Beyond Surface Similarity: Detecting Subtle Semantic Shifts in Financial Narratives

Anonymous ACL submission

Abstract

In this paper, we introduce the Financial-STS task, a financial domain-specific NLP task designed to measure the nuanced semantic similarity between pairs of financial narratives. These narratives originate from the financial statements of the same company but correspond to different periods, such as year-over-year comparisons. Measuring the subtle semantic differences between these paired narratives enables market stakeholders to gauge changes over time in the company's financial and operational situations, which is critical for financial decision-making. We find that existing pre-trained embedding models and LLM embeddings fall short in discerning these subtle financial narrative shifts. To address this gap, we propose an LLM-augmented pipeline specifically designed for the Financial-STS task. Evaluation on a human-annotated dataset demonstrates that our proposed method outperforms existing methods trained on classic STS tasks and generic LLM embeddings.

1 Introduction

In accordance with the U.S. Securities and Exchange Commission (SEC) regulations, publicly listed companies are mandated to disclose financial reports. These reports, carefully prepared by the companies, offer a wealth of information about their business operations and financial performance. A substantial amount of natural language processing (NLP) research has focused on this rich dataset to extract insights beneficial for investors and regulators (Cohen et al., 2020; Hoberg and Phillips, 2018; Kogan et al., 2009; Tsai and Wang, 2017; He et al., 2018; Agrawal et al., 2021; Lin et al., 2021; Chun et al., 2023)

An intriguing aspect of corporate financial reporting is the subtle variation in language used to convey information. In corporate communication, companies deliberately select nuanced wording in their communications. For instance, one company's report states, *we report a year of strong performance, with revenues exceeding our targets. Our innovative strategies have driven substantial market growth.* The following year, another statement from the company reads, *we report a year of solid performance, with revenues meeting our targets. Our innovative strategies have led to consistent market share growth.* At first glance, these year-over-year statements appear similar. However, a closer analysis reveals significant differences: the first statement suggests rapid expansion, while the second implies a more steady and moderate growth trajectory. Numerous anecdotal evidences have shown that a company's choice of words can have a huge impact on the company's stock performance (Bochkay et al., 2020; Cohen et al., 2020; Willhite, 2015)

Measuring the similarity in financial narratives resembles the classic Semantic Textual Similarity (STS) task (Mueller and Thyagarajan, 2016; Ranasinghe et al., 2019; Shao, 2017; Tai et al., 2015). However, a significant distinction in the financial narrative STS is that paired financial statements often exhibit a high level of overlap in superficial words. While the semantics appear largely similar on the surface, it is crucial to detect subtle semantic differences that are relevant to market stakeholders. Therefore, we define this task as the *Financial-STS* task, emphasizing its unique characteristics in detecting subtle semantic shifts within the financial domain.

We find that existing pre-trained embedding models or LLM embeddings do not perform satisfactorily for the Financial-STS task. In a preliminary study, we construct a dataset comprising over four thousand paired financial statements. Utilizing OpenAI's Ada embedding[1] and SentenceBERT embedding (Reimers and Gurevych, 2019), we ob-

[1] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

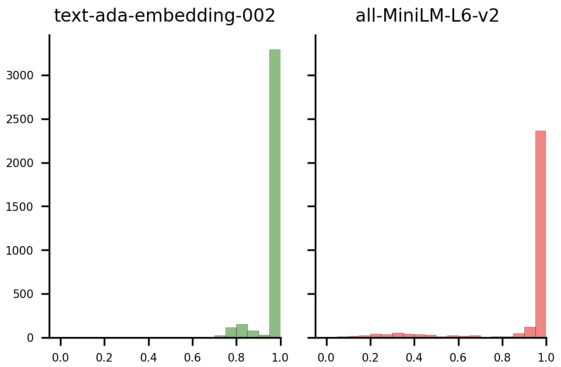

Figure 1: Cosine similarity between 4,027 paired financial statements encoded by OpenAI's Ada embedding ('text-ada-embedding-002') and SentenceBERT ('all-MiniLM-L6-v2'). The pairs were obtained from the annual reports of the Dow Jones Index component companies from year 2018 to 2019.

serve that both models yield excessively high similarity scores for the financial narrative pairs, as illustrated in Figure 1. This suggests a deficiency in pre-trained embedding models in discerning nuanced semantic shifts in financial sentences that are superficially similar. Consequently, for financial-related domain tasks, such as comparing year-over-year financial report similarity (Cohen et al., 2020), the performance of these pre-trained embeddings is unsatisfactory.

This paper proposes a novel method for the Financial-STS task. First, we define four types of subtle semantic changes that convey informational content potentially impactful to financial market stakeholders: *intensified sentiment*, *elaborated details*, *plan realization*, and *emerging situations*. For example, *intensified sentiment* measures situations in which one sentence employs stronger positive or negative phrases compared to another. This can occur when a company's operations improve or exacerbate. Second, inspired by recent NLP advancements using large language models for data augmentation (Dai et al., 2023; Kumar et al., 2020; Yang et al., 2020; Anaby-Tavor et al., 2020; Hu et al., 2023; Schick and Schütze, 2021), we prompt large language models (such as ChatGPT and Llama-2 (Touvron et al., 2023)) to generate financial sentences with no or minimal subtle semantic shifts in one of the four defined categories. Next, with the LLM-augmented dataset, we train a classical Triplet network. This network is capable of distinguishing subtle semantic shift pairs from pairs exhibiting no semantic shift, thus generating meaningful similarity scores for pairs of financial statements. In the evaluation, we manually annotate a dataset with human-judged similarity scores for pairs of financial statements. Results show that our method significantly outperforms existing STS approaches trained on classic STS task, such as SentenceBERT, SimCSE (Gao et al., 2021), and Contriver (Izacard et al., 2022), as well as generic LLM embeddings such as OpenAI's Ada embedding.

Our research makes two significant contributions. First, we introduce a novel financial NLP task, Financial-STS, which focuses on financial sentence pairs that are superficially similar but may differ subtly in semantics. Second, we present a comprehensive pipeline designed to effectively detect subtle semantic shifts in financial narratives. This pipeline, accompanied by datasets comprising both LLM-augmented sentence pairs and a manually annotated dataset, will be made publicly available to facilitate further research and application in this field.

2 STS in Financial Narratives

In the realm of corporate communication, financial documents such as annual reports and press releases play a pivotal role for companies in communicating with capital markets. Due to the significance of financial texts, companies carefully craft their narratives. For instance, research in financial economics has shown that managers tend to use more positive words, such as 'tremendous' or 'extremely well', to convey good news. Conversely, for negative news, they are inclined to use moderately negative terms like 'limitation', 'unexpected', or 'complexity' (Bochkay et al., 2020). More importantly, analyzing year-over-year narrative changes, particularly in sentiment and modifier words, enables investors to accurately assess fundamental shifts in a company. Cohen et al. (2020); Brown and Tucker (2011) find that annual modifications in company reports, especially sentiment shifts, correlate with the company's future stock returns and trading volume. Hence, evaluating the semantic similarity between financial narratives is an essential yet under-explored task.

We identify and summarize four categories of semantic shifts, focusing on those that may potentially provide informational content for stakeholders. In addition to the four identified types of semantic shifts, a considerable number of financial statements exhibit no semantic shift, but only se-

	Sentence (Year: 2018)	Sentence (Year: 2019)
Intensified Sentiment	The Company is subject to laws and regulations worldwide, changes to which could increase the Company's costs and individually or in the aggregate adversely affect the Company's business.	Company is subject to **complex and changing** laws and regulations worldwide, which exposes the Company to potential liabilities, increased costs and other adverse effects on the Company's business.
Elaborated Details	If the other businesses on whose behalf we perform inventory fulfillment services deliver product to our fulfillment centers in excess of forecasts, we may be unable to secure sufficient storage space and may be unable to optimize our fulfillment network.	**Our failure** to properly handle such inventory or the inability of the other businesses on whose behalf we perform inventory fulfillment services to accurately forecast product demand may result in us being unable to secure sufficient storage space or to optimize our fulfillment network **or cause other unexpected costs and other harm to our business and reputation.**
Plan Realization	JPMorgan Chase **expects that** under CECL, it will need to, among other things, increase the allowance for credit losses related to its loans and other lending-related commitments, which **may have** a negative impact on its capital levels.	The allowance for credit losses related to JPMorgan Chase's loans and other lending-related commitments increased **as a result of** the implementation of CECL, which **has** a negative impact on JPMorgan Chase's capital levels.
Emerging Situations	On the other hand, a low interest rate environment may cause: 1)net interest margins to be compressed, which could reduce the amounts that JPMorgan Chase earns on its investment securities portfolio to the extent that it is unable to reinvest contemporaneously in higher-yielding instruments, and 2) a reduction in the value of JPMorgan Chase's mortgage servicing rights ("MSRs") asset, thereby decreasing revenues.	On the other hand, a low **or negative** interest rate environment may cause: 1) net interest margins to be compressed, which could reduce the amounts that JPMorgan Chase earns on its investment portfolio to the extent that it is unable to reinvest contemporaneously in higher-yielding instruments **2) unanticipated or adverse changes in depositor behavior, which could negatively affect JPMorgan Chase's broader asset and liability management strategy,** and 3) a reduction in the value of JPMorgan Chase's mortgage servicing rights ("MSRs") asset, thereby decreasing revenues.
No Semantic Shift	**Many of our competitors are companies that are larger than we are**, with greater financial and operational resources than we have.	**We compete with many larger companies** that have greater financial and operational resources than we have.

Table 1: Nuanced semantic shifts examples. Financial statement pairs are extracted from the annual reports of 2018 and 2019, respectively. Words and phrases that are indicative of semantic shifts are boldface. Last row shows an example of *No Semantic Shift*.

mantic paraphrasing. We provide examples of the four categories of semantic shifts, as well as an example of no semantic shift, in Table 1.

- *Intensified Sentiment*: One sentence employs stronger positive or negative phrases compared to another. This occurs when a company's operations improve or exacerbate.
- *Elaborated Details*: One sentence offers significantly more details about a business situation than another. This may happen when new regulations or changes in existing ones necessitate more detailed disclosures in financial statements.
- *Plan Realization*: One sentence forecasts a future event, while another mentions that this event has already occurred or is currently happening. Under specific regulations, companies must disclose potential future risks or events, and then they are required to update stakeholders on these issues.
- *Emerging Situations*: One sentence introduces completely new information compared to another. This occurs when a company releases new information to address an emerging change in market conditions.

Financial-STS task: Formally, given two **paired** financial narratives, s_i and s_j, Financial-STS aims to develop a mapping function Φ. This function is designed to measure the level of semantic similarity between the statements, represented as $\Phi(s_i, s_j) \rightarrow \mathbb{R}$. A lower similarity score signifies a more significant semantic shift.

It is noteworthy that the Financial-STS task focuses on paired narratives. Two financial narratives are considered paired if they meet the following requirements: First, both s_i and s_j must originate from the *same* company's financial statements, but from *different* periods (such as year-over-year or quarter-over-quarter). Second, they should focus on the same aspect of business operations and exhibit a high level of overlap in superficial words.

3 Proposed Pipeline

As demonstrated in Figure 1, existing pretrained embedding models fall short in measuring the sub-

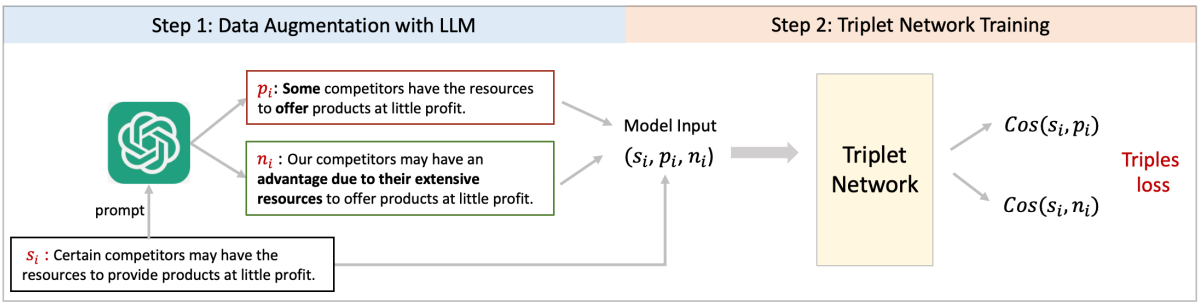

Figure 2: We propose to prompt LLM to generate financial narrative pairs that exhibit either no semantic shift or minimal shift, based on the identified semantic shift categories.

tle semantic shifts in financial statements for two main reasons. First, financial statements often overlap in superficial words, yet the semantic shifts in these statements occur in sophisticated ways, posing a challenge. For example, a shift might involve a change from a modal verb phrase such as *likely to affect revenue* to a past participle like *affected revenue*. Second, existing pretrained embedding models, such as SentenceBERT (Reimers and Gurevych, 2019), are fine-tuned on classic STS benchmarks (Cer et al., 2017; Agirre et al., 2012), which are not adequately suited for the Financial-STS task.

In this section, we propose a pipeline for addressing the Financial-STS task, as shown in Figure 2. In essence, we utilize large language models (LLMs) to generate an augmented dataset, in which examples exhibit different types of semantic shifts as well as instances of no semantic shift (Section 3.1). Then, we use the LLM-augmented dataset to train a classic Triplet network capable of differentiating between pairs of subtle semantic shifts and pairs showing no semantic shift (Section 3.2).

3.1 Financial Semantic Shift Data Augmentation with LLM

To the best of our knowledge, no publicly available semantic similarity dataset specifically designed for paired financial narratives exists. Recent literature has demonstrated that large language models (LLMs) can generate high-quality datasets benefiting specific NLP tasks (Dai et al., 2023; Hu et al., 2023; Schick and Schütze, 2021). Motivated by this, we prompt LLMs to generate specific pairs of financial narratives.

Based on the four nuanced semantic shift types defined in Table 1, we develop different prompts corresponding to each type, as illustrated in Table 2. In these prompts, we present a focal financial narrative example to an LLM and request the generation of a semantically similar sentence. Each generated sentence should exhibit a subtle shift aligned with one of the four categories. For instance, in the *intensified sentiment* category, we instruct the LLM to express a stronger negative sentiment using negative words while maintaining the rest of the sentence's semantic unchanged. Additionally, we include a one-shot example in each prompt to guide the LLM more effectively in adhering to the instruction.

Additionally, we prompt the LLM to generate a paraphrased sentence that maintains the original semantic content. Thus, for a given financial narrative, we obtain two examples: one without a semantic shift, considered a positive example, and another with a specific type of semantic shift, considered a negative example. We denote the LLM-augmented dataset as $\mathcal{S} = \{(s_i, p_i, n_i)\}_{i=1}^{N}$, where s_i is a focal financial narrative, p_i is the corresponding LLM-augmented positive example (no semantic shift), and n_i is the corresponding LLM-augmented negative example (semantic shift).

3.2 Triplet network for Measuring Financial Narrative Similarity

Our objective is to train a network capable of taking a pair of financial narratives and producing a numerical similarity score. According to the LLM-augmented dataset, we anticipate that the similarity between a financial narrative and its corresponding augmented positive pair will be higher than that between the narrative and its augmented negative pair. For this purpose, we employ a classic Triplet network, an effective network for classic STS task (Reimers and Gurevych, 2019).

Specifically, for a triplet input (s_i, p_i, n_i), we feed each of its sentence using the BERT model (Vaswani et al., 2017), resulting to a 768-

Intensified Sentiment
You are required to finish the task: Restating the given sentence so that the resulting sentence is semantically similar to the original sentence, but with much stronger negative sentiment by using more negative words. **### Example:** The given sentence is: Changes in laws, regulations and policies and the related interpretations and enforcement practices may alter the landscape in which we do business and may significantly affect our cost of doing business. Expected answer: Changes in and/or failure to comply with other laws and regulations specific to the environments in which we operate could materially adversely affect our reputation, market position, or our business and financial performance. **### Question:** The given sentence is: SENTENCE. Expected answer:
Elaborated Details
You are required to finish the task: Restating the given sentence so that the resulting sentence is semantically similar to the original sentence, but with much stronger negative sentiment by using more detailed description about the unfavorable situation. **### Example:** The given sentence is: We also have outsourced elements of our operations to third parties, and, as a result, we manage a number of third-party vendors who may or could have access to our confidential information. Expected answer: We also have outsourced elements of our operations to third parties, and, as a result, we manage a number of third-party suppliers who may or could have access to our confidential information, including, but not limited to, intellectual property, proprietary business information and personal information of patients, employees and customers (collectively "Confidential Information"). **### Question:** The given sentence is: SENTENCE. Expected answer:
Plan Realization
You are required to finish the task: Restating the given sentence so that the resulting sentence is semantically similar to the original sentence, but with much stronger negative sentiment by changing the tense (from going to influence to have influenced). **### Example:** The given sentence is: Although these attacks and breaches have not had a direct, material impact on us, we believe these incidents are likely to continue and we are unable to predict the direct or indirect impact of future attacks or breaches to our business. Expected answer: Such attacks and breaches have resulted, and may continue to result in, fraudulent activity and ultimately, financial losses to Visa's clients, and it is difficult to predict the direct or indirect impact of future attacks or breaches to our business. **### Question:** The given sentence is: SENTENCE. Expected answer:
Emerging Situations
You are required to finish the task: Restating the given sentence so that the resulting sentence is semantically similar to the original sentence, but with much stronger negative sentiment by adding some unfavorable circumstances. **### Example:** The given sentence is: These tariffs, and any additional tariffs imposed by the U.S., China or other countries or any additional retaliatory measures by any of these countries, could increase our costs, reduce our sales and earnings or otherwise have an adverse effect on our operations. Expected answer: While the U.S. and China signed what is being known as the Phase One Deal in January 2020, which included the suspension and rollback of tariffs, any new tariffs imposed by the U.S., China or other countries or any additional retaliatory measures by any of these countries, could increase our costs, reduce our sales and earnings or otherwise have an adverse effect on our operations. **### Question:** The given sentence is: SENTENCE. Expected answer:
No Semantic Shifts
You are required to finish the task: Restating the sentence so that the resulting sentence is semantically and sentimentally similar to the given sentence. **### Example:** The given sentence is: Many of our competitors are companies that are larger than we are, with greater financial and operational resources than we have. Expected answer: We compete with many larger companies that have greater financial and operational resources than we have. **### Question:** The given sentence is: SENTENCE. Expected answer:

Table 2: Prompts used to generate the augmented dataset for each semantic shift category, as well as for the no semantic shift category

dimensional embedding triplet $(\overrightarrow{s_i}, \overrightarrow{p_i}, \overrightarrow{n_i})$ for (s_i, p_i, n_i) respectively.

We further fine-tune the BERT model so that similarity between a financial narrative embedding $\overrightarrow{s_i}$ and its augmented positive pair embedding $\overrightarrow{p_i}$ is higher than that between $\overrightarrow{s_i}$ and its augmented negative pair embedding $\overrightarrow{n_i}$. Thus, we use the triplet loss:

$$\max\left(cos(\overrightarrow{s_i}, \overrightarrow{n_i}) - cos(\overrightarrow{s_i}, \overrightarrow{p_i}) + \epsilon, 0\right) \quad (1)$$

where ϵ is the margin hyperparameter that defines how far apart negative examples should be from the positive examples. We use the LLM-augmented dataset to fine-tune the BERT model.

4 The FinSTS dataset

In this section, we present the **FinSTS** dataset, which comprises two distinct subsets: an LLM-augmented dataset and a human-annotated dataset. We use the LLM-augmented dataset to train the Triplet network and use the human-annotated dataset for evaluation.

4.1 LLM-augmented FinSTS Dataset

As described in Section 3.1, we propose to leverage LLM to generate an augmented dataset for subsequent network training. Specifically, we select annual reports from Dow Jones 30 index firms during the period from 2018 to 2019 as our sample. We parse the annual reports (Item 1A) into sentence-level using the Python NLTK library, resulting in

		GPT-turbo-3.5		Llama-13B-chat	
		(s_i, p_i)	(s_i, n_i)	(s_i, p_i)	(s_i, n_i)
Data Description	Size	8,803	8,803	8,803	8,803
	#Tokens	(36.59, 35.81)	(36.59, 43.40)	(36.59, 30.36)	(36.59, 45.94)
Jaccard Similarity	25%↑	0.821	0.821	0.852	0.853
	50% ↑	0.855	0.867	0.881	0.886
	75% ↑	0.897	0.926	0.946	0.963
Semantic shifts	Transrate↑	0.027	0.248	0.032	0.093

Table 3: Description and Evaluation of FinSTS dataset. Both positive and negative pairs exhibit a high level of superficial overlap, and negative pairs has larger semantic shifts than positive pairs.

a total of 8,803 sentences (4,330 for the year 2018 and 4,473 for the year 2019).

To examine the generalizability of the proposed method, we employ two LLMs: GPT-3.5-turbo and Llama-13B-chat (Touvron et al., 2023), for generating an augmented dataset. In total, we obtain two sets of 8,803 sentence triplets. Each triplet comprises one focal financial narrative from the annual report, one corresponding LLM-augmented positive example (without semantic shift), and one corresponding negative example (with semantic shift). The dataset description is shown in Table 3.

4.1.1 Dataset Assessment

We now quantitatively assess the quality of the LLM-augmented FinSTS dataset.

LLM-augmented financial narrative pairs exhibit a high level of superficial overlap. We calculate the Jaccard similarity for paired sentences in both LLM-augmented datasets at the token level. As shown in Table 3, the 25th percentile of Jaccard similarities reaches 0.821, and the 75th percentile is as high as 0.963. These figures indicate a significant level of superficial overlap in augmented financial narratives.

Semantic Shifts in the augmented dataset are as expected. We also study whether the sentence pairs in the LLM-augmented FinSTS dataset meet our requirements. Specifically, we measure the mutual information between the two sets of samples and their labels using the TrasRate score (Huang et al., 2022; Dai et al., 2023). A low TransRate score suggests difficulty in differentiating the examples from the two sets. Our analysis reveals that the TransRate between (s_i, p_i) is very low, indicating minimal semantic shift between the sets. Furthermore, the average TransRate for positive pairs is lower than that for negative pairs, confirming that the LLM-augmented FinSTS meets the desired requirements.

4.2 Human-annotated FinSTS Dataset

In addition to the LLM-augmented FinSTS Dataset, we have also manually annotated another dataset with financial narrative pairs which can serve as ground truth for evaluation. To obtain paired narratives, we use a different group of S&P 500 companies' annual reports in the year 2018 and 2019. Since paired financial narratives should focus on the same aspect of business operations with a high level of overlap in superficial words, we treat it as an assignment problem. We first employ BERT to encode each sentence from the annual report into an embedding. Then, we calculate the pairwise similarity between sentences in 2018 and sentences in 2019. Finally, we utilize the Hungarian algorithm (Kuhn, 1955) to match each 2018 sentence with a corresponding 2019 sentence, ensuring that the overall sum of similarities is maximized. In total, we select 370 paired financial narratives.

Annotators are instructed to label each pair on two dimensions: 1) Semantic shift score, where 1 indicates no shift and -1 indicates the presence of a shift; 2) If the score is -1, they are further asked to identify the specific type from four predefined categories. Each sentence pair is independently annotated by two human annotators. In cases of inconsistency between annotations, a third annotator will discuss the issue with the initial annotators to resolve any differences. The Cohen's kappa coefficient (Cohen, 1960) for our annotation process is 0.9183, demonstrating a high level of inter-rater reliability. Detailed annotation guidelines are provided in Appendix B.

The **FinSTS** dataset, including the LLM-augmented and a human-annotated dataset will be made publicly available for future research.

5 Evaluation of Financial-STS task

In this section, we test our model performance on Financial-STS task. We use 85% of examples in

	GPT-turbo-3.5		Llama-13B-chat	
Dataset	Splited Test set	Annotated dataset	Splited Test set	Annotated dataset
SBERT(all-MiniLM-L6-v2)	$0.737_{35.06\%\uparrow}$	$0.599_{18.46\%\uparrow}$	$0.658_{49.59\%\uparrow}$	$0.599_{26.46\%\uparrow}$
SBERT(all-mpnet-base-v2)	$0.814_{22.27\%\uparrow}$	$0.626_{13.39\%\uparrow}$	$0.741_{32.96\%\uparrow}$	$0.626_{21.04\%\uparrow}$
SimCSE(sup)	$0.885_{12.43\%\uparrow}$	$0.590_{20.33\%\uparrow}$	$0.805_{22.33\%\uparrow}$	$0.589_{28.45\%\uparrow}$
SimCSE(unsup)	$0.781_{27.42\%\uparrow}$	$0.563_{26.14\%\uparrow}$	$0.746_{31.99\%\uparrow}$	$0.563_{34.65\%\uparrow}$
Contriver	$0.679_{46.48\%\uparrow}$	$0.616_{15.28\%\uparrow}$	$0.577_{70.77\%\uparrow}$	$0.616_{23.06\%\uparrow}$
ADA(text-embedding-ada-002)	$0.779_{27.88\%\uparrow}$	$0.624_{13.80\%\uparrow}$	$0.690_{42.78\%\uparrow}$	$0.624_{21.48\%\uparrow}$
Ours	**0.995**	**0.710**	**0.985**	**0.758**

Table 4: Evaluation result of Financial-STS task. AUC on two evaluation dataset using GPT-3.5-turbo and Llama-13B-chat as data augmentation models.

LLM-augmented FinSTS for training. We use a batch size of 64, margin of $\epsilon = 0.2$, Adam optimizer with a learning rate of 2e-5, and implement linear learning rate warm-up over 10% of the training data. The pooling strategy we choose is mean pooling.

5.1 Financial-STS Benchmarks

We consider the following benchmarks:

- SBERT (or SentenceBERT) (Reimers and Gurevych, 2019): It is an state-of-the-art model that is fine tuned on classic STS tasks. We employ both "all-MiniLM-L6-v2" and "all-mpnet-base-v2" from SentenceTransformer library[2].

- SimCSE (Gao et al., 2021): SimSCE uses contrastive learning for sentence encoding, which achieves significant improvement on classic STS tasks. We consider both unsupervised and supervised versions based on BERT.("sup-simcse-bert", "unsup-simcse-bert").

- Contriver (Izacard et al., 2022): Contriver also uses contrastive learning for sentence encoding which works well for information retrieval tasks. We use the pretrained Contriver. [3]

- ADA[4]: We also consider the state-of-the-art LLM embedding provided by OpenAI, named "text-embedding-ada-002," which demonstrates strong performance on the classic STS task.

[2]https://www.sbert.net/

[3]https://github.com/facebookresearch/contriever

[4]https://openai.com/blog/new-and-improved-embedding-model

5.2 Evaluation Dataset and Metrics

Benchmarks are evaluated on the following datasets. **LLM-augmented FinSTS test set:** 15% of the examples from the LLM-augmented FinSTS is held out as the test set. **Human-annotated FinSTS:** all of the human-annotated dataset are used for testing.

We use Area Under the ROC Curve (**AUC**) as the evaluation metric to assess the quality of identified semantic simiarlity between a pair of financial narratives. A high AUC means that a model ranks the positive (similar) pairs higher than the negative (dissimilar) pairs consistently.

5.3 Experiment results

Our method exhibits superior performance in the Financial-STS task. As shown in Table 4, our method significantly outperforms all benchmarks on both the LLM-augmented FinSTS test set and the human-annotated FinSTS datasets. The largest improvement is observed in the LLM-augmented FinSTS test set, where our method achieves a near-perfect AUC score. This is not surprising, considering our method is fine-tuned using the LLM-augmented FinSTS training set. However, the performance improvement on the human-annotated FinSTS data is quite encouraging. For instance, our method achieves an AUC of 0.7576 on the human-annotated FinSTS dataset, marking a 21.48% improvement compared to the ADA embedding. This underscores our method's capability in discerning nuanced semantic similarities between pairs of financial narratives. It also demonstrates that a model trained on the LLM-augmented dataset can perform effectively on real-world financial narrative pairs.

The utilization of category-specific prompts has demonstrated the potential to enhance model

performance. To further investigate this, we conduct an experiment examining the impact of including different prompts designed for specific types of semantic shift. For each type, we remove its corresponding data from the LLM-augmented FinSTS and train a distinct Triplet model. Subsequently, we assess the model's performance on examples of the respective semantic shift type within the human-annotated FinSTS dataset.

The results are presented in Table 5. Consider the first column, C1 (*intensified sentiment*), as an example: our model is trained on the LLM-augmented FinSTS dataset, excluding the data for each specific semantic shift type. Notably, the model train without the C1 augmented data shows the lowest performance in identifying the intensified sentiment semantic shift, with an AUC of 0.796. This pattern is consistent across other semantic shift types, except for C4 (*Emerging Situations*), where the lowest performance is observed in the model trained without C2 (*Elaborated Details*) examples. This could be attributed to the similarities between C2 and C4 types, as both involve providing certain details.

	C1	C2	C3	C4
Train w/o C1	**0.796**	0.752	0.809	0.762
Train w/o C2	0.826	**0.692**	0.810	**0.691**
Train w/o C3	0.848	0.771	**0.774**	0.758
Train w/o C4	0.808	0.728	0.808	0.722

Table 5: AUC performance on certain semantic shift type. C1: Intensified Sentiment; C2: Elaborated Details; C3: Plan Realization; C4: Emerging Situations. Boldface indicates the model with the lowest performance

6 Related work

Semantic Textual Similarity: Our Financial-STS task, which aims to measure the semantic similarity between paired financial narratives, is related to the classic Semantic Textual Similarity (STS) task (Majumder et al., 2016; Wang and Dong, 2020; Mueller and Thyagarajan, 2016; Ranasinghe et al., 2019; Reimers and Gurevych, 2019). However, in the Financial-STS task, the paired financial narratives originate from the same company's financial statements but from different periods, and they exhibit a high level of superficial overlap. This makes it challenging to discern subtle semantic differences. Consequently, existing embedding models trained for the classic STS task do not perform satisfactorily on the Financial-STS task.

Data augmentation with LLMs: Data augmentation has been widely adopted to enhance classification tasks by generating semantically similar texts (Wei and Zou, 2019; Feng et al., 2021; Shorten et al., 2021; Dai et al., 2023). This study explores leveraging large language models to synthesize more nuanced data - samples that exhibit no or minimal semantic changes in a specific domain.

Financial NLP tasks: A growing body of literature has benchmarked LLMs for NLP tasks in the financial domain (Guo et al., 2023; Shah et al., 2022; Xie et al., 2023). However, a benchmark for financial narrative similarity has not yet been established. This work introduces Financial-STS, a classic yet domain-specific task, and presents the FinSTS dataset, a financial semantic similarity dataset.

7 Conclusion

In this paper, we study a new financial domain NLP task that measures the subtle semantic shift between a pair of financial narratives with high superficial similarity, which we call Financial-STS. We find that existing pretrained embedding models fall short in discerning the nuanced semantic shifts between these narratives. As a result, financial market practitioners face challenges in accurately gauging a company's financial and operational changes.

To address this problem, we identify four types of subtle semantic shifts commonly occurring in companies' financial narratives. Based on these identified shifts, we prompt a large language model to generate sentence pairs that exhibit subtle or no semantic shift. We then use the LLM-augmented dataset to train a classic triplet network capable of measuring the semantic similarity between pairs of financial narratives. Experimental results on a human-annotated dataset demonstrate that our proposed approach outperforms existing pre-trained embedding models specialized for the classic STS task, as well as LLM embeddings such as OpenAI's Ada.

By introducing this task and publicly releasing a new FinSTS dataset, we aim not only to establish a new benchmark for evaluating financial domain LLMs (Wu et al., 2023; Yang et al., 2023) but also to complement the classic STS task, thereby fostering further advancement in this area.

8 Limitations

This paper has several limitations for improvement in future research. Firstly, our experiments are restricted to annual reports, disregarding alternative financial data sources such as earnings conference calls and quarterly reports that exhibit a quarter-to-quarter structure. Subsequent investigations can adopt our methodology to delve into these additional data sources. Secondly, due to limitations of computation resources, we only employ the 13B version of open-source Llama for data generation. Future research can explore larger-scale language models and compare the quality of the generated data accordingly. Lastly, in this study, we solely identify sentence pairs exhibiting shifts in semantics, without distinguishing whether these shifts imply positive or negative financial signals relevant to market stakeholders. We leave this distinction for future investigations.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Yash Agrawal, Vivek Anand, S Arunachalam, and Vasudeva Varma. 2021. Hierarchical model for goal guided summarization of annual financial reports. In *Companion Proceedings of the Web Conference 2021*, pages 247–254.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Khrystyna Bochkay, Jeffrey Hales, and Sudheer Chava. 2020. Hyperbole or reality? investor response to extreme language in earnings conference calls. *The Accounting Review*, 95(2):31–60.

Stephen V Brown and Jennifer Wu Tucker. 2011. Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, 49(2):309–346.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ye Eun Chun, Sunjae Kwon, Kyunghwan Sohn, Nakwon Sung, Junyoup Lee, Byungki Seo, Kevin Compher, Seung-won Hwang, and Jaesik Choi. 2023. Cr-copec: Causal rationale of corporate performance changes to learn from financial reports. *arXiv preprint arXiv:2310.16095*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Lauren Cohen, Christopher Malloy, and Quoc Nguyen. 2020. Lazy prices. *The Journal of Finance*, 75(3):1371–1415.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yue Guo, Zian Xu, and Yi Yang. 2023. Is chatgpt a financial expert? evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585, Melbourne, Australia. Association for Computational Linguistics.

Gerard Hoberg and Gordon M Phillips. 2018. Text-based industry momentum. *Journal of Financial and Quantitative Analysis*, 53(6):2355–2388.

Xuming Hu, Aiwei Liu, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, and Philip S. Yu. 2023. GDA: Generative data augmentation techniques for relation extraction tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10221–10234, Toronto, Canada. Association for Computational Linguistics.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. 2022. Frustratingly easy transferability estimation. In *International Conference on Machine Learning*, pages 9201–9225. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Ting-Wei Lin, Ruei-Yao Sun, Hsuan-Ling Chang, Chuan-Ju Wang, and Ming-Feng Tsai. 2021. Xrr: Explainable risk ranking for financial reports. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part IV 21*, pages 253–268. Springer.

Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4):647–665.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2019. Semantic textual similarity with Siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, Varna, Bulgaria. INCOMA Ltd.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Yang Shao. 2017. HCTI at SemEval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 130–133, Vancouver, Canada. Association for Computational Linguistics.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

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*.

Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

James Willhite. 2015. How to ace an earnings conference call. *The Wall Street Journal*. Available at: https://www.wsj.com/articles/BL-CFOB-7597 (Accessed: Dec 15th, 2023).

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*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

A Examples of FinSTS

We present various examples of LLM-augmented FinSTS Dataset in Table 6.

B Annotation Guidelines

All annotators are master's or doctoral students with backgrounds in both Natural Language Processing and finance. They are presented with the annotation interface where the difference between two sentences are highlighted. And for each sentence pair, the annotators use the following instructions (Figure 3) to provide their labels on two dimensions (score and category). Each sentence pair is annotated by two individuals. For sentences where there are discrepancies in the annotations, a third annotator will discuss the issue with the initial annotators to resolve any differences.

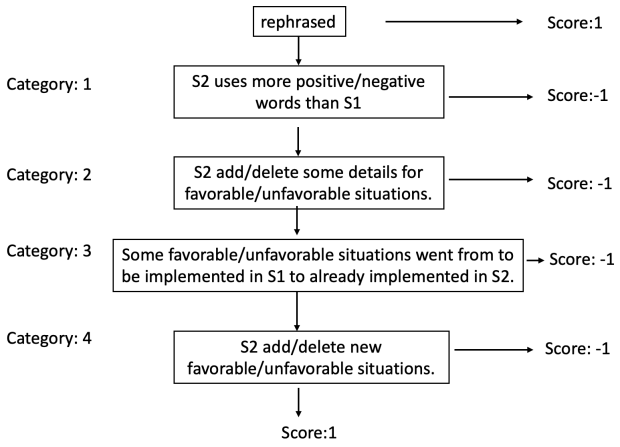

Figure 3: Annotation instructions.

Category: Intensified Sentiment	
Anchor sentence	The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not limited to those described below, any one or more of which could, directly or indirectly, cause the Company's actual financial condition and operating results to vary materially from past, or from anticipated future, financial condition and operating results.
Positive sentence	The Company's financial condition and operating results can be influenced by various factors, both known and unknown, such as those described below, which could cause material variations in our actual financial condition and operating results compared to our past or anticipated future performance.
Negative sentence	The Company's financial performance and stability could be **gravely jeopardized** by a variety of unforeseen and uncontrollable factors, such as but not limited to the ones mentioned below, which could cause a significant decline in our financial health and market position.
Category: Elaborated Details	
Anchor sentence	With respect to Insurance Services' investment activities, changes in the equity and bond markets could result in a decline in value of its investment portfolio, resulting in an unfavorable impact to earnings.
Positive sentence	The performance of Insurance Services' investment activities could be negatively impacted by changes in the equity and bond markets, leading to a decrease in the value of its investment portfolio and unfavorable earnings.
Negative sentence	With respect to Insurance Services' investment activities, drastic fluctuations in the equity and bond markets could result in a precipitous decline in value of its investment portfolio, **potentially leading to a catastrophic impact on earnings, resulting in a devastating blow to the company's financial performance and stability.**
Category: Plan Realization	
Anchor sentence	Larger retailers may seek lower prices from us and our bottling partners, may demand increased marketing or promotional expenditures, and may be more likely to use their distribution networks to introduce and develop private label brands, any of which could negatively affect the Coca-Cola system's profitability.
Positive sentence	Larger retailers may exert pressure on us and our bottling partners by seeking lower prices, demanding more marketing or promotional expenditures, and using their distribution networks to introduce and develop private label brands, which could negatively impact the profitability of the Coca-Cola system.'
Negative sentence	Large retailers **have already demanded** lower prices from us and our bottling partners, and have pressured us to increase marketing and promotional expenditures, and have successfully introduced and developed their own private label brands, all of which have negatively impacted the Coca-Cola system's profitability.
Category: Emerging Situations	
Anchor sentence	While fixed-price contracts enable us to benefit from performance improvements, cost reductions and efficiencies, they also subject us to the risk of reduced margins or incurring losses if we are unable to achieve estimated costs and revenues.
Positive sentence	We can reap the benefits of fixed-price contracts, such as improved performance, lower costs, and greater efficiency, but we also face the risk of lower profit margins or incurring losses if we fail to meet estimated costs and revenues.'
Negative sentence	While fixed-price contracts enable us to benefit from performance improvements, cost reductions and efficiencies, they also subject us to the risk of significantly reduced margins or incurring substantial losses if we are unable to achieve estimated costs and revenues **due to unforeseen circumstances such as material defects, design flaws, or changes in project scope, which could lead to a significant decline in our profitability and financial stability.**

Table 6: Examples of augmented data by Llama-13b-chat.