Price trend prediction using analyst report sentiment extracted via language models: Evidence from the IBEX

ARTICLE INFO

Keywords:
Natural Language Processing
Large Language Models
Stock Market Prediction
Analyst Recommendations

ABSTRACT

This study investigates the utility of language models in extracting sentiment from sell-side equity analyst reports and their potential as predictors of stock price trends using the IBEX index as a case study. RoBERTa, FinBERT, and GPT natural language processing models are employed to analyze a corpus of equity research reports from the 2016-2022 period. The results indicate that the extracted sentiment can serve as a valuable tool for forecasting stock price movements, avoiding potential analyst bias when assigning a target price. Clearly, 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 attracted significant attention owing to the potential benefits of successful strategies. Market unpredictability, coupled with the vast amount of information, large number of variables that can potentially affect stock values, and unanticipated noise, makes forecasting stock markets difficult (Henrique et al., 2019) and for investors to make informed decisions (Bernales et al., 2021).

Price evolution is essentially the confluence of buyers and sellers in which economic decisions are driven by expectations. The Efficient Market Hypothesis (EMH) (Fama, 1970) states that the available information is already reflected in the price and investors are rational. The semi-strong form of EMH specifically suggests that all publicly available information has already been incorporated into stock prices. Sell-side analyst research does not violate the semi-strong form of the EMH. Instead, it plays a role in the mechanism via which public information is incorporated into stock prices, supporting the notion of market efficiency.

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, demonstrated by the development of consistently profitable factors based on market anomalies (Azevedo and Hoegner, 2022). Investor expectations may differ from rational forecasts, as they are built on personal beliefs that are subject to social influence. Traditional models for predicting market behavior are based on either fundamental (company evolution) or technical analysis (price evolution) (McMillan, 2016).

Although EMH asserts that all relevant information has been incorporated into stock prices, analyst forecasts frequently challenge this premise. This is because analysts may have access to proprietary information or provide expert interpretations, which are not yet publicly available or fully comprehended by the market. Consequently, their forecasts can anticipate a broader market understanding and cause price movements that the EMH may not completely account for, indicating potential inefficiency. Essentially, analyst forecasts may affect the EMH through the speed of information assimilation into stock prices. We can distinguish three different effects before or after the date company releases its results:

- Company Earnings Management. Analyst expectations can significantly shape corporate behavior, particularly regarding 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 introduce short-term distortions in price formation, which may deviate from the EMH. These practices demonstrate that analyst-driven expectations are not mere reflections of public information but actively influence managerial decisions, potentially distorting the mechanism via which prices adjust to fundamentals.

- Information Dissemination Speed. The speed at which information is absorbed into stock prices often hinges on analysts' expertise and access. Analysts with domain-specific knowledge (Bradley et al., 2017) and proprietary company interactions (Brown et al., 2015) can interpret complex data more rapidly than the average market participant. This accelerated information dissemination reduces asymmetry and facilitates quicker price adjustments, aligning with the EMH in theory. However, temporary inefficiencies may arise when analyst 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) discuss, revisions based on evolving interpretations of data can cause price shifts which the market does not immediately predict. This challenges the EMH by demonstrating that even publicly available information can generate unforeseen price movements due to analysts'

ORCID(s):

reinterpretations. Moreover, the markets reliance on these revisions underscores a vulnerability to biases or inaccuracies in analyst forecasts, further complicating the efficiency narrative.

Analysts' influence on stock prices is highlighted by *Bloomberg* rankings of the analysts covering a certain stock.

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

Equity analysts typically analyze a company's financial statements, management teams, industry trends, and other factors to determine the target price for the company's stock. They issue recommendations such as "overweight", "neutral," or "underweight" (or any other analogous terminologies) based on their analysis and beliefs about future price evolution.

The effectiveness of equity analysts' recommendations is particularly important for portfolio management (Markowitz, 1991). Moreover, the distribution of stock recommendations is skewed towards the positive side (Morgan and Stocken, 2003). This may be because of the following factors:

- An investment bank issuing a negative recommendation on a company may see some influence on its other businesses with the said company.
- An analyst issuing a non-positive recommendation may not have access to the top 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 price evolution. Similarly, Bandyopadhyay et al. (1995) show that while stock price evolution is determined by profits in the long-term (60% explainability of price variation), short-term stock price evolution is determined by the sentiment of non-earnings variables (Nyakurukwa and Seetharam, 2023).

Recently, artificial intelligence (AI) and machine learning (ML) are being 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, such as positive, negative, or neutral. The business potential of conversational AI technologies in finance is yet to be discovered (Yue et al., 2023). Still, finance has involved AI since its early stages (Bickley et al., 2022).

NLP research shows that *transformer* models have achieved remarkable performance in language modeling, surpassing previous dictionary-based algorithms. The release of large language models (LLM), such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and generative pre-trained transformers (GPT) (Radford et al., 2019), represented a substantial leap in NLP. Sentiment extraction through LLM is far more insightful than previous sentiment analysis techniques such as bag-of-words. A seminal study by Vaswani et al. (2017) on the attention mechanism signaled the start of a race on the size of transfer learning models based on transformer architectures. Some model 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 META OPT-66B language model (Zhang et al., 2022).

Reviewing the literature on sentiment analysis in finance shows that the majority of studies employ lexicon-based approaches that concentrate on extensive word corpora. Wisniewski and Yekini (2015) review 1,262 annual reports (2006-2012) of 209 UK listed firms to extract sentiment using a lexicon-based approach. The authors count word frequency under three categories (praise, concreteness, and activity) to forecast future stock performance, and find weak correlation coefficients with stock price evolution (0.0708-0.0803). Next, Loughgran et al. (2011) study 50,155 annual reports during the 1994-2008 period for US based companies and find that lexicon-based sentiment classification may not extract sentiment properly when applied to the financial domain, thereby reducing their meaningfulness. Li (2006) investigates 10,000 US annual report filings using word frequency counting and reveals that reports with negative sentiment were followed by negative share price performance on a 12-month horizon.

Sul et al. (2017) suggest that sentiment analysis has a greater impact on small companies. 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.

Next, Olof (2019) processes equity analyst reports using bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec to extract sentiment features. The author finds that analyst upgrade or downgrade recommendations are the most informative labels (neutral labels do not convey much information). Subsequently, the author applies logistic regressions jointly with convolutional neural networks to classify the sentiment. Next, Schlaubitz

(2021) has trained a DistilBERT model using a financial phrase databank 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 expressed negative sentiments, even when companies reported lower earnings. Conversely, news articles were more balanced between classes. The author argues that text sentiments predictive ability of future stock returns is weak, as only 4 out of 15 companies under study showed a significant connection under a linear regression analysis.

Meanwhile, the financial domain LLM StonkBERT (Pasch and Ehnes, 2022) showed that the models predictive capability depends on the informational value of the underlying text data and that the news sample outperforms both the blog and annual reports samples. Firms predicted to be "good," "average," and "bad" by StonkBERT showed an average performance of 16.83%, 4.72%, and -3.17% in the 12-month post-prediction period, respectively. Experiments with PIXIU (Xie et al., 2023) showed that the quality of the instructions, rather than the model size, is critical for LLM performance. Models that are not fine-tuned using financial prediction datasets exhibit limited performance in stock movement prediction. van Binsbergen et al. (2021) find that FinBERT (Araci, 2019; Yang et al., 2020) did not perform well when applied to hedge fund reports. This is because 95% of the reports on short-selling included words as "fraud" and "misconduct," while FinBERT was trained on companies annual reports and conference calls with a language not mentioning those words. Fatouros et al. (2023) finds that compared with FinBERT, ChatGPT exhibits an approximately 35% enhanced performance in sentiment classification and a 36% higher correlation with market returns in the short term.

ChatGPT (Yue et al., 2023) is an LLM developed by OpenAI. ChatGPT's latest version is based on the GPT-4. It is trained on a vast corpus of text data, and can generate coherent and meaningful responses to diverse questions and prompts. ChatGPT-4 exhibits distinct behavior compared with the older GPT-3 model. For instance, ChatGPT-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 ChatGPT-4 model operates on a conversation-in and message-out basis.

This study seeks to advance research showing that sentiment embedded in analyst reports has some price evolution explanatory power (Suzuki et al., 2022). Among related studies, Rybinski (2020b) reveals that the forecasting 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 analysts 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 current traction in applying LLMs in different research fields and their relative scarcity in financial studies, this study aims to answer the following research questions:

- RQ1: Is sell-side analyst sentiment consistent with their recommendation?
- RQ2: To what end do the sentiments identified in financial reports by LLMs improve their capability to forecast the stock price trend?

To assess analysts' potential human biases and determine if the analyst means what they meant, we checked whether the explicit sell-side analyst report sentiment (explicitly stated in the report) was consistent with the implicit sentiment (extracted through language models).

We then checked the forecasting capacity of analysts and language models for different price movement ranges and forecasting horizons. Specifically, we compared the sentiment extracted from a) Fine-tuned RoBERTa language model, b) financial domain-specific FinBERT, and c) general GPT-4 plus the explicit analyst recommendation.

2. Data description

Our corpus comprised sell-side research reports in English extracted from the *Bloomberg* professional database for the 2016-2022 period for 10 IBEX companies listed in Table 2.

We selected sell-side rather than buy-side research reports because of 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 due to their wide dissemination. Buy-side reports are proprietary.
- 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.

Our selected 10-IBEX listed companies comply with: a) market/company size (Tables 1 and 2), as the literature shows that small company size is a factor in analyst recommendations impact due to lower liquidity and less coverage (Lo, 2017); b) sufficient sell-side coverage; and c) industry diversification. After the implementation of MiFID II

(Markets in Financial Instruments Directive II), which stated direct and explicit sell-side analyst payments, the number of reports available fell drastically: 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 index, as the companies under study account for more than 50% of the index market capitalization and cover different sectors despite the index banking concentration. This makes the analysis robust and non-data-dependent.

Table 1
European indexes size. 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

Our training data are unusual for LLM training because they include data from reliable sources rather than web scraped data, which is the usual case. The model entry is the text on the first page of the sell-side equity research report, which contains a detailed summary of the report. We performed preprocessing to eliminate upper cases and any mention of the label assigned by the sell-side analyst to avoid forward-looking bias. These labels are Sell/Underweight/Underperform (UP), Maintain/Neutral (N), and Buy/Overweight/Outperform (OP).

Table 2
IBEX companies studied

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	Telecom
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	Steel producers

After extracting the available research reports issued by Barclays (BAR), Deutsche Bank (DB), JP Morgan (JPM), and Credit Suisse (CS) provided by *Bloomberg*, and discarding those that gave rise to errors, we studied a dataset of 379 reports (Tables 3 and 4).

Table 3
Equity reports dataset issued by banks BAR, DB, JPM and CS

Ticker	# docs	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

The dataset/corpus was divided into the following two groups:

- Training: 303 reports (80%)
- Testing: 76 reports (20%)

A bag-of-words analysis of the reports examined the most frequently used terms in the equity reports of the studied companies, with the results listed in Table 5. Top four most-mentioned words had a frequency of 18-26%.

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

3. Methodology

The overall equity research sentiment is typically assessed for the report in its entirety. Specifically, positive, neutral, and negative sentiments are associated with upward (OP), neutral (N), and downward movements (UP), respectively. Studies analyzing the sentiments of analyst reports have not distinguished between current company situations and forward-looking perspectives, as they are closely intertwined. The LLM understanding of the current situation is critical for assessing sentiments (Yang et al., 2020), (Yang et al., 2023a).

The size of the corpus of reports from different sectors and different time frames considered smooth the potential noise of one-off effects such as macro data and political events.

As mentioned above, the LLM categorized each reports sentiment as positive, neutral, or negative, with the respective sentiments indicating an upward trend (OP), no change (N), and a downward trend (UP), respectively. The stock price movement was then analyzed from the day before each

Table 4
Equity reports dataset yearly distribution (# docs)

Ticker	2016	2017	2018	2019	2020	2021	2022	# docs
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
Ranking of most mentioned terms on a bag-of-words analysts' reports analysis 2016-2022

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

report's release across various futures time windows: 2, 8, 16, 30, and 60 days.

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

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

$$\frac{Price_t - Price_0}{Price_0} > X\% \rightarrow OP \tag{1}$$

$$\frac{Price_t - Price_0}{Price_0} < -X\% \rightarrow UP \tag{2}$$

$$[-X\%, +X\%] \rightarrow N \tag{3}$$

We framed the multiclass classification problem into three classes: Outperform (OP), Neutral (N), and Underperform (UP). We compared the price evolution during the selected horizon $t \in [2, 8, 16, 30, 60]$ days for each range of price movement $X \in [1\%, 3\%, 5\%, 7\%]$ to qualify the movement as OP (Equation 1), UP (Equation 2), or NEUTRAL (Equation 3) if the movement is above X, below -X, or inside +/-X, respectively.

The adopted F1 weighted metric can help evaluate the model's performance by balancing the F1 scores across the different classes based on their support (i.e., the number of instances of each class). We followed the four-step method depicted below, where the first two steps represent the practical implementation, and the other two provide the explanation and interpretation:

1. Calculating the F1 score for each class:
   - For each class (OP, N, and UP), compute Precision (Equation 4) and Recall (Equation 5) as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Table 6
Most mentioned words analysis out of 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

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

- Then, calculate the F1 score for each class as follows:

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

2. Weighting by support:
   - In an F1 weighted metric, each class's F1 score was weighted by the number of instances (support) in that class. Thus, classes with more samples will contribute more to the overall score.
   - The F1 weighted score was 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 count of samples in that class.
3. Interpreting the F1 weighted score:
   - The F1 weighted score provides an overall metric of the model's performance, considering both the performance on each class and distribution of samples across classes.
   - This is especially useful if the classes are imbalanced (i.e., more Neutral samples than Outperform or Underperform), as it prevents the metric from being overly influenced by a class with fewer samples.
4. Comparing Analyst versus Model:
   - By comparing the analyst classification (ground truth) with the models predictions using the F1 weighted score, we can get an indication of how well the model performs across all classes, accounting for any imbalances in the dataset.
   - This metric will highlight if the model struggles with certain classes or if it aligns closely with the analysts judgments across all classes.

In summary, the adopted F1 weighted metrics become effective because they provide a balanced view of the models performance across all categories (OP, N, and UP) while adjusting for class imbalances. This ensures a comprehensive assessment of the model compared with the analysts classifications.

Using the BERT models transfer learning capabilities (Devlin et al., 2018), we fine-tuned our models using a sample of sell-side reports as a training set to link sentiment with analyst recommendations. Using these fine-tuned models, we ingested text from the unseen analysts reports to test the sample for the models to return a sentiment: positive, stock goes up; negative, stock goes down; and neutral, stock remains flat. Price was never used to fine-tune the model. Therefore, look-forward bias was excluded (Sun et al., 2019). This methodology is similar to those of (Yang et al., 2023b) and (Kirtac and Germano, 2024).

3.1. RoBERTa

RoBERTa means the Robustly optimized BERT Pretraining approach (Liu et al., 2019), and has been pretrained on a massive corpus of text and code. It is based on Google's BERT model but makes several improvements, including dynamic masking at each epoch, longer sequences, larger batches, and Byte Pair Encoding (BPE). This helps increase the text handling efficiency of RoBERTa versus BERT.

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

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

3.2. FinBERT

FinBERT stands for Financial BERT (Araci, 2019; Yang et al., 2020) and is a pre-trained model with financial texts in English. It is a specialized language model designed for financial text analysis, and has been fine-tuned specifically to understand financial languages and concepts. Given that itis trained on a large corpus of financial documents, it is adept at tasks such as sentiment analysis, entity recognition, and other financial NLP tasks. FinBERT is particularly valuable for applications in the finance industry, including stock market sentiment analysis and financial news sentiment tracking, because it has state-of-the-art sentiment scoring on financial PhraseBank.

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

3.3. GPT

Since its release in November 2022, ChatGPT has revolutionized NLP. GPT-4, the most recent version at the time of writing this paper, is even 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 specific stocks or sectors, and answer questions related to investment strategies. ChatGPT can generate responses that include information on a company's financials, recent news, and market trends.

The GPT-4 training set is publicly available web-based information. Meanwhile, analyst reports are client-only data that are not available on the web, which prevents forward-looking bias.

We performed inferences on ChatGPT-4 to categorize each report. We uploaded the analyst reports through the ChatGPT-4 API and asked ChatGPT-4 to assign the above-mentioned labels: OP, N, or UP.

4. Results

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

Table 7
Distribution of analyst and LLM extracted sentiment recommendation (%)

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

Once the report text sentiments are extracted, we perform the following:

- Checked the consistency of the label assigned by the author versus that extracted from the report text (Table 8). F1 weighted metric was used to check the coincidence between the sentiment tags assigned by the analyst and each of the 3 different models. Market capitalization, industry, volatility and analyst background were considered, being this last factor the most influential in the consistency of text sentiment and explicit recommendation.
- Considering different price range movements for different timeframes since the report issue date, we checked the forecasting capacity of the report recommendation versus the extracted label through the LLMs applied (Table 13).

Analysts are not particularly good at price movement predictions for any horizon, as per the precision (Equation 4) shown in Tables 9, 10, 11, 12.

Table 8
Analyst vs LLM extracted recommendation (Equation 7)

	An. vs RoBERTa	An. vs FinBERT	An. vs GPT
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

In the short-term (2d-8d) for 1%, 3%, 5% and 7% price ranges (Table 13), RoBERTa demonstrates higher precision than other models as it seems to anticipate better the market reaction after report issuance.

Over time, the stock market may experience increasingly pronounced fluctuations (5%-7%). In these cases, our fine-tuned RoBERTa model performs better because it tends to make estimates able to detect upward or downward trends rather than adopting a neutral position (Tables 11 and 12). GPT-4 shows good performance in the medium term (30d-60d) with flat markets (1%-3%), likely due to its generalist training and reliance on neutral predictions. Within the 3% price range (Table 10), RoBERTa equals FinBERT in precision, particularly at the 16d horizon, where its predictions outperform analysts and other language models. The improvement in precision for RoBERTa and FinBERT at this range suggests that these models are better equipped to capture moderate price movements. Analysts exhibit comparable performance to GPT-4 in the 60d horizon under flat markets (1%-3%), highlighting the limitations of general models without fine-tuning for financial contexts. The selected time horizons (2d, 8d, 16d, 30d, and 60d) reflect different trading and forecasting scenarios relevant to market participants. Short horizons (2d, 8d) capture immediate market reactions to new information, while medium and longer horizons (16d, 30d, 60d) assess the sustainability of predictive models over broader timeframes. These horizons are designed to reflect practical investment decision timelines and market dynamics, ensuring relevance across various potential trading strategies.

Table 9
Stock price vs prediction (Equation. 4). 1% price range

	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
Stock price vs prediction (Equation. 4). 3% price range

	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
Stock price vs prediction (Equation. 4). 5% price range

	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
Stock price vs prediction (Equation. 4). 7% price range

	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
Best language model according to price evolution and forecasting horizon
(AN: Analyst, 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

5. Conclusions

Language models provide a new approach to behavioral finance by enabling advanced sentiment analysis. This study analyzes the analyst reports of IBEX companies during the 2016-22 period, studying the consistency of the recommendation in an analyst report with its sentiment. We find that companies with higher capitalization and fewer reports show the greatest alignment between sentiment and recommendation, regardless of industry and market beta. UK-based analysts drive implicit language sentiment versus explicit recommendation consistency. Considering different LLMs, we find the following across time frames:

- Short term: The RoBERTa model leads in terms of forecasting but with little margin over the rest.
- Medium term: GPT-4 tends to outperform the remaining models if markets tend to stay flat; during bigger price movements, the RoBERTa model is more precise.

This study demonstrates the superiority of smaller but finance-related pretrained LLMs like RoBERTa versus bigger general models like GPT-4 which are not fine-tuned. This is similar to the findings of Xie et al. (2023). LLMs predictive power with no additional data is low, which is consistent with the results of Rybinski (2020a). LLMs pretrained in the finance domain, such as FinBERT, but with no finetuning do not perform as well. Larger general models such as GPT-4 do not discriminate, as they tend to be on the safe side with neutral tags that match with flattish markets (i.e., 1%-3%).

GPT-4 is not specifically trained on analyst reports but has more extensive generalist training, which is consistent with Gururangan et al. (2020). This may be one of the reasons why RoBERTa is better at identifying trends on longer forecasting horizons.

Clearly, while language models can provide valuable insights and analyses (Lopez-lira and Tang, 2023), (Pelster and Val, 2024), they cannot replace the experience and knowledge of human investors. GPT-4 is suitable for summarizing information overload for retail investors (Kim et al., 2023a). Meanwhile, by combining the insights generated by sentiment analysis with human expertise, investors can make more informed decisions, and reduce the potential for errors or biases (Cao et al., 2021). In any case, NLP sentiment price evolution accuracy hardly exceeds 50%, indicating low prediction capabilities. These results are similar to those obtained in Rybinski (2020b).

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

Our research contributes to the literature in the following ways:

- It shows 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 forecast. These results are consistent with Li et al. (2023).
- Crucially, our work is novel in its use of analyst reports as its dataset (i.e., IBEX listed company research reports from the 2016-2022 period). Similar studies have only covered the Korean stock market (Kim et al., 2023b; Cho et al., 2021).

- We also use a novel methodology, wherein 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 forecasting feature importance (Sidogi et al., 2022).

.

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

- Secrecy and limited access to the training corpus of the LLMs. Controversy continues regarding 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 the development of LLMs in the finance domain jointly with datasets like The Pile Biderman et al. (2022), C4, and Wikipedia.

- Noise and instability. LLM strategies, especially when utilized by major institutional investors or hedge funds, may influence on the wider financial markets. Substantial capital allocation into specific securities or sectors because of LLM strategies can sway prices and market sentiment, potentially inciting herd behavior or unforeseen repercussions that can engender systemic risks.

Overall, this study provides novel evidence on the predictive capabilities of sentiment obtained from state-of-the-art NLP models, such as FinBERT, RoBERTa, and GPT-4, using sell-side equity analyst reports. Focusing on companies in the IBEX 35 index, this study demonstrates the potential of such NLP models in enhancing the accuracy of stock price forecasts, even in markets with relatively high efficiency. The results show that sentiment extracted from analysts reports explains a wide portion of both short and medium-term stock price movements. Therefore, it is valuable for the literature on behavioral finance and ML applications in equity markets. This study also compares explicit recommendations with the implicit sentiments extracted via language models 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 surrounding 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 explicit analyst recommendations. Finally, our insights can be helpful for policymakers, 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 can consider more general applications of this approach to other markets, particularly to emerging markets where information asymmetry and imperfections are higher. Meanwhile, extending sentiment analysis by incorporating new alternative data sources, such as social media or macroeconomic indicators, may improve this predictive power. Overall, the main value of this study, which shows the efficiency of using NLP models in stock analyst reports, is that it closes the gap between traditional financial analysis and advanced AI technologies. These findings contribute not only to the academic literature, but also to actionable knowledge relevant to market participants striving for better forecasting accuracy and optimized investment strategies.

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

References

Abarbanell, J. S. and Bushee, B. J. (1997). Fundamental Analysis, Future Earnings, and Stock Prices. *Journal of Accounting Research*, 35(1):1.

Amin, M. M., Cambria, E., and Schuller, B. W. (2023). Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT. *IEEE Intelligent Systems*, 38(2):15–23.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv*.

Azevedo, V. and Hoegner, C. (2022). *Enhancing stock market anomalies with machine learning*. Number 0123456789. Springer US.

Bandyopadhyay, S. P., Brown, L. D., and Richardson, G. D. (1995). Analysts' use of earnings forecasts in predicting stock returns: Forecast horizon effects. *International Journal of Forecasting*, 11(3):429–445.

Beaver, W., Cornell, B., Landsman, W. R., and Stubben, S. R. (2008). The Impact of Analysts' Forecast Errors and Forecast Revisions on Stock Prices. *Journal of Business Finance & Accounting*, 35(5-6):709–740.

Bernales, A., Valenzuela, M., and Zer, I. (2021). Effects of Information Overload on Financial Market Returns: How Much Is Too Much? *SSRN Electronic Journal*, 2500(1372).

Bickley, S. J., Chan, H. F., and Torgler, B. (2022). Artificial intelligence in the field of economics. *Scientometrics*, 127(4):2055–2084.

Biderman, S., Bicheno, K., and Gao, L. (2022). Datasheet for the Pile. pages 1–22.

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., and 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*, pages 95–136, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bradley, D., Gokkaya, S., Liu, X., and Xie, F. (2017). Are all analysts created equal? Industry expertise and monitoring effectiveness of financial analysts. *Journal of Accounting and Economics*, 63(2-3):179–206.

Brown, E. (1999). A NonRandom Walk Down Wall Street. *Journal of Economic Surveys*, 13(4):477–478.

Brown, L. D., Call, A. C., Clement, M. B., and Sharp, N. Y. (2015). Inside the Black Box of SellSide Financial Analysts. *Journal of Accounting Research*, 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., and Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 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., and Yang, B. (2021). From Man vs. Machine to Man + Machine: The Art and Ai of Stock Analyses. *SSRN Electronic Journal*.

CFA Institute (2019). MIFID II: ONE YEAR ON Assessing the Market for Investment Research. Technical report.

Cho, P., Park, J. H., and 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., and Cummins, M. (2015). Analyst recommendations and volatility in a rising, falling, and crisis equity market. *Finance Research Letters*, 15:187–194.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Duz Tan, S. and Tas, O. (2021). Social Media Sentiment in International Stock Returns and Trading Activity. *Journal of Behavioral Finance*, 22(2):221–234.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2):383.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., and Kyriazis, D. (2023). Transforming Sentiment Analysis in the Financial Domain with ChatGPT. *Machine Learning with Applications*, 14(October):100508.

Frijns, B. and Garel, A. (2021). The effect of cultural distance between an analyst and a CEO on analysts earnings forecast performance. *Economics Letters*, 205:109957.

Gao, K., Lin, W., Yang, L., and Chan, K. C. (2020). The impact of analyst coverage and stock price synchronicity: Evidence from brokerage mergers and closures. *Finance Research Letters*, 33(May 2019):101190.

Givoly, D., Hayn, C., and Yoder, T. R. (2011). What do Analysts Really Predict? Inferences from Earnings Restatements and Managed Earnings. *SSRN Electronic Journal*, (November).

Graaf, J. (2023). Can Sell-side Analysts Compete Using Public Information? Analysts as Frame-makers Revisited. *European Accounting Review*, 32(1):141–167.

Grant, A., Jarnecic, E., and Su, M. (2015). Asymmetric effects of sell-side analyst optimism and broker market share by clientele. *Journal of Financial Markets*, 24:49–65.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Dont Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Stroudsburg, PA, USA. Association for Computational Linguistics.

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124:226–251.

Hou, Y. and Hu, C. (2023). Understanding the role of aggregate analyst attention in resolving stock market uncertainty. *Finance Research Letters*, 57(May):104183.

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., and Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, 61(May):215–234.

Karmaziene, E. (2023). The greater the volume, the greater the analyst. *Finance Research Letters*, 51(August 2022):103377.

Kim, A. G., Muhn, M., Nikolaev, V. V., and Kim, A. G. (2023a). Bloated Disclosures: Can ChatGPT Help Investors Process Information? *SSRN Electronic Journal*, (23).

Kim, S., Lee, W.-J., Park, S., and Sunwoo, H.-Y. (2022). Busy analysts in uncertain times. *Finance Research Letters*, 46(PB):102488.

Kim, S. S. S., Kim, S. S. S., Kim, Y., Park, J., Kim, S. S. S., Kim, M., Sung, C. H., Hong, J., and Lee, Y. (2023b). LLMs Analyzing the Analysts: Do BERT and GPT Extract More Value from Financial Analyst Reports? In *4th ACM International Conference on AI in Finance*, pages 383–391, New York, NY, USA. ACM.

Kirtac, K. and Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62(PB):105227.

Ko, H. and Lee, J. (2023). Can ChatGPT Improve Investment Decision? From a portfolio manager perspectice.

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., and Villanova, A. (2022). The BigScience ROOTS Corpus: A 1 . 6TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems*, number 35, pages 31809–31826.

Le, T. D. and Trinh, T. (2022). Distracted analysts and earnings management. *Finance Research Letters*, 49(June):103038.

Li, E. X., Tu, Z., and Zhou, D. (2023). The Promise and Peril of Generative AI: Evidence from ChatGPT as Sell-Side Analysts. *SSRN Electronic Journal*, (June):1–31.

Li, F. (2006). Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? *SSRN Electronic Journal*.

Li, T. (2022). Analyst's stock views and revision actions. *Finance Research Letters*, 44(April 2021):102033.

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., and Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and 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. *Accounting and Finance Research*, 6(4):181.

Lopez-lira, A. and Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements ? Return Predictability and Large Language Models. *SSRN Electronic Journa*.

Loughgran, T., McDonald, B., and Loughran; T.; & McDonald; B (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2020). Textual Analysis in Finance. *Annual Review of Financial Economics*, 12(1):357–375.

Markowitz, H. (1991). Foundations of Portfolio Theory. *Journal of Finance*, 46(2):469–477.

McMillan, D. G. (2016). Which Variables Predict and Forecast Stock Market Returns? *SSRN Electronic Journal*, 44(June):0–28.

Morgan, J. and Stocken, P. C. (2003). An Analysis of Stock Recommendations. *The RAND Journal of Economics*, 34(1):183.

Nyakurukwa, K. and Seetharam, Y. (2023). Can textual sentiment partially explain differences in the prices of dual-listed stocks? *Finance Research Letters*, 58(PC):104529.

Olof, L. (2019). *Sentiment Analysis of Equity Analyst Research Reports using Convolutional Neural Networks*. PhD thesis, Uppsala University.

Pasch, S. and Ehnes, D. (2022). StonkBERT: Can Language Models Predict Medium-Run Stock Price Movements? *arXiv*, pages 1–16.

Pelster, M. and Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, 59(November 2023):104786.

Pursiainen, V. (2018). Cultural Biases in Equity Analysis. *SSRN Electronic Journal*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. v. d., 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., Sottiaux, 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., and Irving, G. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv*.

Reid, B. H., Jessop, S., and 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 Commision.

Rybinski, K. (2020a). Should asset managers pay for economic research? A machine learning evaluation. *The Journal of Finance and Data Science*, 6:31–48.

Rybinski, K. (2020b). The forecasting power of the multi-language narrative of sell-side research: A machine learning evaluation. *Finance Research Letters*, 34(August 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*. PhD thesis, Zurich University of Applied Sciences.

Sidogi, T., Mongwe, W. T., Mbuvha, R., and Marwala, T. (2022). Fusing Sell-Side Analyst Bidirectional Forecasts Using Machine Learning. *IEEE Access*, 10(July):76966–76974.

Sul, H. K., Dennis, A. R., and Yuan, L. I. (2017). Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns. *Decision Sciences*, 48(3):454–488.

Sun, C., Qiu, X., Xu, Y., and 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, pages 194–206.

Suzuki, M., Sakaji, H., Izumi, K., and Ishikawa, Y. (2022). Forecasting Stock Price Trends by Analyzing Economic Reports With Analyst Profiles. *Frontiers in Artificial Intelligence*, 5(June).

Thas Thaker, H. M., Mohamad, A., Mustaffa Kamil, N. K., and Duasa, J. (2018). How useful are the equity analysts' report? Evidence from Malaysia. *Reports on Economics and Finance*, 4(4):221–246.

van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2021). Textual analysis of short-seller research reports, stock prices and real investment. *SSRN Electronic Journal*, d.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *IEEE Industry Applications Magazine*, 8(1):8–15.

Wisniewski, T. P. and Yekini, L. S. (2015). Stock market returns and the content of annual report narratives. *Accounting Forum*, 39(4):281–294.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). BloombergGPT: A Large Language Model for Finance.

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., and Huang, J. (2023). PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance.

Yang, G., He, P., and Liu, X. (2023a). FundRecLLM: Fund Recommendation Based on Financial News and Research Analyst Report. *Frontiers in Artificial Intelligence and Applications*, 378:515–523.

Yang, H., Liu, X.-Y., and Dan Wang, C. (2023b). FinGPT: Open-Source Financial Large Language Models. *SSRN Electronic Journal*.

Yang, Y., UY, M. C. S., and Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications.

Yue, T., Au, D., Au, C. C., and Iu, K. Y. (2023). Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology. *SSRN Electronic Journal*, pages 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., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv*.