China Accounting and Finance Review

Entity-specific text classification with large language models: two applications on financial and brand news

Laiyi Meng and P. Paul Wang
Faculty of Business, School of Accounting and Finance, The Hong Kong Polytechnic University, Hong Kong, China
Huidi Lu
Saïd Business School, University of Oxford, Oxford, UK, and
W. Yuna Yang and Suhas Vijayakumar
Michael Smurfit Graduate Business School, University College Dublin, Dublin, Ireland

Received 4 December 2025
Revised 16 March 2026
Accepted 2 April 2026

Abstract

Purpose – This paper evaluates the performance of off-the-shelf large language models (LLMs) on two distinct entity-specific classification challenges in empirical economic research: assessing sentiment toward individual financial assets and identifying brand involvement during a product-harm crisis. Multiple assets or brands may appear in the same article, thus rendering document-level analysis of little use when the goal is to extract information at a granular (specific asset or brand) level.

Design/methodology/approach – In this paper, we attempt to benchmark recent Large Language Models (LLMs), such as OpenAI's GPT, on entity-specific text classification tasks. We illustrate their performance in two distinct applications. First, we demonstrate that LLMs can identify asset-specific sentiments in *Wall Street Journal (WSJ)* financial news. Second, we evaluate the effectiveness of these models in identifying brand involvement in a product-harm crisis, using a news corpus of the 2008 Chinese infant milk formula scandal. To assess the performance of these models, we compared machine coding results with human annotations. We calculated the F1 score, which helps measure how well a model works by balancing two key factors: precision (i.e. how often the model's positive predictions are correct) and recall (i.e. how many of the actual positive cases the model correctly identifies), providing a holistic evaluation of the models' overall performance. The two applications jointly illustrate that LLM-based entity-specific text classification works robustly in different scenarios and with different languages.

Findings – Our results demonstrate that LLMs can achieve high accuracy in entity-specific classification while maintaining simplicity and cost-effectiveness.

Originality/value – This research is of interest to various stakeholders. Using these new models, investors can leverage asset-level sentiment to make better decisions in complex markets. Lenders could benefit from improved risk evaluation models that incorporate more precise data related to a specific company. Managers can clinically tease apart sentiment towards their brands vis-à-vis others. For researchers, this research adds to the existing literature on the use of machine learning in text classification, specifically on extracting entity-specific information. Previous studies in this area have shown that machine learning models can be used to analyse texts in finance and psychology at the document level. Our study extends this line of work to an entity-focused case, which yields better

Funding: This work was supported by Research Centre of Digital Economy at The Hong Kong Polytechnic University (Grant no. P0047874).

China Accounting and Finance Review
Emerald Publishing Limited
e-ISSN: 2307-3055
p-ISSN: 1029-807X
DOI 10.1108/CAFR-12-2025-0269

insights. Our findings suggest increasing capabilities of machine learning models and show new opportunities for efficient and large-scale solutions in a field that has suffered from cost and scalability factors.

Keywords Entity identification, Large language models, News, Product-harm crisis, Sentiment, Text classification

Paper type Research article

1. Introduction

Accurate text classification is crucial for extracting meaningful economic insights from news coverage (Bach, Kern, Bonnay, & Kalaora, 2022; Feyzollahi & Rafizadeh, 2025). In finance, positive or negative sentiments conveyed in news reports can influence investor behaviour and shape market trends (Algieri, Leccadito, Sicoli, & Tunaru, 2025; Consoli, Tiozzo Pezzoli, & Tosetti, 2022; Liu & Shi, 2025). Similarly, during a product-harm crisis, the media informs the public about the implicated brands. Unlike the cases dealing with a single entity (e.g. a brand), many news articles simultaneously discuss multiple entities with varying sentiments. For example, a single news report may cover the performance of various financial instruments, such as stocks, bonds, or cryptocurrencies (Tang, Yang, Huang, Tam, & Tang, 2023). In another case, a news article may reference two brands, one of which is directly involved in a crisis and the other that is unaffected. However, existing text classification techniques often lack the precision and efficiency required to extract entity-specific information in such cases (Goyal, Gupta, & Kumar, 2018). Addressing this limitation presents a significant research opportunity, enabling more granular insights into market dynamics for varying parties of interest such as investors, managers, researchers, and regulators.

In this paper, we attempt to validate the usefulness of recent Large Language Models (LLMs), such as OpenAI's GPT, on entity-specific text classification tasks. Unlike previous studies that are domain-specific (Guo, Xu, & Yang, 2023; Wu *et al.*, 2023), we illustrate their performance in two distinct applications. First, we demonstrate that LLMs can identify asset-specific sentiments in *Wall Street Journal (WSJ)* financial news. Second, we evaluate the effectiveness of these models in identifying brand involvement in a product-harm crisis, using a news corpus of the 2008 Chinese infant milk formula scandal (Chan, Griffiths, & Chan, 2008). To assess the performance of these models, we compared machine coding results with human annotations. We calculated the F1 score, which helps measure how well a model works by balancing two key factors: precision (i.e. how often the model's positive predictions are correct) and recall (i.e. how many of the actual positive cases the model correctly identifies), providing a holistic evaluation of the models' overall performance (Guo *et al.*, 2023; Rathje *et al.*, 2024). The two applications jointly illustrate that LLM-based entity-specific text classification works robustly in different scenarios and with different languages.

This research is of interest to various stakeholders. Using these new models, investors can leverage asset-level sentiment to make better decisions in complex markets. Lenders could benefit from improved risk evaluation models that incorporate more precise data related to a specific company. Managers can clinically tease apart sentiment towards their brands vis-à-vis others. For researchers, this research adds to the existing literature on the use of machine learning in text classification, specifically on extracting entity-specific information. Previous studies in this area have shown that machine learning models can be used to analyse texts in finance (Baraniak & Sydow, 2021; Du, Xing, & Cambria, 2023; Guo *et al.*, 2023; Huang, Wang, & Yang, 2023; Wu *et al.*, 2023) and psychology (Rathje *et al.*, 2024) at the document level. Our study extends this line of work to an *entity*-focused case, which yields better insights. Our findings suggest increasing capabilities of machine learning models and show new opportunities for efficient and large-scale solutions in a field that has suffered from cost and scalability factors.

2. First application: classifying asset-specific sentiment

2.1 Background and data

Investors, researchers and market regulators often employ machine learning to extract insights from financial texts (Bochkay, Brown, Leone, & Tucker, 2023). Notably, financial news is

often used as a key proxy for investor sentiment, as media outlets tailor content to align with audience beliefs (Elejalde, Ferres, & Schifanella, 2019). The sentiment embedded in news text can often provide strong predictive signals for market return and trading volume (Garcia, 2013; Tetlock, 2007).

Following previous research (Obaid & Pukthuanthong, 2022), we collected 187,736 articles from the *WSJ* in the following sections: "Business", "Economy", "Markets", "Politics" and "Opinion", published between 1 December 2012 and 31 December 2022. Through the application programming interface (API), we fed raw text as the user prompt and used the system prompt in Supplementary Material Part SM1 to instruct GPT (Auger & Saroyan, 2024) to provide, for each asset type, one of the following sentiment classifications: positive, negative, mixed, or none (if the specific asset was not mentioned), along with the reasoning for that answer. Table 1 shows the breakdown of the sentiment labels provided by the GPT-4o model for each asset class. In general, more articles discuss stocks, compared to the other two asset classes. Importantly, many articles discuss these assets at the same time. We illustrate the accuracy of the GPT models in the following section.

2.2 Benchmarking results

Using financial news, we compared several models from the GPT family on asset-specific sentiment classification. To assess the performance of these models, we randomly selected 500 *WSJ* articles and asked two research assistants to independently code the assets mentioned as well as their sentiments in each article. The initial agreement rate between the coders was above 80%. Any inconsistencies were resolved through discussion. This forms the "ground truth" of our performance evaluation.

Because we have multiple assets and sentiment labels for each asset type, we calculate the weighted F1 score, the micro F1 score, and the macro F1 score, for each asset type, to incorporate a holistic and robust evaluation of the model's sentiment classification performance (Uysal, 2016).

We first calculated the precision and recall (Ting, 2010) for each of the four sentiment labels (i.e. positive, negative, mixed, and none) i and asset type j. Precision measures the proportion of correctly (i.e. consistent with human annotations) classified instances (TP) of sentiment label i for asset type j out of all instances classified as i. Some of these predictions could be false positives (FP), meaning they were incorrectly classified as label i:

$$\text{precision}_{ij} = \frac{\text{TP}_{ij}}{\text{TP}_{ij} + \text{FP}_{ij}} \tag{1}$$

Recall measures the proportion of correctly identified instances (TP) out of all actual instances of that label. This includes cases that were falsely classified as a different label (FN):

Table 1. GPT-4o Coded sentiments of news articles from wall street journal

	Bond	Stock	Crypto
Positive	4,965	48,349	329
Negative	5,026	72,624	213
Mixed	282	6,544	222
None	177,463	60,219	186,972

Note(s): $N = 187{,}736$

$$\text{recall}_{ij} = \frac{\text{TP}_{ij}}{\text{TP}_{ij} + \text{FN}_{ij}} \tag{2}$$

Based on these values, we calculate the F1 score for each individual label *i* using the following formula:

$$\text{F1}_{ij} = 2 \times \frac{\text{precision}_{ij} \times \text{recall}_{ij}}{\text{precision}_{ij} + \text{recall}_{ij}} \tag{3}$$

We then calculated the weighted F1 score, the micro F1 score and the macro F1 score that averages the four sentiment labels for each asset category. The weighted F1 score is calculated by taking the weighted average of the F1 scores for each label, where the weights w_{ij} correspond to the frequency of each label in the dataset:

$$\text{Weighted } F1_j = \frac{\sum_i \left(F1_{ij} \times w_{ij}\right)}{\sum_i w_{ij}} \tag{4}$$

The micro F1 score sums up the results of all classes and then calculates the precision and recall of all samples and then the F1 score. This method is more suitable when label imbalances exist, as it gives equal weight to all instances:

$$\text{Micro } F1_j = 2 \times \frac{\text{global precision}_j \times \text{global recall}_j}{\text{global precision}_j + \text{global recall}_j} \tag{5}$$

where

$$\text{global precision}_j = \frac{\sum_i \text{TP}_{ij}}{\sum_i \text{TP}_{ij} + \sum_i \text{FP}_{ij}} \tag{6}$$

and

$$\text{global recall}_j = \frac{\sum_i \text{TP}_{ij}}{\sum_i \text{TP}_{ij} + \sum_i \text{FN}_{ij}} \tag{7}$$

Macro F1 score computes the F1 score for each label and then takes the mean of all the scores. This method treats all classes equally, regardless of their frequency in the dataset:

$$\text{Macro F1}_j = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_{ij} \tag{8}$$

Here, $N = 4$, which equals the total number of available labels (i.e. positive, negative, mixed, and none).

Table 2 presents benchmark results for sentiment classification of different financial assets (bonds, stocks, and cryptocurrencies), comparing performances of various models: GPT-3.5 (gpt-3.5-turbo-0125), GPT-4 (gpt-4-0125-preview), and GPT-4o (gpt-4o-2024-08-06), using different input contexts (title and first paragraph vs. full text). Overall precision and recall metrics are reported in Supplementary Material Part SM2.

While GPT-3.5 (Panel a of Table 2) performs well for bonds (weighted-F1: 0.920) and crypto classifications (0.977), its performance on stock-related sentiment is notably poor (weighted-F1: 0.364, micro-F1: 0.299). This suggests that GPT-3.5 has difficulty handling the complexity of stock sentiment, possibly due to higher variability in stock-related financial

Table 2. Benchmarks for sentiment classification of different GPT models

(a) GPT-3.5 using title and first paragraph	Bond	Stock	Crypto
weighted-F1	0.920	0.364	0.977
micro-F1	0.899	0.299	0.980
macro-F1	0.340	0.269	0.745

(b) GPT-4 using title and first paragraph	Bond	Stock	Crypto
weighted-F1	0.944	0.724	0.994
micro-F1	0.916	0.586	0.996
macro-F1	0.364	0.389	0.749

(c) GPT-4o using title and first paragraph	Bond	Stock	Crypto
weighted-F1	0.967	0.823	0.995
micro-F1	0.962	0.721	0.999
macro-F1	0.422	0.424	0.750

(d) GPT-4o using full text	Bond	Stock	Crypto
weighted-F1	0.950	0.766	0.995
micro-F1	0.938	0.667	0.999
macro-F1	0.355	0.401	0.750

news, where context-dependent nuances (e.g. earnings reports, macroeconomic factors) play a crucial role. GPT-4 shows significant improvement over GPT-3.5 (Panel b of Table 2), particularly in stock-specific sentiment (weighted-F1 increases to 0.724), nearly doubling performance compared to GPT-3.5. This suggests that GPT-4 has a better ability to capture sentiment from stock-related news, likely benefiting from enhanced contextual understanding. GPT-4o performs best overall (Panel c of Table 2), which is expected as this is the most advanced model and can disambiguate complex sentiment signals. Its performance matches that of earlier models that were specifically designed for such tasks but were also complex to implement (Tang *et al.*, 2023). Interestingly, allowing GPT-4o to process the full text of articles rather than just the title and first paragraph results in slightly worse performance across all asset classes (Panel d of Table 2). This indicates that the most critical sentiment information in a news article is often contained in the beginning.

Across all versions, the models achieved good performance. More advanced models lead to better sentiment analysis performance. Notably GPT-4o using only title and first paragraph outperforms full-text processing, emphasizing the importance of concise, high-signal text in financial sentiment classification. This has practical implications for building efficient financial AI tools, as shorter text processing reduces computational costs while maintaining high accuracy.

2.3 Model performance changes by number of entities

This section examines how relative model accuracy changes with the number of entities, in order to show that text classification indeed becomes increasingly difficult when using less advanced models and when multiple entities are mentioned in the same article.

We estimated a simple linear regression (N = 500) where the dependent variable is the relative number of errors in sentiment classifications between GPT-3.5 and GPT-4o, both

applied to the title and first paragraph of *each article*, as described in the previous section. For each article, we counted the number of errors made by GPT-3.5 and GPT-4o and took their difference. GPT-3.5 serves as a benchmark for more traditional models, while GPT-4o represents state-of-the-art performance, expected to handle nuanced language more effectively. The independent variable is the manually coded number of entities in each article. We also included the quadratic term to capture potential non-linear effects.

Our regression results are attached in the Supplementary Material Part SM3. Figure 1 plots the fitted relationship. The results show a clear upward trend: as the number of entity increases, GPT-3.5 makes disproportionately more errors relative to GPT-4o. In other words, the complexity of multi-entity classification highlights where recent LLMs offer the largest improvements.

2.4 Forecasting experiment

We performed a forecasting experiment (Algieri *et al.*, 2025; Tang *et al.*, 2023) to illustrate the benefits of incorporating asset-level sentiment in asset return predictions. For each asset class, we used common proxies for their returns. For bonds, we used the daily returns of the bond ETF under the symbol BOND. For stocks, we used the daily returns of S&P 500 ETF under the symbol SPY. We obtained the return data between 1 December 2012 and 31 December 2022 of both ETFs from WRDS. For cryptocurrency, we used the Bitcoin daily returns available on Investing.com (https://www.investing.com/crypto/bitcoin/historical-data). Returns are more useful than the absolute price of the asset because it is a primary driver for the investor to decide whether to invest. For bonds and stocks, returns are calculated for each working day; for cryptocurrency, the returns are calculated for each calendar day because it can be traded 24/7.

We estimated regressions of contemporaneous asset returns against four different predictor sets. First, we established a baseline auto-regressive model where the returns of an asset at day t depend solely on their lagged values at day $t-1$. Second, we extended this specification by including aggregated document-level sentiment of day $t-1$, computed as the sum of individual document-level sentiments. Each article's overall sentiment is determined by the majority classification of its asset-level sentiments: positive if most are positive, negative if most are negative, and neutral otherwise. Sentiments are assigned numerical values: 1 for positive, -1 for negative, and 0 for neutral. The daily news sentiment is then computed as the sum of document-level sentiments across all articles. This simulates the case where one only has the tool to identify document-level sentiment. Third, we replaced this general sentiment measure with asset-specific sentiment, identified using GPT-4o. Here, the sentiment score for a given asset is obtained by summing its sentiment across all articles published on day $t-1$. Finally, we refined this model further by separately quantifying the occurrences of positive and negative sentiments on day $t-1$.

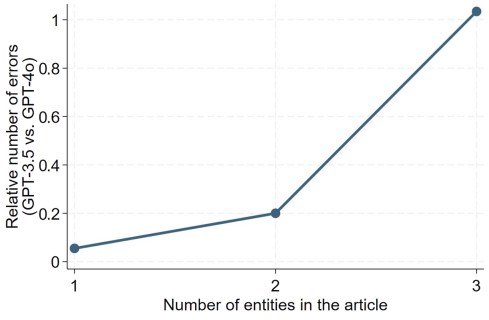

Figure 1. Changes in relative model performance by number of entities

Figure 2 shows the forecasting root mean square error (RMSE) of each approach by asset type. Lower values of RMSE represent a more accurate prediction (Tang *et al.*, 2023). The different levels of RMSE for each asset reflect economic magnitude of the different assets, with bonds having the lowest value of RMSE compared with stocks and crypto due to generally more stable and predictable returns. In all cases, incorporating asset-specific sentiments separated into positives and negatives improved forecast accuracy the most. This further emphasizes that mere document-level overall sentiment is insufficient due to the presence of multiple entities within the same news article, highlighting the practical value of asset-specific sentiment information.

3. Second application: classifying product-harm crisis involvement of specific brands

3.1 Background and data

The previous section focused on sentiment analysis using financial texts, demonstrating the capability of LLMs to identify asset-specific sentiments in news articles that may discuss multiple assets simultaneously. In this section, we evaluate the performance of the models on a different classification application, brand social listening, where a focal news article may concurrently reference multiple brands.

Brand social listening is a critical task, particularly during product-harm crises, as brands must address consumer complaints promptly (Chung *et al.*, 2022a, b; Tang & Guo, 2015). In such situations, news reports often mention several brands within the same article. Therefore, an approach capable of disentangling information related to each individual brand is essential. To assess model performance in these situations, we took the 2008 Chinese infant milk formula scandal as a case study and analysed Chinese news collected from WiseNews. Beyond demonstrating the models' performances, this example also illustrates its robustness when applied to a non-English dataset.

The 2008 Chinese infant milk formula scandal represents one of the most serious product-harm crises of recent decades, involving multiple Chinese brands (particularly Sanlu, Yili, Guangming, Nanshan, Scient, and Yashili) that adulterated their products with melamine to artificially boost protein content (Chung *et al.*, 2022a, b). In line with previous research leveraging news archives to assess media coverage during crises, we conducted keyword-based searches using the Chinese names of each brand in the WiseNews database (Qin, Strömberg, & Wu, 2018). The search period was restricted to March 2008 through February 2009, encompassing six months before and six months after the crisis. We retrieved a total of 1,829 news articles. Two native Chinese-speaking research assistants independently read and coded each article for reports of quality issues associated with each brand (Kopalle, Fisher,

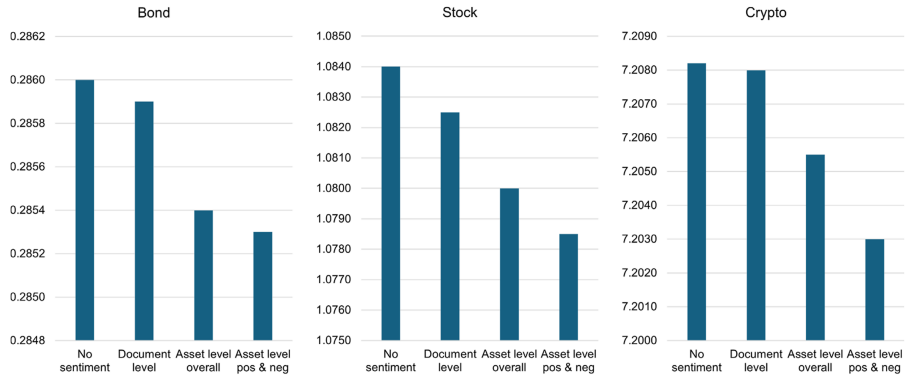

Figure 2. Forecasting errors (RMSE) using different levels of news sentiments (%)

Sud, & Antia, 2017). The initial inter-coder agreement rate was close to 90%. Discrepancies were resolved through discussion.

3.2 Benchmarking results

Similar to the previous study, we fed raw news text as the user prompt and used the system prompt provided in the Supplementary Material (Part SM4) to instruct GPT to identify whether the article mentioned quality problems for each of the six brands, along with the reasoning for that answer. We calculated micro F1 scores for each brand to benchmark the model's performance (Guo *et al.*, 2023).

Table 3 presents the micro-F1 scores for classifying brand-related news during the product-harm crisis using GPT models. Overall precision and recall metrics are reported in Supplementary Material Part SM5. Consistent with the first application, GPT-4o demonstrated strong performance across brands (Column 1), with Sanlu achieving the highest micro-F1 score (0.813), followed by Guangming (0.754) and Yili (0.690). The other brands, Scient (0.667), Yashili (0.588) and Nanshan (0.571), exhibited slightly lower classification performance. Notably, these brands were less prominently featured in the crisis, with few news articles reporting their quality issues (Nanshan: 3, Scient: 6, Yashili: 35), rendering their scores more susceptible to idiosyncratic errors.

To assess efficiency trade-offs, we compared GPT-4o to its smaller counterpart, GPT-4o mini (Hurst *et al.*, 2024). While GPT-4o mini (Column 2) consistently yielded lower micro-F1 scores, its performance was comparable, underscoring its potential utility in cost-sensitive situations. Overall, the results suggest GPT models provide robust classification performances.

4. Discussion

Our findings revealed that LLMs can perform entity-level text classification with remarkable accuracy. This capability opens up new avenues for efficient solutions in fields traditionally constrained by high costs and scalability. By leveraging LLMs, organizations can significantly enhance precision, rendering these tools invaluable across a variety of domains. For instance, investors can gain more nuanced insights into market sentiments; Managers can monitor brand perceptions with unprecedented granularity; Researchers across diverse fields, such as social sciences and computational linguistics, can uncover patterns and trends that were previously inaccessible. More importantly, by lowering barriers to advanced text analysis, LLMs democratize access to sophisticated tools, empowering smaller organizations and individual researchers to compete globally (Garrett, 2024). By evaluating performance across distinct empirical settings and documenting key design choices, we also offer some methodological guidance for researchers considering the use of large language models in entity-level applications.

Table 3. Micro-F1 scores for classifying brand news during a product-harm crisis

Brand	(1) GPT-4o	(2) GPT-4o mini
Sanlu	0.813	0.792
Guangming	0.754	0.707
Yili	0.690	0.652
Scient	0.667	0.600
Yashili	0.588	0.644
Nanshan	0.571	0.364

Although we have showcased the utility of these models, some limitations and future improvements should be acknowledged. First, the models used in our studies were off-the-shelf. Although these models are highly capable, fine-tuning them on domain-specific corpora could further enhance their performances and tailor their outputs to particular use cases (Abdurahman *et al.*, 2024). Secondly, as LLMs are projected to evolve at a rapid pace, their accuracy, interpretability, and potential applications are likely to expand. This underscores the importance of ongoing validation and exploration as new versions of LLMs emerge. Third, the finding that classification based on the title and first paragraph outperforms classification using the full text is intriguing, and it merits deeper theoretical and empirical investigation. Several mechanisms may plausibly explain this effect, including noise dilution, prompt-length constraints, context-window trade-offs, and differences in informational or sentiment density.

It would also be crucial to address potential ethical and operational considerations (Head, Jasper, McConnachie, Raftree, & Higdon, 2023), such as data privacy, bias in training data, and the environmental costs of large-scale computations. By addressing these challenges responsibly, the development and deployment of LLMs could align with ethical best practices while maximizing their potential impact. The use of LLMs for entity-level text classification is just one example of their transformative potential. With ongoing advancements and responsible implementation, these models will stand poised to revolutionize how we interact with and extract meaning from textual data across industries and disciplines (Garrett, 2024).

Supplementary material

The supplementary material for this article can be found online.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., . . . Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), 245. doi: 10.1093/pnasnexus/pgae245.

Algieri, B., Leccadito, A., Sicoli, D., & Tunaru, D. (2025). Combining density forecast accuracy tests: An application to agricultural, energy, and metal commodities. *Journal of the Royal Statistical Society - Series C: Applied Statistics*, *74*(3), 598–616. doi: 10.1093/jrsssc/qlae069.

Auger, T., & Saroyan, E. (2024). Overview of the openai apis. In *Generative ai for web development: Building web applications powered by openai apis and next.js* (pp. 87–116). Berkeley, CA: Apress.

Bach, R. L., Kern, C., Bonnay, D., & Kalaora, L. (2022). Understanding political news media consumption with digital trace data and natural language processing. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, *185*(Supplement_2), S246–S269. doi: 10.1111/rssa.12846.

Baraniak, K., & Sydow, M. (2021). A dataset for sentiment analysis of entities in news headlines (sen). *Procedia Computer Science*, *192*, 3627–3636. doi: 10.1016/j.procs.2021.09.136.

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next?. *Contemporary Accounting Research*, *40*(2), 765–805. doi: 10.1111/1911-3846.12825.

Chan, E., Griffiths, S., & Chan, C. (2008). Public-health risks of melamine in milk products. *The Lancet*, *372*(9648), 1444–1445. doi: 10.1016/s0140-6736(08)61604-9.

Chung, S., Shin, D., & Park, J. (2022a). Predicting firm market performance using the social media promoter score. *Marketing Letters*, *33* (4), 545–561, 10.1007/s11002-022-09615-w.

Chung, T., Tam, I. Y. S., Lam, N. Y. Y., Yang, Y., Liu, B., He, B., Xu, J., Yang, Z., Zhang, L., Cao, J. N., & Lau, L. T. (2022b). Non-targeted detection of food adulteration using an ensemble machine-learning model. *Scientific Reports*, *12*(1), 20956. doi: 10.1038/s41598-022-25452-3.

Consoli, S., Tiozzo Pezzoli, L., & Tosetti, E. (2022). Neural forecasting of the Italian sovereign bond market with economic news. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*(Supplement_2), S197–S224. doi: 10.1111/rssa.12813.

Du, K., Xing, F., & Cambria, E. (2023). Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Transactions on Management Information Systems*, *14* (3), 1–24. doi: 10.1145/3580480.

Elejalde, E., Ferres, L., & Schifanella, R. (2019). Understanding news outlets' audience-targeting patterns. *EPJ Data Science*, *8*(1), 1–20. doi: 10.1140/epjds/s13688-019-0194-8.

Feyzollahi, M., & Rafizadeh, N. (2025). The adoption of large language models in economics research. *Economics Letters*, *250*, 112265. doi: 10.1016/j.econlet.2025.112265.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267–1300. doi: 10.1111/jofi.12027.

Garrett, A. D. (2024). The devil, the detail, and the data. *Journal of the Royal Statistical Society - Series A: Statistics in Society. 187* (4), 857–878. doi: 10.1093/jrsssa/qnae063.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, *29*, 21–43. doi: 10.1016/j.cosrev.2018.06.001.

Guo, Y., Xu, Z., & Yang, Y. (2023). Is ChatGPT a financial expert? Evaluating language models on financial natural language processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 815–821). Available from: https://aclanthology.org/2023.findings-emnlp.58/

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, (178-179), 33–46. doi: 10.1002/ev.20556.

Huang, A. H., Wang, H., & Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841. doi: 10.1111/1911-3846.12832.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., . . . others (2024). GPT-4o system card. arXiv preprint arXiv:2410.21276.

Kopalle, P. K., Fisher, R. J., Sud, B. L., & Antia, K. D. (2017). The effects of advertised quality emphasis and objective quality on sales. *Journal of Marketing*, *81*(2), 114–126. doi: 10.1509/jm.15.0353.

Liu, T., & Shi, Y. (2025). News sentiment and investment risk management: Innovative evidence from the large language models. *Economics Letters*, *247*, 112124. doi: 10.1016/j.econlet.2024.112124.

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, *144*(1), 273–297. doi: 10.1016/j.jfineco.2021.06.002.

Qin, B., Strömberg, D., & Wu, Y. (2018). Media bias in China. *The American Economic Review*, *108* (9), 2442–2476. doi: 10.1257/aer.20170947.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. doi: 10.1073/pnas.2308950121.

Tang, C., & Guo, L. (2015). Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (ewom) communication. *Marketing Letters*, *26*(1), 67–80. doi: 10.1007/s11002-013-9268-8.

Tang, Y., Yang, Y., Huang, A., Tam, A., & Tang, J. (2023). Finentity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15465–15471).

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62* (3), 1139–1168. doi: 10.1111/j.1540-6261.2007.01232.x.

Ting, K. M. (2010). Precision and recall. In *Encyclopedia of machine learning* (p. 781). Boston, MA: Springer US.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications*, *43*, 82–92. doi: 10.1016/j.eswa.2015.08.050.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., . . ., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint*, arXiv:2303.17564 .

Corresponding author
P. Paul Wang can be contacted at: paul.wang@polyu.edu.hk