FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading

Giorgos Iacovides
Imperial College London
UK
giorgos.iacovides20@imperial.ac.uk

Thanos Konstantinidis
Imperial College London
UK
a.konstantinidis16@imperial.ac.uk

Mingxue Xu
Imperial College London
UK
m.xu21@imperial.ac.uk

Danilo Mandic
Imperial College London
UK
d.mandic@imperial.ac.uk

Abstract

Online sources of financial news have a profound influence on both market movements and trading decisions. Standard sentiment analysis employs a lexicon-based approach to aid financial decisions, but struggles with context sensitivity and word ordering. On the other hand, Large Language Models (LLMs) are powerful, but are not finance-specific and require significant computational resources. To this end, we introduce a finance specific LLM framework, based on the Llama 2 7B foundational model, in order to benefit from its generative nature and comprehensive language manipulation. Such a generator-discriminator scheme, referred to as FinLlama, both classifies sentiment valence and quantifies its strength, offering a nuanced insight into financial news. The FinLlama model is fine-tuned on supervised financial sentiment analysis data, to make it handle the complexities of financial lexicon and context, and is equipped with a neural network-based decision mechanism. The subsequent parameter-efficient fine-tuning optimises trainable parameters, thus minimising computational and memory requirements without sacrificing accuracy. Simulation results demonstrate the ability of FinLlama to increase market returns in portfolio management scenarios, yielding high-return and resilient portfolios, even during volatile periods.

Keywords

Large language models, sentiment analysis, algorithmic trading, portfolio construction, parameter-efficient fine-tuning

ACM Reference Format:
Giorgos Iacovides, Thanos Konstantinidis, Mingxue Xu, and Danilo Mandic. 2024. FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3677052.3698696

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698696

1 Introduction

The ever increasing prominence of algorithmic trading in quantitative finance has necessitated the need for reliable and actionable AI-aided domain knowledge from vast streams of data with multiple modalities. Of particular interest is generative AI, owing to its ability to distill insights from non-numerical sources such as news, earnings calls, financial reports, and other textual sources. In this context, sentiment analysis from text promises to bridge the gap between market movements caused by geopolitical and socioeconomic events, human actions, and quantitative trading.

The sentiment contained in on-line textual sources can drive market movements; such information harbours intrinsic advantages and gives a competitive edge to those equipped with the tools to harness it. Sentiment analysis rests upon the quantification of opinions present in unlabeled textual data, and aims to categorize whether the overall perspective is positive, negative, or neutral. When applied to large-scale information sources, this promises to enhance the understanding for the overall direction of macroscopic trends, a task which is both challenging and time-consuming for human analysts.

Despite conceptual benefits, the diverse, nuanced, and vast nature of financial text presents unique challenges when it comes to extracting sentiment in a manner that is both accurate and actionable. For example, the words 'bull' and 'bear' are neutral in the general vocabulary, but in financial markets, their respective connotations are strictly positive or negative [20]. This highlights the need for context-aware sentiment extraction, and underpins the complexities of employing natural language processing (NLP) in financial applications.

To address these issues, we consider the following fundamental questions:

- Can large language models (LLMs), which have already revolutionized manifold areas of NLP, be specifically tailored for sentiment analysis in the finance domain, particularly for enhancing algorithmic trading?
- Can this be achieved in a way which does not require vast computational resources, typically associated with NLP models, thus making the approach accessible to anyone equipped with standard computational resources?

Our proposed solution, termed *FinLlama*, is obtained by fine-tuning a pre-trained LLM (namely Llama 2 7B [22]) on specialised, labelled and publicly available financial news datasets. The ultimate goal of FinLlama is to enhance the performance of financial sentiment analysis, whilst leveraging on parameter-efficient fine-tuning (PEFT) and 8-bit quantization, through LoRA [9], to minimise resource requirements.

The main contributions of this work are:

- **Targeted fine-tuning**: Rather than utilising one general LLM for financial tasks, our approach capitalizes on the foundational pre-trained Llama 2 model, whereby fine-tuning is performed specifically for the purpose of sentiment classification through a SoftMax classification layer at its output.
- **Efficient resource utilization**: Our approach ensures that even standard computational resources, with no high-end GPUs, can be employed. By virtue of the pre-trained Llama 2 model and through targeted parameter-efficient fine-tuning, computational demands are dramatically reduced compared to the existing methods, thus bridging the gap between academic benchmarks and practical utility.
- **Benchmarking and real-world application**: The success of fine-tuned LLMs for finance has also highlighted that these have not yet adequately addressed the domain of portfolio construction. To this end, we integrate the extracted sentiment signals by FinLlama into a long-short portfolio, which allows us to obtain finance-specific real-world metrics including cumulative returns and the Sharpe ratio.

2 Related Work

The potential of sentiment analysis in finance was first recognised in 1970 by Eugene Fama who introduced the Efficient Market Hypothesis (EMH) [7], which states that stock prices change in response to unexpected fundamental information and news. In this context, before the introduction of advanced machine learning tools, the financial sector has employed lexicon-driven approaches [20]. These methods analyse textual content, sourced from news articles or financial disclosures, based on specific keywords, which are then linked to established sentiment ratings [11, 12]. However, an exponential increase in the volume and richness of online available information posed significant challenges for lexicon-based analysis, but has opened a fertile ground for machine learning strategies, including techniques such as Naive Bayes and Support Vector Machines [5], as summarised in Figure 1.

In parallel, the advances in deep learning have become instrumental for NLP research and have spurred pioneering works that sought to harness the power of neural networks for NLP tasks. Recently, the introduction of the attention mechanism and transformer networks has enabled a significant shift away from recurrent and convolutional methods, traditionally used in deep-learning tasks [27]. This has led to the development of transformer-based models, such as BERT [6], which owing to its contextual comprehension of language has been used extensively for sentiment analysis. However, the performance of BERT in the financial domain has encountered limitations, primarily because it is not specifically trained on financial datasets. Moreover, its requirement for substantial amounts of data for fine-tuning purposes poses a considerable challenge for financial applications, where such data may not be readily available.

More recently, FinBERT [1], a version of BERT which is fine-tuned on financial text, has shown promising results for the task of financial sentiment analysis. However, FinBERT still suffers from limitations such as insensitivity to numerical values, while due to its relatively small size (110 million parameters) its classification accuracy deteriorates with sentence complexity [4]. The FinGPT [14, 26] and Instruct-FinGPT [28] aim to enhance their expressive power by using the Llama 7B as their base model. However, FinGPT is not optimized for the task of financial sentiment analysis whilst Instruct-FinGPT only classifies the sentiment valence but is not capable of quantifying the strength of a sentiment class.

To the best of our knowledge, BloombergGPT [24] is the only pre-trained finance-specific LLM, as Bloomberg was able to train the model using data collected over a span of 40 years. Despite the impressive performance of the model on financial sentiment analysis, the resources required to train such a model are substantial (1.3M GPU hours at a cost of $5M) whilst much of the training data is confidential and not publicly available. This is different from our proposed methodology, which focuses on achieving a high classification accuracy whilst minimizing the training corpus and computational resources, and utilizing publicly available training data. This is achieved by fine-tuning a pre-trained general-purpose LLM on a smaller-scale financial data corpus.

3 Methodology

Our work aims to embark upon the immense expressive power and contextual understanding of general-purpose LLMs in order to make them finance-specific. This is achieved by fine-tuning the state-of-the-art (SOTA) Llama 2 7B model on a finance-specific corpus of online data. The effectiveness of our approach is demonstrated on financial sentiment analysis through a new set of benchmarks that align closely with end portfolio construction – the ultimate goal of financial analysis.

3.1 Fine-tuning the Llama 2 model

Even though pre-trained LLMs offer a range of capabilities such as reasoning, translation, summarising and text generation, they often struggle when applied to a specific task of interest, such as sentiment analysis. This limitation becomes even more critical in the finance domain, where the nuanced language, media hype and extensive length of financial news articles pose significant challenges.

To tackle these challenges, our work revisits the first principles of LLMs in order to align them to the task of financial sentiment analysis. This is achieved by using four labelled financial text datasets as training data to fine-tune the Llama 2 model. Such finance-specific training equips the model with the ability to understand the linguistic nuances present in the financial domain. Furthermore, a three-class SoftMax classification layer is employed at the output of the foundational model. This made it possible to alter the primary function of the LLM from text generation to sentiment classification. In this way, the proposed fine-tuned FinLlama model acts as a generator-discriminator and produces sentiment decision outputs for three labels: positive, negative or neutral.

3.1.1 Training datasets. The training data was a combination of four labelled publicly available financial news datasets, namely

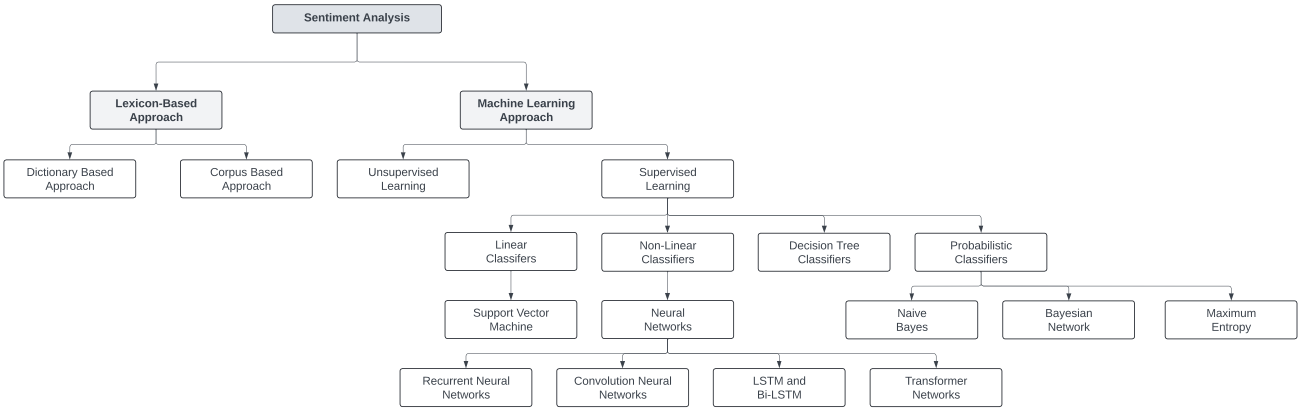

Figure 1: Overview of sentiment analysis methods.

the Financial PhraseBank (FPB) dataset [19], FiQA dataset [18], Twitter Financial News dataset [23] and GPT-labelled Financial News dataset [17]. This resulted in a comprehensive collection of 34,180 labelled samples, as outlined below.

- **Financial PhraseBank (FPB) Dataset.** This dataset, accessible via HuggingFace, consists of 4,840 samples which are randomly extracted from financial news articles. In order to ensure high quality annotation, the samples were annotated by 16 experts with backgrounds in finance and business. Each sample was annotated with one of the three labels: positive, negative, and neutral.
- **FiQA Dataset.** This dataset is also accessible via HuggingFace and consists of 1,210 labelled sentences. Each sentence was annotated with one of the three labels: positive, negative, and neutral.
- **Twitter Financial News Sentiment.** This dataset, accessible via HuggingFace, includes 11,930 tweets with content from the financial domain. Each tweet was annotated as positive, negative, and neutral.
- **GPT-labelled Financial News.** This dataset, accessible via HuggingFace, consists of 16,200 financial news articles labelled by GPT-3.5. Each article was annotated with one of the five labels: strongly negative, mildly negative, neutral, mildly positive, and strongly positive. To align this dataset with the three-class output of our FinLlama model, the strongly and mildly negative classes were combined into a single negative class, and similarly, the strongly and mildly positive classes were combined into a single positive class.

3.1.2 Model Training. The proposed FinLlama model was first initialised with the Llama 2 7B model, followed by fine-tuning over 5 epochs. The training process utilised the AdamW optimizer [15], as it effectively decouples the weight decay from the optimization steps, leading to more effective training. The initial learning rate was deliberately kept small as the Llama 2 7B model is already pre-trained on a large corpus of data, whilst the warm-up ratio and weight decay served as key regularisation techniques to prevent overfitting, a crucial aspect given the limited size of our fine-tuning dataset.

Moreover, the LoRA implementation was employed in the fine-tuning process with a rank, $r = 8$, a scaling factor, $\alpha = 16$, and a dropout of 0.05, in order to minimize the number of trainable parameters whilst achieving high and robust end performance. Through the LoRA implementation, the number of trainable parameters was set to 4.2M, amounting to just 0.0638% of the total number of parameters in the Llama 2 7B model. This made it possible for our fine-tuning process to be **implemented on a single A100 (40 GB) GPU**, thus avoiding the need for excessive computational resources. A summary of the most important training parameters used in the fine-tuning process is given in Table 1.

3.2 Proposed Framework

After establishing the proposed fine-tuned Llama 2 model, we followed the framework shown in Figure 2, with the aim of assessing the performance of our FinLlama model against other established sentiment analysis methods, using finance-specific real-world metrics.

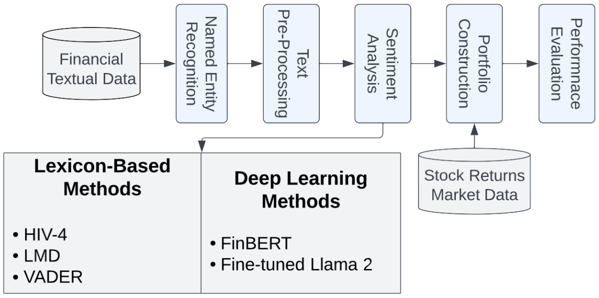

Figure 2: Framework for sentiment analysis.

Data Collection. Both textual and market data were analysed in order to construct appropriate long-short (L/S) portfolios. Regarding the textual data, 204,017 articles dating between 2015 to 2021 were

Table 1: Training parameters used in the fine-tuning process of the proposed FinLlama.

Parameter	Definition	Value
Learning rate	Determines the step size at each iteration of gradient descent	0.0003
Weight decay	Regularization technique to prevent overfitting by penalizing large weights	0.01
Batch size	Number of training samples used in one iteration of gradient descent	128
Training epochs	A full training pass over the entire training set	5
LR scheduler	Framework that adjusts the learning rate between iterations	Cosine Annealing
Warmup ratio	Increases the learning rate gradually over a certain number of epochs	0.1
GPUs	Number of GPUs used	1 A100 (40GB)
LoRA rank	Defines the dimensions of low-rank matrices	8
LoRA alpha	Scaling factor for the weight matrices within LoRA	16
LoRA dropout	Proportion of randomly deactivated neurons during training	0.05

collected from online sources such as Reuters, The Motley Fool and MarketWatch. These sources were selected due to their reliability, reputation, lack of bias and focus on major corporations. Financial market data were collected for the same time period from Yahoo Finance. These market data contained daily stock returns for the 500 companies in our Investable Universe (S&P 500), resulting in 1,672 days of stock returns data for each company.

Named Entity Recognition. To ensure that news articles are accurately linked to the correct organizational entity, each article must be associated with at least one relevant stock [8]. This step reduces the issue of irrelevant articles being connected to a particular stock [3]. For robustness, in our study we utilised the BERT-base-NER model [13], which is capable of recognising four types of entities: location, organization, person, and miscellaneous, and provides a confidence score for each identified entity. For each article in our corpus, if the confidence score for the entity associated with the company was above 98%, the article was retained; otherwise, it was rejected. Table 2 presents the results of running the NER filtering on the initially scraped dataset, with respect to the number of articles. Observe that in this way the total number of articles was reduced by 24.1%. The effect is similar for MarketWatch and The Motley Fool, with both exhibiting reductions of just above 25%, while Reuters experienced a smaller reduction of approximately 6.3%.

Table 2: Total number of articles scraped per source before and after NER filtering.

News Source	No. of articles pre processing	No. of articles post NER filtering
MarketWatch	309,187	236,214
Reuters	38,141	35,741
The Motley Fool	205,270	147,413
Total	**552,598**	**419,368**

Text Pre-Processing: The news article was represented as a 'bag-of-words' and the following steps were subsequently performed:
a) Tokenization, b) Stop-Word Removal, c) Lemmatization, d) (Lower) Case Normalization, e) Feature Selection. Regarding the Feature Selection step, the frequency of each word was used as a feature, especially for the lexicon-based approaches.

Sentiment Analysis. In total, five sentiment analysis methods were applied. For the lexicon-based approaches (see Appendix A.1), LMD [16] and HIV-4 [21] were implemented using the pysentiment2 Python library, while VADER [10] was implemented using the NLTK library. Regarding the deep learning methods (see Appendix A.2), both the FinBERT model and our FinLlama model were obtained through HuggingFace, and were utilised via the Transformers library.

The considered methods were evaluated on every article within each corpus for a given company. In cases where multiple articles were published on the same day for a given company, the average sentiment for that day was calculated as

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} S_{it} \tag{1}$$

Here, S_t represents the average sentiment for the t-th day, N_t denotes the number of news articles published on that same t-th day for a given company, while S_{it} designates the sentiment strength of the i-th news article on a particular t-th day. The daily sentiment outputs for each company were merged to arrive at the final sentiment data that were utilised as a parameter in the portfolio construction stage.

Portfolio Construction. Once the sentiment for each method was defined for every company, the long-short portfolio was constructed. We used the sentiment as a parameter to determine which companies should be in a long or a short position, aiming to maximise returns from both positions. The long-short portfolio was constructed using the following procedure:

- *Define the Investable Universe:* Even though the S&P 500 comprises 500 companies, the financial textual data collected did not contain articles associated to some of the companies for the test period of February 2015 to June 2021. Consequently, 417 companies were considered.
- *Define the long and short position*: The sentiment signal obtained from each of the five methods was used to construct five distinct portfolios. For each method, companies were ranked daily according to their sentiment. Companies that did not have sentiment data on a particular day were omitted from the ranking. As the daily sentiment score for each company ranges between -1 and 1, those with the highest

positive sentiment were placed in a long position, whilst those with the strongest negative sentiment were placed in a short position.

- *Allocation:* An equally-weighted portfolio strategy was considered in our portfolio construction as this strategy is mostly utilised by hedge funds [11]. The percentage of companies in a long and short position was fixed at 35%. Consequently, the top 35% of companies in terms of performance were allocated to long positions, while the bottom 35% were allocated to short positions.
- *Determine daily returns:* The daily return for each company that was held in a long or short position was obtained by the market data on that particular day. The average daily return of companies that were held in a long position, r_{Long}, was defined as

$$r_{Long} = \frac{1}{N_{Long}} \sum_{i=1}^{N_{Long}} r_{Long}(i) \tag{2}$$

Similarly, the average daily return of companies that were held in a short position, r_{Short}, was defined as

$$r_{Short} = \frac{1}{N_{Short}} \sum_{i=1}^{N_{Short}} r_{Short}(i) \tag{3}$$

For each particular day, the number of companies that were held in either a long position (N_{Long}) or a short position (N_{Short}) were equal. Consequently, the total portfolio return on a particular day was the difference between the daily long return, $r_{Long}(i)$, and daily short return, $r_{Short}(i)$, and is given by

$$r_{daily}(i) = r_{Long}(i) - r_{Short}(i) \tag{4}$$

Portfolio Evaluation. The performance of the portfolio constructed using our fine-tuned model was assessed against the portfolios constructed using other SOTA sentiment methods. To this end, the employed real-world financial metrics were: cumulative returns, r_{cum}, annualized return, R_p, annualized volatility, σ_p, and the Sharpe ratio, S_a [2], defined as

$$r_{cum} = \sum_{i=1}^{N} r_{daily}(i) \tag{5}$$

$$R_p = \frac{1}{N} \sum_{i=1}^{N} r_{log}(i) \times 252 \tag{6}$$

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^{N} (r_{log}(i) - \bar{r})^2}{N-1}} \times \sqrt{252} \tag{7}$$

$$S_a = \frac{R_p - R_f}{\sigma_p} \tag{8}$$

where N is the total number of investing days, totaling 1,672, $r_{log}(i)$ represents the logarithmic daily return, $\bar{r}$ denotes the average daily logarithmic return, R_f designates the annualized risk-free rate of return, and 252 is the number of business days in a year. The risk-free return, R_f, typically represents the yield of the 10-Year Treasury Note; however, due to its prolonged low yield [25] during the analysed period, a 0% rate is commonly used and was adopted in our analysis.

Table 3: Difference in cumulative returns between our FinLlama model and the best-performing existing method (among LMD, HIV-4, VADER, and FinBERT) on the first day of each year, along with the daily average number of companies traded during the previous year. A negative difference in returns indicates that the cumulative returns of our model are lower than those of the best existing method at that date.

Date	Daily Companies Traded	Return Difference	Best existing method
1/1/2016	14.7	-8.1	LMD
1/1/2017	19.0	40.1	FinBERT
1/1/2018	20.0	59.3	FinBERT
1/1/2019	20.0	54.7	FinBERT
1/1/2020	28.0	73.2	FinBERT
1/1/2021	49.2	98.5	FinBERT

4 Experimental Results

The performances of the five portfolios which were constructed as described in Section 3 are illustrated in Figure 3. Notice that the deep learning approaches outperformed the lexicon-based approaches in terms of cumulative returns, particularly those relying on general-purpose dictionaries (HIV-4 and VADER). This was to be expected, given that lexicon-based approaches often fail to capture the contextual meaning of sentences, whilst the nuanced nature of financial text significantly reduces the accuracy of general-purpose dictionaries.

Moreover, observe from the top-right panel of Figure 3 and Table 3 that the difference in cumulative returns between our model and the best performing method among the considered ones increased over time. The significant advantage of our FinLlama from 2019 onwards can be explained by a significant rise in the daily average number of companies traded, as a result of an increasingly more diverse set of articles in our news corpus over the years. Indeed, this difference in returns exhibits a positive correlation of 0.81 with the daily average number of companies invested, with a P-value of 0.048, indicating the statistical significance of the trend (significant if P-value < 0.05). The summary of the difference in cumulative returns between our model and the best performing existing method on the first day of each year, along with the daily average number of companies traded during the *previous* year, is shown in Table 3.

It is important to note that the increase in the daily average number of companies traded coincides with a rise in the number of articles used to calculate the daily sentiment of each company from 2018 onwards. This behaviour is attributed to Reuters first starting to produce digital content in 2018, followed by a dramatic increase from 2020 onwards, when MarketWatch began producing AI-generated articles on stock price updates, as shown in Figure 4. Additionally, there has been a natural increase in the amount of digital articles produced by all three sources since 2019.

The increased returns resulting from more informed trading decisions, along with the growing gap between the returns of our model and those of the best existing method, highlight the superior ability of our model to achieve accurate financial sentiment valence and strength quantification, compared to existing methods. This is because, the accuracy of sentiment parameters becomes increasingly important with the rise in the number of companies traded and the volume of articles used to make trading decisions. Such

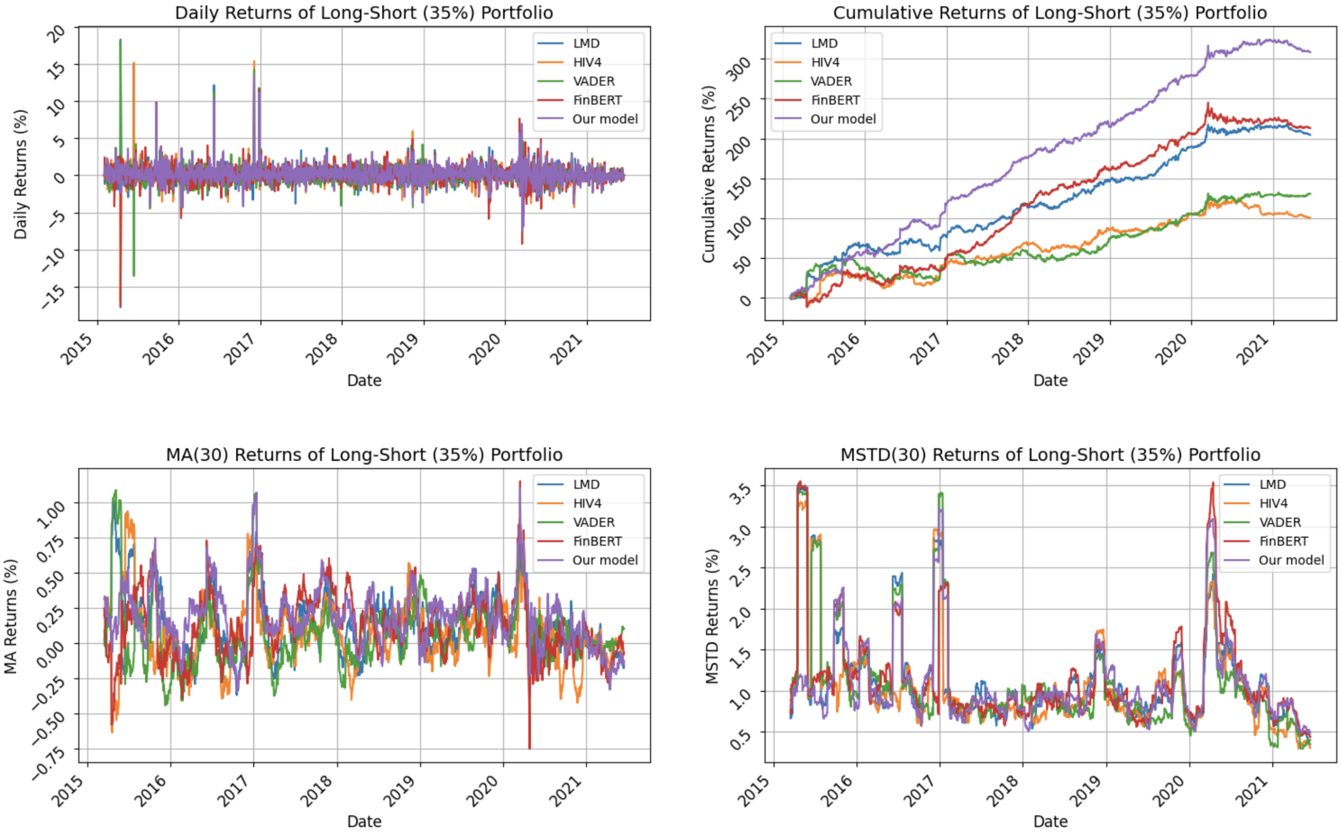

Figure 3: Comparison of the performance of the 35% long-short portfolios which were constructed using the five considered sentiment analysis methods, for the time period of February 2015 to June 2021. The MA(30) and MSTD(30) represent, respectively, the moving average and the moving standard deviation of the returns calculated over a 30-day rolling window.

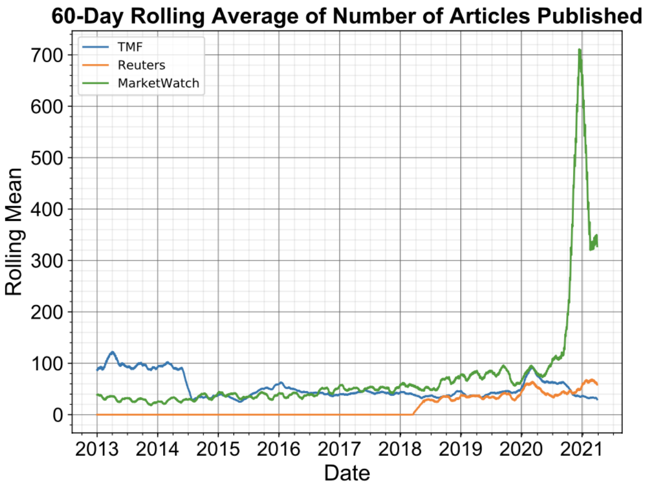

Figure 4: The 60-day rolling average of total number of articles published on each of The Motley Fool, Reuters and MarketWatch from 01/01/2013 to 31/05/2021

trend has been observed over time due to the expanding corpus of financial news articles used during the trading stage.

The improved sentiment classification accuracy exhibited by our model also leads to more robust trading decisions, as indicated in the bottom two panels of Figure 3. In particular, a comparison of our FinLlama model with FinBERT, the current best performing model in the literature, shows that during turbulent economic periods caused by unexpected events or economic changes, the standard deviation of our model was lower than that of FinBERT, while achieving similar or higher returns. The enhanced robustness of FinLlama is evident across a range of socio-economic and geopolitical events that caused significant movements in the S&P 500, identified through the business information database Factiva, most notably:

- New trading regulations in China, renewed worries about the Greek economy running out of money, and tepid US corporate earnings in April 2015.
- Concerns about the Federal Reserve increasing interest rates, uncertainty about Greece defaulting on their debt, and geopolitical events and tensions, including the Saint-Quentin-Fallavier attack in June 2015.
- Apprehension about the economic impact of the 2016 US elections, including potential changes in trade policies, tax

Table 4: Statistical comparison between the performances of the five considered sentiment analysis methods using a 35% long-short portfolio. For Cumulative Returns, Annualized Return and Sharpe Ratio, higher is better. For Annualized Volatility, lower is better.

	LMD	**HIV-4**	**VADER**	**FinBERT**	**FinLlama (Ours)**	**S&P 500**
Cumulative Returns (%)	204.6	100.4	130.6	213.0	**308.2**	83.1
Annualized Return (%)	29.1	13.5	17.9	30.3	**45.0**	11.3
Sharpe Ratio	1.5	0.7	0.9	1.5	**2.4**	0.62
Annualized Volatility (%)	19.5	18.9	19.6	20.3	18.6	**18.5**

reforms, regulatory adjustments, and shifts in domestic and international economic relations in January 2017.

- Significant fears about the economic effects of the COVID-19 pandemic, including concerns about a severe economic downturn, increased unemployment rates, corporate bankruptcies, and a dramatic decline in consumer spending and business investments in March 2020.

The quantitative results, displayed in Table 4, support the qualitative observations mentioned above and suggest that the 35% long-short portfolio, constructed using our fine-tuned Llama-2 model, was the most successful.

Overall, our FinLlama model successfully generated significantly higher returns for investors compared to all other considered methods, and most importantly FinBERT, whilst simultaneously reducing portfolio risk and being more robust to turbulent economic periods, as indicated by the higher Sharpe ratio and lower annualized volatility.

5 Conclusion and Future Work

We have introduced an innovative approach to financial sentiment analysis which rests upon the fine-tuning of a general-purpose LLM. The proposed method has capitalised on the extensive knowledge base and generative nature of LLMs, combining their inherent text generation with the classification ability. In addition, such an approach has enabled the LLMs to become more attuned to the nuanced language of the finance sector, whilst minimising their resource utilisation and computational demands.

Our fine-tuned Llama2 7B model, termed FinLlama, has been used to construct a long-short portfolio, yielding results that have surpassed those of the existing methods in the field. The FinLlama has achieved cumulative returns which have outperformed the currently leading FinBERT model by 44.7%, while achieving a significantly higher Sharpe ratio and lower annualized volatility. This demonstrates that fine-tuning an LLM can yield superior results, even with a small amount of task-specific data. In addition, the present work has set a new benchmark in the field, transcending traditional measures such as the accuracy and F1-score, which are commonly used in the literature. It is our hope that such an approach is a step towards narrowing down the divide between academic research and practical applications within quantitative finance.

Our future research will aim to enhance both the sentiment classification accuracy and efficiency of fine-tuned LLM models by incorporating additional techniques to produce a tractable and interpretable platform to facilitate the application of artificial intelligence (AI) in the finance sector. Moreover, we will explore enhancing our training dataset with a non-English language corpus and incorporating trading costs and stop-loss limits into our portfolio construction, to enhance its utility.

Disclaimer: Nothing herein is financial advice, and NOT a recommendation to trade real money. Please use common sense and always first consult a professional before trading or investing.

References

[1] Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).

[2] J .B. Berk and P. M. DeMarzo. 2019. Corporate Finance. *Pearson* 5.

[3] Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew Richardson. 2013. Which News Moves Stock Prices? A Textual Analysis. *SSRN Electronic Journal.*

[4] Ziwei Chen, Sandro Gössi, Wonseong Kim, Bernhard Bermeitinger, and Siegfried Handschuh. 2023. FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes. Proceedings of the 4th ACM International Conference on AI in Finance, 357–364.

[5] Nello Cristianini and John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics.* https://api.semanticscholar.org/CorpusID:52967399

[7] Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25, 2 (1970), 383–417.

[8] Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011. The Stock Sonar - Sentiment Analysis of Stocks Based on a Hybrid Approach. *Proceedings of the National Conference on Artificial Intelligence.*

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[10] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, Vol. 08. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, 216–225.

[11] Zheng Tracy Ke, Bryan T. Kelly, and Dacheng Xiu. 2019. *Predicting Returns With Text Data.* NBER Working Papers 26186. National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:26186

[12] Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems* 69 (2014), 14–23.

[13] D. S. Lim. 2021. BERT-base-NER. https://huggingface.co/dslim/bert-base-NER.

[14] Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023. FinGPT: Democratizing Internet-scale Data for Financial Large Language Models. *arXiv preprint arXiv:2307.10485* (2023).

[15] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *ArXiv* abs/1711.05101 (2017). https://api.semanticscholar.org/CorpusID:3312944

[16] Tim Loughran and Bill Mcdonald. 2011. When Is a Liability NOT a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66 (02 2011), 35 – 65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[17] Neural Magic. 2022. Twitter Financial News Sentiment. http://precog.iiitd.edu.in/people/anupama

[18] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. *Companion Proceedings of the The*

Web Conference 2018 (2018). https://api.semanticscholar.org/CorpusID:13866508

[19] 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 (2014), 782–796.

[20] Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. 2020. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* 8 (07 2020), 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626

[21] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press.

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[23] Oliver Wang. 2023. News with GPT instructions. https://huggingface.co/datasets/oliverwang15/news_with_gpt_instructions

[24] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *ArXiv* abs/2303.17564 (2023). https://api.semanticscholar.org/CorpusID:257833842

[25] Yahoo Finance. 2023. Treasury yield 10 years historical data. https://finance.yahoo.com/quote/%5ETNX/history

[26] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint arXiv:2306.06031* (2023).

[27] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. 1480–1489. https://doi.org/10.18653/v1/N16-1174

[28] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *ArXiv* abs/2306.12659 (2023). https://api.semanticscholar.org/CorpusID:259224880

A Existing Sentiment Analysis Methods

A.1 Lexicon-Based Approaches

A.1.1 Harvard IV-4 Psychological Dictionary (HIV-4). The HIV-4 is one of the oldest manually constructed lexicons, and is used for objectively identifying specified characteristics of messages in areas involving social science, political science, and psychology. The latest version of the HIV-4 dictionary contains over 11,000 words which are classified into one or more of 183 categories. In this work, we focus on the 1,045 words labelled as positive and the 1,160 words labelled as negative.

A.1.2 Loughran and McDonald (LMD) Dictionary. Loughran and McDonald evaluated standard dictionaries and found that these frequently misclassify terms within financial texts. This insight led to the development of the LMD dictionary, which is specifically tailored for the financial sector. The dictionary categorizes words into six distinct sentiment categories: negative, positive, uncertainty, litigious, strong modal, and weak modal. It was constructed using data from 50,115 10-K filings from 8,341 firms listed on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ), covering the period from 1994 to 2008. Overall, the LMD dictionary contains 2,355 negative financial words and 353 positive financial words.

A.1.3 Valence Aware Dictionary for Sentiment Reasoning (VADER). The VADER dictionary combines lexical features, derived from micro-blog contexts, with the grammatical and syntactical conventions that humans typically employ to express or emphasize sentiment intensity. This enables VADER to accurately quantify the sentiment strength of text. The model contains approximately 9,000 token features, which are each assigned a sentiment score ranging from -4 (indicating extremely negative sentiment) to +4 (indicating extremely positive sentiment). The overall polarity score for a text is calculated by summing the sentiment scores of each word present in the lexicon, with the final score normalized to fall within the range of -1 to +1.

A.2 Deep Learning Approaches

A.2.1 FinBERT. FinBERT leverages the BERT model architecture, and is specifically tailored for financial contexts. It was pre-trained on a substantial financial text corpus consisting of 1.8M news articles sourced from the Thomson Reuters Text Research Collection (TRC2) dataset, spanning the years between 2008 to 2010. Further refinement was achieved through fine-tuning on the Financial Phrasebank (FPB) dataset, thus enhancing its capabilities in financial sentiment classification. FinBERT generates SoftMax outputs for three labels: positive, negative, and neutral.