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

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

A Retrieval-Augmented Multi-Agent System for Financial Sentiment Analysis

Kelvin Du, *Nanyang Technological University, Singapore*

Yazhi Zhao, *Visa Inc, Singapore*

Rui Mao, *Nanyang Technological University, Singapore*

Frank Xing, *National University of Singapore, Singapore*

Erik Cambria, *Nanyang Technological University, Singapore*

Abstract—Financial Sentiment Analysis (FSA) has seen substantial advancements with the use of Large Language Models (LLMs). Prior research highlighted the effectiveness of Retrieval-Augmented Generation (RAG) and Multi-Agent LLMs for FSA, as these approaches alleviate the problems of hallucination, 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 Multi-Agent System (RAMAS) to enhance LLM-based FSA performance. RAMAS is specifically designed to deepen the understanding of critical factors inherent in FSA and mimic human-like consensus-making processes by adaptively learning from semantically similar few-shot samples and engaging in conversations between the generator, discriminator, and arbitrator agents. Our evaluation of RAMAS 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 a 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, the utilization of highly specialized language [2], and the presence of distinctive cognitive patterns across different market environments [3]. FSA differs significantly from general sentiment analysis in various aspects. One key difference is its frequent encounter 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. Secondly, the financial domain places a premium on precision and brevity. 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

1541-1672 © 2025 IEEE
Digital Object Identifier 10.1109/MIS.2025.XXXXXXX

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. Thirdly, 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 predominantly focuses 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 to 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 negatively. 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 LLMs is largely due 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 the financial texts. We propose a Retrieval-Augmented Multi-Agent System (RAMAS) designed to strategically select few-shot learning examples, enabling LLMs to adaptively learn and perform FSA. Additionally, RAMAS orchestrates conversational agents, including a generator, discriminator, and arbitrator, which mimic human dialogue. The agents engage in interactions 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 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 RAMAS achieves even better performance than vanilla GPT-4o. We also demonstrate the effectiveness of various modules in our ablation study.

The contributions can be summarized as follows:

1) We conducted an extensive study from the zero-shot 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 can be further improved by identifying semantically similar learning examples.
2) We proposed a retrieval-augmented multi-agent system that includes a semantic retriever, an adaptive learner, and generator-discriminator-arbitrator conversable agents to enhance the capabilities of LLMs in 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.
3) We demonstrated that both retrieval-augmented generators and generator-discriminator-arbitrator conversable agents can enhance the performance of FSA using LLMs, 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, [5] adopted a zero-shot prompting approach to evaluate various ChatGPT prompts on a carefully curated dataset of forex-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, the in-context learning is adopted with a focus on 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, [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. Lastly, researchers are exploring other techniques such as retrieval-augmented generation from financial knowledge sources to enhance the performance of FSA. For example, [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 has introduced BloombergGPT [4], an LLM tailored for financial contexts, which has demonstrated superior performance in financial NLP tasks, including sentiment analysis, question answering, 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]. Drawing inspiration from the human annotation process for the FSA dataset as outlined by [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 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, which mimic human interactions, enhancing the learning process.

METHODOLOGY

The architecture and algorithm of RAMAS with adaptive few-shot learning are shown in Fig. 1 and Table 1, respectively. Comprising three principal components, e.g., Retrieval-Augmented Generation, Prompt Engineering, and Conversational Agents with LLM, RAMAS provides explicit instructions to LLMs for conducting sentiment analysis.

TABLE 1. Algorithm for RAMAS.

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

Retrieval Augmented Generation
    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.
```

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, RAMAS performs few-shot learning from effectively selected samples and makes decisions via agent conversations.

Retrieval Augmented Generation

The RAG module includes a query encoder, embedding model, vector database, and semantic search function. Each financial text is vectorized using OpenAI's text-embedding-3-large model and then stored in Chroma DB. Simultaneously, the query undergoes embedding, and a semantic similarity search is conducted 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 have selected $k = 6$. The first three samples are allocated to the generator agent, while the remaining three are designated for the discriminator.

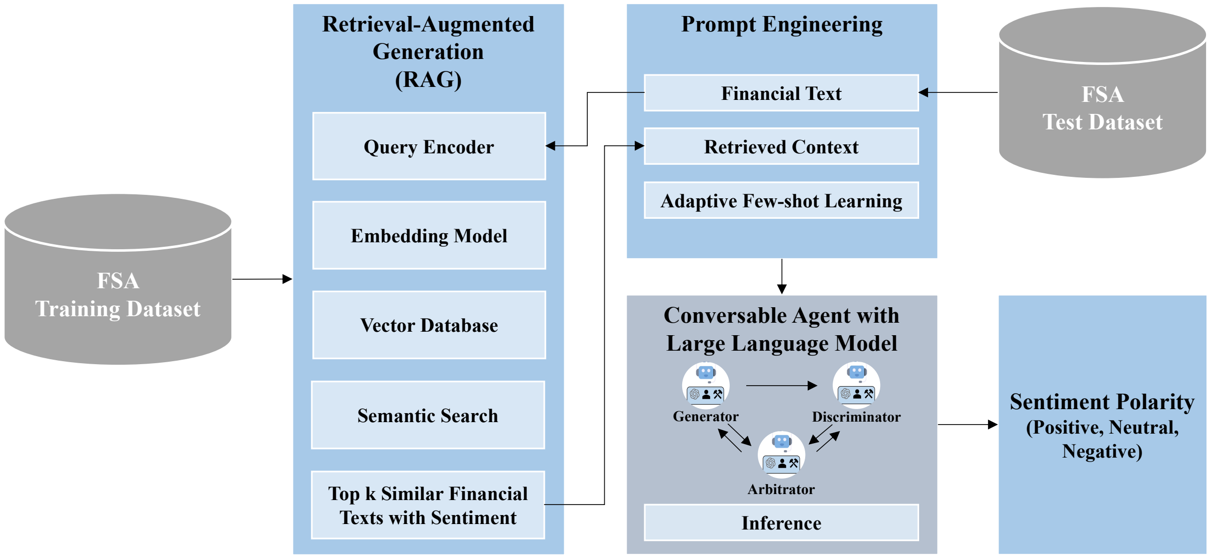

FIGURE 1. Proposed Retrieval-Augmented Multi-Agent System (RAMAS) for Financial Sentiment Analysis.

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 using the multi-agent 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 between 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 [2], includes 4,846 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 HFD, 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, primarily aimed 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 2,355 negative words, 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 introduced in a recent study by [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. The Linearized Phrase-Structure (LPS) [2], Hierarchical Sentiment Classifier (HSC) [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 the 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 presented by [13] and [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 0. 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. While 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 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, evidenced by a substantial increase in accuracy scores from 0.7757 to 0.9417 on the PhraseBank 100% Agree dataset, from 0.6668 to 0.8230 on the PhraseBank 50% Agree dataset, and from 0.5518 to 0.8041 on the Twitter financial news dataset. Similar trends are observed with GPT-4o, showcasing notable improvements in accuracy rising from 0.9284 to 0.9505, from 0.7894 to 0.8360, and from 0.5979 to 0.7682 on the PhraseBank 100% and 50% Agree datasets, as well as Twitter financial news dataset respectively. Notably, GPT-3.5-turbo with 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

In order to ascertain the efficacy of various elements within the proposed framework, an ablation study was undertaken, with the corresponding results are presented in Table 3. First, conversable agents have improved the FSA performance across datasets. Furthermore, consistently observed 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 random-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 0.9470, 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 con-

TABLE 2. Comparison with baseline methods on FSA benchmark datasets. Boldface indicated 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	0.5020	0.5971	0.4604
	FinSenticNet [11]	0.7619	0.7216	0.6624	0.6215	0.6000	0.5269
Learning-based Method	LPS [2]	0.7900	0.8000	0.7100	0.7100	-	-
	HSC [12]	0.8300	0.8600	0.7100	0.7600	-	-
	ULMFit [13]	0.9300	0.9100	0.8300	0.7900	-	-
	FinBERT^a [13]	**0.9700**	**0.9500**	**0.8600**	**0.8400**	-	-
	FinBERT^b [14]	0.9169	0.8970	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 (w/ FAP) [7]	0.9187	0.9174	0.7783	0.7718	0.7324	0.7057
	GPT-3.5-turbo (RAMAS)	0.9417	0.9263	0.8230	0.8050	**0.8041**	**0.7645**
	GPT-4o	0.9284	0.9275	0.7894	0.7960	0.5979	0.6045
	GPT-4o (RAMAS)	**0.9505**	**0.9387**	**0.8360**	**0.8163**	**0.7682**	**0.7436**

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 (w/ Conversable agent only)	0.8114	0.8263	0.7835	0.7756	0.6259	0.6269
GPT-3.5-turbo (w/ Random few-shot only)	0.7924	0.7269	0.7125	0.6310	0.6878	0.5449
GPT-3.5-turbo (w/ RAG few-shot only)	0.9192	0.9015	0.7995	0.7815	0.7949	0.7530
GPT-3.5-turbo (w/ RAMAS)	0.9417	0.9263	0.8230	0.8050	**0.8041**	**0.7645**
GPT-4o	0.9284	0.9275	0.7894	0.7960	0.5979	0.6045
GPT-4o (w/ Conversable agent only)	0.9465	0.9355	0.8298	0.8135	0.7150	0.6942
GPT-4o (w/ Random few-shot only)	0.9302	0.9219	0.8074	0.7891	0.7426	0.7223
GPT-4o (w/ RAG few-shot only)	**0.9470**	**0.9414**	**0.8362**	**0.8278**	0.7656	0.7448
GPT-4o (w/ RAMAS)	**0.9505**	**0.9387**	**0.8360**	**0.8163**	**0.7682**	**0.7436**

versational agents that mimic human dialogue can enhance the performance of FSA. Agents, by engaging in interactions 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 have conducted a series of case studies to demonstrate the functionality of 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 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 concluded that the sentiment of the sentence “The recent troubles simply make NETeller cheaper” is negative.

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 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. Lastly, 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 retrieval-augmented multi-agent system is introduced to deepen the understanding of critical factors within FSA and enhance LLMs’ performance in this domain. This system strategically leverages retrieval-augmented generation to select semantically similar samples for adaptive few-shot learning.

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: **1.** The company also said that it would lower the price of development projects by about one third compared with last November. Neutral **2.** The company's advantage in pricing can be seen in the significantly lower prices offered by the new unmanned stations. Positive **3.** The company said it estimates to make a slight profit thanks to cost-cutting measures. Positive **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 **5.** A realignment of interests in the sector is clearly under way . Positive **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	**Top-k Financial Text**: **1.** The contract involves refurbishing the bathrooms of 189 units as well as re-plumbing their water and sewer pipes. Neutral **2.** Finnish plastic pipe maker Uponor is selling its Irish and UK infrastructure businesses to private equity company 3i. Neutral **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 UK and internationally. Neutral **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.** Kershaw takes up the position with immediate effect from her previous role as marketing manager of Sankey Home & Garden Products. Neutral **6.** Finnish plumbing and heating systems supplier Uponor Corporation (OMX Helsinki: UNR1V) said on Tuesday (12 August) that its board of directors has authorised the management to proceed with a company-wide cost reduction program. Positive
GPT-3.5-turbo: Positive	**GPT-4o**: Positive
GPT-3.5-turbo (w/ RAG few-shot only): Neutral	**GPT-4o (w/ RAG few-shot only)**: Neutral
GPT-3.5-turbo (w/ RAMAS): **Generator**: The sentiment of the sentence "The recent troubles simply make NETeller cheaper." 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. **Arbitrator**: I stand corrected. Thank you for pointing that out. The sentiment of the sentence "The recent troubles simply make NETeller cheaper." is negative.	**GPT-4o (w/ RAMAS)**: **Generator**: The sentiment of the sentence "A PLUMBING business has announced it is sponsoring a professional darts player." is neutral. **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**: The new sentence is neutral, as it states a fact without conveying a clear positive or negative sentiment.

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 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, RAMAS surpasses transformer encoder architectures like FinBERT in terms of generalization and overall performance. Moreover, the GPT-3.5-turbo with RAMAS achieves even better performance than the vanilla GPT-4o.

In summary, RAMAS emerges as an innovative and potent multi-agent framework empowering LLMs to excel in FSA, offering superior performance compared to existing methods.

REFERENCES

1. K. Du et al., "Financial Sentiment Analysis: Techniques and Applications," *ACM Computing Surveys*, vol. 56, no. 9, 2024.
2. P. Malo et al., "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, 2014, pp. 782–796.
3. R. Mao et al., "Discovering the Cognition behind Language: Financial Metaphor Analysis with MetaPro," *IEEE ICDM*, 2023, pp. 1211–1216.

4. S. Wu et al., "BloombergGPT: A large language model for finance," *arXiv:2303.17564*, 2023.
5. G. Fatouros et al., "Transforming Sentiment Analysis in the Financial Domain with ChatGPT," *Machine Learning with Applications*, vol. 14, 2023, p. 100508.
6. S. Fatemi and Y. Hu, "A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis," *arXiv:2312.08725*, 2023.
7. K. Du et al., "An evaluation of reasoning capabilities of large language models in financial sentiment analysis," *IEEE CAI*, 2024.
8. B. Zhang et al., "Enhancing financial sentiment analysis via retrieval augmented large language models," *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 349–356.
9. E. Henry, "Are investors influenced by how earnings press releases are written?" *The Journal of Business Communication*, vol. 45, no. 4, 2008, pp. 363–407.
10. T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, 2011, pp. 35–65.
11. K. Du et al., "FinSenticNet: A Concept-Level Lexicon for Financial Sentiment Analysis," *IEEE SSCI*, 2023, pp. 109–114.
12. S. Krishnamoorthy, "Sentiment analysis of financial news articles using performance indicators," *Knowledge and Information Systems*, vol. 56, 2018, pp. 373–394.
13. D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *arXiv:1908.10063*, 2019.
14. A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, 2023, pp. 806–841.
15. J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *Proceedings of ACL*, 2018, pp. 328–339.

Kelvin Du is currently working toward the Ph.D. degree with Nanyang Technological University, Singapore. His research interests include financial sentiment analysis, natural language based financial forecasting and NLP in finance. Contact him at zidong001@e.ntu.edu.sg.

Yazhi Zhao is a Senior Data Scientist at Visa Singapore. Her research interests lie in NLP in finance. Zhao was awarded a Master's degree in Intelligent Systems from the National University of Singapore. Contact her at yazzhao@visa.com.

Rui Mao is a Research Scientist at Nanyang Technological University. His research focuses on cognitive computing and quantitative finance. Rui received his Ph.D. degree in Computing Science from the University of Aberdeen. Contact him at rui.mao@ntu.edu.sg.

Frank Xing is an Assistant Professor at the National University of Singapore. His research focuses on AI for finance, FinTech, and information systems design. He is an early advocate for the systematic use of natural language processing in quantitative investing. Contact him at xing@nus.edu.sg.

Erik Cambria is a Professor at Nanyang Technological University, where he also holds the appointment of Provost Chair in Computer Science and Engineering. His research focuses on neurosymbolic AI for interpretable, trustworthy, and explainable affective computing. Contact him at cambria@ntu.edu.sg