International Review of Financial Analysis 94 (2024) 103291

Contents lists available at ScienceDirect

International Review of Financial Analysis

journal homepage: www.elsevier.com/locate/irfa

FinSentGPT: A universal financial sentiment engine?

Aref Mahdavi Ardekani [a], Julie Bertz [a], Cormac Bryce [b], Michael Dowling [a,c,*], Suwan(Cheng) Long [d,e]

[a] DCU Business School, Dublin City University, Ireland
[b] Bayes Business School, London, UK
[c] Rennes School of Business, France
[d] Trinity Business School, Trinity College Dublin, Ireland
[e] Judge Business School, Cambridge University, UK

ARTICLE INFO

JEL classification:
C22
F47

Keywords:
ChatGPT
Large language models
Financial sentiment
Monetary policy
Fine-tuning

ABSTRACT

We present FinSentGPT, a financial sentiment prediction model based on a fine-tuned version of the artificial intelligence language model, ChatGPT. To assess the model's effectiveness, we analyse a sample of US media news and a multi-language dataset of European Central Bank Monetary Policy Decisions. Our findings demonstrate that FinSentGPT's sentiment classification ability aligns well with a prominent English-language finance sentiment model, surpasses an established alternative machine learning model, and is capable of predicting sentiment across various languages. Consequently, we offer preliminary evidence that advanced large-language AI models can facilitate flexible and contextual financial sentiment determination, transcending language barriers.

1. Introduction

Sentiment analysis has developed as a potent tool for comprehending and forecasting market dynamics in finance. Sentiment, which is a collective set of investor beliefs, is widely known to affect asset pricing and influence markets. For this reason, integrating a suitable measure of sentiment in classical finance models has become a critical task in recent years (Zhou, 2018). One major issue has been how to accurately measure sentiment. While financial analysis has traditionally relied primarily on quantitative information and economic indicators to make wise investment decisions, sentiment analysis has become more popular in the financial industry due to the development of sources of big data and improvements in natural language processing (NLP).

Our study proposes that Generative AI models can be a game-changing development with far-reaching consequences for sentiment analysis. The underlying generative AI models use deep learning to develop material that can, with limitations, mimic human creativity and judgement (Guo et al., 2023; Wiegreffe, Hessel, Swayamdipta, Riedl, & Choi, 2021). Generative AI models in finance might be capable of producing synthetic financial documents, market scenarios, and investment strategies. These models contribute to the synthesis of novel insights and views that supplement quantitative investigations by learning patterns from large datasets. The combination of generative AI with sentiment analysis opens up new options for investigating subtle market attitudes, improving decision-making processes. This nexus of generative AI and sentiment analysis adds a transformational layer to the landscape of financial analysis by allowing the development of original material that reflects the complexities of human emotion in financial discourse.

Textual analysis has gained traction in studying financial phenomena (Ash & Hansen, 2022; Kearney & Liu, 2014), but standard text analysis can encounter difficulties in capturing language nuances in financial contexts. Enter ChatGPT, an artificial intelligence language model developed by OpenAI that has revolutionized the field of NLP models with claims of demonstrating 'sparks' of artificial general intelligence (Bubeck et al., 2023). The model shows strong potential in contextual analysis (Korinek, 2023) and text simulation (Dowling & Lucey, 2023), offering new opportunities for finance research. With the incorporation of ChatGPT into the domain of financial sentiment analysis, a crucial step towards addressing the intricate linguistic nuances of financial conversation is being taken.

ChatGPT is based on an underlying style of model, a Generative Pre-training Transformer or GPT, and is part of a suite of models offered by OpenAI tailored to specific tasks and price points. This suite of large-language models (LLM) corresponds with the emerging environment of AI-powered research and promises to improve comprehension and

* Correspondence to: DCU Business School, Dublin City University, Glasnevin, Dublin 9, Ireland.
E-mail address: michael.dowling@dcu.ie (M. Dowling).

https://doi.org/10.1016/j.irfa.2024.103291
Received 9 September 2023; Received in revised form 15 March 2024; Accepted 12 April 2024
Available online 17 April 2024

prediction of financial market dynamics through a comprehensive examination of both quantitative and qualitative textual data. We use ChatGPT and the various OpenAI models as examples of LLMs, but there are many competing models — most notably the Llama 2 models of Meta Group and the Gemini models of Alphabet Group.

This study presents a finance sentiment model that uses the capability of a selection of OpenAI models, to recognize financial sentiment. Our focus is on harnessing these models generative capabilities to enhance the accuracy and depth of sentiment analysis in the finance domain. We focus on central bank related news in this study. Central banks hold significant authority over financial markets, with their policy announcements closely scrutinized for insights about economic prospects and interest rate adjustments (Correa, Garud, Londono, & Mislang, 2021). The sentiments expressed in their statements influence market reactions and economic estimates. Similarly, investment decisions and market perceptions are deeply influenced by debates on the inflation news, whether motivated by anxieties about price escalation or confidence about continuous growth. As market players negotiate uncertainty, these sentiments shape financial landscapes.

Financial sentiment is a rapidly growing research area with recent studies such as Barbaglia, Consoli, and Manzan (2022) and Picault, Pinter, and Renault (2022) emphasizing the importance of contextual understanding to properly gauge sentiment. However, existing means of capturing context require cumbersome human labelling (Li, 2020) and struggle with lexicon (Mishev, Gjorgjevikj, Vodenska, Chitkushev, & Trajanov, 2020), and language context determination and multi-language applicability (Degani & Tokowicz, 2010).

A first interest of our study is the capability of a model fine-tuned in one language to accurately detect sentiment in another language. Most sentiment models primarily focus on the English language, leaving other languages under-served. Understanding textual sentiment across many languages has become critical yet difficult for investors and decision-makers due to the increased globalization of risk factors (Ejara, Krapl, O'Brien, & Ruiz de Vargas, 2020). Determining sentiment in the financial environment demands a deep understanding of context, which is not an easy process. The complex interplay of linguistic nuances, market emotions and cross-cultural differences further complicates issues. While quantitative measures may be evaluated objectively, sentiments are fundamentally subjective, as they are influenced by perceptions, emotions and biases.

Adoption of financial sentiment analysis across several languages raises numerous issues that must be addressed in order to get accurate and trustworthy results (Lo, Cambria, Chiong, & Cornforth, 2017). Finance's global, multilingual, and multicultural character complicates sentiment analysis. Because the financial sphere crosses languages, nations, and cultures, a one-size-fits-all strategy is ineffective. Understanding other languages requires not just language translation but also context retention. Language nuances, cultural references, and linguistic idioms can all dramatically influence sentiment perception. As a result, the capacity to undertake sentiment analysis across many languages is critical for gaining a thorough understanding of the global financial ecosystem.

Furthermore, financial sentiment analysis has difficulty comprehending both structured and less-structured textual input. While structured data such as financial reports and market indicators give clear and organized information it tends to have sentiment removed from its presentation, and extracting sentiment from more complicated textual content requires the aforementioned contextual understanding (Gu, Zhang, Hou, & Song, 2018). Less-structured textual data, such as news articles, social media and online forums, provide particular hurdles due to their unstructured nature. Because these sources commonly utilize informal language, acronyms, sarcasm, and colloquial idioms, sentiment analysis becomes more complex. The second purpose of our study is to show how a GPT model can handle this unstructured form of textual data.

Our study exploits GPT's contextual understanding abilities (Tur & Traum, 2022) and multi-language capabilities (Hendy et al., 2023). A challenge is GPT's generalization tendency when domain knowledge is limited (Wang, Yao, Kwok, & Ni, 2020), and we use fine-tuning with private data to mitigate this issue. FinSentGPT, our fine-tuned GPT model trained on textual sentiment in finance, demonstrates proficiency in identifying sentiment across languages.

The fundamental contribution of our research is the pioneering development of large-language AI models for financial sentiment analysis. We show that it is possible to estimate sentiment from both structured (central bank policy statements) and less-structured (financial media articles) textual data. Our FinSentGPT model outperforms simpler sentiment models contextually and across languages, demonstrating that fine-tuning a GPT model on English sentiment samples suffices for applying knowledge to unseen text in other languages. This revelation has major implications for establishing a universal financial sentiment engine, as well as minimizing existing research biases that favour the English language. Furthermore, the incorporation of generative AI models into sentiment analysis opens up new opportunities for content creation and novel data-driven initiatives in the financial realm.

As finance advances and AI models improve, our research serves as a link between the two fields. This study highlights how powerful AI technologies may be used to reshape sentiment analysis inside the intricate fabric of finance and banking. The next parts go into our methodology, empirical analysis, and implication of our findings. we seek to contribute to the growing body of knowledge unravelling financial sentiments' mysteries and guiding decision-makers and investors in the dynamic financial landscape.

2. Methodology

2.1. Data

The research design encompasses two principal studies, each capitalizing on distinct datasets. The initial study is predominantly anglophone-centric, utilizing a dataset constituting 2226 articles from the New York Times which referenced domestic US inflation throughout the year 2022. This dataset was selected to evaluate the model's efficacy in predicting sentiment within the confines of a single language. We generated word clouds for New York Times news articles and Monetary Policy Decisions statements in English (see Figs. 1 and 2).

These visualizations clearly describe the keywords, main themes and financial sentiment of the respective content of the New York Times article and the ECB monetary policy statement.

The subsequent study advances upon the preceding one, harnessing monthly European Central Bank (ECB) Monetary Policy Decisions statements promulgated in English, French, German, Spanish, and Portuguese from the years 2017 to 2021. This dataset was judiciously selected to enable exploration into the GPT's multi-lingual capabilities.

Furthermore, the previously collated New York Times data was amalgamated into the training dataset utilized for the second study's model, FinSentGPT_ECB2. The intention was to mitigate the scarcity of relevant ECB statement sentences in our training dataset and assess whether the supplementary data could enhance the predictive performance of the model.

2.2. Methodology

The research methodology involves the scrutiny of various FinSentGPT versions to ascertain the GPT's performance in the prediction of financial sentiment. FinSentGPT is our novel financial sentiment analysis model designed to leverage the capabilities of large-language AI models, specifically the GPT models of OpenAI. The model is developed as a response to the intricate linguistic nuances and multilingual challenges present in financial discourse. At its core, FinSentGPT is a

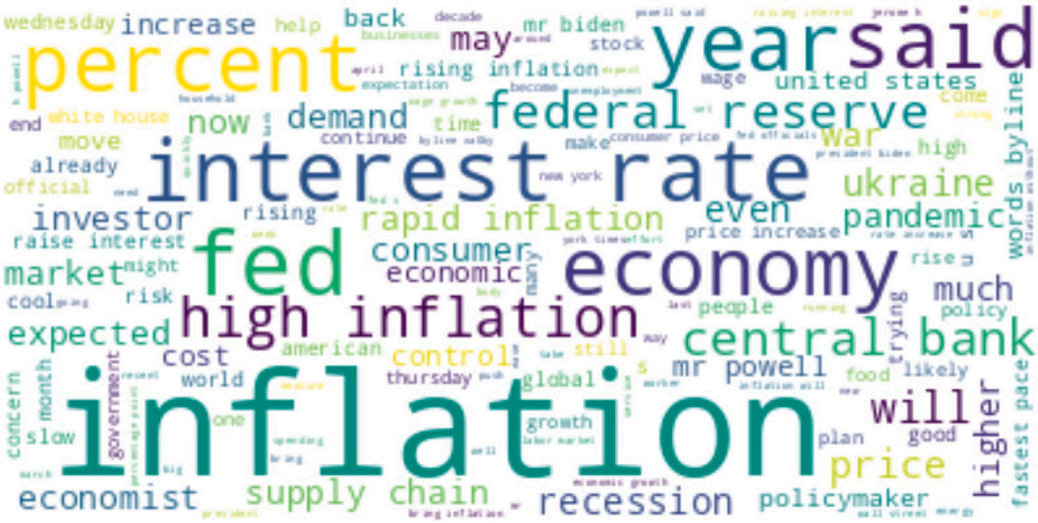

Fig. 1. New York Times articles wordcloud.

Fig. 2. Monetary Policy Decisions statements wordcloud.

fine-tuned version of a general GPT model (fine-tuning is explained later in this section), with a specific focus on recognizing and interpreting financial sentiment across multiple languages. Our training focuses on enabling the model to compute financial sentiment scores for sentences containing the term "*inflation*" in both news media and central bank statements. By financial sentiment we mean the method delineated by Barbaglia et al. (2022).

Subsequently, OpenAI's ADA GPT3 model is fine-tuned[1] to engender FinSentGPT_NYT, utilizing pertinent sentences from New York Times and sentiment scores from the first half of 2022. We briefly discuss the fine-tuning process here. The base ADA GPT-3 model was trained by OpenAI on a broad spectrum of texts, which makes it adept at comprehending and producing text across various topics. However, its training is broad-based rather than focused on a specific area. Our fine-tuning involves enhancement of the ADA GPT-3 model's capabilities in accurately understanding and analysing financial sentiment in textual data. To achieve this, we introduced a targeted dataset consisting of New York Times sentences from the first half of 2022 concerning domestic US inflation. Accompanying these sentences were their financial sentiment scores, which we computed using the methodology outlined by Barbaglia et al. (2022). This process effectively tailored the ADA GPT-3 model to our focused dataset, enabling it to learn and adapt to the unique patterns, terms, and nuances inherent in financial sentiment.

We follow the standard fine-tuning process on OpenAI; that is, in the absence of other information to the contrary, we do not adjust the parameters from their defaults. The fine-tuning process involved several key steps. First, we preprocessed the New York Times dataset. For this we extracted out whole sentences that contained the word 'inflation', calculating the sentence's sentiment score as outlined in the previous paragraph, and storing these sentence-sentiment pairs in a format of: "prompt":"['SENTENCE']","completion":"SENTIMENT_SCORE". Next, we use the default optimization algorithms and tuned hyperparameters such as learning rate, batch size, and the number of epochs (3) to generate the fine-tuned model from these training examples. We also allow the standard regularization techniques, such as dropout or early stopping, to be implemented as per standard OpenAI fine-tuning procedures. Our training loss (1.0139) and validation loss (0.9466) at the last step indicate a reasonable fit and generalizability to unseen data.

[1] https://platform.openai.com/docs/guides/fine-tuning At the time of this analysis, higher-level models, such as GPT3.5 and GPT4 were not fine-tunable.

After fine-tuning we then set about appraising this model, we introduced unseen sentences encompassing the term "*inflation*" from the latter half of the year and compared the model's sentiment scores with those calculated employing (Barbaglia et al., 2022)'s approach. In total, we subject 600 sentences to this testing process, derived from a random sampling of 100 sentences per month. This approach ensures a comprehensive temporal representation, capturing the variations in financial sentiment across different months. Such a distribution guarantees that the dataset reflects a wide range of conditions that could influence financial sentiment, like market fluctuations or economic events, thereby enhancing the generalizability of the findings. The uniformity of maintaining a consistent number of sentences from each month allows for an equitable comparison across various time periods, ensuring that no single month unduly affects the overall results. The choice of 100 sentences per month struck a balance between having a statistically significant sample size — large enough to detect patterns and trends, yet manageable in terms of data processing and analysis. Random sampling of these sentences helped reduce selection bias, ensuring that the dataset accurately represented the spectrum of financial sentiment within each period. This randomness captures a diverse array of sentiment expressions and contexts, offering a comprehensive overview of the sentiment landscape. Furthermore, by utilizing a consistent and representative sample across months, we could more precisely assess the performance of the FinSentGPT_NYT model.

The accuracy of our model is evaluated through a multifaceted approach incorporating four key metrics — Spearman's Rank Correlation, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Average Error (MAE). Furthermore, we implement the creation of a text-based linear regression machine learning model, using the same dataset, to compare its results with those produced by the FinSentGPT_NYT model. This model construction follows the methodology set forth by Hasan, Maliha, and Arifuzzaman (2019), employing a TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer to transform word usage into numerical counts which are then utilized to fit a linear regression model. The selection of TF-IDF vectorization is driven by its capability to effectively represent the importance of words within the corpus, accounting for their frequency across various documents. Differing from basic count-based approaches, TF-IDF emphasizes words that are more unique to individual documents, thus offering a more precise reflection of their significance in financial sentiment analysis. Our preprocessing strategy involves a comprehensive approach to ensuring data quality and consistency. This includes cleaning up text by removing special characters, numbers, and irrelevant symbols, converting all text to lowercase for consistency, tagging text into individual words for frequency analysis, removing common words that lack emotional relevance, and converting Words are lemmatized to their word base forms. This comprehensive approach enables us to ascertain whether the findings generated by FinSentGPT_NYT are merely reflecting the general advantages conferred by machine learning in the field of financial text analysis, or if there are distinct accuracy benefits attributable to the FinSentGPT_NYT model.

In our second study, we move towards a more linguistically diverse approach, concentrating on the multilingual capacity of the GPT. The central dataset employed in this study comprises monthly Monetary Policy Decisions statements issued by the European Central Bank (ECB) in five languages — English, French, German, Spanish, and Portuguese. The English-language statements spanning from 2017 to 2021 are utilized for the training of both ECB models, with financial sentiment scores calculated as per the method delineated by Barbaglia et al. (2022). Capitalizing on OpenAI's DaVinci GPT model, we create an initial fine-tuned model, termed FinSentGPT_ECB1, using this training data.

The model that ensues, FinSentGPT_ECB2, is designed to mitigate the challenges posed by the limited number of pertinent sentences in our ECB statements training dataset. To enhance the model, we incorporate the New York Times economic news training dataset, thereby enabling us to evaluate whether the inclusion of additional data boosts the predictive capacity of the model. In our efforts to test these multilingual models, we predict sentiment scores for unseen sentences from the year 2022, containing non-English translations of the term "inflation". These predicted scores are then juxtaposed against English-language sentiment scores, following the Barbaglia et al. (2022) approach. Our research is predicated on the model's inherent capability to interpret and understand varying language usage contexts, all the while not necessitating any fine-tuning for non-English languages.

Table 1
US inflation sentiment prediction.

	Overall	Month					
		Jul	Aug	Sep	Oct	Nov	Dec
Linear regression model							
Correlation	0.28	0.27	0.28	0.52	0.10	0.37	0.24
MSE	0.19	0.21	0.17	0.13	0.24	0.17	0.21
RMSE	0.43	0.46	0.41	0.36	0.49	0.42	0.46
MAE	0.33	0.32	0.30	0.29	0.38	0.33	0.36
FinSentGPT_NYT model							
Correlation	0.61	0.62	0.58	0.63	0.59	0.75	0.53
MSE	0.09	0.09	0.07	0.10	0.08	0.06	0.13
RMSE	0.30	0.30	0.27	0.32	0.28	0.24	0.36
MAE	0.17	0.17	0.16	0.19	0.17	0.14	0.20

Table reports US inflation financial sentiment model accuracy findings. Reported are the results of a linear regression model with TF-IDF vectorization of text (top panel), and a fine-tuned GPT model — FinSentGPT_NYT (bottom panel). Training for both models is on news stories mentioning US inflation in the New York Times from January to June 2022. Testing is on news stories between July 2022 and December 2022. Correlation is Spearman's Rank Correlation Test, MSE is Mean Squared Errors, RMSE is Root Mean Squared Errors. MAE is Mean Average Errors.

3. Results

We start with the finance news study. Fig. 3 displays the monthly average financial sentiment towards inflation using the Barbaglia et al. (2022) method. Sentiment is notably negative at the year's beginning and remains subdued throughout.

Table 1 presents the FinSentGPT_NYT model findings for this dataset, with Fig. 4 providing a visual representation. The table's top panel shows the text-based linear regression machine learning model results, while the bottom panel displays FinSentGPT_NYT findings. Overall and monthly accuracy scores for the six tested months are reported, revealing no decrease in accuracy over time.

Focusing on overall results, FinSentGPT_NYT outperforms the linear regression model on all measures. The Spearman Rank Correlation is 0.61, indicating good correlation with the underlying financial sentiment model, and the MAE score demonstrates an average 0.17 difference between FinSentGPT and the underlying model for average sentiment scores.

Examining the individual datapoints reveals skewness in score accuracy, with over 40% of individual sentiment scores exhibiting less than a 0.05 absolute difference between the Barbaglia et al. (2022) method and FinSentGPT_NYT. Removing the top 10% of score differences would increase the correlation to 0.79. This outlier finding suggests the model could be iterated for closer correlation by examining sentences with large differences and providing additional training examples to rectify underlying errors. GPT fine-tuning allows the addition of new training examples.

Our judgement of the ability of the FinSentGPT_NYT model to contextually understand financial sentiment in sentences relies primarily on prior studies showing that it does contextually understand text (Tur & Traum, 2022). However, our FinSentGPT_NYT model itself only returns sentiment scores. By way of additional investigation, albeit constrained by article space, we ask GPT to explain the reasoning it would use for a selection of financial sentiment use case examples.

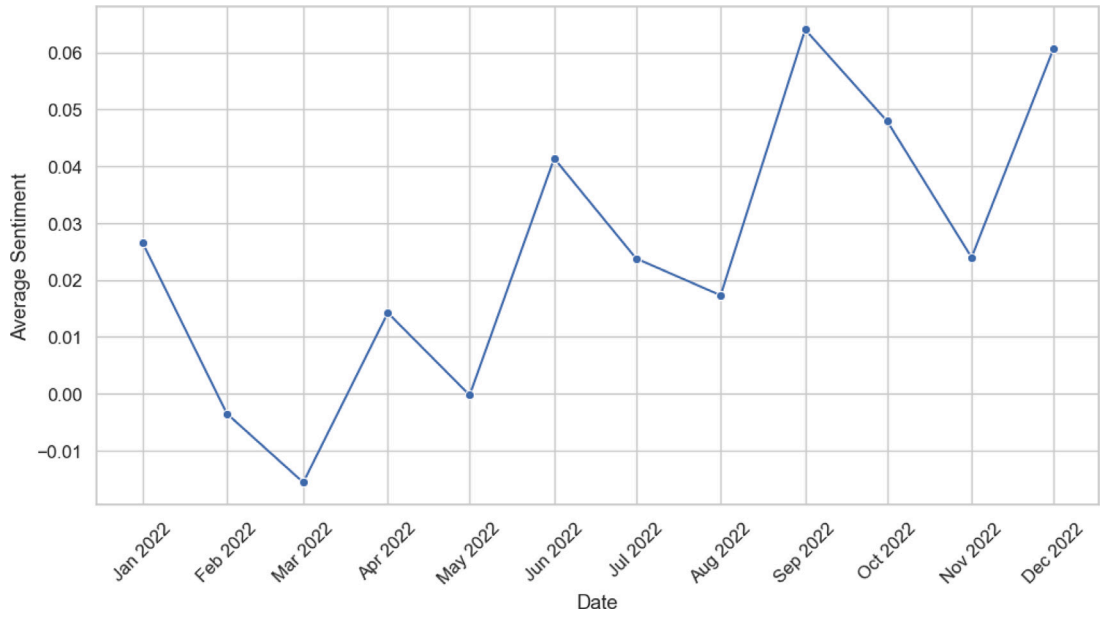

Fig. 3. Financial sentiment towards US domestic inflation for 2022 calculated from New York Times news articles.

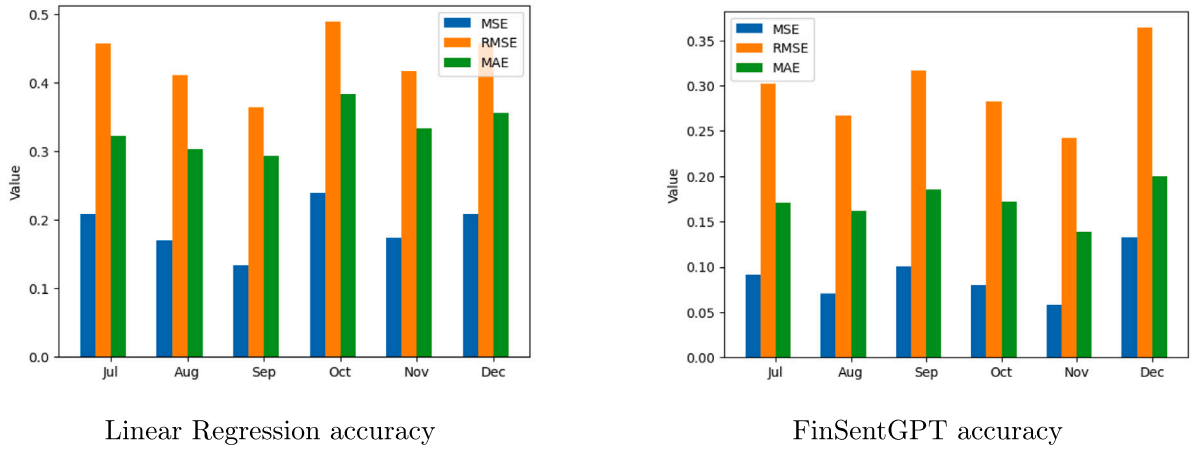

Fig. 4. Comparison of Linear regression vs. GPT accuracy for US finance news sentiment prediction.

Table 2
Inferring contextual understanding by FinSentGPT.

Sentiment	Input sentence	GPT contextual explanation
Negative (−0.97)	'Shelter inflation picked up sharply in august, with a rent measure climbing at its fastest pace since the 1980s'.	'Rent prices increasing rapidly suggests a negative outlook for inflation'.
Neutral (0.00)	'The decrease on overall inflation came as some durable goods, like household appliances, televisions and luggage, became cheaper, and as prices for financial services and insurance eased'.	'The sentiment score of 0 likely indicates that the overall impact of the decrease in inflation was neutral, as the decrease in prices for durable goods and financial services was balanced out'.
Positive (+0.82)	'Bank officials have said they are determined to bring inflation down to their 2 percent target and will use higher interest rates to do so'.	'The score likely reflects the model's positive sentiment towards the bank's determination to bring inflation down to their target'.

Table reports GPT3 output explaining the context of sample financial sentiment score determination from the FinSentGPT_NYT model. The actual sentiment score for each sentence is in brackets in the Sentiment column. The prompt to GPT includes the sentence and the prompt and the request "what is the likely reason for such a score based on your understanding of GPT3 reasoning?"

These are reported in Table 2 and do appear to show contextual understanding. Ideally, future versions of the GPT model would allow a more definitive investigation of these aspects of model generation.

The multi-language study results are reported in Table 3, with Spearman correlation scores visualized in Fig. 5 using the FinSentGPT_ECB2 model. Results are reported for French, German, Spanish, and Portuguese.

FinSentGPT_ECB2's performance is impressive considering its lack of exposure to non-English training text, suggesting the limited training examples for FinSentGPT_ECB1 hindered performance. However, FinSentGPT_ECB2's accuracy is lower than the English-to-English sentiment prediction in the economic news study. The MAE for FinSentGPT_ECB2 indicates an average 0.34 score variation from the underlying sentiment model.

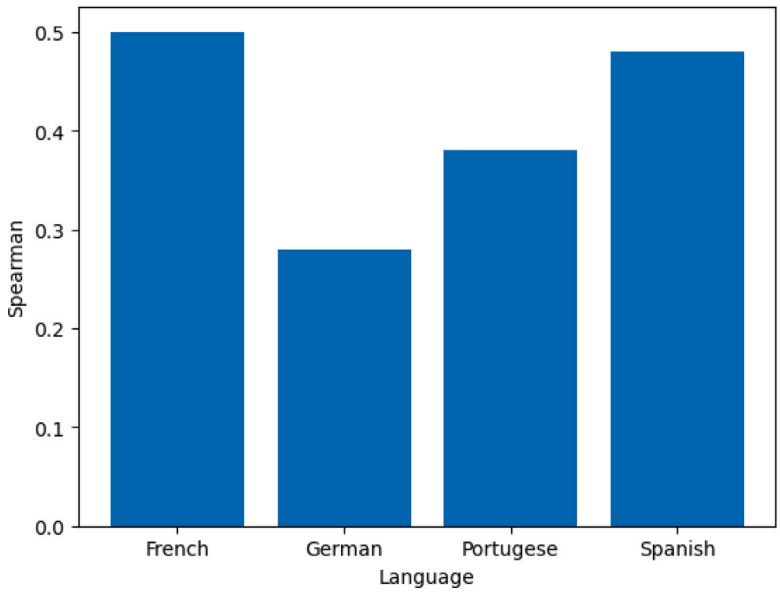

Fig. 5. Spearman correlations for ECB non-English language financial sentiment prediction.

Table 3
ECB multi-language sentiment prediction.

	French	German	Portugese	Spanish
EconSentGPT_ECB1 model				
Correlation	0.22	0.34	0.33	0.36
MSE	0.14	0.17	0.15	0.19
RMSE	0.38	0.41	0.39	0.44
MAE	0.23	0.22	0.22	0.22
EconSentGPT_ECB2 model				
Correlation	0.50	0.28	0.38	0.48
MSE	0.19	0.21	0.19	0.17
RMSE	0.43	0.46	0.44	0.41
MAE	0.31	0.36	0.36	0.32

Table reports ECB Monetary Policy Decisions economic sentiment model accuracy findings for unseen non-English ECB sentences containing translated versions of the keyword 'inflation' for 2022. All variables as defined in Table 1. Top panel shows results for FinSentGPT_ECB1 - a fine-tuned GPT model based only on prior ECB English-language sentence sentiment. Bottom Panel shows results for FinSentGPT_ECB2 - a model trained on both prior ECB English-language sentence sentiment and New York Times inflation news sentence sentiment.

With FinSentGPT_ECB2, French and Spanish both yield around 0.50 for Spearman's correlation, while German and Portuguese are lower. This discrepancy may result from GPT's higher use of training data in Spanish and French, the 4th and 5th most spoken languages globally, compared to Portuguese (9th) and German (12th).[2] This observation bodes well for multi-language sentiment prediction from more advanced GPT models, such as GPT4, when they allow fine-tuning.

4. Conclusions

This initial exploration of large-language AI models' effectiveness in financial sentiment prediction demonstrates their capability for accurate sentiment understanding, essential for policymakers and industry. Our FinSentGPT models offer insight into how we can derive contextual sentiment within a single language and understand economic sentiment across languages, promising a new era of global sentiment analysis.

FinSentGPT, or similar models, hold significant potential for diverse applications within the financial landscape. For instance, real-time market sentiment tracking across news and social platforms could equip investors, traders, and financial institutions with valuable insights for data-driven decision-making. Further, FinSentGPT's ability to identify sentiment shifts around specific assets, sectors, or broader economic conditions could aid in risk assessment, potentially acting as an early warning system for market volatility or emerging opportunities. Additionally, the capacity to generate synthetic financial text with targeted sentiment profiles could revolutionize areas such as report writing, scenario simulation, or the creation of tailored marketing materials. More broadly, it appears that a modern GPT model can approximate understanding of an underlying sentiment model based on merely being given the outcomes of that model (the sentiment scores attached to sentences, without being told the underlying generation process).

It is important to acknowledge the limitations inherent in our study. Firstly, FinSentGPT models represent an early-stage exploration with room for further refinement. Due to technical constraints, we relied on the GPT-3 class of model, whereas fine-tuning with the more powerful GPT-3.5 (or even GPT-4 class models, promised on the development roadmaps of current major LLM creators) could potentially yield even greater accuracy in sentiment analysis. We know that higher-general intelligence models, such as GPT-4 tend to generally outperform on most tasks (OpenAI, 2023), so fine-tuning on more powerful models should introduce greater sentiment intelligence. Secondly, our training data relies on an underlying sentiment generative model, potentially introducing biases. Future research should investigate techniques enabling direct sentiment perception by GPT models. Additionally, a more extensive study encompassing a broader range of languages is necessary to confirm the true generalizability of our findings. Lastly, our research focused primarily on monetary policy-related sentiment, leaving a wider landscape of financial sentiment dimensions, such as investor sentiment or sector-specific analyses (Fatouros, Soldatos, Kouroumali, Makridis, & Kyriazis, 2023; Zhang, Yang, Zhou, Ali Babar, & Liu, 2023), open for further exploration.

This study offers several promising avenues for future research. Exploration into mitigating biases within training data or developing techniques for direct sentiment understanding by large-language models will be vital for model accuracy and fairness. Investigating methods to tailor models such as FinSentGPT to specific financial subdomains such as commodities, equities, or foreign exchange has the potential to significantly enhance its precision in practical sentiment understanding. A large-scale exploration of FinSentGPT's sentiment analysis capabilities across a diverse range of languages is crucial for establishing a truly global financial sentiment engine. Finally, combining FinSentGPT

[2] Source: https://www.ethnologue.com/insights/ethnologue200/.

with structured financial data or knowledge graphs could unlock richer insights and novel applications.

Data availability

No.

References

Ash, E., & Hansen, S. (2022). Text algorithms in economics. *Annual Review of Economics*.

Barbaglia, L., Consoli, S., & Manzan, S. (2022). Forecasting with economic news. *Journal of Business & Economic Statistics*, 1–12.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2021). Sentiment in central banks' financial stability reports. *Review of Finance*, *25*(1), 85–120.

Degani, T., & Tokowicz, N. (2010). Semantic ambiguity within and across languages: An integrative review. *Quarterly Journal of Experimental Psychology*, *63*(7), 1266–1303.

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, Article 103662.

Ejara, D. D., Krapl, A. A., O'Brien, T. J., & Ruiz de Vargas, S. (2020). Local, global, and international CAPM: For which countries does model choice matter? *Journal of Investment Management*, 18–04.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications*, *14*, Article 100508.

Gu, S., Zhang, L., Hou, Y., & Song, Y. (2018). A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics* (pp. 774–784).

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., et al. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597.

Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019). Sentiment analysis with NLP on Twitter data. In *2019 international conference on computer, communication, chemical, materials and electronic engineering* IC4ME2, (pp. 1–4). IEEE.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., et al. (2023). How good are GPT models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Korinek, A. (2023). *Language models and cognitive automation for economic research*: *National bureau of economic research working paper 30957*, National Bureau of Economic Research.

Li, X. (2020). When financial literacy meets textual analysis: A conceptual review. *Journal of Behavioral and Experimental Finance*, *28*, Article 100402.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, *48*, 499–527.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE Access*, *8*, 131662–131682.

OpenAI (2023). GPT-4 technical report.

Picault, M., Pinter, J., & Renault, T. (2022). Media sentiment on monetary policy: Determinants and relevance for inflation expectations. *Journal of International Money and Finance*, *124*, Article 102626.

Tur, A., & Traum, D. (2022). Comparing approaches to language understanding for human-robot dialogue: An error taxonomy and analysis. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5813–5820).

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*(3), 1–34.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., & Choi, Y. (2021). Reframing human-AI collaboration for generating free-text explanations. arXiv preprint arXiv:2112.08674.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X.-Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance* (pp. 349–356).

Zhou, G. (2018). Measuring investor sentiment. *Annual Review of Financial Economics*, *10*, 239–259.