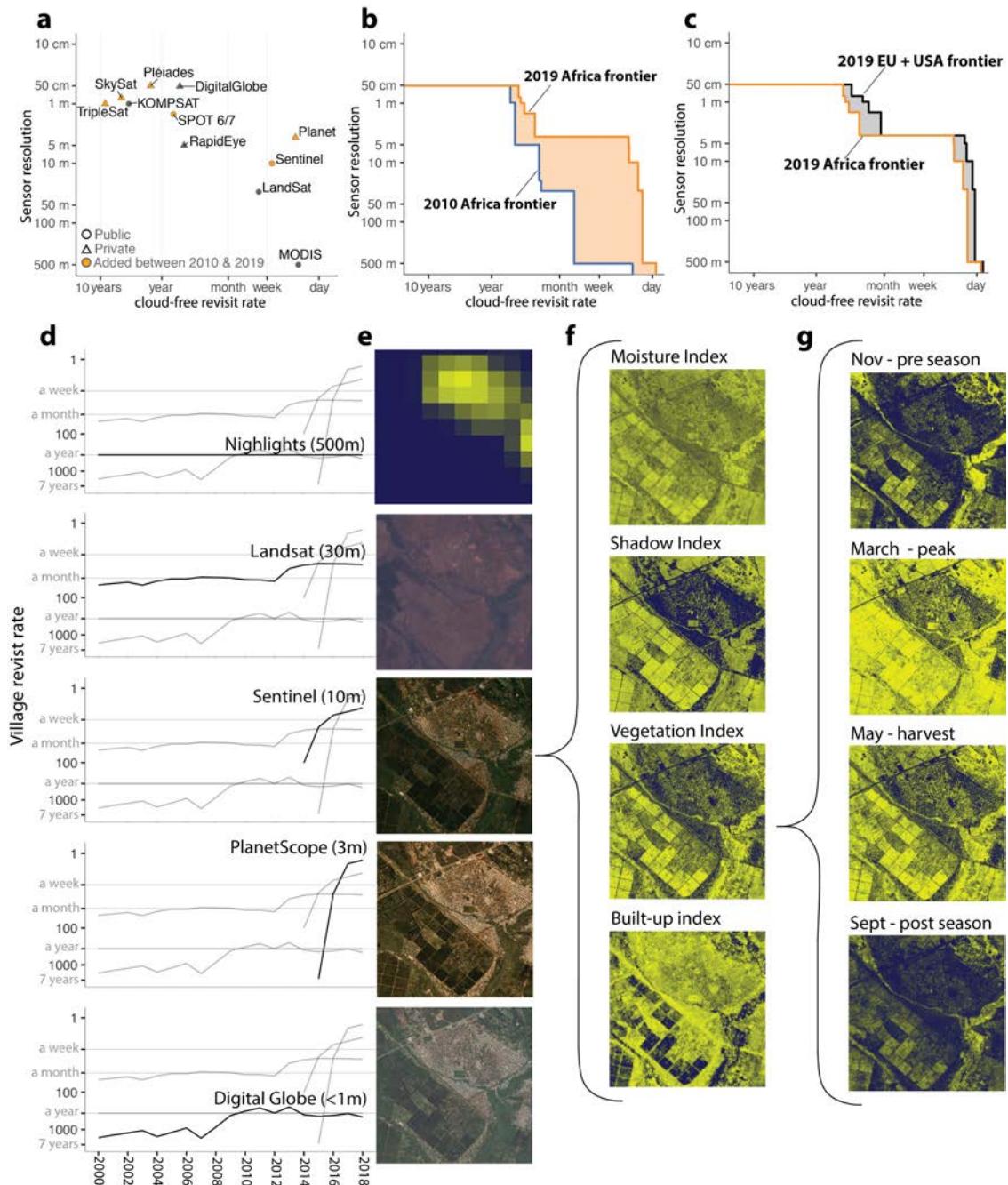


Figure 3: Spatial resolution, temporal frequency, and spectral availability of satellite imagery have increased substantially since 2000. **a** Average revisit rate and sensor resolution of cloud-free optical imagery in 2019, averaged across 100 populated African locations. **b** Blue line (“frontier”) shows overall revisit rate across all available sensors at a given spatial resolution in 2010 for same 100 locations (e.g. at 1m, line denotes average cloud-free revisit rate using all sensors $\leq 1\text{m}$); orange line shows same for 2019. Orange area denotes the new combinations of temporal and spatial resolution available by 2019, which expanded greatly at resolutions $>1\text{m}$. **c** Average 2019 coverage in Africa (orange line) vs 100 locations in US/EU (grey line; locations randomly sampled proportional to population). Grey shaded area depicts inequalities in coverage between wealthy and developing regions in 2019, which are larger for imagery $<3\text{m}/\text{px}$. **d** Calculated revisit periods for several satellites over 500 randomly selected survey locations in Africa since 2000. Nightlights is set to a one year revisit rate given the stable yearly product. **e** Example imagery corresponding to each sensor in a single location in central Zambia. Images are real color except for NL. **f** Indices generated from various bands can convey different information, as depicted here using Sentinel 2 data (yellow colors indicate higher values of the index). **g** Frequent revisit rates of new public sensors capture temporal variation in human activity, including rapid changes throughout the main agricultural season shown here.

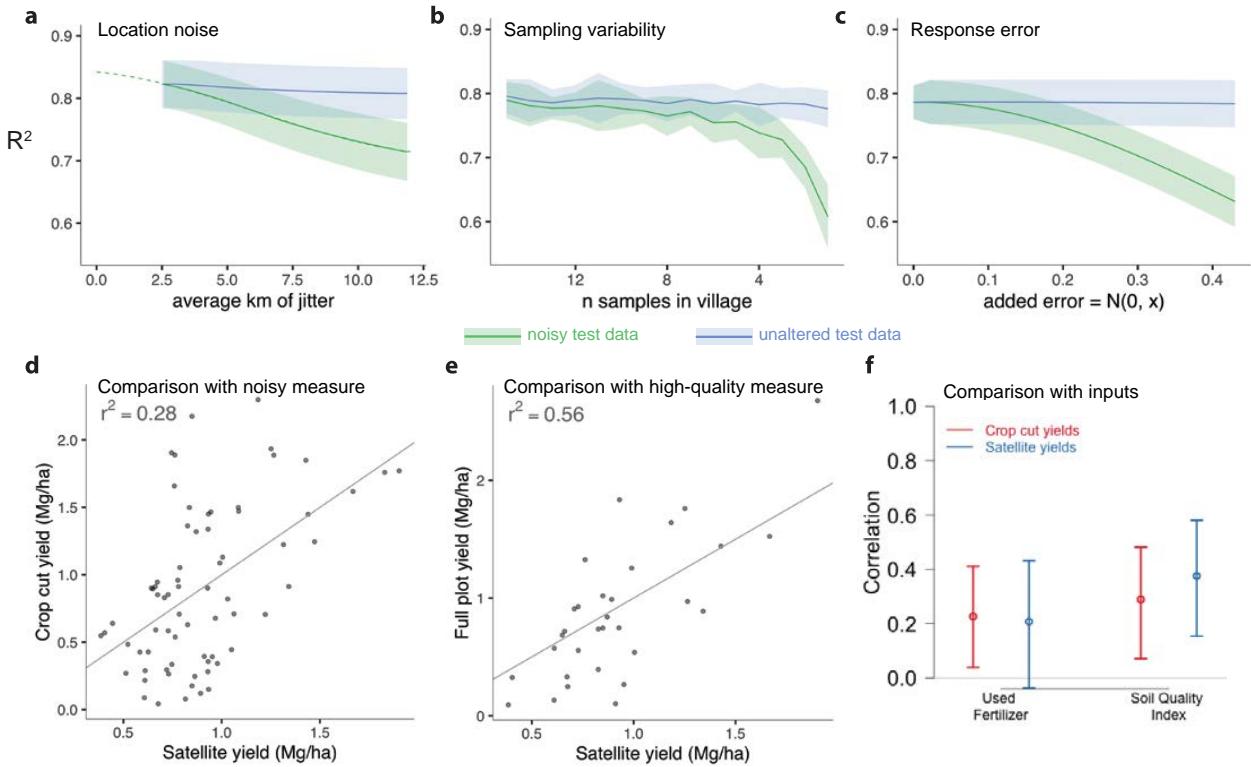


Figure 4: The role of noise in model performance and evaluation. **a-c** Performance of wealth prediction model as noise is added to train and test data. Model trained to predict asset wealth from nightlights imagery across 4000 African villages, using the dataset from ref.^[30] Performance is evaluated as three different types of noise are added to training data: **a** random noise in village geo-coordinates (starting from 2.5km, the actual noise in the survey data), **b** noise from constructing village-level wealth estimates from decreasing numbers of households within the village to represent sampling variability, and **c** random noise added to village-level wealth estimates, representing random response error from respondents. Green lines show performance evaluated on test data where similar noise has been added, blue lines show performance on test data where noise has not been added. Shaded areas indicate confidence intervals across 200 runs at a given level of added noise. As all types of training noise increase, model performance degrades when evaluated against similarly noisy test data but does not degrade when evaluated against unaltered test data. **d-f** Example from a study of maize yields in Uganda^[51] in which both ground-based and satellite-based measurements can have noise, and multiple approaches can help adjudicate which is noisier. **d** Imperfect correlation between ground- and satellite-based yield measure does not reveal source of noise. **e** Comparison of satellite measure with available gold-standard ground measure from full plot harvest shows higher correlation, indicating ground measure in (d) responsible for at least some of the noise. **f** Comparison of satellite measure and ground measure with independent third measures expected to correlate with yields (here, fertilizer use and soil quality) suggests that the two yield measures in (d) are roughly equally noisy.

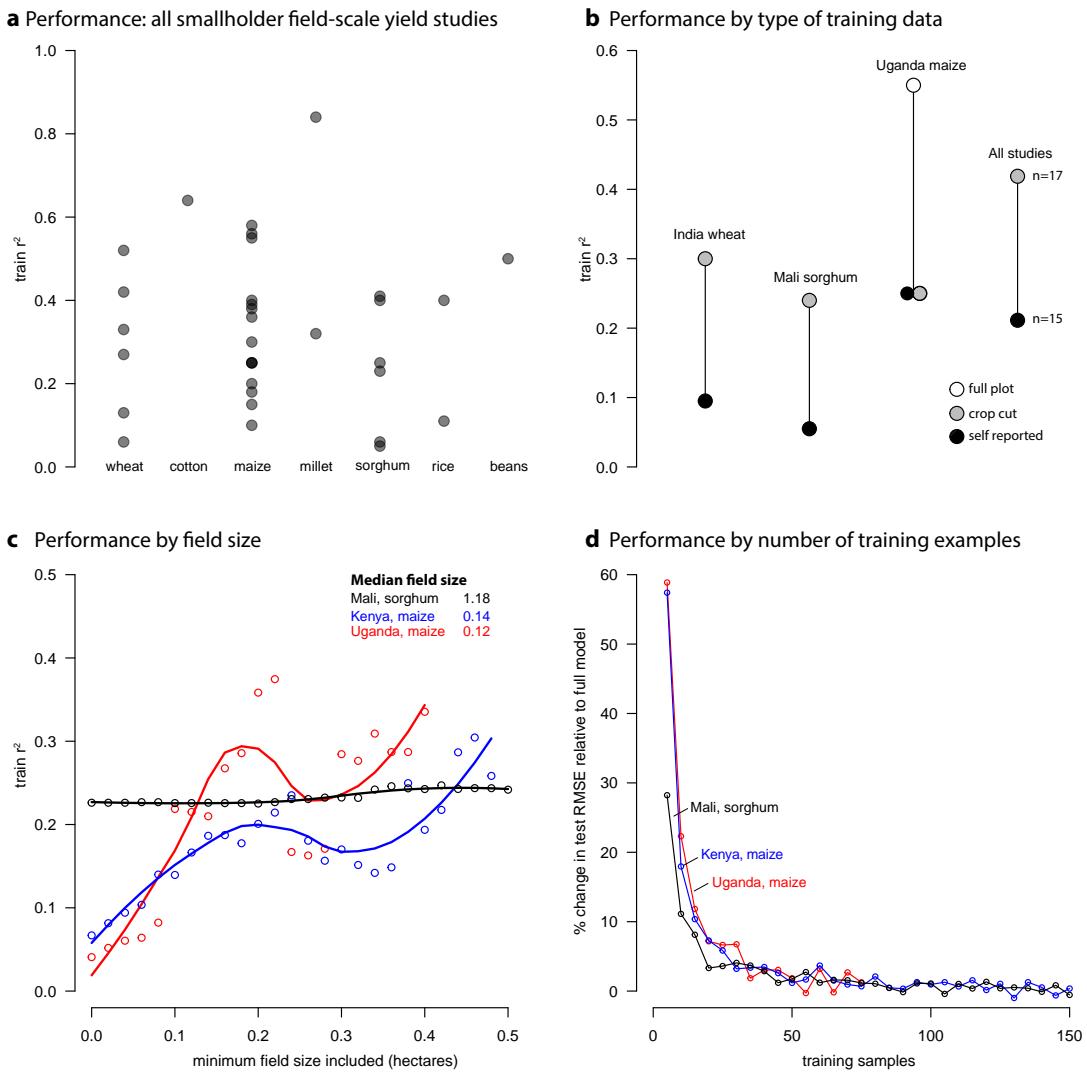


Figure 5: Performance of satellite-based approaches to measuring smallholder yield at field scale. **a** Performance across all known published studies where coefficient of determination (r^2) was reported (32 estimates across 11 studies); r^2 estimates are “in-sample”, i.e. for data on which model was trained. **b** Difference in performance for models trained and evaluated on crop-cut, self-reported, or full-plot harvest data suggest that more objective crop measures improve performance. First three estimates are for studies that compared at least two types of ground data in the same setting. “All studies” estimates pool across estimates in (a). **c** Performance generally increases when sample is restricted to larger fields, particularly in East African settings where field sizes are very small. **d** Performance on test data improves rapidly with additional training examples up to ~ 30 data points, and then improves more gradually thereafter. Performance measured as average root mean squared error between predicted and observed yields in the test set, averaged over 100 different random subsets of training samples at each size of the training set.

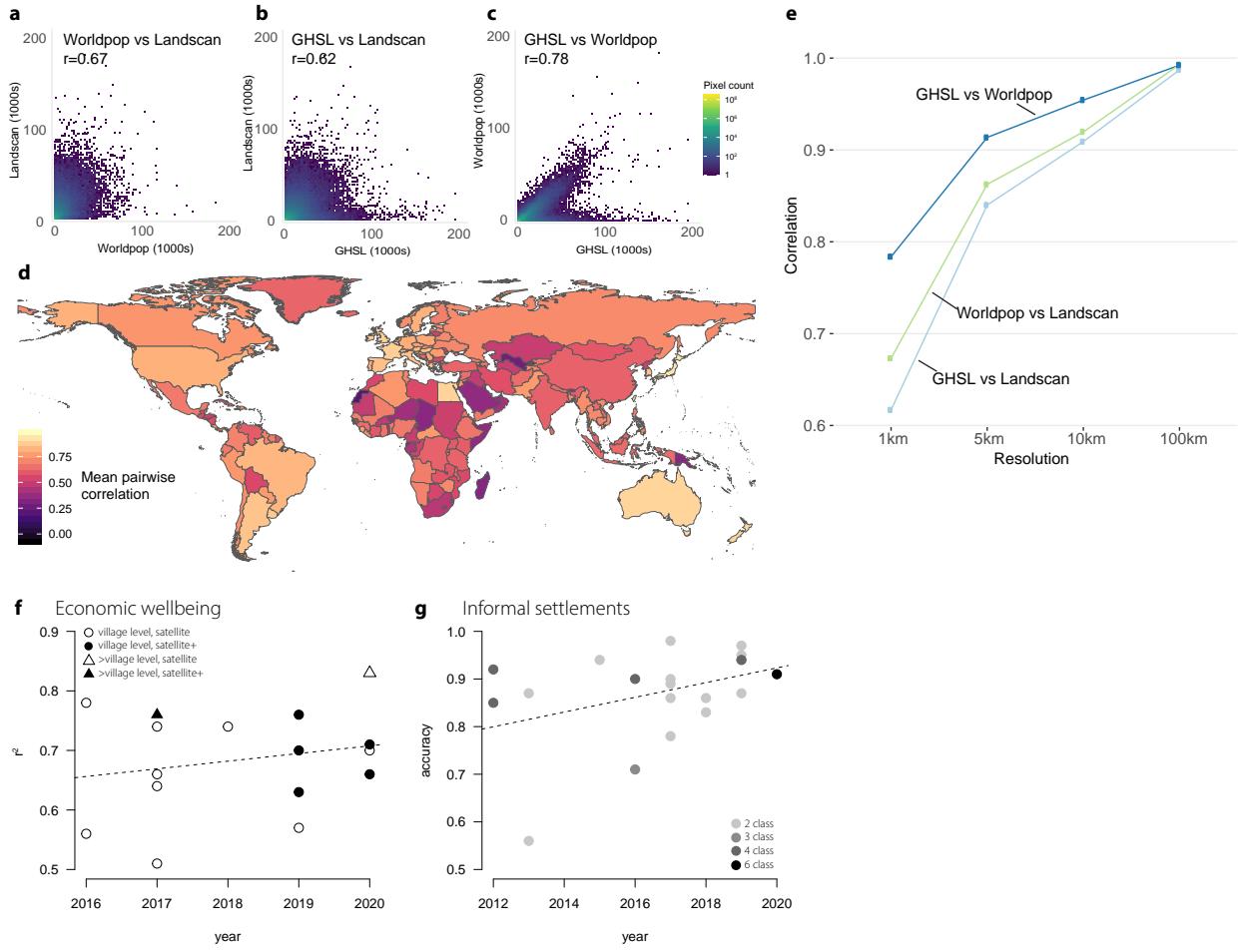


Figure 6: Performance of satellite-based approaches to measuring population, wealth, and informal settlements. **a-c** Comparison of three different satellite-informed global population datasets (Landscan, WorldPop, and GHSL population) datasets at 1km resolution globally (colors correspond to scale at right). **d** Average pairwise correlation within each country at 1km resolution. Comparisons show modest correlation between datasets at global scale and often poor correlation in many developing countries. **e** Correlations across datasets improve when data are spatially aggregated. All comparisons are made for pixels that were not missing and not zero across all three datasets. **f** Performance in predicting asset wealth in various developing countries from satellite data (16 estimates from 12 papers), as measured by coefficient of determination on test data. Filled markers are estimates that combine satellite information with other data (cell phone data, social media data, or Wikipedia). Circles indicate estimates at the village level, triangles are estimates at more aggregate spatial scale (sub-district or district). **g** Performance in predicting the location of informal settlements from imagery (20 estimates from 17 papers)

Supplementary Information

Collecting Satellite Revisit Data

Construction of Figure 3 involved acquiring data from several sources. First we use Gridded Population of the World (GPW) data raster data to create a population weighted sample of 100 locations in Africa, as well as 100 locations across the EU and the USA. These 200 locations are then buffered with an approximately 10 meter radius and used to query for satellite imagery for the years 2010 and 2019.

Planet products (SkySat, PlanetScope and RapidEye) were downloaded from the Planet API.^[11] Footprints for other private satellites were downloaded from LandInfo,^[12] while footprints from public satellites (Landsat, Sentinel, MODIS) were downloaded using Google Earth Engine.^[13]

We attempted to maintain consistency in filtering across data sources, but filtering works slightly differently in each system. LandInfo (though it lacks in depth documentation confirming this) appears to filter solely over the area of interest (AOI) rather than over the entire footprint. Public data was processed to match this, using only the cloud cover percentage over the buffered polygon, but Planet is filtered at the footprint level. All image footprints were filtered to be <30% cloud cover and off-nadir <|20|.

Sensors in Figure 3a are grouped slightly to make the figure easier to process visually. Landsat 7 and 8 are combined; WorldView-1 through 4, GeoEye-1, QuickBird-2, and IKONOS are grouped as “DigitalGlobe”; KOMPSAT-3, KOMPSAT-3A, and KOMPSAT-2 are grouped; SPOT-4 and 5 are grouped as well as SPOT-6 and 7. As sensor resolution varies within these groups, we use the mode of the resolutions in the group to represent the group as a whole. This does compress the range of resolutions significantly, for example “DigitalGlobe” is recorded as a resolution of 51cm, where the true resolutions range from 31cm to 91cm. “KOMPSAT” ranges from 70cm to 1m, “SkySat” ranges from 70cm to 1m, “SPOT 4/5” ranges from 2.5m to 10m.

To calculate the average revisit rate, we sum up the total number of images collected in each group and calculate (number of locations*365)/number of images. For the frontier, we calculate the revisit rate by summing the total number of images collected for all satellites with resolution less than or equal to the resolution of interest and run the same calculation as above. As this is an average of time between images, a number below 1 does not necessarily indicate that there is a cloudless picture on every day.

Table S1: Performance of studies using satellites to predict smallholder yields at the plot level. Year=study year, res=sensor resolution, n=number of observations, r2=squared correlation, data=training data (CC=crop cuts, SR=self reported, FP=full plot). In the sensor column, "Skysat (c)" refers to Skysat data that has been coarsened to lower resolution.

	study	year	location	crop	sensor	res	n	r2	data
1	Jain et al 2016 ^[14]	2014	India	wheat	Skysat	2	50	0.27	CC
2	Jain et al 2016 ^[14]	2015	India	wheat	Skysat	2	37	0.33	CC
3	Jain et al 2016 ^[14]	2014	India	wheat	Skysat	2	52	0.13	SR
4	Jain et al 2016 ^[14]	2015	India	wheat	Skysat	2	29	0.06	SR
5	Jain et al 2016 ^[14]	2014	India	wheat	Landsat	30	50	0.52	CC
6	Jain et al 2016 ^[14]	2015	India	wheat	Landsat	30	37	0.42	CC
7	Lambert et al 2017 ^[15]	2016	Mali	cotton	S2	10	9	0.64	CC
8	Lambert et al 2017 ^[15]	2016	Mali	maize	S2	10	9	0.58	CC
9	Lambert et al 2017 ^[15]	2016	Mali	millet	S2	10	8	0.84	CC
10	Lambert et al 2017 ^[15]	2016	Mali	sorghum	S2	10	9	0.41	CC
11	Guan et al 2017 ^[16]	2015	Vietnam	rice	Landsat	30	71	0.40	CC
12	Karst et al 2020 ^[17]	2018	Burkina Faso	beans	S2	10	31	0.50	CC
13	Karst et al 2020 ^[17]	2018	Burkina Faso	maize	S2	10	31	0.40	CC
14	Karst et al 2020 ^[17]	2018	Burkina Faso	sorghum	S2	10	57	0.40	CC
15	Karst et al 2020 ^[17]	2018	Burkina Faso	millet	S2	10	45	0.32	CC
16	Jin et al 2017 ^[18]	2016	Kenya	maize	S2	10	41	0.36	CC
17	Schulthess et al 2013 ^[19]	2010	Bangladesh	maize	rapideye	5	30	0.56	SR
18	Zhao et al 2017 ^[20]	2009	China	rice	formosat-2	8	22	0.11	SR
19	Lobell et al 2020 ^[21]	2017	Mali	sorghum	S2	10	575	0.23	CC
20	Lobell et al 2020 ^[21]	2017	Mali	sorghum	Planet	3	575	0.25	CC
21	Lobell et al 2020 ^[21]	2017	Mali	sorghum	S2	10	575	0.05	SR
22	Lobell et al 2020 ^[21]	2017	Mali	sorghum	Planet	3	575	0.06	SR
23	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat	1	72	0.39	SR
24	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	5	72	0.38	SR
25	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	10	72	0.30	SR
26	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	30	72	0.25	SR
27	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat	1	386	0.20	SR
28	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	5	386	0.18	SR
29	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	10	386	0.15	SR
30	Burke and Lobell 2017 ^[16]	2014	Kenya	maize	Skysat (c)	30	386	0.10	SR
31	Lobell et al 2019 ^[54]	2016	Uganda	maize	S2	10	252	0.55	FP
32	Lobell et al 2019 ^[54]	2016	Uganda	maize	S2	10	252	0.25	SR
33	Lobell et al 2019 ^[54]	2016	Uganda	maize	S2	10	252	0.25	CC

Table S2: Performance of efforts to predict economic wellbeing with imagery and ML. "Samples" reports the number of training examples when available. "Geo" reports the geographic level at which models were evaluated, with (e.g.) "geo2" equal to county or district.

	Year	Target	Metric	Result	Samples	Geo	Location
[30]	2020	Asset Wealth Index	r2	0.83	19,669 clusters	geo2	23 African countries
[122]	2016	Asset Wealth Index	r2	0.78		cluster	Nigeria
[29]	2017	Wealth Index	r2	0.76		voroni polygons	Bangladesh
[27]	2016	Asset Wealth Index	r2	0.75	1,411 clusters	cluster	Rwanda
[123]	2017	Asset Wealth Index	r2	0.66		cluster	Africa
[124]	2017	Income <= National 10th percentile	r2	0.61	1,291 villages	geo4	Sri Lanka
[57]	2020	Consumption	r2	0.54	320 clusters	cluster	Uganda
[125]	2017	Multi-Dimensional Poverty Index	cor	0.91	552 communes	geo4	Senegal
[126]	2016	Multi-Dimensional Poverty Index	cor	0.88	416 sectors	geo3	Rwanda
[127]	2019	Asset Wealth Index	cor	0.57	4,839 clusters	cluster	Africa
[128]	2017	If below Comparative Wealth Index poverty line or not	accuracy	0.83	eval on 636,448 hholds	cluster	36 countries
[129]	2017	Reported living condition good/neutral/bad	accuracy	0.83		cluster	Botswana, Kenya, Zimbabwe
[130]	2019	If county is "non-poverty" or not	accuracy	0.82	192 counties	geo2	China
[131]	2016	Above or below poverty line	accuracy	0.72	643 clusters	cluster	Uganda
[132]	2018	Bottom 40%, middle 40%, top 20% classification	accuracy	0.62	330 hholds	hhold	Kenya
[133]	2016	Welfare Index quintiles	accuracy	0.36	14,000 clusters	cluster	India

Table S3: Performance of efforts to predict the location of informal settlements (slums) with imagery and ML

	Year	Target	Metric	Result	Samples	Location
I ³⁴	2018	MajiData spatial extent of slums	recall	0.95		Kenya
I ³⁵	2011	Slum delineations	recall	0.85	70km ² classified	Morocco
I ³⁶	2008	Manual slum delineations	recall	0.68		Brazil
I ³⁷	2012	Slum index	r2	0.4	eval on 1,724 EAs	Ghana
I ³⁸	2018	Slum Delineations	IoU	0.9		India
I ³⁹	2017	Slum delineations	accuracy	0.98	12,398 100m cells	Colombia
I ⁴⁰	2019	Annotated ground truth points for slums	accuracy	0.97		India
I ⁴¹	2015	Accra Metropolitan Assembly slum dichotomy map	accuracy	0.94	3,000 samples	Ghana
I ⁴²	2019	Point locations of squatter settlements	accuracy	0.94		Jamaica
I ⁴³	2019	Slum delineations	accuracy	0.94		India
I ⁴⁴	2018	Slum delineations	accuracy	0.94		Indonesia
I ⁴⁵	2012	?	accuracy	0.92	12,000 points	Afghanistan
I ⁴⁶	2020	Slum delineations	accuracy	0.91		Saudi Arabia
I ³⁶	2017	Slum delineations	accuracy	0.9	3,000 samples	Tanzania
I ³⁵	2017	Slum delineations	accuracy	0.9	3,060 samples	Tanzania
I ⁴⁷	2018	Regular/irregular settlement 1km grid	accuracy	0.9		Afghanistan
I ⁴⁸	2017	Slum areas from visual image interpretation	accuracy	0.89	1,159,662 pixels	India
I ⁴⁹	2016	Municipality provided slum location data	accuracy	0.88	80 points	India
I ⁵⁰	2013	Rule based classification, using 7 ground truth locations	accuracy	0.87	eval on 7 points	India
I ⁵¹	2017	Manual classification map	accuracy	0.78	eval on region	South Africa
I ⁵²	2016	Slum delineations	accuracy	0.71	eval on 250 pts	India
I ⁵³	2013	Manual slum delineations	accuracy	0.56	eval on city	India