SILC-EFSA: Self-aware In-context Learning Correction for Entity-level Financial Sentiment Analysis

Senbin Zhu[1*], Chenyuan He[1*], Hongde Liu[1], Pengcheng Dong, Hanjie Zhao[1], Yuchen Yan[1], Yuxiang Jia[1†], Hongying Zan[1], Min Peng[2]

[1]School of Computer and Artificial Intelligence, Zhengzhou University, China
[2]School of Computer Science, Wuhan University, China
{nlpbin,hechenyuan_nlp,lhd_1013,dongpc}@gs.zzu.edu.cn, pengm@whu.edu.cn
Correspondence: ieyxjia@zzu.edu.cn

Abstract

In recent years, fine-grained sentiment analysis in finance has gained significant attention, but the scarcity of entity-level datasets remains a key challenge. To address this, we have constructed the largest English and Chinese financial entity-level sentiment analysis datasets to date. Building on this foundation, we propose a novel two-stage sentiment analysis approach called Self-aware In-context Learning Correction (SILC). The first stage involves fine-tuning a base large language model to generate pseudo-labeled data specific to our task. In the second stage, we train a correction model using a GNN-based example retriever, which is informed by the pseudo-labeled data. This two-stage strategy has allowed us to achieve state-of-the-art performance on the newly constructed datasets, advancing the field of financial sentiment analysis. In a case study, we demonstrate the enhanced practical utility of our data and methods in monitoring the cryptocurrency market. Our datasets and code are available at https://github.com/NLP-Bin/SILC-EFSA.

1 Introduction

The importance of sentiment analysis in the financial domain has increasingly become apparent. As early as 1970, Fama recognized the potential of sentiment analysis in finance and introduced the concept of the Efficient Market Hypothesis (EMH) (Fama, 1970). The EMH suggests that stock prices respond to unexpected fundamental information, supporting the use of sentiment analysis in finance. With the rapid growth of the internet and the financial sector, numerous stock reports, research papers, and investor opinions have become valuable for assessing companies and events, playing a key role for both investors and regulators.

Currently, most sentiment analysis corpora in the financial domain use sequence-level annotation.

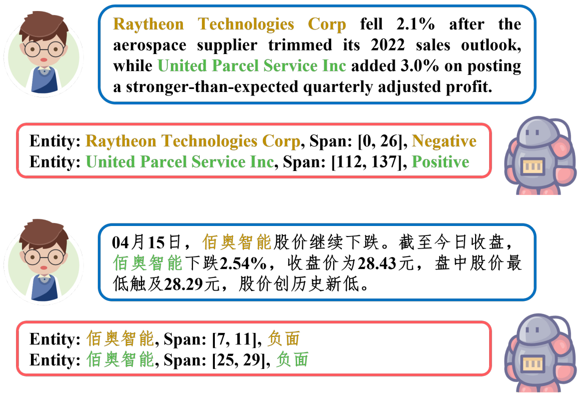

Figure 1: Examples of financial entity-level sentiment analysis data in English and Chinese. In the task, the objective is to identify financial entities within the text and analyze their sentiment within the context. Specifically, this involves annotating each entity along with its position (span) in the text and determining its sentiment polarity (e.g., positive, negative, or neutral).

While sequence-level sentiment analysis has enhanced understanding of financial dynamics, many financial texts (such as news articles, analyst reports, and social media data) often contain multiple entities with differing sentiments (Malo et al., 2014; Huang et al., 2023; Sinha and Khandait, 2021; Shah et al., 2023a). Consequently, the need for fine-grained entity-level sentiment analysis in the financial domain has emerged. This task requires extracting the entities in the text and their location information, as well as the sentiment polarity corresponding to the entities. Task examples are illustrated in Figure 1.

Tang et al. (2023) construct the first dataset with financial entity annotations and sentiment labels, named FinEntity, which clarifies the task requirements for entity-level sentiment analysis in the financial domain. However, the FinEntity dataset is relatively small in scale. To conduct a more comprehensive study of entity-level financial sentiment analysis, we thoroughly investigate current

[*]Equal contribution
[†]Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 4980–4992
January 19–24, 2025. ©2025 Association for Computational Linguistics

open-source financial datasets containing entity information. We select the SEntFiN dataset (Sinha et al., 2022) and a Chinese financial event-level sentiment analysis dataset (Chen et al., 2024), re-screen and reconstruct them, ultimately creating the largest (English and Chinese) entity-level financial sentiment analysis database to date. The construction methods will be detailed in Section 3.

Currently, large language models (LLMs) have achieved significant research milestones in the financial domain, such as BloombergGPT (Wu et al., 2023) and FinLlaMA (Konstantinidis et al., 2024). Most research focuses on further pre-training base models on large-scale financial data to enhance their expertise in the financial domain. However, methods based on LLMs for entity-level sentiment analysis remain relatively few. On the dataset we constructed, we design a two-stage strategy to leverage large models and achieve performance improvements.

In the first stage, we fine-tune a base model to perform entity-level sentiment analysis, and we find that its understanding of knowledge needs to be enhanced. Even after training, models still generate some incorrect predictions on the data. This is analogous to the human learning process, where initial learning often lacks comprehensive knowledge, and self-checking and error correction can further enhance performance. Based on this observation, we design an error-correction strategy in the second stage, training a self-correcting mechanism to improve model performance. Specifically, we first obtain and filter erroneous sample data, then train a graph neural network (GNN) to retrieve relevant examples, and subsequently fine-tune the base model to function as an error-corrector to rectify the predictions from the first stage. Experiments conducted on the dataset we constructed demonstrate that this approach achieves state-of-the-art performance. Leveraging the sentiment analysis capabilities of our model, we perform information extraction on time-series financial texts from the cryptocurrency market, and achieve more accurate price predictions using an LSTM network.

The main contributions of this paper are as follows:

- We restructure existing English and Chinese datasets to build the largest financial entity-level sentiment analysis datasets to date, providing a rich and reliable data resource for future research.
- We propose a novel two-stage self-correction approach, covering Initial Response Generation and Self-correction Steps, which significantly improves the model's predictive accuracy and reliability.
- We achieve state-of-the-art performance on our customized datasets and obtain better predictive results in real-world cases, providing substantial support for financial sentiment analysis.

2 Related Work

2.1 Entity-level Sentiment Analysis of Financial Texts

NLP techniques have gained widespread adoption in financial sentiment classification (Kazemian et al., 2016; Yang et al., 2022; Chuang and Yang, 2022; Xing et al., 2020). Recent studies have achieved state-of-the-art performance in SemEval 2017 Task 5 and FiQA Task 1 (Du et al., 2023). With advances in NLP, sentiment analysis has shifted from coarse- to fine-grained approaches (Du et al., 2024). However, entity-level sentiment analysis in financial texts, a key fine-grained task, remains in its early stages (Zhu et al., 2020) and faces several challenges (Tan et al., 2023).

Existing financial sentiment classification datasets, such as Financial Phrase Bank (Malo et al., 2014), SemEval2017 (Cortis et al., 2017), AnalystTone dataset (Huang et al., 2023), Headline News dataset (Sinha and Khandait, 2021), and Trillion Dollar Words (Shah et al., 2023a), are based on entire text sequences (sentences or articles). FiQA[1] is an open challenge dataset with aspect-level sentiment. However, it does not include entity annotations. For financial entity annotation datasets, FiNER (Shah et al., 2023b) and FNXL (Sharma et al., 2023) have been created for financial entity recognition and numerical span annotation, respectively, but both lack sentiment annotations. The FinEntity dataset is a dataset with entity spans and sentiment information.

Gururangan et al. (2020); Zhang et al. (2023a) suggest that re-training general PLMs on domain-specific corpora enhances performance on specialized tasks. However, entity-level financial sentiment analysis requires further research due to the unique complexity of financial entities compared to general text entities (Zhang et al., 2023b). Tang et al. (2023) have achieved preliminary results in entity-level sentiment analysis tasks using a combination of FinBERT and CRF.

[1] https://sites.google.com/view/fiqa/home

	FinEntity	SEntFiN-Span	FinEntCN
Number of Texts	979	10753	10832
Single Entity Texts	390 (39.84%)	7897 (73.44%)	8194 (75.65%)
Multiple Entity Texts	589 (60.16%)	2856 (26.56%)	2638 (24.35%)
Average Text Length by Tokens	37.01	9.91	145.23
Max Text Length by Tokens	300	23	518
Min Text Length by Tokens	21	3	12
Positive Entities	503 (23.60%)	5084 (35.21%)	8037 (53.89%)
Negative Entities	498 (23.37%)	3828 (26.51%)	5040 (33.79%)
Neutral Entities	1130 (53.03%)	5527 (38.28%)	1838 (12.32%)
All Entities	2131	14439	14915
Average Entity Num Per Text	2.18	1.34	1.38

Table 1: The statistics of the constructed datasets.

2.2 LLMs in Finance

Large Language Models (LLMs) are considered a technological breakthrough in the field of natural language processing, as exemplified by GPT-3 and GPT-4 (Brown et al., 2020). LLMs have been applied to various tasks in the financial domain (Dredze et al., 2016; Araci, 2019; Bao et al., 2022; DeLucia et al., 2022; Konstantinidis et al., 2024; Ahmed et al., 2024), from predictive modeling to generating insightful narratives from raw financial data.

An early example of a financial LLM is BloombergGPT (Wu et al., 2023). Lan et al. (2024) further apply large models to specific tasks in enterprise alert systems and constructed the FinChina SA dataset, achieving meaningful results. Chen et al. (2024) obtain desirable performance on event-level datasets using their proposed four-hop reasoning chains. However, it is worth noting that experiments have shown that the CoT (Chain-of-Thought) framework negatively impacts entity-level financial sentiment analysis tasks, likely due to the complex reasoning processes involved.

Self-correction techniques aim to improve the accuracy of LLM outputs by enabling models to revise their initial predictions (Pan et al., 2023; Kamoi et al., 2024). The fundamental problem with existing methods is that LLMs cannot reliably assess the correctness of their inferences (Huang et al., 2024a). Recent studies have shown that incorporating examples with feedback into the context can improve response quality (Xu et al., 2024). These findings highlight the significant research potential for correction strategies in entity-level sentiment analysis using LLMs.

3 Datasets Construction

The FinEntity[2] dataset provides annotations for financial entity spans and their associated sentiments within text sequences, with sentiment labels categorized as positive, neutral, or negative (Tang et al., 2023). It treats entity-level sentiment classification as a sequence labeling task using the BILOU annotation scheme. The dataset contains 979 example sentences, with 2,131 entities in total. Additionally, approximately 60% of the financial texts contain multiple entities. However, due to the small size of this dataset, we construct two additional datasets—one in English and one in Chinese—to provide a more comprehensive analysis of this task. Detailed information about the constructed dataset is shown in Table 1, with 20% of the data randomly selected as the test set for experiments.

3.1 English Dataset

SEntFiN[3] is a manually annotated dataset designed for fine-grained financial sentiment analysis of news headlines, with sentiment labels linked to financial entities (Sinha et al., 2022). The sentiment labels are defined as positive, neutral, and negative. In total, the dataset includes 14,404 entity and sentiment annotations.

We apply a rule-based approach to add entity location tags to the annotations, aligning the label format with the FinEntity dataset for ease of subsequent work. This restructured dataset is named SEntFiN-Span.The reconstructed dataset remains relatively balanced in terms of sentiment distribution.

[2] https://github.com/yixuantt/FinEntity
[3] https://github.com/pyRis/SEntFiN

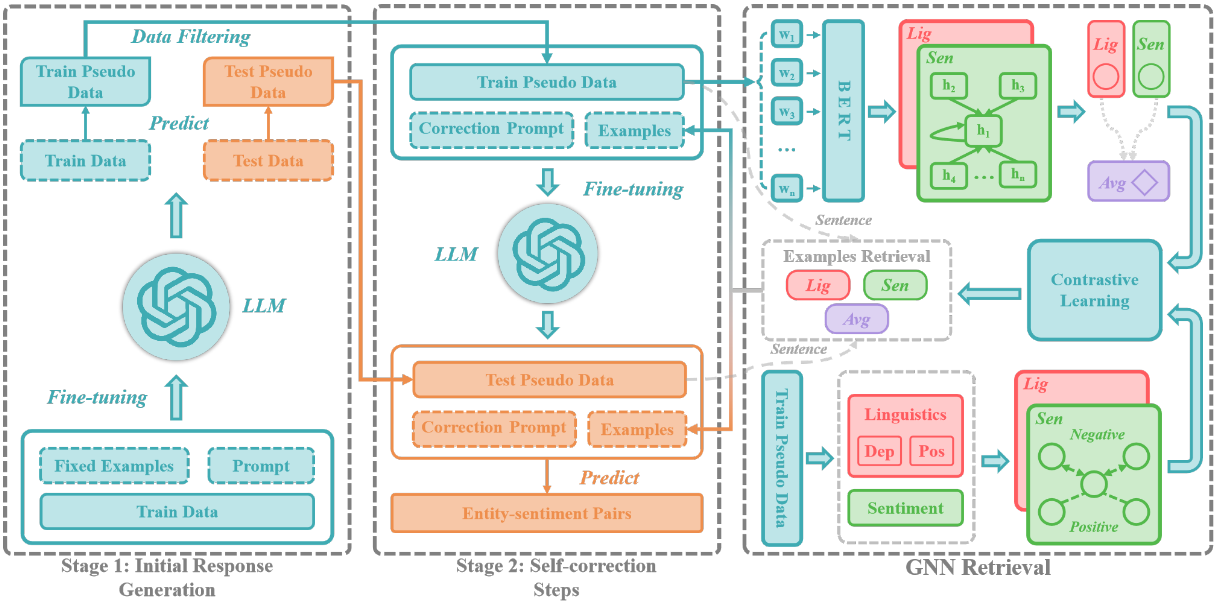

Figure 2: Overview of our framework.

3.2 Chinese Dataset

To address the shortage of Chinese financial sentiment analysis datasets, Chen et al. (2024) pioneer a new task called Event-Level Financial Sentiment Analysis, which involves predicting a five-tuple (company, department, coarse-grained event, fine-grained event, sentiment). To support this task, they construct a large-scale publicly available dataset containing 12,160 news articles and 13,725 five-tuples[4].

To adapt this dataset for entity-level sentiment analysis tasks, we first keep data containing only single-type entity label information. Such data amount to 10,832 instances, accounting for 89% of the dataset. We then use a rule-based approach to annotate the spans of financial entities while ignoring event labels, simplifying the task to purely entity-level financial sentiment analysis. The processed dataset has an average text length of 145.23 words, with 75.65% of the data containing a single entity and 24.35% containing multiple entities. This restructured dataset is named FinEntCN.

4 Methodology

The methodology of our study consists of two stages: base model fine-tuning and error correction model training. The methodological framework is illustrated in Figure 2.

[4] https://anonymous.4open.science/r/EFSA-645E

4.1 Stage 1: Initial Response Generation

In the first stage, we aim to fine-tune the base LLMs for entity-level sentiment analysis in the financial domain. We use two models for different datasets: LlaMA2-7b-hf-finance for the English dataset and Baichuan2-7b for the Chinese dataset.

During the fine-tuning process, we develop multiple versions of instruction templates to ensure their general applicability. Each instruction is designed to comprehensively describe and clearly convey the task requirements, including financial entity recognition, sentiment classification, and the details of result annotation. In the output phase, the model must accurately identify entity names, clearly delineate their boundaries within the text, and classify their sentiment polarity. To enhance the effectiveness of the fine-tuning, we incorporate three manually selected fixed examples into the instructions. These examples are carefully chosen to cover a range of scenarios within the task, providing strong representativeness and comprehensiveness. The instruction template for this stage is shown in Figure 6 in Appendix. After the fine-tuning process, we perform predictions on both the training and test sets, generating initial labels, which we refer to as pseudo-labels.

The problem can be formalized as follows: Let D_{train} and D_{test} represent the training and test datasets, respectively. For each text $t \in D_{train} \cup D_{test}$, we want to predict the sentiment label y_s for each entity e within the text. The fine-tuned model

M_{ft} predicts the sentiment labels:

$$\{(e_i, \text{start}_i, \text{end}_i, \hat{y}_i)\}_{i=1}^{n} = M_{\text{ft}}(t) \quad (1)$$

Where: e_i represents the i-th entity recognized in the text t, start_i and end_i denote the starting and ending positions of entity e_i in t, $\hat{y}_i$ is the sentiment label for entity e_i (positive, negative, or neutral), n is the total number of entities identified in the text.

The pseudo-labels are generated for both the training and test sets. Despite the higher accuracy on the training set, some pseudo-labels are erroneous, reflecting the model's limitations on this specific task. This approach provides foundational data for training the subsequent error correction model.

4.2 Stage 2: Self-correction Steps

The second stage focuses on training an error correction model to identify and rectify errors in the pseudo-labels generated during the first stage. The specific steps are as follows: We begin by filtering the pseudo-labeled data D_{train}. Let $C_{correct}$ and $C_{incorrect}$ represent the correctly and incorrectly labeled samples, respectively. To emphasize the model's attention on erroneous samples, we filter the training data by retaining all incorrectly predicted samples $C_{incorrect}$, while removing a portion of the correctly predicted samples $C_{correct}$:

$$D_{filtered} = C_{incorrect} \cup S(C_{correct}) \quad (2)$$

where S is a sampling function that selects a subset of correctly predicted samples. Next, we fine-tune an error correction model $M_{correct}$ using the filtered dataset $D_{filtered}$.

We reference the GNN-based context example retriever proposed by Yang et al. (2024), as shown in Figure 2. The GNN example retriever uses a graph attention network (GAT) as the base model, with two GAT layers designed for linguistic and sentiment features. It outputs feature representations rich in linguistic and sentiment information, along with sentence-level average representations. During training, the GNN model employs contrastive learning, where linguistic features (such as syntactic dependencies and part-of-speech tags) and sentiment features (such as sentiment polarity) are extracted from the training set based on heuristic rules for comparison. This encourages the GNN's output to be optimized in both linguistic and sentiment dimensions. By encoding the dataset using the trained GNN model, three feature representations enriched with linguistic and sentiment characteristics are obtained. Finally, for prediction, approximate nearest neighbor search is used to retrieve the most similar examples from the encoded training set.

We incorporate the retrieved examples into the context of the fine-tuning instructions, providing information on whether the pseudo-labels are correct. This allows the model to learn how to judge the accuracy of pseudo-labels. Since the model from the first stage may produce similar errors on similar examples, this retrieval method helps the correction model identify and correct these errors. The trained error correction model is then used to detect and correct the pseudo-labels generated by the first stage on the test set. By employing this two-stage approach, we further enhance prediction accuracy and reliability. The corrected-labels are formally consistent with pseudo-labels.

It is important to note that during the fine-tuning of the error correction model, we do not include additional instructions related to entity-level sentiment analysis in the financial domain but focused solely on the error correction task. This design avoids potential negative impacts on model performance from the comprehensiveness of task instructions, ensuring that the model remains concentrated on the error correction task itself. The correction task prompt template can be found in Figure 7 in Appendix.

5 Experiments

5.1 Baselines

FinBERT-CRF: FinBERT[5] is a BERT variant for finance, and FinBERT-CRF adds a CRF layer for token label dependencies (Yang et al., 2020; Tang et al., 2023).

SpanABSA: SpanABSA is a span-based extract-then-classify framework (Hu et al., 2019).

Instruct ABSA: InstructABSA (Scaria et al., 2024) is an aspect-based sentiment analysis method based on instruction learning.

T5: T5 is a general text generation model proposed by Google Research (Raffel et al., 2020).

Ch_finT5_base[6]: Pre-trained Language Model in the Chinese financial domain (Lu et al., 2023).

[5] https://huggingface.co/yiyanghkust/finbert-tone

[6] https://huggingface.co/SuSymmertry/BBT/tree/main/Model/1b

Method	FinEntity			SEntFiN-Span		
	Precision	Recall	F1	Precision	Recall	F1
BERT	-	-	0.8000*	-	-	-
BERT-CRF	-	-	0.8100*	-	-	-
FinBERT	-	-	0.8300*	-	-	-
FinBERT-CRF	-	-	0.8400*	-	-	-
SpanABSA	0.6635	0.5210	0.5837	0.7922	0.7029	0.7449
T5-base	0.8578	0.8477	0.8527	0.7639	0.7546	0.7592
BGCA	0.8625	0.8524	0.8574	0.7626	0.7670	0.7648
InstructABSA	0.8152	0.7471	0.7797	0.7881	0.7826	0.7853
LlaMA2-7B	0.8920	0.8237	0.8565	0.7677	0.7719	0.7698
GPT-3.5$_{(0-shot)}$	0.5253	0.6723	0.5902	0.5143	0.6923	0.5902
GPT-3.5$_{(3-shot)}$	0.7265	0.6347	0.6775	0.5644	0.6525	0.6053
GPT-4o$_{(0-shot)}$	0.7989	0.6714	0.7296	0.5354	0.6721	0.5952
GPT-4o$_{(3-shot)}$	0.7751	0.7588	0.7669	0.6679	0.6299	0.6484
Stage 1 LlaMA2-7B-Finance	0.8983	0.8399	0.8681	0.7833	0.7789	0.7811
Stage 2 (fix)	0.9082	0.8632	0.8826	0.7881	0.7887	0.7884
Stage 2 (gnn)	**0.9104**	**0.8724**	**0.8910**	**0.7895**	**0.7898**	**0.7896**

Table 2: The experimental results on two English datasets (FinEntity, SEntFiN-Span). The BGCA's base model is T5. The reported scores are F1 scores over three runs. '-' denotes that the corresponding results are not available. '*' Indicates that the results are derived from previous studies. The best results are bolded. 'fix' indicates that the in-context examples used are fixed, while 'gnn' indicates they are retrieved using the GNN-based retriever.

Method	FinEntCN		
	Precision	Recall	F1
Ch_finT5_base	0.3582	0.3627	0.3604
BGCA	0.3712	0.3804	0.3757
GPT-3.5$_{(0-shot)}$	0.1564	0.3723	0.2203
GPT-3.5$_{(3-shot)}$	0.4281	0.4858	0.4551
GPT-4o$_{(0-shot)}$	0.2097	0.4468	0.2854
GPT-4o$_{(3-shot)}$	0.2908	0.5177	0.3724
Stage 1	0.8574	0.8546	0.8560
Stage 2 (fix)	0.8625	0.8582	0.8604
Stage 2 (gnn)	**0.8671**	**0.8681**	**0.8675**

Table 3: The experimental results on the Chinese dataset FinEntCN. The Stage 1 results are obtained by fine-tuning Baichuan2-7B-Base-LLaMAfied, and BGCA's base model is Ch_finT5_base.

BGCA: BGCA is a unified bidirectional generation framework for cross-domain ABSA tasks (Deng et al., 2023).

LlaMA2-7B-Finance[7]: This model is fine-tuned on a financial dataset based on the LlaMA2-7B language model.

Baichuan2-7B-Base-LLaMAfied[8]: This is the LLaMAfied version of the Baichuan2-7B-Base model (Huang et al., 2024b).

ChatGPT: To compare with state-of-the-art generative LLMs, we use the OpenAI API[9] to evaluate ChatGPT's zero-shot and few-shot in-context learning performance. Detailed prompts are provided in Appendix A.

5.2 Experimental Settings

The detailed experimental settings of our method and parameter configurations of methods such as SpanABSA, BGCA, and InstructABSA can be found in Appendix B.

5.3 Overall Performance

Table 2 and Table 3 present the main experimental results, demonstrating that our proposed method outperforms all benchmark models on most metrics across three datasets. On the FinEntity dataset, our approach improves the F1 score by 5.1% compared to the previous best method, and it also performs well on the other two datasets. In our comparative study, we explore pre-trained models that have been successfully applied to aspect-based sentiment analysis tasks. The results show that the fine-tuning approach for LLMs demonstrates excel-

[7]https://huggingface.co/cxllin/Llama2-7b-Finance

[8]https://huggingface.co/hiyouga/Baichuan2-7B-Base-LLaMAfied

[9]https://platform.openai.com

lent performance, especially on the Chinese dataset. Our method further achieves the best results.

We also investigate the performance of the GPT series on three datasets, reporting the experimental results of GPT-3.5 and the latest GPT-4o versions in zero-shot and few-shot settings. Due to cost constraints, we select 200 data points from the test set for the experiments. The results indicate that on two English datasets, GPT-4o outperforms GPT-3.5 by 5.21% and 4.31% in F1 scores, respectively, under the 3 in-context examples setup, demonstrating stronger performance. Interestingly, however, both models perform poorly on the Chinese dataset, with GPT-4o even underperforming GPT-3.5. Analysis shows that GPT-4o extracts too much non-financial information, like person entities, indicating that its general extraction capabilities interfere with task-specific understanding.

5.4 Ablation Study

At this stage, we explore the contributions of various components within our framework. Table 4 presents the results of different model variants.

To evaluate the effectiveness of the second stage, we compare the experimental results with those from the first stage, which involves only basic fine-tuning. The results demonstrate that the correction strategy improves performance across all three datasets. Furthermore, we examine the role of the GNN-based example retriever in the second stage. We replace the in-context examples with the same number of fixed examples and report the corresponding results. Overall, the absence of any feature typically results in a decrease in F1 scores compared to the complete model.

Method	FinEntity	SEntFiN-Span	FinEntCN
all	89.10	78.96	86.75
w/o gnn	(-0.84)	(-0.12)	(-0.71)
w/o stage2	(-2.29)	(-0.85)	(-1.15)

Table 4: The Macro-F1 score(%) of the ablation experiments.Values in green indicate the drop in performance after removing a module.

5.5 Key Parameters Analysis

5.5.1 The Number of In-context Examples

Figure 3 presents the experimental results of fine-tuning the LlaMA2-7B model on the FinEntity dataset using different numbers of randomly selected in-context examples. Compared to the task fine-tuning template without examples, the inclusion of such examples leads to a significant improvement in model performance. As the number of in-context examples increases, the results indicate that the model achieves optimal performance with three examples. This finding provides a basis for selecting three examples in the above main experiments.

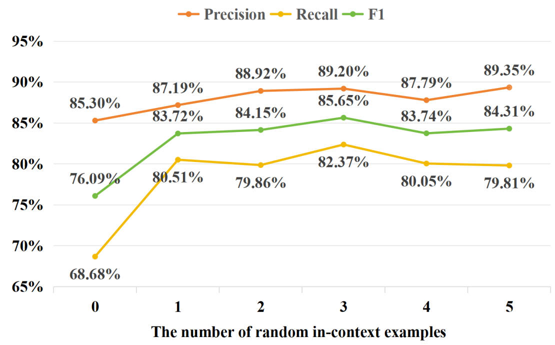

Figure 3: Performance impact of different numbers of in-context examples on the FinEntity dataset.

5.5.2 The Positive-to-negative Sample Ratio

After obtaining the pseudo-labeled data from the first stage, we filter the training set and retain different proportions of correct examples to investigate the impact of the positive-to-negative sample ratio on the performance of the correction model in the second stage. The experimental results are shown in Figure 4. As the proportion of correct samples retained increases, the performance of the fine-tuned correction model initially improves and then declines. This indicates that when training a correction model, an excessively high or low ratio of correct to incorrect samples in the training set can negatively affect correction performance, potentially leading to worse outcomes.

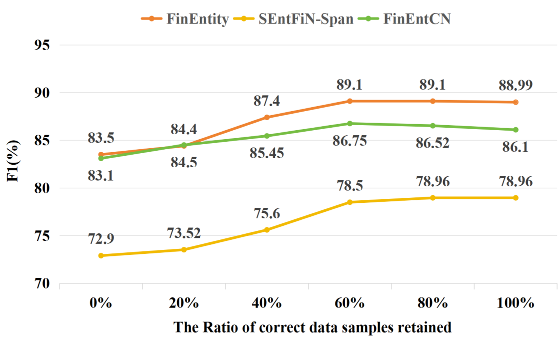

Figure 4: Impact of the ratio of correct data samples retained.

We observe that retaining about 60% or 80% of the correct samples yields the best performance for the correction model. At this point, the overall ratio of correct pseudo-labeled samples to incorrect pseudo-labeled samples in the training data may better align with the true accuracy, enabling the model to have a more accurate perception of its own error rate. This results in improved performance and aligns with human intuition.

5.6 Case Study

To better demonstrate the effectiveness of our framework, we conduct case studies for both English and Chinese texts, as shown in Table 7 in Appendix.

In the English case 1, the model fine-tuned in the first stage accurately identifies two financial entities in the text: Twitter Inc and Tesla Inc. The description of Twitter Inc's stock increase is correctly classified as positive sentiment. However, the model incorrectly classifies the sentiment for Tesla Inc as positive as well, which is a common issue in multi-entity sentiment analysis where the model erroneously assumes uniform sentiment across multiple entities. In the second stage, we introduce three correction examples to guide the model in evaluating and adjusting the pseudo-labels, which leads to successful results. Table 8 in Appendix presents the experimental results of our method, showing the F1 scores for entity and sentiment classification. Both scores improved after our correction stage, with a greater increase in the sentiment polarity score.

Additionally, we observe that LLMs also produce significant errors similar to those in the Chinese case 1, likely due to the lack of financial domain optimization in the Chinese base models. Our correction strategy proves effective in these instances as well. English case 2 and Chinese case 2 demonstrate the limited ability of our method to correct the model's persistent misjudgments regarding entity boundaries and categories.

6 Case Application: Cryptocurrency Market Prediction

Studies have shown a positive contemporaneous correlation between Bitcoin prices and entity-level sentiment scores, with the maximum information coefficient (MIC) between cryptocurrency prices and sentiment indicating a moderate positive correlation. Furthermore, entity-level sentiment demonstrates higher correlations than sequence-level sentiment (Tang et al., 2023), suggesting that market sentiment plays a positive role in regulating price volatility.

Features	RMSE
OHLC + ELS (**SILC**)	0.07936
OHLC + ELS (Bert_CRF)*	0.08502
OHLC + SLS*	0.09549
OHLC only*	0.11218

Table 5: Bitcoin price prediction performance. ELS refers to entity-level sentiment, and SLS refers to sequence-level sentiment. '*' indicates that the data is sourced from the research of Tang et al. (2023).

Based on a dataset of 15,290 timestamped articles from May 20, 2022, to February 1, 2023, we conduct a Bitcoin price prediction task. Sentiments are labeled as positive (+1), neutral (0), and negative (-1), and daily sentiment scores are calculated and normalized. For the price feature, we use the Open-High-Low-Close (OHLC) price, which provides information on the market price movements during a specific time period. A long short-term memory (LSTM) network is used for prediction, with a time step of 10 and a hidden size of 128.

Table 5 reports the RMSE (Root Mean Squared Error) of the model predictions. RMSE measures the average error between the predicted values and the true values, with a smaller RMSE indicating higher predictive accuracy. The results indicate that the LSTM model using only OHLC prices performs the worst, while the model incorporating entity-level sentiment features outperforms both the sequence-level model and the model without sentiment features. The sentiment score features derived from SILC method achieve the best performance.

7 Conclusions

Our research focuses on the task of entity-level sentiment analysis in the financial domain, for which we have constructed the largest English and Chinese datasets. Moreover, we propose an innovative strategy called "Self-aware In-context Learning Correction" (SILC). The SILC framework consists of two stages and significantly improves the accuracy by enabling the model to learn correction examples relevant to the current instance. Experimental results demonstrate that the proposed SILC strategy effectively enhances model performance, achieving state-of-the-art results. Additionally, the case study in the cryptocurrency market demonstrates the practical utility of our datasets and methods, which we believe are valuable resources for financial sentiment analysis.

Limitations

The proposed method involves multiple training stages, which, while enhancing model refinement, also increase training time and computational requirements. This could impact scalability and resource use. To mitigate these issues, optimization techniques such as model pruning and knowledge distillation, as well as cloud computing and distributed training, can be employed.

Ethics Statement

This study adheres to the ACL Code of Ethics. The data collected comes entirely from publicly accessible datasets. Our constructed datasets do not disseminate personal information and do not contain content that could potentially harm any individual or community.

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments. This work is mainly supported by the Key Program of the Natural Science Foundation of China (NSFC) (No.U23A20316), the Key R&D Project of Hubei Province (No.2021BAA029), the Key Research and Development Program of Henan Province (No.241111212700) and the Key Lab of Information Network Security, Ministry of Public Security (No.C23600-04).

References

Rabbia Ahmed, Sadaf Abdul Rauf, and Seemab Latif. 2024. Leveraging large language models and prompt settings for context-aware financial sentiment analysis. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–9. IEEE.

D Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, et al. 2022. Plato-xl: Exploring the large-scale pre-training of dialogue generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 107–118.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tianyu Chen, Yiming Zhang, Guoxin Yu, Dapeng Zhang, Li Zeng, Qing He, and Xiang Ao. 2024. EFSA: Towards event-level financial sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7455–7467, Bangkok, Thailand. Association for Computational Linguistics.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for twitter. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6191–6205.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Mark Dredze, Prabhanjan Kambadur, Gary Kazantsev, Gideon Mann, and Miles Osborne. 2016. How twitter is changing the nature of financial news discovery. In *proceedings of the second international workshop on data science for macro-modeling*, pages 1–5.

Kelvin Du, Frank Xing, and Erik Cambria. 2023. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Transactions on Management Information Systems*, 14(3):1–24.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.

Eugene F Fama. 1970. Efficient capital markets. *Journal of finance*, 25(2):383–417.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Siavash Kazemian, Shunan Zhao, and Gerald Penn. 2016. Evaluating sentiment analysis in the context of securities trading. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2094–2103.

Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G Constantinides, and Danilo Mandic. 2024. Finllama: Financial sentiment classification for algorithmic trading applications. *arXiv preprint arXiv:2403.12285*.

Yinyu Lan, Yanru Wu, Wang Xu, Weiqiang Feng, and Youhao Zhang. 2024. Chinese fine-grained financial sentiment analysis with large language models. *Neural Computing and Applications*, pages 1–10.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

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):782–796.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. Instructabsa: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

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*.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Hongling Xu, Qianlong Wang, Yice Zhang, Min Yang, Xi Zeng, Bing Qin, and Ruifeng Xu. 2024. Improving in-context learning with prediction feedback for sentiment analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3879–3890, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Songhua Yang, Xinke Jiang, Hanjie Zhao, Wenxuan Zeng, Hongde Liu, and Yuxiang Jia. 2024. FaiMA: Feature-aware in-context learning for multi-domain aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7089–7100, Torino, Italia. ELRA and ICCL.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *Preprint*, arXiv:2006.08097.

Yi Yang, Kunpeng Zhang, and Yangyang Fan. 2022. Analyzing firm reports for volatility prediction: A knowledge-driven text-embedding approach. *INFORMS Journal on Computing*, 34(1):522–540.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM at IJCAi*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023b. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 349–356, New York, NY, USA. Association for Computing Machinery.

Qinglin Zhu, Bin Liang, Liuyu Han, Yi Chen, Ruifeng Xu, and Ruibin Mao. 2020. Attention-based recurrent network combined with financial lexicon for aspect-level sentiment classification. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 676–687, Haikou, China. Chinese Information Processing Society of China.

A Instruction Template

A.1 The Instruction Template for ChatGPT

This section provides a directive template for utilizing the GPT model API, designed to perform entity-level financial sentiment analysis and consisting of four key components: System Message, Guidelines, Examples, and Task Text.

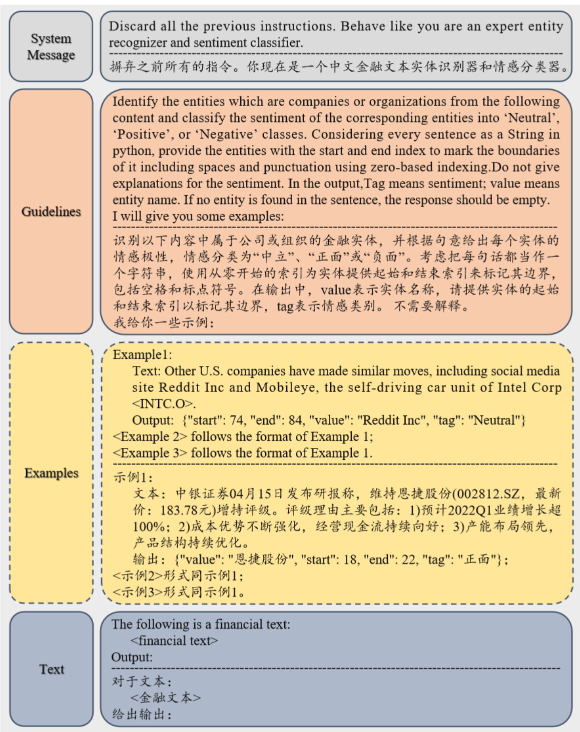

Figure 5: The English and Chinese instruction template for ChatGPT.

A.2 The Instruction Template for SILC

This section introduces the fine-tuning instruction templates for the two stages of SILC. We design multiple similar fine-tuning instruction templates, and one is randomly selected during use. One of them is presented here as an example.

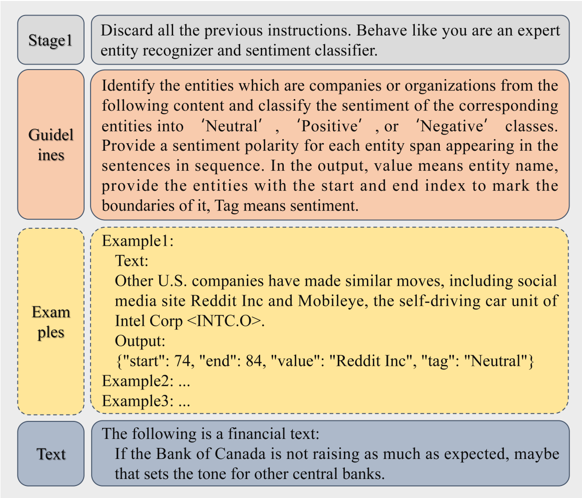

Figure 6: Fine-tuning instruction template for the initial response generation.

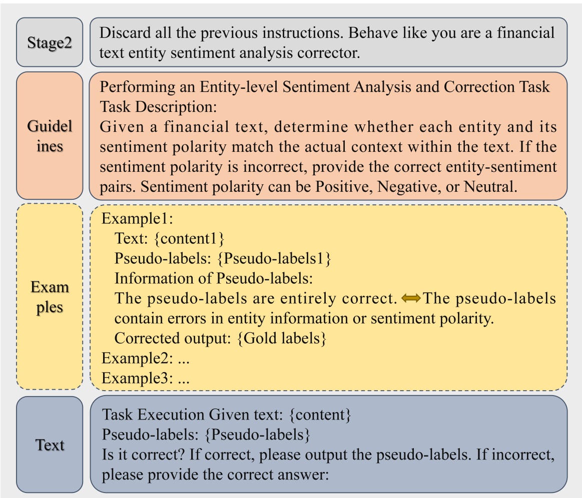

Figure 7: Fine-tuning instruction template for the self-correction steps. The bidirectional arrow indicates that the real usage is either the preceding or the following.

B Experimental Parameter Settings

Our method involves multiple stages. In the first stage of supervised fine-tuning, the learning rate and number of epochs are set to 8×10^{-5} and 5, respectively. For the GNN training phase, we define different heuristic rules for θ_{Lig} and θ_{Sen} to distinguish between linguistic and sentiment feature similarities. The linguistic (θ_{Lig}) and sentiment (θ_{Sen}) parameters are illustrated in Table 6. We use a BERT-based uncased tokenizer as the token encoder, with an initial learning rate of 1×10^{-4}, running for 10 epochs. In the fine-tuning stage of the correction model, we insert the top 3 most relevant examples (k = 3), ranked by similarity (Liu et al., 2022), including one linguistic example, one sentiment example, and one average sample. We leverage Low-Rank Adaptation (LoRA)(Hu et al., 2021) for efficient parameter tuning. All methods utilize the AdamW optimizer(Loshchilov and Hutter, 2019), incorporating gradient decay, dynamic learning rate scheduling, and gradient clipping techniques.

	FinEntity	SEntFiN-Span	FinEntCN
θ_{Lig}	0.37	0.46	0.15
θ_{Sen}	0.8	0.8	0.7

Table 6: The gnn retrieval model training parameters.

All experiments are conducted on an Ubuntu 20.04.5 server equipped with a V100-32G GPU. We randomly split 10% of the training set as the validation set and select the best-performing model, using Macro-F1 as the primary evaluation metric. Each experiment is repeated three times with different random seeds, and we report the average results.

BGCA Method Experimental Settings: BGCA is a data augmentation strategy based on the T5 model. Originally a cross-domain data augmentation method, it is modified to enhance the training set specifically within the financial domain. The experimental parameters are set as follows: num_train_epochs is set to 15, learning_rate is 3×10^{-4}, data_gene_epochs is 20, and whether to use the same model is set to use_same_model.

SpanABSA Method Experimental Settings: The task is set to run_joint_span, train_batch_size is 8, and logit_threshold is 8.0.

InstructABSA Method Experimental Settings: num_train_epochs is set to 4 and learning_rate is 5×10^{-5}.

Task prompt template: """Definition: The output will include each financial entity (both implicit and explicit) along with their sentiment polarity. If there are no financial entities, the output should be noaspectterm: none. The output should follow the order in which the financial entities appear in the text. Two positive examples, Two negative examples, Two neutral examples, Now complete the following example input: """

C Case Study

English case study		
Case#1	Sentence	Micro-blogging site Twitter Inc <TWTR.N> gained 3.8%, adding to its 27% surge in the previous session, after saying it will name top shareholder and Tesla Inc <TSLA.O> CEO Elon Musk to its board.
	Gold labels	{value: Twitter Inc, start: 20, end: 31, tag: Positive} {value: Tesla Inc, start: 149, end: 158, tag: Neutral}
	Stage1 Predict (Pseudo-labels)	{value: Twitter Inc, start: 20, end: 31, tag: Positive} ✓ {value: Tesla Inc, start: 149, end: 158, tag: Positive}✗
	Stage2 Predict (Corrected-labels)	{value: Twitter Inc, start: 20, end: 31, tag: Positive}✓ {value: Tesla Inc, start: 149, end: 158, tag: Neutral}✓
Case#2	Sentence	Rupiah leads Asia FX losses after solid US data, weekly slides seen.
	Gold labels	{value: Rupiah, start: 0, end: 6, tag: negative}
	Stage1 Predict (Pseudo-labels)	{value: Rupiah, start: 0, end: 6, tag: negative}✓ {value: Asia FX, start: 13, end: 20, tag: negative}✗
	Stage2 Predict (Corrected-labels)	{value: Rupiah, start: 0, end: 6, tag: negative}✓ {value: Asia FX, start: 13, end: 20, tag: negative}✗
Chinese case study		
Case#1	Sentence	2月3日,荣联科技(维权)收深交所关注函。公司此前披露,因涉嫌信息披露违法违规,证监会决定对公司立案。深交所要求公司说明非公开发行股票进展情况,并说明立案调查对公司非公开发行股票事项的影响。值得注意的是,2月3日当日,荣联科技还录得涨停。
	Gold labels	{value: 荣联科技, start: 5, end: 9, tag: 负面} {value: 荣联科技, start: 120, end: 124, tag: 正面}
	Stage1 Predict (Pseudo-labels)	{value: 荣联科技, start: 5, end: 9, tag: 中立}✗ {value: 荣联科技, start: 120, end: 124, tag: 中立}✗
	Stage2 Predict (Corrected-labels)	{value: 荣联科技, start: 5, end: 9, tag: 负面}✓ {value: 荣联科技, start: 120, end: 124, tag: 正面}✓
Case#2	Sentence	东风集团股份在港交所公告，2021年12月销售乘用车227782辆，上年同期为293747辆；全年累计销售乘用车2252496辆，同比下降2.6%。其中，2021年12月销售新能源汽车26383辆，上年同期为12661辆；全年累计销售新能源汽车160641辆，同比增长263.3%。
	Gold labels	{value: 东风集团, start: 0, end: 4, tag: 负面}
	Stage1 Predict (Pseudo-labels)	{value: 东风集团股份, start: 0, end: 6, tag: 负面}✗
	Stage2 Predict (Corrected-labels)	{value: 东风集团股份, start: 0, end: 4, tag: 负面}✗

Table 7: English and Chinese case studies.

Method	FinEntity		SEntFiN-Span		FinEntCN	
	Entity	Sentiment	Entity	Sentiment	Entity	Sentiment
stage1	0.9415	0.9469	0.8960	0.8664	0.9321	0.9201
stage2	0.9433	0.9563	0.8994	0.8758	0.9382	0.9308

Table 8: The F1 scores for entity and sentiment polarity in two stages on our dataset. The calculation method deems a prediction correct as long as the predicted entity or sentiment appears in the ground truth, without requiring sequence or full matching.