Enhanced Financial Sentiment Analysis and Trading Strategy Development Using Large Language Models

Kemal Kirtac and Guido Germano
Department of Computer Science, University College London,
66–72 Gower Street, London WC1E 6EA, United Kingdom
kemal.kirtac.21@ucl.ac.uk, g.germano@ucl.ac.uk

Abstract

This study proposes a novel methodology for enhanced financial sentiment analysis and trading strategy development using large language models (LLMs) such as OPT, BERT, FinBERT, LLAMA 3 and RoBERTa. Utilizing a dataset of 965,375 U.S. financial news articles from 2010 to 2023, our research demonstrates that the GPT-3-based OPT model significantly outperforms other models, achieving a prediction accuracy of 74.4% for stock market returns. Our findings reveal that the advanced capabilities of LLMs, particularly OPT, surpass traditional sentiment analysis methods, such as the Loughran-McDonald dictionary model, in predicting and explaining stock returns. For instance, a self-financing strategy based on OPT scores achieves a 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. This study highlights the superior performance of LLMs in financial sentiment analysis, encouraging further research into integrating artificial intelligence and LLMs in financial markets.

1 Introduction

The integration of text mining into financial analysis represents a significant shift in how researchers approach market predictions. Utilizing a diverse array of text data—from financial news to social media posts—this new wave of research aims to extract insights that traditional data sources might overlook (Loughran and Mcdonald, 2011; Malo et al., 2014; Loughran and McDonald, 2022). Despite the complexity and the lack of structured information within text data, advancements in LLMs such as BERT (Devlin et al., 2019), OPT (Zhang et al., 2022), LLAMA 3 (Touvron et al., 2023) and RoBERTa (Liu et al., 2019), have opened new avenues for in-depth analysis and understanding of financial markets. These models have shown a notable ability to outperform traditional sentiment analysis methods, demonstrating the untapped potential of text data in predicting market trends and stock returns (Jegadeesh and Wu, 2013; Baker et al., 2016; Manela and Moreira, 2017).

Our research harnesses the power of LLMs to create refined representations of news text, aiming to bridge the gap in sentiment analysis at the individual stock level—an aspect often overlooked by macro- or market-level sentiment indicators (Baker and Wurgler, 2006; Lemmon and Ni, 2014; Shapiro et al., 2022). By employing a two-step analytical process that first converts text into numerical data and then models economic patterns, we explore the predictive accuracy of these models against traditional dictionary-based methods (Tetlock, 2007; Devlin et al., 2019). We contribute to the ongoing dialogue on the role of text analysis in finance, advocating for a broader adoption of LLMs in economic forecasting and investment strategy development (Acemoglu et al., 2022; Hoberg and Phillips, 2016; Garcia, 2013; Ke et al., 2020; Tetlock, 2007; Campbell et al., 2014; Baker et al., 2016; Calomiris and Mamaysky, 2019; Ashtiani and Raahemi, 2023a; Kirtac and Germano, 2024).

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 trading on major U.S. stock exchanges. It includes details like stock prices, trading volumes, and market capitalization. We use this data to analyse the link between stock market returns and sentiment scores derived from LLMs.

Our analysis includes companies from the Amer-

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 1–10

ican 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 (a cosine similarity score of 0.8 or more) to an older article published within the past 20 days, we exclude it. This approach helps us 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 with 6,214 unique companies. After applying our filters, we were left with 965,375 articles. Our sample dataset is summarised in Table 1.

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

Table 1: Summary statistics of our U.S. news articles sample, showing the count of total news, news for a single stock, and unique news after filtering for redundancy. This data set forms the basis for our sentiment analysis and subsequent stock return prediction model.

Table 2 presents descriptive statistics of our dataset. We find that the daily mean return is 0.37%, with a standard deviation of 0.18%. The sentiment scores derived from the OPT, BERT, FinBERT, LLAMA 3 and RoBERTa LLMs 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.

2.2 Methods

This study begins with the fine-tuning of pre-trained language models, specifically OPT, BERT, LLAMA 3, and RoBERTa, 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 forecasting stock returns through textual analysis. This necessity enforces the adaptation phase, where the models are recalibrated post their original training on extensive data, preparing them for specific analytical functions (Radford et al., 2018).

Besides OPT, BERT, LLAMA 3 and RoBERTa, our analysis incorporates FinBERT, a variant of BERT pre-trained specifically for financial texts, and the Loughran and McDonald dictionary. FinBERT and the Loughran and McDonald dictionary do not necessitate the fine-tuning process because 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).

We present a unique approach that integrates fine-tuning pre-trained LLMs with financial text data. This section outlines our process of adapting LLMs for the financial domain, including the steps of fine-tuning and the specific features used in our sentiment analysis. Our methodology involves the systematic adaptation of models such as OPT, BERT, FinBERT, LLAMA 3 and RoBERTa, focusing on domain-specific nuances by fine-tuning them on a comprehensive dataset of financial news. This process not only improves the models' understanding of financial sentiment but also enhances their predictive accuracy regarding stock market movements. By leveraging the advanced capabilities of LLMs and tailoring them specifically for financial text, our approach presents a robust framework for sentiment-based financial forecasting.

The use of LLMs such as OPT, BERT, FinBERT, LLAMA 3 and RoBERTa in financial sentiment analysis offers distinct advantages over traditional methods, particularly in handling the complexity and unstructured nature of financial text data. Traditional techniques, such as the Loughran-McDonald dictionary, rely on predefined word lists that may not capture the nuanced and evolving language used in financial news. In contrast, LLMs leverage deep learning to understand context, sentiment, and subtle linguistic cues within text, leading to more accurate sentiment predictions. Our study demonstrates that LLMs, through their ability to fine-tune on domain-specific data, significantly outperform traditional methods in predicting stock returns. The fine-tuning process involves training these models

Variable	Mean	StdDev	Minimum	Median	Maximum	N
Daily return (%)	0.37	0.18	-64.97	-0.02	237.11	965,375
OPT score	0.53	0.24	0	0.5	1	965,375
BERT score	0.48	0.25	0	0.5	1	965,375
FinBERT score	0.44	0.23	0	0.5	1	965,375
LLAMA 3 score	0.45	0.29	0	0.5	1	965,375
RoBERTa score	0.51	0.24	0	0.5	1	965,375
LM dictionary score	0.68	0.32	0	0.5	1	965,375

Table 2: Descriptive statistics for daily stock returns and sentiment scores derived from the OPT, BERT, FinBERT, LLAMA 3 and RoBERTa LLMs, alongside the Loughran-McDonald dictionary. It includes the mean, standard deviation, minimum, median, maximum values, and the total count of observations for each variable.

on a vast corpus of financial news, allowing them to learn and adapt to the specific language and sentiment indicators pertinent to financial markets. Additionally, the use of LLMs facilitates the development of a robust investment strategy, as evidenced by the superior performance metrics achieved in our experiments. Future research could focus on optimizing these models further, exploring efficient training algorithms and model compression techniques to enhance their practicality and application in real-time trading scenarios.

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, BERT, LLAMA 3 and RoBERTa 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. Acknowledging the common practice in economics and finance of studying events that span multiple days, we establish sentiment labels using three-day returns (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 nuanced 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.

Our analysis focused on the ability of OPT, BERT, LLAMA 3, RoBERTa, 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.

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

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

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, respectively.

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 with respect to stock returns of different LLMs, including OPT, BERT, FinBERT, LLAMA 3 and RoBERTa, plus the Loughran and McDonald dictionary.

Our choice of the linear regression model corresponds to a standard panel regression approach where article features $x_{i,n}$ are directly translated into the expected return $E(r_{i,n+1})$ of the corresponding stock for the next period. 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, LLAMA 3, RoBERTa and 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 with 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.

2.2.1 Training and Inference Process

The training and inference process involves several key steps as presented in Algorithm 1. Initially, we collect financial news articles and the corresponding stock return data. These articles are preprocessed to remove irrelevant and similiar information and ensure consistency. Following this, we fine-tune LLMs using the training news dataset. After fine-tuning, the fine-tuned LLMs are utilized to calculate sentiment scores for the news articles in the test dataset. Based on these sentiment scores, we implement a portfolio investment strategy for the test period. This strategy includes creating three distinct portfolios: a long portfolio consisting of stocks with the top 20 percentile positive sentiment scores, a short portfolio with stocks having the top 20 percentile negative sentiment scores, and a self-financing long-short portfolio that incorporates both the top 20 percentile negative and positive scores. Additionally, we include benchmark comparisons with value-weighted and equal-weighted market portfolios that do not consider sentiment scores. The performance of these portfolios is then evaluated using key financial metrics, including the Sharpe ratio, mean daily returns, standard deviation of daily returns and maximum drawdown.

We update the portfolios 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.

2.2.2 Computational Cost and Comparative Analysis

Computational Cost The training and inference processes for fine-tuning LLMs are computationally intensive. Specifically, the fine-tuning phase involves extensive preprocessing of financial news articles, training on large datasets and continuous updating of models based on new data. In our experiments, we utilized high-performance computing resources, including GPUs and TPUs, to manage these tasks efficiently. The training time varied significantly depending on the model size and the volume of data processed. For instance,

Algorithm 1 Training and Inference Process

Require: Pre-trained language model (PLM), financial news articles $\{A_i\}$, three-day aggregated stock returns $\{R_i\}$

Ensure: Updated sentiment portfolios

Training Phase:
for each article A_i in the training set **do**
  Associate A_i with its three-day aggregated return R_i
  Fine-tune the PLM on the paired data $\{A_i, R_i\}$
end for
Save the fine-tuned model as FTM
Forming Sentiment Portfolios:
for each stock i **do**
  Use FTM to predict sentiment score S_i from recent news articles
  Rank all stocks by their sentiment scores S_i
  Form top 20% highest sentiment portfolio P_{high}
  Form bottom 20% lowest sentiment portfolio P_{low}
end for
Updating Portfolios:
for each new day **do**
  for each stock i **do**
    Update sentiment score S_i with new articles using FTM
    Re-rank all stocks by updated sentiment scores S_i
    Update P_{high} and P_{low} with the new rankings
  end for
end for

fine-tuning BERT and OPT models required approximately 48 hours on a cluster of 4 NVIDIA V100 GPUs for our dataset of 965,375 articles. The computational cost also encompasses storage and memory requirements, which were substantial given the need to handle large volumes of text data and model parameters. Despite these costs, the enhanced performance of dialogue-level augmentation techniques justifies the computational investment. Future work could explore more efficient training algorithms and model compression techniques to mitigate these costs while retaining performance gains.

Comparative Analysis with Existing Techniques We included a variety of existing individual utterance-level augmentation methods. They include back-translation, synonym replacement and noise injection, which are commonly used in text augmentation. Our comparative analysis highlights several key findings. Firstly, dialogue-level augmentation techniques consistently outperformed individual utterance-level methods across multiple evaluation metrics. Specifically, our dialogue-level approach yielded higher sentiment prediction accuracy and improved stock return forecasting capabilities. For example, the OPT model with dialogue-level augmentation achieved an accuracy of 74.4%, compared to 68.9% with utterance-level back-translation. Additionally, our approach demonstrated better robustness and generalization, particularly in handling nuanced financial texts. This superiority is attributed to the ability of dialogue-level augmentation to capture contextual dependencies and sentiment flows across multiple utterances, which is often lost in utterance-level methods. To substantiate these findings, we refer to recent studies by Ashtiani and Raahemi (2023b) and Ke et al. (2020) which also emphasize the limitations of traditional text augmentation techniques in complex domains like financial forecasting. These studies provide a benchmark for our results, reinforcing the effectiveness of the methods we propose. In conclusion, the dialogue-level augmentation not only enhances model performance but also aligns more closely with real-world applications where understanding the flow of information and sentiment over a series of interactions is crucial.

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 six models: OPT, BERT, FinBERT, LLAMA 3, RoBERTa and the Loughran-McDonald dictionary. Their performance in sentiment analysis is shown in Table 3.

The results show that the OPT model was the most accurate, followed closely by BERT and

Metric	OPT	BERT	FinBERT	LLAMA 3	RoBERTa	Loughran-McDonald
Accuracy	0.744	0.725	0.722	0.632	0.671	0.501
Precision	0.732	0.711	0.708	0.681	0.673	0.505
Recall	0.781	0.761	0.755	0.663	0.632	0.513
Specificity	0.711	0.693	0.685	0.642	0.701	0.522
F1 score	0.754	0.734	0.731	0.691	0.678	0.508

Table 3: Language model performance metrics. The table presents accuracy, precision, recall, specificity and the F1 score for each model.

FinBERT. The Loughran-McDonald dictionary, a traditional finance text analysis tool, had significantly lower accuracy. This indicates that language models like OPT, BERT, FinBert, LLAMA 3 and RoBERTa 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, outlined in 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, LLAMA 3, RoBERTa and the Loughran-McDonald dictionary.

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. LLAMA 3 and RoBERTa models also exhibit significant predictive capabilities. In contrast, the Loughran-McDonald dictionary model exhibits the least predictive power among the models examined .

In addressing the differential performance observed among OPT, BERT, FinBERT, RoBERTa and LLAMA 3, our analysis suggests that several factors contribute to this variance, notably model design, parameter scale and the specificity of training data. OPT's expanded parameter space, exceeding that of BERT, FinBERT, LLAMA 3 and RoBERTa, alongside its advanced training methodologies, likely underpins its superior forecasting accuracy in stock returns and portfolio management. The nuanced performance of FinBERT, despite its financial domain specialization, raises intriguing considerations. LLAMA 3 and RoBERTa, while demonstrating significant predictive capabilities, also highlight the importance of model architecture and training data diversity. Our exploration posits that the broader pre-training data diversity of BERT and RoBERTa, coupled with the potential for overfitting in highly specialized models such as FinBERT, might elucidate these unexpected outcomes. LLAMA 3's performance suggests that advancements in language model architectures continue to enhance predictive accuracy. 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^2 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.

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

Regression	1	2	3	4	5	6
OPT score	0.254***					
	(4.871)					
BERT score		0.129*				
		(2.334)				
FinBERT score			0.181***			
			(4.674)			
LLAMA 3 score				0.191**		
				(2.992)		
RoBERTa score					0.199***	
					(3.129)	
LM dictionary score						0.083
						(1.871)
Observations	965,375	965,375	965,375	965,375	965,375	965,375
R2	0.195	0.145	0.174	0.168	0.147	0.087
R2 adjusted	0.195	0.145	0.174	0.168	0.147	0.087
R2 within	0.017	0.009	0.016	0.011	0.008	0.002
R2 within adj.	0.017	0.009	0.016	0.011	0.008	0.002
AIC	62,345	97,473	67,345	77,842	73,934	135,783
BIC	115,655	114,746	109,272	121,232	123,393	123,382
RMSE	4.21	14.12	9.75	11,21	14,23	23.54
FE: date	X	X	X	X	X	X
FE: firm	X	X	X	X	X	X
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$						

Table 4: Regression of stock returns on LLM sentiment scores. The table presents the results of regressions done with Eq. (1), which includes firm and time-fixed effects represented by a_i and b_n respectively. The independent variable $x_{i,n}$ includes prediction scores from the language models. This analysis compares scores from OPT, BERT, FinBERT, LLAMA 3, RoBERTa 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.

value-weighted portfolios. These portfolios are developed using sentiment scores derived from different language models, including OPT, BERT, FinBERT, LLAMA 3, RoBERTa and the Loughran-McDonald dictionary. The investment strategies employed in our analysis are described as follows: each LLM is used 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 capitalization of each stock, while equal-weighted portfolios allocate investments equally to all stocks, regardless of market capitalization. We evaluate these strategies using key financial metrics, including the Sharpe ratio, mean daily returns, standard deviation of daily returns and maximum drawdown.

As indicated in Table 5, the long-short OPT strategy demonstrated the most robust risk-adjusted performance, as evidenced by its superior Sharpe ratio. On the other hand, the Loughran-McDonald dictionary model-based strategy (L-S LM dictionary) lagged 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.

	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 (%)	2.91	2.49	2.59	2.49	3.19	2.68	2.18	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			LLAMA 3			RoBERTa		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Sharpe ratio	0.87	0.66	1.23	1.37	1.11	1.44	1.04	1.18	1.51
MDR (%)	0.12	0.13	0.22	0.14	0.16	0.22	0.20	0.19	0.29
StdDev (%)	3.54	4.13	3.74	3.01	3.12	3.41	2.99	3.13	3.33
MDD (%)	-35.47	-45.39	-38.29	-29.13	-22.21	-22.85	-23.46	-28.44	-30.24
	EW			VW					
	Long	Short	L-S	Long	Short	L-S			
Sharpe ratio	1.25	1.05	1.40	1.28	1.08	1.45			
MDR (%)	0.18	0.15	0.33	0.19	0.16	0.35			
StdDev (%)	2.90	3.70	3.20	2.95	3.75	3.25			
MDD (%)	-31.13	-42.21	-32.87	-28.76	-38.95	-31.87			

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, LLAMA 3, RoBERTa and the Loughran-McDonald (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.

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 an application of OPT, BERT, FinBERT, LLAMA 3 and RoBERTa LLMs, we enhance the understanding of LLM capabilities 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 compare the performance of OPT, BERT, FinBERT, LLAMA 3 and RoBERTa scores to sentiment scores derived from conventional methods, such as the Loughran-McDonald dictionary model. Our analysis reveals that basic models exhibit 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 predictability. 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 behavior, information dissemination and price formation. Our results add insights to the dialogue on 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. Such evi-

dence 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 discussion about the role of AI in finance, particularly through our investigation into how well LLMs can 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.

5 Limitations

Despite the promising results of our study, several limitations should be acknowledged.

The fine-tuning of LLMs such as OPT, BERT, FinBERT, LLAMA 3 and RoBERTa requires substantial computational resources and time. This includes the need for high-performance computing resources such as GPUs and TPUs, and extensive preprocessing of financial news articles. The significant computational cost may limit the accessibility and scalability of these models for smaller organizations or individual researchers.

LLMs like FinBERT that are specialized for financial texts have a higher risk of overfitting due to their specificity. Overfitting can limit the model's ability to generalize to new, unseen data, especially in rapidly changing financial markets. Conversely, the broader pre-training data diversity of models like BERT and RoBERTa might introduce noise that affects their performance in specialized domains such as finance.

Our analysis is based on a dataset of 965,375 U.S. financial news articles spanning from 2010 to 2023. This dataset, although extensive, may not fully capture global financial trends and sentiments. Moreover, the quality and reliability of the financial news sources can vary, potentially impacting the accuracy of the sentiment analysis.

The evaluation metrics used in our study, such as accuracy, precision, recall and the Sharpe ratio, while robust, may not comprehensively capture all aspects of model performance in real-world trading scenarios. Market conditions, investor behavior and external economic factors are dynamic and can influence the effectiveness of sentiment-based trading strategies.

The integration of LLMs in financial markets raises important regulatory and ethical questions. The impact of algorithmic trading on market stability, the potential for market manipulation and the need for transparency and accountability in AI-driven decision-making are critical areas that require further exploration and regulatory oversight.

There is a need for ongoing research to address these limitations. Exploring more efficient training algorithms, model compression techniques and the integration of additional data sources can help mitigate computational costs and improve model performance. Studying the impact of LLMs in diverse and global financial contexts will enhance the generalizability and applicability of these models.

By acknowledging these limitations, we aim to provide a balanced perspective on the potential and challenges of using LLMs for financial sentiment analysis and trading strategy development. Future work should continue to refine these models and address the outlined challenges to fully realize their potential in financial markets.

The parameters of the trading algorithm should be justified by exploring alternatives. For instance, the lag or correlation time between news and returns has not been determined, and there are several other parameters in the algorithm that would benefit from an explanation or the testing of values above or below the ones used.

We tested only passive trading strategies; it would be beneficial to test active trading strategies as well. Furthermore, these strategies are based solely on sentiment, whereas sentiment-augmented strategies could further enhance the trading performance.

Funding

We acknowledge the EPSRC Doctoral Training Partnership EP/R513143/1.

References

Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1):S293–S340.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*.

M. N. Ashtiani and B. Raahemi. 2023a. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, 217:119509.

Mohammad N. Ashtiani and Bijan Raahemi. 2023b. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, 217:119509.

Malcolm Baker and Jeffrey Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4):1645–1680.

Scott R. Baker, Nicholas Bloom, and Steven J. Davis. 2016. Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4):1593–1636.

Charles W. Calomiris and Harry Mamaysky. 2019. How news and its context drive risk and returns around the world. *Journal of Financial Economics*, 133(2):299–336.

John L. Campbell, Hsinchun Chen, Dan S. Dhaliwal, Hsin-min Lu, and Logan B. Steele. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diego Garcia. 2013. Sentiment during recessions. *Journal of Finance*, 68(3):1267–1300.

Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

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

Hugging Face. 2023. Hugging Face's transformer models.

Narasimhan Jegadeesh and Di Wu. 2013. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Yanbo Ke, Bryan T. Kelly, and Dacheng Xiu. 2020. Predicting returns with text data. *Review of Financial Studies*, 33(11):5104–5144.

Kemal Kirtac and Guido Germano. 2024. Sentiment trading with large language models. *Finance Research Letters*, 62(B):105227.

Michael Lemmon and Sophie X. Ni. 2014. The impact of investor sentiment on the market's reaction to stock splits. *Review of Financial Studies*, 27(5):1367–1401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

TIM Loughran and BILL Mcdonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Tim Loughran and Bill McDonald. 2022. Master Loughran-MacDonald Word Dictionary.

A. C. MacKinlay. 1997. Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Asaf Manela and Alan Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI Blog.

Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. 2022. Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *arXiv:2205.01068*.