Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models

Conference Paper · November 2023

DOI: 10.1145/3604237.3626866

Tianyu Zhou
Brown University

Xiao-Yang Liu
Columbia University

Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models

Boyu Zhang[*]
boyu.zhang01@adelaide.edu.au
The University of Adelaide
Australia

Hongyang (Bruce) Yang[*]
hy2500@columbia.edu
Columbia University
USA

Tianyu Zhou[*]
zhoutianyu0426@gmail.com
Brown University
USA

Ali Babar
ali.babar@adelaide.edu.au
The University of Adelaide
Australia

Xiao-Yang Liu[†]
xl2427@columbia.edu
Rensselaer Polytechnic Institute &
Columbia University
USA

ABSTRACT

Financial sentiment analysis is critical for valuation and investment decision-making. Traditional NLP models, however, are limited by their parameter size and the scope of their training datasets, which hampers their generalization capabilities and effectiveness in this field. Recently, Large Language Models (LLMs) pre-trained on extensive corpora have demonstrated superior performance across various NLP tasks due to their commendable zero-shot abilities. Yet, directly applying LLMs to financial sentiment analysis presents challenges: The discrepancy between the pre-training objective of LLMs and predicting the sentiment label can compromise their predictive performance. Furthermore, the succinct nature of financial news, often devoid of sufficient context, can significantly diminish the reliability of LLMs' sentiment analysis. To address these challenges, we introduce a retrieval-augmented LLMs framework for financial sentiment analysis. This framework includes an instruction-tuned LLMs module, which ensures LLMs behave as predictors of sentiment labels, and a retrieval-augmentation module which retrieves additional context from reliable external sources. Benchmarked against traditional models and LLMs like ChatGPT and LLaMA, our approach achieves 15% to 48% performance gain in accuracy and F1 score.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

Sentiment Analysis, Large Language Models, Instruction Tuning, Retrieval Augmented Generation

ACM Reference Format:
Boyu Zhang, Hongyang (Bruce) Yang[*], Tianyu Zhou[*], Ali Babar, and Xiao-Yang Liu. 2023. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604237.3626866

1 INTRODUCTION

Financial sentiment analysis is a critical tool that extracts, quantifies, and studies the affective states and subjective information within financial documents, news articles, and social media content [6]. Its significance lies in its potential to forecast market movements and provide valuable insights into investors' behaviors. Given that market reactions are often influenced by news sentiments, which can be positive, negative, or neutral, financial sentiment analysis plays a pivotal role in aiding traders and financial institutions in making informed decisions. It helps manage risks and identify potential investment opportunities by providing a nuanced understanding of the market's emotional undercurrents.

In recent years, numerous studies have turned to Natural Language Processing (NLP) models to enhance the accuracy and efficiency of financial sentiment analysis [1, 5, 21, 29–31]. Traditional NLP models, constrained by the limitations of their model parameters and the scale of their training corpora, often lack the capability to comprehensively understand intricate financial news, thereby limiting the efficacy of financial sentiment analysis [1, 5, 21, 31]. These limitations have sometimes resulted in suboptimal outcomes in financial sentiment analysis tasks. In contrast, the advent of large language models (LLMs) [15, 24, 29, 30] has ushered in a new era in the NLP domain. These LLMs, having been pre-trained on vast and diverse corpora, boast formidable zero-shot learning abilities. As a result, they are gradually outperforming many other models across various NLP tasks, owing to their ability to generalize from their extensive training and derive meaningful insights even from previously unseen financial data.

However, directly applying LLMs for financial sentiment analysis poses two notable challenges. Firstly, the discrepancy between the objective function used in LLMs' pre-training and the goal of predicting financial sentiment may result in LLMs' inability to consistently output labels for financial sentiment analysis as expected

[*] Authors contributed equally to this research.
[†] Corresponding author.

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11…$15.00
https://doi.org/10.1145/3604237.3626866

[15, 23]. Secondly, the typical subjects of financial sentiment analysis, such as news flashes and tweets, are characteristically concise and often lack adequate background information. The scarcity of information has not only interfered with the judgment of human experts [13] but also poses a significant challenge to the accurate prediction of large language models.

To address the aforementioned challenges, in our study, we present a retrieval-augmented large language model framework for financial sentiment analysis. This framework consists of two key components. 1) instruction-finetuned LLMs [15], which refines LLMs using a limited set of instruction-following examples crafted specifically for financial sentiment analysis, aligning LLMs' predictions with user intentions and significantly boosting their prediction accuracy. 2) retrieval-augmented component [7], which introduces additional context to brief statements from news flashes or tweets. It employs search engines and verified financial sources to gather relevant background information from external sources. This enriched context is then passed to the instruction-tuned LLMs for prediction, resulting in more accurate and nuanced results.

Through extensive evaluations on multiple financial sentiment analysis benchmarks, we demonstrate that compared to traditional smaller-scale sentiment analysis models [1] and general-purpose LLMs, such as ChatGPT [15] and LLaMA [24], our approach markedly outperforms them. The primary contributions of this paper can be summarized as follows:

- We introduce a novel retrieval-augmented large language model framework tailored for financial sentiment analysis. By integrating external knowledge retrieval, we optimize the depth and context of the information feeding into the LLMs, ensuring more nuanced and informed predictions.
- Our method of instruction tuning leverages a unique set of instruction-following examples. This fine-tuning process realigns LLMs to respond more accurately to user-intended financial sentiment analysis tasks, markedly enhancing their predictive accuracy.
- Through extensive evaluations on established benchmarks, we demonstrate that our approach outperforms traditional sentiment analysis models and notable general-purpose LLMs, achieving a 15% to 48% performance gain in accuracy and F1 score.

The remainder of this paper is organized as follows. Section 2 briefly reviews the backgroun and related work. In Section 3, we describe the retrieval augmented method that consists two modules. In Section 4, we present the performance evaluation from three aspects. Section 5 concludes this work and points out directions for future work.

2 BACKGROUND AND RELATED WORK

2.1 Financial Sentiment Analysis

Financial sentiment analysis has been a prominent area of research in NLP, and deep learning has found widespread application due to its effective feature representation. Early approaches [1, 5, 21, 31] involved fine-tuning pre-trained models on financial sentiment analysis datasets, but they faced challenges in understanding complex financial news, especially those with numerical information or lacking background context [1].

Recently, Large Language Models (LLMs) have emerged as an attractive option in NLP. With increasing model size and training data, LLMs demonstrate impressive abilities in in-context learning and chain-of-thought reasoning, allowing them to make predictions in a zero-shot manner. However, LLMs like BloombergGPT [29] and FinGPT [30], tailored for the financial domain, face difficulty in generating expected sentiment labels due to the mismatch between their training objective, typically Causal Language Modeling, and the objective of financial sentiment analysis. Additionally, financial sentiment analysis often deals with brief subjects like news flashes and tweets, lacking sufficient background information. This brevity and contextual deficiency pose a significant challenge for LLMs, making the task of reliable sentiment analysis more difficult.

2.2 Instruction Tuning

The latest LLMs such as GPT-3 [2], LLaMA [24], and others have been trained using Causal Language Modeling, which involves predicting the next token given the previous content. However, this training approach introduces randomness in LLMs' outputs, leading to results that may not always align with the desired expectations.

To address this issue and make LLMs follow specific instructions, researchers have proposed a technique known as instruction tuning [10, 15, 19, 27]. It involves fine-tuning pre-trained LLMs on a collection of formatted instances presented in natural language, aiming to guide the LLMs to follow user instructions. These instances typically take the form of task descriptions along with their corresponding desired output, often labeled by humans [15] or semi-automatically constructed [26]. Through this process, LLMs can be fine-tuned to understand and execute specific instructions effectively, making them more reliable for various applications that require controlled and directed behavior.

2.3 Retrieval Augmented Generation

Retrieval-augmented generation (RAG) [3, 7] is a technique that combines the strength of context retrieval and LLMs for language generation. RAG operates in a two-step process. First, it retrieves relevant documents using the retrieval module based on the input prompt. These documents are typically sourced from external knowledge bases like news sources, research publications, and social media and provide additional context for the subsequent generation step. Next, the retrieved documents are combined with the original input prompt and fed into the LLMs, generating the final output. The combination of retrieval and generation in RAG allows it to utilize two distinct sources of knowledge: the parametric memory stored in the LLMs' parameters and the nonparametric memory obtained from the corpus of retrieved documents. This dual-knowledge approach enables RAG to effectively guide the generation process and produce more accurate and contextually relevant responses. RAG has been widely used in the areas like open world QA [14] and code summarization [8, 16].

3 METHOD

3.1 Overview

Our proposed framework, as shown in Fig. 1, consists two modules of the instruction-tuned LLM and the RAG module.

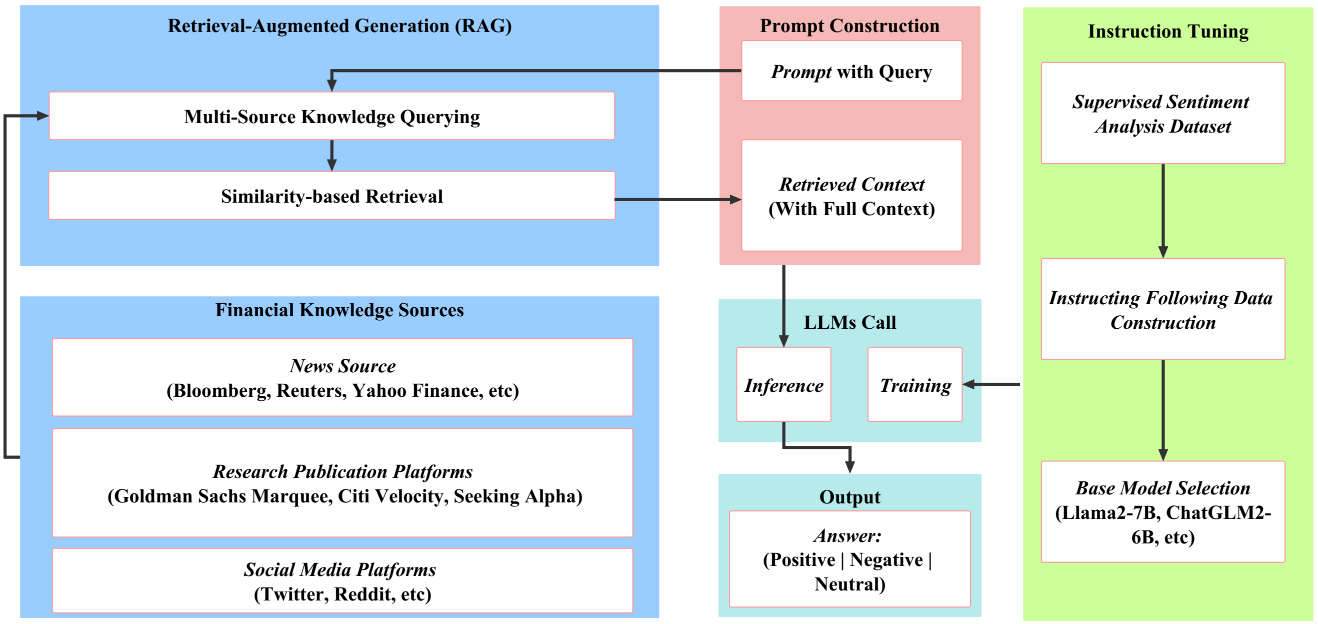

Retrieval-Augmented Generation (RAG)
Multi-Source Knowledge Querying
Similarity-based Retrieval
Financial Knowledge Sources
News Source
(Bloomberg, Reuters, Yahoo Finance, etc)
Research Publication Platforms
(Goldman Sachs Marquee, Citi Velocity, Seeking Alpha)
Social Media Platforms
(Twitter, Reddit, etc)
Prompt Construction
Prompt with Query
Retrieved Context
(With Full Context)
LLMs Call
Inference
Training
Output
Answer:
(Positive | Negative | Neutral)
Instruction Tuning
Supervised Sentiment Analysis Dataset
Instructing Following Data Construction
Base Model Selection
(Llama2-7B, ChatGLM2-6B, etc)

Figure 1: Framework of retrieval-augmented large language model for financial sentiment analysis.

In the first module, we apply instruction tuning to fine-tune an open-source pretrained LLM, such as LLaMA [24, 24] and ChatGLM [32], to align their behavior with predicting financial sentiment labels when provided with financial news or tweets. This process involves constructing an instruction-following dataset specific to the task of financial sentiment analysis and using it to fine-tune the pretrained LLM.

The RAG module plays a crucial role in the framework by retrieving pertinent background information from external sources related to the input query. These external sources include well-authenticated news platforms like Bloomberg and Reuters, research publications from institutions like Goldman Sachs and Citi Velocity, and social media platforms such as Twitter and Reddit. We employ a multi-source query and similarity-based retrieval approach to locate the most relevant information from these sources.

Subsequently, the retrieved context is combined with the original query to construct the final query. The instruction-tuned LLM is then called upon to generate a sentiment prediction based on this augmented query. In this way, the missing background knowledge is provided to the LLM, enabling it to make more accurate predictions.

The implementation of these steps will be further detailed in the following sections, showcasing how our framework effectively incorporates instruction tuning and retrieval-augmented generation to enhance the accuracy and precision of financial sentiment analysis.

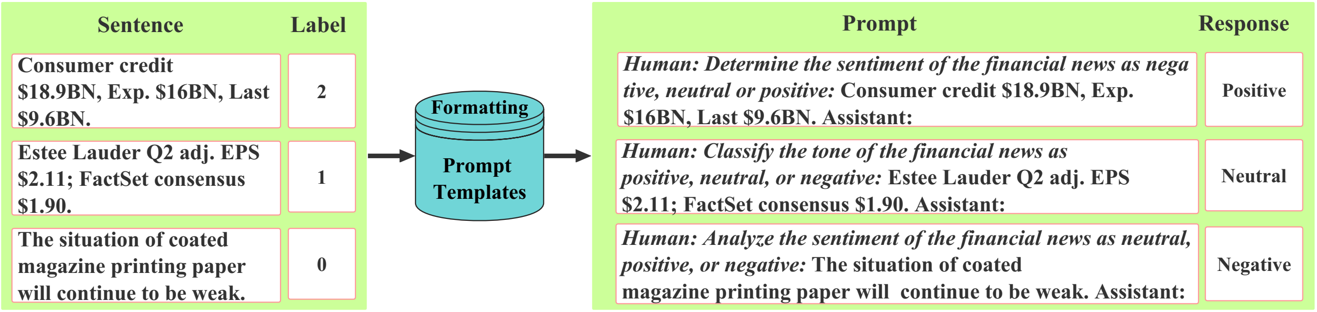
3.2 Instruction-tuned LLMs

Instruction tuning proves to be a highly effective approach to align the behavior of LLMs with user instructions, particularly in our study, where we aim to predict financial sentiment labels. Encouraging results from recent studies [4, 22, 33] demonstrate that limited instruction following data, when used in instruction tuning, allows the resulting LLMs to adhere remarkably well to user instructions.

There are typically three steps to apply instruction tuning in the domain of financial sentiment analysis. Firstly, we construct an instruction-following dataset, consisting of paired instructions and their corresponding expected responses – essentially the labels of sentiment. This dataset serves as the foundation for guiding the LLMs to understand the user's instructions effectively. The second step involves fine-tuning the LLMs on the constructed dataset. Through this fine-tuning process, the model learns to generate the expected response accurately when provided with instructions to predict sentiment labels. The final step is to map the generated outputs from the LLMs back into predefined sentiment classes. This step further aligns the prediction with predefined sentiment classes and allows the model's performance to be measurable. We detail these steps in the following.

3.2.1 Formatting Financial Sentiment Instruction Following Dataset. Creating a financial sentiment instruction-following dataset through manual labeling requires the expertise of specialized financial professionals, which can be costly. An alternative approach is to convert existing supervised financial sentiment analysis datasets into instruction-following datasets at a lower cost [15, 26, 27, 34]. These datasets are often formatted as text classification tasks where the **inputs** are the financial news or headlines and the **outputs** are integer-type labels representing *positive*, *negative* and *neutral* sentiments.

Following [34], we create 10 human-written **instructions** describing the task of financial sentiment analysis, and formulate



Sentence	Label
Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN.	2
Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90.	1
The situation of coated magazine printing paper will continue to be weak.	0

Prompt	Response
Human: Determine the sentiment of the financial news as negative, neutral or positive: Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN. Assistant:	Positive
Human: Classify the tone of the financial news as positive, neutral, or negative: Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90. Assistant:	Neutral
Human: Analyze the sentiment of the financial news as neutral, positive, or negative: The situation of coated magazine printing paper will continue to be weak. Assistant:	Negative

Figure 2: Formatting sentiment analysis dataset into instruction-following dataset.

each sample from the original dataset by combining one randomly selected **instruction** with the **input** and **output** in the format of "Human: [**instruction**] + [**input**], Assistant: [**output**]". This process is shown in Fig. 2.

3.2.2 Instruction Tuning. Instruction tuning involves fine-tuning pre-trained LLMs using the instruction-following dataset. We tokenize the texts into tokens with the byte-pair encoding (BPE) algorithm [20] first. Then we fine-tune the LLMs with the causal language modeling (CLM) objective, which maximizes the likelihood of predicting the next token in a sequence given the preceding context. This process is achieved by minimizing the following negative log-likelihood,

$$\mathcal{L}_{\text{CausalLM}} = -\sum_{t=1}^{T} \log P(w_t | w_1, w_2, \ldots, w_{t-1}; \theta), \tag{1}$$

where T is the length of the input sequence, w_i represents the i-th token in the sequence, θ denotes the model parameters, and $P(w_i | w_1, w_2, \ldots, w_{i-1}; \theta)$ represents the conditional probability of predicting the target token x_i given the preceding tokens $x_1, \ldots, x_{i-1}$. By optimizing this objective function, the model learns to maximize the probability of generating the expected financial sentiment label given the instruction. We use the gradient-based optimization method to minimize the above objective where the gradient is represented as,

$$\nabla_\theta \mathcal{L}_{\text{CausalLM}} = -\sum_{t=1}^{T} \frac{\partial \log P(w_t | w_1, w_2, \ldots, w_{t-1}; \theta)}{\partial \theta}. \tag{2}$$

This gradient is typically computed by backpropagation through time [28]. It involves computing the gradients at each time step t and then propagating them back through the entire sequence of time steps.

3.2.3 Mapping the Generated Outputs into Sentiment Classes. Since the instruction-finetuned LLM is an autoregressive generative model, even though we train it using instruction-following dataset to guide its output towards the desired sentiment label, it still has the possibility of generating free-style text. Therefore, we need to map the model's output back to the specified three emotions for proper evaluation. Our approach is as follows: we sequentially check if the output results contain "negative", "neutral", or "positive". Once a term is found, we map it to the corresponding label. Otherwise, we consider it a "neutral" sentiment.

3.3 RAG Module

RAG is an effective method for injecting external knowledge into LLM to enhance the accuracy of response generation. The implementation of the RAG module involves several steps. Firstly, we set up external knowledge sources which are highly likely to contain relevant financial background information. Next, we perform a two-step knowledge retrieval process, consisting of Multi-Source Knowledge Querying and Similarity-Based Retrieval. These steps enable us to gather relevant context related to the input query. Finally, we combine the original input query with the retrieved context, creating input data for the instruction-tuned LLM, which generates the final result.

3.3.1 Setup External Knowledge Sources. When retrieving relevant financial context based on the query, our objective is to access authentic, pertinent, insightful, and comprehensive data, as opposed to random internet searches. To achieve this, we first identify the following sources of information:

News Sources. Esteemed outlets such as Bloomberg, Yahoo Finance, Reuters, CNBC, and Market Screener supply information that is inherently consistent and crucial for financial interpretations. These sources tend to have stringent internal guidelines for their writers and reporters, ensuring reliable and verified content. Furthermore, owing to the nature of their operations, these outlets often offer the earliest reports on a variety of financial news.

Research Publication Platforms. Centralized, as well as crowd-based research publication platforms, provide a wealth of financial insights.

- **Centralized Publishers** Renowned institutions such as Goldman Sachs and Citi proffer exclusive research services, Marquee and Velocity respectively, predominantly to their institutional clients. Given their direct applicability, these researches provide highly consistent, systematic, and authenticated insights.
- **Crowd-based Publishers** Platforms like Seeking Alpha serve as a repository for diverse insights from independent contributors. They cover a broad spectrum of financial information that includes a vast array of price movement analyses,

transcripts of earnings calls and conferences, and investment research pertaining to companies of all sizes.

All of these sources provide retrieval APIs, enabling us to access and retrieve information.

Social Media Platforms: Social media platforms like Twitter and Reddit have become significant sources of financial information. These platforms offer real-time updates and discussions, which can be insightful for understanding market sentiment and trends. However, the information on these platforms can be highly volatile and unverified, necessitating careful analysis and cross-referencing with other sources.

3.3.2 Two-Step Knowledge Retrieval. We retrieve contextual financial information for a given query through a two-step process.

Multi-Source Knowledge Query. Financial news headlines or tweets are typically short and often include irrelevant content like tickers. To address this, our first step involves using regular expressions to preprocess the text and remove irrelevant tickers or symbols. Subsequently, we utilize various knowledge sources' retrieval APIs to extract relevant information. If the news item contains time information, we perform searches within that specific time range. The search returns a list of relevant context snippets from identified financial sources. For each context snippet, we gather the original headline, editorial bullet points, article body paragraphs, posts, and reposts as the full context. This query strategy allows us to capture a wide spectrum of information related to financial news.

Similarity-Based Retrieval. Even after the initial retrieval, the content obtained may still contain a considerable amount of irrelevant information, which could potentially hinder sentiment prediction accuracy. To address this issue, we propose an advanced retrieval algorithm based on similarity. This algorithm aims to further filter and extract the most relevant content from the results obtained in the first step. Specifically, we use a modified overlap coefficient as the similarity measure for retrieval and empirically select those context with a similarity higher than 0.8 to the input query. The overlap coefficient, also known as the Szymkiewicz-Simpson coefficient [25], is used to measure the similarity degree between two samples. In the task of sentence-context pair similarity evaluation, this coefficient measures the number of words in the intersection divided by the union of the pair. The specific formula given as follows:

$$\text{overlap}(\mathbf{X}, \mathbf{Y}) = \frac{|\mathbf{X} \cap \mathbf{Y}|}{min(|\mathbf{X}|, |\mathbf{Y}|)}, \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ represent sets of financially relevant tokens from a queried sentence and its context, respectively.

We prefer the Szymkiewicz-Simpson coefficient over the semantic similarity for two major reasons. In financial news, the need for exact matches, especially for tickers, is paramount. This coefficient emphasizes this hard matching, minimizing irrelevant retrievals. In contrast, semantic similarity can sometimes miss the intricacies of specific financial terms. Moreover, the challenge of short-to-long text matching, highlighted in [18], is adeptly managed by the Szymkiewicz-Simpson coefficient, ensuring relevant news are not overshadowed by length. The overall two-step knowledge retrieval algorithm is shown in Alg. 1.

Algorithm 1 Financial Knowledge Retrieval

```
Input: Query $\mathbf{Q}$
Output: Context $\mathbf{C}$
Search $\mathbf{Q}$ using a search engine and obtain a list of documents, $\mathbf{D}$, where each contains some phrases of $\mathbf{Q}$
$\mathbf{C} \leftarrow \emptyset$
for document $\mathbf{d} \in \mathbf{D}$ do
    if $\text{overlap}(\mathbf{Q}, \mathbf{d}) > 0.8$ then
        Split $\mathbf{d}$ into syntactic units, $\mathbf{u}_i$ for $\{i | 1 \le i \le n + 1\}$, which are separated by $n$ separative tokens (i.e., paragraphs, bullet points)
        for $i = 1$ to $n + 1$ do
            if $\text{overlap}(\mathbf{Q}, \mathbf{u}_i) > 0.7$ then
                $\mathbf{C}$=Concat($\mathbf{C}$, $\mathbf{u}_i$)
            end if
        end for
    end if
end for
```

4 PERFORMANCE EVALUATION

In this section, we evaluate the effectiveness of instruction fine-tuning and RAG. To validate our method's performance, we compare it against state-of-the-art sentiment analysis models and the general-purpose LLMs. Our experimental results validate the effectiveness of our approach. With only a small amount of instruction-following data, our model consistently outperforms other baselines in sentiment analysis and its performance can be further enhanced with the RAG module. The code of this experiment is available at Github[1].

4.1 Datasets

4.1.1 Training Datasets. Our training data is an amalgamation of the Twitter Financial News dataset [11] and FiQA dataset [12], resulting in a collection of 10,501 samples.

- **Twitter financial news sentiment training**: This dataset [2] is a corpus of news tweets that pertain to the financial sector. Its primary purpose is the classification of financial sentiment within the context of Twitter discussions. The dataset comprises 9,540 samples for training, each annotated with one of three labels: Bearish, Bullish, or Neutral.
- **FiQA dataset**: This dataset [3] includes 961 samples. Each sample has been annotated with one of three labels: positive, neutral, or negative, denoting the sentiment conveyed in the corresponding text.

4.1.2 Testing Datasets.

- **Twitter financial news sentiment validation (Twitter Val)**: This is the validation split of the Twitter dataset which contains 2,388 samples. It can validate how well the model can predict the financial sentiment from the social media. It's important to note that the platform often lacks clear sources and context for news items.

[1] https://github.com/AI4Finance-Foundation/FinGPT/tree/master/fingpt/FinGPT_RAG/instruct-FinGPT

[2] https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment

[3] https://huggingface.co/datasets/pauri32/fiqa-2018

- **Financial PhraseBank (FPB) dataset**: This dataset [4] [13] comprises 4,840 samples randomly extracted from financial news articles available on the LexisNexis database. The samples were carefully annotated by a team of 16 annotators with backgrounds in finance and business, ensuring high quality annotations.

For all the above datasets, we use the approach mentioned in Section 3.2 to format them as instruction-following datasets before training and testing.

4.2 Model Training

We initialize our model with Llama-7B and perform instruction tuning over 10 epochs. The training process utilizes the AdamW optimizer [9], with a batch size of 32, an initial learning rate of $1e^{-5}$, and a weight decay of 0.1. To maintain efficiency, we set a maximum input text length of 512 tokens. We utilize DeepSpeed [17] for the fine-tuning process on 8×A100 (40GB) GPUs, resulting in a total training time of 58 minutes.

4.3 Baseline Models

BloombergGPT [29]. BloombergGPT is a 50 billion parameter language model that is trained on a wide range of financial data. As it's a closed-source model, we directly use their reported performance on the FPB dataset.

ChatGPT [15]. ChatGPT is the cutting-edge closed-source LLM developed by OpenAI. The use of OpenAI's ChatGPT API for sentiment analysis involves four steps: API setup, data preparation using the instruction-following dataset, making requests using the GPT-4.0 API, and interpreting the direct sentiment output from the response.

Llama-7B [24]. Llama-7B is an open-source LLM created by Meta, with the majority of the training corpus being in English. We acquired the Llama-7B[5] model from Meta and maintain the identical inference configuration as our base model.

ChatGLM2-6B. [32] ChatGLM2-6B is an open-source LLM crafted by Tsinghua University, supporting both English and Chinese. We acquired the ChatGLM2-6B model from Hugging Face Model Hub.

FinBERT. [1] FinBERT is a financial sentiment analysis model which is fine-tuned on the pretrained BERT language model. The FinBERT model is also accessible through the Hugging Face Model Hub.

4.4 Evaluation and Analysis

To evaluate the performance of our model, we first compare our instruction-tuned LLM with the sentiment analysis model FinBERT and the general-purpose LLM to verify the effectiveness of the instruction tuning. We then compare the LLMs including ours and the baselines with and without RAG to further validate the efficacy of RAG.

4.4.1 *Performance metrics.*

The performance metrics for our model include accuracy, and F1-score. Accuracy measures the proportion of correct predictions, and the F1-score represents the harmonic mean of precision and recall.

4.4.2 *Assessment of Instruction Finetuning.*

In this experiment, we aim to verify the effectiveness of instruction tuning, denoted as "Ours" in the presented Table 1. The comparative analysis is performed against all the baseline models. The evaluation is conducted on the Financial PhaseBank (FPB) and Twitter Val. We excluded FinBERT from the comparison on FPB as it uses the exact dataset for training.

Dataset	FPB		Twitter Val	
Metrics	Acc	F1	Acc	F1
FinBERT [1]	-	-	0.725	0.668
BloombergGPT [29]	-	0.510	-	-
ChatGLM2-6B [32]	0.474	0.402	0.482	0.381
Llama-7B [24]	0.601	0.397	0.544	0.363
ChatGPT 4.0 [15]	0.643	0.511	0.788	0.652
Ours	**0.758**	**0.739**	**0.863**	**0.811**

Table 1: Comparison between our model and the baselines on the datasets of financial phaseBank (FPB) and Twitter Val.

The outcomes in Table 1 suggest that our instruction-tuned Llama-7B model outperforms the others, achieving the highest accuracy and F1 score. The process of fine-tuning with instruction-following data enhances the model's ability to discern sentiment in financial phrases, leading to superior performance compared to both ChatGPT 4.0 and the original Llama-7B model. From these findings, it is evident that the instruction tuning method significantly improves the model's performance on financial sentiment analysis.

4.4.3 *Performance of RAG Module.*

We verify the effectiveness of RAG module on both our instruction-tuned model and the ChatGPT 4.0 on the Twitter Val dataset in this experiment. From the results presented in Table 2, it demonstrates the introduction of RAG context will universally improve the performance of the LLMs which verifies that the retrieved context enhances the information and allow the LLMs to make more accurate prediction. Specially, our model with RAG again achieves the best performance among all the methods.

To better highlight the RAG module's effectiveness, we present a case study in Table 3. Initially, the statement's ambiguity causes our instruction-tuned model to misclassify it as "neutral." With RAG, we augment the context using information from Seeking Alpha, clarifying the phrase "shakes off" to indicate a rating upgrade, which helps our model correctly reclassify the statement as "positive." This showcases RAG's ability to enhance model comprehension and provide a more nuanced understanding of the sentiment in the news headline.

Metrics	Acc	F1
ChatGPT 4.0 w/o RAG	0.788	0.652
ChatGPT 4.0 w/ RAG	**0.813**	**0.708**
Ours w/o RAG	0.863	0.811
Ours w/ RAG	**0.881**	**0.842**

Table 2: Experimental results on the Twitter Val dataset.

[4] https://huggingface.co/datasets/financial_phrasebank

[5] We utilize Llama-7B for research and education purposes.

	Text	Result
w/o RAG	$ENR - Energizer shakes off JPMorgan's bear call.	Neutral
w/ RAG	*"Energizer shakes off JPMorgan's bear call. JPMorgan **hikes Energizer Holdings (NYSE:ENR) to a Neutral rating from Underweight**... We came away **encouraged** by some of the company's initiatives and believe their focus on innovation and brand investment can lead to relative outperformance going forward... Shares of Energizer are **0.46% premarket to $50.44**."*	**Positive**

Table 3: Case study: before and after using RAG.

5 CONCLUSION AND FUTURE WORK

In conclusion, this paper unveils a novel retrieval-augmented large language model framework tailored for financial sentiment analysis. Our unique instruction tuning method has realigned LLMs to respond more accurately to user-intended financial sentiment analysis tasks, significantly enhancing their predictive accuracy. The integration of external knowledge retrieval has further enriched the depth and context of the information fed into the LLMs, enabling more nuanced predictions.

However, a limitation of our approach is its exclusive reliance on textual similarity to retrieve relevant information. This method overlooks crucial macroeconomic information related to the timing of the news and microeconomic information concerning the financial and operational status of the related enterprise. Incorporating such economic data could provide a more holistic view, allowing LLMs to make more accurate judgments. Future work could explore amalgamating these additional economic dimensions with textual data, to further improve the precision and reliability of financial sentiment analysis performed by large language models.

REFERENCES

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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 3417–3419.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

[5] Min-Yuh Day and Chia-Chou Lee. 2016. Deep learning for financial sentiment analysis on finance news providers. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* IEEE, 1127–1134.

[6] Gartner Glossary. 2023. Definition of Sentiment Analysis - Finance Glossary - Gartner.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[8] Shangqing Liu, Yu Chen, Xiaofei Xie, Jingkai Siow, and Yang Liu. 2020. Retrieval-augmented generation for code summarization via hybrid gnn. *arXiv preprint arXiv:2006.05405* (2020).

[9] Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101* (2017).

[10] Renze Lou, Kai Zhang, and Wenpeng Yin. 2023. Is prompt all you need? no. A comprehensive and broader view of instruction learning. *arXiv preprint arXiv:2303.10475* (2023).

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

[12] Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra. Balahur. 2018. WWW '18: Companion Proceedings of the The Web Conference 2018. In *International World Wide Web Conferences Steering Committee* (Lyon, France). Republic and Canton of Geneva, CHE.

[13] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.

[14] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553* (2020).

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[16] Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. *arXiv preprint arXiv:2108.11601* (2021).

[17] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Association for Computing Machinery* (Virtual Event, CA, USA) *(KDD '20)*. New York, NY, USA, 3505–3506.

[18] Vipula Rawte, Aparna Gupta, and Mohammed J Zaki. 2020. A comparative analysis of temporal long text similarity: Application to financial documents. In *Workshop on Mining Data for Financial Applications.* Springer, 77–91.

[19] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).

[20] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[21] Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data* 5, 1 (2018), 1–25.

[22] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[23] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[25] M.K. Vijaymeena1 and K. Kavitha. 2016. A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications* (2016).

[26] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560* (2022).

[27] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations.*

[28] Paul J Werbos. 1988. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks* 1, 4 (1988), 339–356.

[29] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

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

[31] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* (2020).

[32] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

[33] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *arXiv preprint arXiv:2306.12659* (2023).

[34] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]