EDITOR: Erik Cambria, cambria@ntu.edu.sg

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

A Retrieval-Augmented Multiagent System for Financial Sentiment Analysis

Kelvin Du, *Nanyang Technological University, 639798, Singapore*
Yazhi Zhao, *Visa Inc Singapore, 068895, Singapore*
Rui Mao, *Nanyang Technological University, 639798, Singapore*
Frank Xing, *National University of Singapore, 117418, Singapore*
Erik Cambria, *Nanyang Technological University, 639798, Singapore*

Financial sentiment analysis (FSA) has seen substantial advancements with the use of large language models (LLMs). Previous research highlighted the effectiveness of retrieval-augmented generation (RAG) and multiagent LLMs for FSA as these approaches alleviate the problems of hallucination, a lack of factual knowledge, and limited complex problem-solving capability. Despite this, the interplay and potential synergies between these two methods remain largely unexplored. This study presents a notable leap forward by introducing a retrieval-augmented multiagent system (RAMAS) to enhance LLM-based FSA performance. An RAMAS is specifically designed to deepen understanding of the critical factors that are inherent in FSA and mimic human-like consensus-making processes by adaptively learning from semantically similar few-shot samples and engaging in conversations among the generator, discriminator, and arbitrator agents. Our evaluation of RAMASs demonstrates improved accuracy and F1-score across multiple established FSA benchmark datasets.

The advent of Web 2.0 has led to a dramatic increase in the quantity and diversity of available information resources in the last decade. Thus, there is the pressing challenge of transforming this vast reservoir of information into computationally manageable formats. Financial sentiment analysis (FSA) has risen to prominence over the past decade, presenting a more dynamic and robust approach compared to conventional survey-based methods. FSA has emerged as a potent tool for understanding investor sentiment and forecasting financial markets.[1]

It is noteworthy that sentiment analysis exhibits domain specificity, which is particularly pronounced within the domain of finance due to factors such as the concentration of financial topics, utilization of highly specialized language,[2] and presence of distinctive cognitive patterns across different market environments.[3]

FSA differs significantly from general sentiment analysis in various aspects, the first of which is its frequent encounters with metaphorical expressions in financial communications, where figurative language is used to express emotions or describe market scenarios. For instance, a common metaphor like "the market is riding a bull" symbolically describes a strong and rising market trend, adding complexity to the sentiment analysis of financial texts. Second, the financial domain places a premium on precision and brevity.

Digital Object Identifier 10.1109/MIS.2025.3544912
Date of current version 11 April 2025.

Professionals in this arena employ concise language to efficiently convey intricate information. Rather than resorting to lengthy descriptions such as "The company experienced a substantial increase in revenue and a corresponding improvement in profitability," financial analysts often opt for succinct statements like "The company posted robust revenue growth, driving higher profits." This demand for brevity necessitates that FSA discerns sentiments embedded within compact sentence structures. Third, the financial industry employs a unique lexicon replete with specialized terminology and jargon, each bearing specific connotations.

A comprehensive understanding of these terms is indispensable for the accurate interpretation and analysis of financial texts in the context of sentiment analysis. For instance, the price-to-earnings (P/E) ratio represents a fundamental financial metric employed to assess a company's valuation, where a high P/E ratio may signify elevated expectations for future earnings. Moreover, unlike general sentiment analysis, which focuses predominantly on textual content, financial texts often integrate qualitative and quantitative data. This requires FSA to not only parse the language used in financial texts but also analyze and interpret the numerical data in the context of the surrounding text, enabling a comprehensive understanding of sentiment. Additionally, FSA often depends on the directionality of events or trends, highlighting the need for contextual awareness. For example, the word *profit* can have positive or negative connotations depending on the context. A rise in profit usually indicates positive sentiment, whereas a decline is typically seen as negative. From this perspective, models developed for general purposes cannot be effectively applied to the finance sector without undergoing domain-specific adaptation.

The recent rise in research interest surrounding large language models (LLMs) is due largely to their sophisticated capabilities in natural language understanding and generation. LLMs are more commonly trained for general purposes, and training domain-specific LLMs like BloombergGPT[4] requires substantial resources. Hence, harnessing general-purpose LLMs to comprehend and identify distinctive knowledge within financial texts, particularly those conveying sentiment, is pivotal in the realm of FSA. We posit that to fully leverage LLMs' potential for FSA, it is essential to design an approach for selecting learning examples and designing prompts that facilitate a deeper understanding of financial texts.

We propose a retrieval-augmented multiagent system (RAMAS) that is designed to strategically select few-shot learning examples, enabling LLMs to adaptively learn and perform FSA. Additionally, the RAMAS orchestrates conversational agents, including a generator, discriminator, and arbitrator, which mimic human dialogue. The agents engage in interactions that are similar to human conversations, effectively interpreting and analyzing the nuances and complexities of financial texts, aiming to enhance the performance of FSA. The efficacy of our proposed framework is validated through extensive experimentation on two widely recognized benchmark datasets. On average, our system exceeds the baseline LLM performance by 29% for Generative Pre-trained Transformer 3.5 (GPT-3.5) Turbo and 10% for GPT-4o in accuracy across datasets, showcasing its superior performance compared to existing approaches. GPT-3.5 Turbo with the RAMAS achieves even better performance than the plain GPT-4o. We also demonstrate the effectiveness of various modules in our ablation study.

The contributions of this article are summarized as follows:

- We conducted an extensive study from both zero- and few-shot learning perspectives to evaluate the efficacy of LLMs in the context of FSA. Our investigation revealed that few-shot learning significantly enhances the performance of FSA and that it can be further improved by identifying semantically similar learning examples.
- We proposed an RAMAS that includes a semantic retriever, an adaptive learner, and generator-discriminator-arbitrator conversable agents to enhance the capabilities of LLMs while performing FSA. This system showcases competitive performance on publicly available datasets, highlighting its effectiveness. Notably, our proposed framework with GPT-3.5 Turbo achieved better performance than that of plain GPT-4o.
- We demonstrated that both retrieval-augmented generators and generator-discriminator-arbitrator conversable agents can enhance the performance of FSA by using LLMs, as evidenced by our ablation study and a series of case studies.

RELATED WORK

The potential and adaptability of LLMs in the context of FSA have garnered increasing attention. Generally, the first type of study focuses on assessing LLMs in FSA. In recent studies, Fatouros et al.[5] adopted a zero-shot prompting approach to evaluate various ChatGPT prompts on a carefully curated dataset of forex (foreign exchange)-related news headlines. The performance was assessed using several metrics, including precision, recall, and F1-score, and the results demonstrated superior performance compared to FinBERT.[6] conducted a thorough comparative analysis

to examine the effectiveness of zero shot, fine-tuning LLMs, and few-shot learning techniques in the context of FSA. In particular, in-context learning is adopted with a focus on the GPT-3.5 Turbo model, and the fine-tuning is performed on Flan-T5.

The study highlights the remarkable capabilities of LLMs, even smaller models, in both fine-tuning and in-context learning for FSA task. Another type of study is to evaluate the reasoning capabilities of LLMs in performing FSA. Specifically, Du et al.[7] conducted an empirical study to evaluate the reasoning capabilities of LLMs in performing FSA. Specifically, six key financial attributes related to semantic, numerical, temporal, comparative, causal, and risk related are identified. This study revealed shortcomings in the reasoning capabilities of LLMs concerning these attributes for FSA. Finally, researchers are exploring other techniques such as retrieval-augmented generation (RAG) from financial knowledge sources to enhance the performance of FSA. For example, Zhang et al.[8] presented a framework that integrates a retrieval-augmented mechanism with LLMs specifically for FSA. The framework consists of two key components: instruction-finetuned LLMs and a retrieval-augmented component. The performance metrics, specifically accuracy and F1 score, show an enhancement ranging between 15% and 48%, underscoring the efficacy of the framework in FSA. Bloomberg introduced BloombergGPT,[4] an LLM that is tailored to financial contexts and has demonstrated superior performance in financial natural language processing (NLP) tasks, including sentiment analysis, question answering, and named entity recognition, among others, further advancing the capabilities of FSA.

Previous research has primarily explored the potential and adaptability of LLMs in FSA tasks and evaluated the reasoning capabilities of LLMs in this context.[7] It draws inspiration from the human annotation process for the FSA dataset as outlined by Malo et al.,[2] in which even humans may disagree on certain text's sentiment. The final annotations were based on financial experts' consensus. Similarly, we advance this research stream by proposing an RAMAS with adaptive few-shot learning to enhance the capabilities of LLMs for FSA tasks. This system is designed to strategically select few-shot examples for in-context learning, drawing knowledge from financial experts. It facilitates conversations between agents that mimic human interactions, enhancing the learning process.

METHODOLOGY

The architecture and algorithm of the RAMAS with adaptive few-shot learning are shown in Figure 1 and Table 1, respectively. Comprising three principal components, e.g., RAG, prompt engineering, and conversational agents with LLM, the RAMAS provides explicit instructions to LLMs for conducting sentiment analysis. To enhance the capabilities of LLMs in understanding the overtly expressed sentiments and the more subtle cues that might indicate a particular sentiment in financial texts, the RAMAS performs few-shot learning from effectively selected samples and makes decisions via agent conversations.

RAG

The RAG module includes a query encoder, embedding model, vector database, and semantic search function. Each financial text is vectorized by using the text-embedding-three-large model from OpenAI embeddings and then stored in Chroma DB, an open source vector database. Simultaneously, the query undergoes

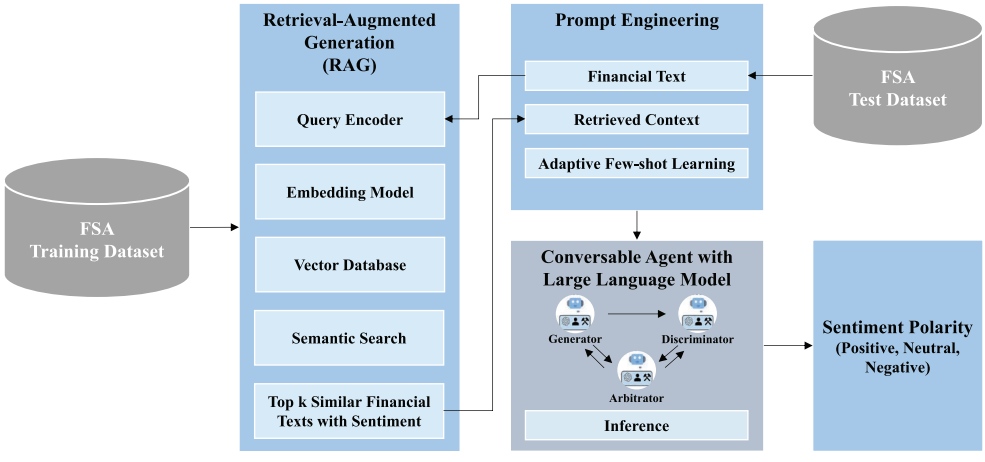

FIGURE 1. Proposed RAMAS for FSA.

TABLE 1. Algorithm for RAMAS.

Require: Query q, Corpus D
Ensure: Sentiment Outcome S

RAG

Initialize `query_encoder`, `embedding_model`, `vector_database`, `semantic_search`.
`vector_database` ← Load FSA Training Dataset D.
`encoded_query` ← `query_encoder.Encode` (q).
`top_k_documents` ← `semantic_search` (`encoded_query`, `vector_database`, $k = 6$, metric="Euclidean").

Adaptive Learning

Initialize `adaptive_learning_llm`.
For each `sentence` in `documents` do
  `predicted_polarity` ← `adaptive_learning_llm.PredictPolarity` (`sentence`).
  `actual_polarity` ← GetPolarity(`sentence`).
  If `predicted_polarity` ≠ `actual_polarity` then
    `adaptive_learning_llm.LearnFromMistake` (`sentence`, `actual_polarity`).
  end if
end for
`sentiment` ← `adaptive_learning_llm.PredictPolarity` (`new_sentence`).

Conversable Agents

Initialize `generator_agent`, `discriminator_agent`, `arbitrator_agent`.
`generator_results` ← [ ].
`discriminator_results` ← [ ].
`generator_results` ← `generator_agent.adaptive_learning_llm` (`top_k_documents` [0 : 3], q).
`discriminator_results` ← `discriminator_agent.adaptive_learning_llm` (`top_k_documents` [4 : 6], q, `generator_results`).
`S` ← `arbitrator_agent.DetermineFinalSentiment` (`generator_results`, `discriminator_results`).
Output `S`.

embedding, and a semantic similarity search is conducted by using Euclidean distance to retrieve the top k-similar financial texts with sentiment. Specifically, the retrieval mechanism fetches the top k-relevant documents based on a query q. The relevance scores $s(d, q)$ are computed for each document d in a corpus D, and the top k documents are selected by $\text{semantic\_search}_k(D, q) = \text{argmax}_k \{s(d, q) \mid d \in D\}$. In our setup, we selected $k = 6$ The first three samples are allocated to the generator agent, while the remaining three are designated for the discriminator.

Adaptive Learning

The adaptive learning process functions by systematically reviewing each sentence within the provided examples, assessing their sentiment polarity and comparing it with the actual sentiment polarity provided. Through this iterative process, any mistakes made are identified and learned from, allowing for repeated attempts until achieving 100% accuracy. Following this, the system then applies its learned knowledge to determine whether the sentiment of a new sentence is positive, neutral, or negative.

Conversable Agents

The generator-discriminator-arbitrator conversation module is constructed by using the multiagent conversation framework, in which the generator agent is tasked with conducting FSA using retrieval-augmented adaptive few-shot learning, generating sentiment polarity along with explanations. Meanwhile, the discriminator agent's role is to review and validate the FSA from the generator, ensuring its accuracy against its defined evaluation criteria, using an additional set of samples selected by the retrieval-augmented adaptive few-shot learning framework. The arbitrator agent makes a determination of the unified sentiment through discussion and consensus building among the three agents.

EXPERIMENTAL SETUP

Datasets

We conduct experiments using two widely recognized datasets for FSA: PhraseBank and Twitter Financial News. The PhraseBank dataset, developed by Malo et al.,[2] includes 4846 news items classified into positive, neutral, and negative sentiments by 16 financial market experts from an investor's perspective. This dataset is organized into four subsets based on the consensus level among annotators: 100%, 75%, 66%, and 50%. For our study, we used the datasets with 100% and 50% agreement as benchmarks. The Twitter Financial News dataset comprises 11,932 tweets in English related to finance, categorized into bearish, bullish, and neutral sentiments. We split the dataset using an 80/20 train-test ratio and conducted five iterations with different random seeds.

Baseline Models

Lexicon-Based Methods

The financial lexicons used as benchmark resources include *Henry's Financial Dictionary* (*HFD*), *Loughran and McDonald* (*LM*), and FinSenticNet. *HFD* is known as one of the first dictionaries tailored specifically for the financial domain. It comprises 104 positive and 85 negative words, aimed primarily at assessing the tone in earnings press releases, which are crucial in the communication between firms and investors.[9] LM is the most extensively used sentiment word list in FSA, crafted from the analysis of company annual reports. It includes 2355 negative words and 354 positive words, alongside 19 strong modal words, 27 weak

modal words, 297 uncertainty-related words, 904 litigious words, and 184 constraining words.[10] The latest addition to this suite of resources is FinSenticNet, a concept-level lexicon that was introduced in a recent study by Du et al.[11] FinSenticNet has shown superior performance over both general and financial-specific lexicons in various evaluations, highlighting its effectiveness in accurately capturing and analyzing sentiment within the financial domain.

Learning-Based Methods

Linearized phrase-structure,[2] the hierarchical sentiment classifier,[12] and ULMFit[15] are adopted as benchmark learning-based models. In addition, recent progress in the field of FSA has been greatly propelled by the introduction of transformer-based encoder architectures like BERT. The finance domain-specific version of BERT, known as *FinBERT*,[13,14] is trained on a diverse array of financial texts from sources such as Reuters Corpora, Yahoo Finance, Reddit Finance, corporate reports, earnings call transcripts, and analyst reports, marking a significant advance in FSA research. We adopted the FinBERT that was presented by Araci[13] and Huang et al.,[14] which are publicly available, as the baseline models.

LLM-Based Methods

We adopted OpenAI's GPT-3.5 Turbo -1106 and the latest flagship model, GPT-4o-2024-05-13, as baseline models, with a temperature of zero. GPT-4o represents a significant advancement by OpenAI, boasting real-time reasoning capabilities across audio, vision, and text. It stands as their most sophisticated system yet, providing responses that are not only safer but also more valuable across various contexts. Although the differences between GPT-3.5 and GPT-4 may not be immediately apparent in casual conversations, they become evident when tackling tasks of considerable complexity. GPT-4 distinguishes itself with superior reliability, creativity, and nuanced instruction handling compared to its predecessor, GPT-3.5 Turbo.

RESULT AND ANALYSIS

Accuracy and the macro-averaged F1-score are adopted as primary evaluative criteria for FSA, and the results are presented in Table 2. A thorough analysis reveals the effectiveness of the RAMAS framework, surpassing a wide array of lexicons and machine learning techniques. The RAMAS framework has attained results that are not only competitive but also comparable to those achieved by sophisticated transformer encoder architectures, such as FinBERT, highlighting its efficacy and potential in the field. Furthermore, our findings demonstrate that the RAMAS framework significantly boosts the performance of GPT-3.5, as evidenced by a substantial increase in accuracy scores from 0.7757 to 0.9417 on the PhraseBank 100% Agree dataset, from 0.6668 to 0.823 on the PhraseBank 50% Agree dataset, and from

TABLE 2. Comparison with baseline methods on FSA benchmark datasets. Boldface indicates the top two results.

Method	Model	PhraseBank: 100% Agree		PhraseBank: 50% Agree		Twitter Financial News	
		Accuracy	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*
Lexicon-based Method	*HFD*[9]	0.8105	0.7714	0.6976	0.6266	0.6415	0.5095
	LM[10]	0.6444	0.3688	0.6244	0.502	0.5971	0.4604
	FinSenticNet[11]	0.7619	0.7216	0.6624	0.6215	0.6	0.5269
Learning-based Method	LPS[2]	0.79	0.8	0.71	0.71	—	—
	HSC[12]	0.83	0.86	0.71	0.76	—	—
	ULMFit[13]	0.93	0.91	0.83	0.79	—	—
	FinBERT^{a}[13]	**0.97**	**0.95**	**0.86**	**0.84**	—	—
	FinBERT^{b}[14]	0.9169	0.897	0.7926	0.7514	0.7483	0.6612
LLM-based Method	GPT-3.5 Turbo	0.7757	0.8039	0.6668	0.7021	0.5518	0.5698
	GPT-3.5 Turbo (with FAP)[7]	0.9187	0.9174	0.7783	0.7718	0.7324	0.7057
	GPT-3.5 Turbo (RAMAS)	0.9417	0.9263	0.823	0.805	**0.8041**	**0.7645**
	GPT-4o	0.9284	0.9275	0.7894	0.796	0.5979	0.6045
	GPT-4o (RAMAS)	**0.9505**	**0.9387**	**0.836**	**0.8163**	**0.7682**	**0.7436**

0.5518 to 0.8041 on the Twitter financial news dataset. Similar trends are observed with GPT-4o, which showcased notable improvements in accuracy, rising from 0.9284 to 0.9505, 0.7894 to 0.836, and 0.5979 to 0.7682 on the PhraseBank 100% and 50% Agree datasets as well as the Twitter financial news dataset, respectively. Notably, GPT-3.5 Turbo with the RAMAS achieves better performance than GPT-4o, highlighting the strength of the RAMAS framework in enhancing model capabilities. However, it is noteworthy that the improvement in GPT-4o is comparatively less significant, indicating that GPT-3.5's ability to identify distinctive features within financial texts conveying sentiment is weaker than that of GPT-4o. This observation suggests that the more sophisticated nature of GPT-4o diminishes the relative contribution of the RAMAS framework's enhancement, resulting in a less pronounced effect compared to models with lower inherent reasoning power.

ABLATION STUDY

To ascertain the efficacy of various elements within the proposed framework, an ablation study was undertaken, with the corresponding results presented in Table 3. First, conversable agents improved FSA performance across datasets. Furthermore, consistent observation across benchmark datasets is the substantial improvement in FSA performance attributed to adaptive few-shot learning, regardless of whether samples are selected randomly or through RAG. Furthermore, the learning samples chosen by the RAG module demonstrate a significant performance boost compared to the randomly selected samples. For example, with GPT-3.5 Turbo, the accuracy increases by 15% on average, achieving scores of 0.7924 versus 0.9192, 0.7125 versus 0.7995, and 0.6878 versus 0.7949 on the PhraseBank 100% Agree, 50% Agree and Twitter Financial News datasets, respectively, underscoring the effectiveness of the RAG module within our framework. Similarly, for GPT-4o, there are 3% improvements in accuracy, with scores of 0.9302 versus 0.947, 0.8074 versus 0.8362, and 0.7426 versus 0.7656 on the same datasets, respectively. Finally, the integration of generator-discriminator-arbitrator conversation further elevates performance levels, demonstrating that conversational agents that mimic human dialogue can enhance the performance of FSA. Agents, by engaging in interactions that are similar to human conversations, can more effectively interpret and analyze the nuances and complexities of financial discourse, leading to more accurate assessments of sentiment in financial texts.

CASE STUDY

We conducted a series of case studies to demonstrate the functionality of the RAMAS. The results are presented in Table 4. In the first example provided in Table 4, the sentence, "The recent troubles simply make NETeller cheaper." is negative. However, GPT-3.5 Turbo misclassified it as positive, while GPT-3.5 Turbo with retrieval-augmented few-shot learning labeled it as neutral. In terms of GPT-3.5 Turbo with the RAMAS, the generator produced neutral sentiment, which is the same as GPT-3.5 Turbo with retrieval-augmented few-shot learning. However, the discriminator pointed out that the sentiment is actually negative. The explanation is that the mention of "troubles" implies a negative impact on the company, leading to a negative sentiment. Eventually, the arbitrator stands corrected and concludes that the sentiment of the sentence, "The recent troubles simply make NETeller cheaper." is negative.

TABLE 3. Ablation study on FSA benchmark datasets. Boldface indicated the top two result.

Model	PhraseBank - 100% Agree		PhraseBank - 50% Agree		Twitter Financial News	
	Accuracy	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*
GPT-3.5 Turbo	0.7757	0.8039	0.6668	0.7021	0.5518	0.5698
GPT-3.5 Turbo (with conversable agent only)	0.8114	0.8263	0.7835	0.7756	0.6259	0.6269
GPT-3.5 Turbo (with random few-shot only)	0.7924	0.7269	0.7125	0.631	0.6878	0.5449
GPT-3.5 Turbo (with RAG few-shot only)	0.9192	0.9015	0.7995	0.7815	0.7949	0.753
GPT-3.5 Turbo (with RAMAS)	0.9417	0.9263	0.823	0.805	**0.8041**	**0.7645**
GPT-4o	0.9284	0.9275	0.7894	0.796	0.5979	0.6045
GPT-4o (with conversable agent only)	0.9465	0.9355	0.8298	0.8135	0.715	0.6942
GPT-4o (with random few-shot only)	0.9302	0.9219	0.8074	0.7891	0.7426	0.7223
GPT-4o (with RAG few-shot only)	**0.947**	**0.9414**	**0.8362**	**0.8278**	0.7656	0.7448
GPT-4o (with RAMAS)	**0.9505**	**0.9387**	**0.836**	**0.8163**	**0.7682**	**0.7436**

TABLE 4. Case Studies for RAMAS.

Financial Text: The recent troubles simply make NETeller cheaper.	**Financial Text: A plumbing business has announced it is sponsoring a professional darts player.**
Sentiment polarity: Negative	Sentiment polarity: Neutral
Top *k* financial text:	Top *k* financial text:
1) The company also said that it would lower the price of development projects by roughly one-third compared to last November: neutral.	1) The contract involves refurbishing the bathrooms of 189 units as well as replumbing their water and sewer pipes: neutral.
2) The company's advantage in pricing can be seen in the significantly lower prices offered by the new unmanned stations: positive.	2) Finnish plastic pipe maker Uponor is selling its Irish and U.K. infrastructure businesses to private equity company 3i: neutral.
3) The company said that it estimates making a slight profit thanks to cost-cutting measures: positive.	3) Trading under the name *Velta UK*, a former Uponor brand, the company has been Uponor's long-term partner in supplying Velta-branded systems, particularly for the commercial and industrial building sector in the United Kingdom and internationally: neutral.
4) "We can say that the number of deals has become normalized by the moment, and therefore, we have decided to devalue our apartments now on sale," he added: negative.	4) Wayne Greensmith, Fiskars Brands UK senior brand manager, said, "We have launched this as we want to give something back to the community": neutral.
5) A realignment of interests in the sector is clearly under way: positive.	5) Kershaw takes up the position with immediate effect from her previous role as marketing manager of Sankey Home & Garden Products: neutral.
6) However, the rapidly increasing costs and the strengthening of the euro were challenging for the whole industry, and they impacted on our results: negative.	6) Finnish plumbing and heating systems supplier Uponor Corporation (OMX Helsinki: UNR1V) said on Tuesday (12 August) that its board of directors has authorized management to proceed with a companywide cost-reduction program: positive.
GPT-3.5 Turbo: positive	GPT-4o: positive
GPT-3.5 Turbo (with RAG few-shot only): neutral	GPT-4o (with RAG few shot only): neutral
GPT-3.5 Turbo (with RAMAS):	GPT-4o (with RAMAS):
Generator: The sentiment of the sentence, "The recent troubles simply make NETeller cheaper." is neutral.	Generator: The sentiment of the sentence "A plumbing business has announced it is sponsoring a professional darts player" is neutral.
Discriminator: The sentiment of the sentence, "The recent troubles simply make NETeller cheaper." is actually negative. The mention of "troubles" implies a negative impact on the company, leading to a negative sentiment: "Thank you for your input."	Discriminator: correct. The new sentence is neutral. It simply states a fact about a plumbing business sponsoring a professional darts player without conveying a clear positive or negative sentiment.
Arbitrator: "I stand corrected. Thank you for pointing that out." The sentiment of the sentence, "The recent troubles simply make NETeller cheaper." is negative.	Arbitrator: The new sentence is neutral as it states a fact without conveying a clear positive or negative sentiment.

In the second example provided in Table 4, the sentence, "A plumbing business has announced it is sponsoring a professional darts player." is neutral. However, GPT-4o misclassified it as positive, but GPT-4o with retrieval-augmented few-shot learning correctly labeled it as neutral. As for GPT-4o with the RAMAS, the generator produced neutral sentiment which is the same as GPT-4o with retrieval-augmented few-shot learning. The discriminator agrees with the generator that the sentiment is neutral. The explanation is that the sentence simply states a fact about a plumbing business sponsoring a professional darts player without conveying a clear positive or negative sentiment. Finally, the arbitrator concluded that the sentiment of the sentence, "A plumbing business has announced it is sponsoring a professional darts player." is neutral as it states a fact without conveying a clear positive or negative sentiment.

CONCLUSION

A novel RAMAS was introduced to deepen the understanding of critical factors within FSA and enhance

LLMs' performance in this domain. This system strategically leverages RAG to select semantically similar samples for adaptive few-shot learning. Our findings indicate that choosing such samples produces superior results compared to random selection. Furthermore, adaptive learning, where LLMs learn from mistakes based on provided samples, enhances model performance. Additionally, the inclusion of generator-discriminator-arbitrator conversational agents further improves FSA performance through discussions between agents. Experimental results highlight that the RAMAS significantly boosts the performance of various LLMs across multiple benchmark datasets, underscoring the importance of providing LLMs with a comprehensive guidance framework to effectively apply their capabilities. In particular, the RAMAS surpasses transformer encoder architectures like FinBERT in terms of generalization and overall performance. Moreover, the GPT-3.5 Turbo with the RAMAS achieves even better performance than the plain GPT-4o. In summary, the RAMAS emerges as an innovative and potent multiagent framework empowering LLMs to excel in FSA, offering superior performance compared to existing methods.

REFERENCES

1. K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–42, 2024, doi: 10.1145/3649451.
2. P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, 2014, doi: 10.1002/asi.23062.
3. R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with MetaPro," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2023, pp. 1211–1216, doi: 10.1109/ICDM58522.2023.00150.
4. S. Wu et al., "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.
5. G. Fatouros et al., "Transforming sentiment analysis in the financial domain with ChatGPT," *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art.no. 100508, doi: 10.1016/j.mlwa.2023.100508.
6. S. Fatemi and Y. Hu, "A comparative analysis of fine-tuned LLMs and few-shot learning of LLMs for financial sentiment analysis," 2023, *arXiv:2312.08725*.
7. K. Du, F. Xing, R. Mao, and E. Cambria, "An evaluation of reasoning capabilities of large language models in financial sentiment analysis," in *Proc. IEEE Conf. Artif. Intell. (CAI)*, 2024, pp. 189–194, doi: 10.1109/CAI59869.2024.00042.
8. B. Zhang et al., "Enhancing financial sentiment analysis via retrieval augmented large language models," in *Proc. 4th ACM Int. Conf. AI Finance*, 2023, pp. 349–356, doi: 10.1145/3604237.3626866.
9. E. Henry, "Are investors influenced by how earnings press releases are written?" *Int. J. Bus. Commun.*, vol. 45, no. 4, pp. 363–407, 2008, doi: 10.1177/0021943608319388.
10. T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *J. Finance*, vol. 66, no. 1, pp. 35–65, 2011, doi: 10.1111/j.1540-6261.2010.01625.x.
11. K. Du, F. Xing, R. Mao and E. Cambria, "FinSenticNet: A concept-level lexicon for financial sentiment analysis," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, 2023, pp. 109–114, doi: 10.1109/SSCI52147.2023.10371970.
12. S. Krishnamoorthy, "Sentiment analysis of financial news articles using performance indicators," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 373–394, 2018, doi: 10.1007/s10115-017-1134-1.
13. D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.
14. A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Res.*, vol. 40, no. 2, pp. 806–841, 2023, doi: 10.1111/1911-3846.12832.
15. J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, 2018, pp. 328–339.

KELVIN DU received his Ph.D. degree in computing and data science from Nanyang Technological University, 639798, Singapore. Contact him at zidong001@e.ntu.edu.sg.

YAZHI ZHAO is a senior data scientist at Visa Inc Singapore, 068895, Singapore. Contact her at yazzhao@visa.com.

RUI MAO is a research scientist at Nanyang Technological University, 639798, Singapore. Contact him at rui.mao@ntu.edu.sg.

FRANK XING is an assistant professor at the National University of Singapore, 117418, Singapore. Contact him at xing@nus.edu.sg.

ERIK CAMBRIA is a professor at Nanyang Technological University, 639798, Singapore. Contact him at cambria@ntu.edu.sg.