Financial Risk Relation Identification through Dual-view Adaptation

Wei-Ning Chiu[1,2], Yu-Hsiang Wang[2], Andy Hsiao[2], Yu-Shiang Huang[1,2], Chuan-Ju Wang[2]

[1]National Taiwan University, [2]Academia Sinica

Correspondence: cjwang@citi.sinica.edu.tw

Abstract

A multitude of interconnected risk events—ranging from regulatory changes to geopolitical tensions—can trigger ripple effects across firms. Identifying inter-firm risk relations is thus crucial for applications like portfolio management and investment strategy. Traditionally, such assessments rely on expert judgment and manual analysis, which are, however, subjective, labor-intensive, and difficult to scale. To address this, we propose a systematic method for extracting inter-firm risk relations using Form 10-K filings—authoritative, standardized financial documents—as our data source. Leveraging recent advances in natural language processing, our approach captures implicit and abstract risk connections through unsupervised fine-tuning based on chronological and lexical patterns in the filings. This enables the development of a domain-specific financial encoder with a deeper contextual understanding and introduces a quantitative risk relation score for transparency, interpretable analysis. Extensive experiments demonstrate that our method outperforms strong baselines across multiple evaluation settings.

1 Introduction

Relation identification between entities is valuable across various domains—including healthcare, legal analytics, social networks—and finance is no exception. The financial market is a complex ecosystem shaped by a wide array of factors, including economic indicators, geopolitical events, corporate developments, regulatory changes, and investor sentiment.

Among various types of inter-entity links, risk relations are particularly important due to their implications for financial performance and decision-making. A risk relation exists when two companies are both exposed to the same risk factors—such as new regulations, ongoing lawsuits, supply chain disruptions, or broader economic downturns. In such cases, an adverse event affecting one firm can also influence the other, creating a shared vulnerability. For example, both Nvidia (NVDA) and Wabtec (WAB) faced disruptions from the semiconductor supply chain crisis in 2022.[1] Identifying such relations is vital for informed investment decisions. However, traditional expert-driven analysis is often time-consuming, subjective, and prone to cognitive biases such as overconfidence or herd behavior.

This creates a growing need for objective, scalable methods to extract risk-related connections from financial texts. Structured documents like Form 10-K filings offer a rich, standardized resource for such analysis.[2] Recent advances in natural language processing (NLP), particularly pretrained language models, offer powerful tools for learning semantic representations. Yet most existing relation extraction methods focus on explicit entity-relation tagging and struggle to capture the implicit or abstract connections—like shared risk exposures—prevalent in financial texts.

To overcome these limitations, we adopt a retrieval-based encoding framework—a foundation of modern NLP—that transforms text into dense, semantically rich vector representations. This enables efficient relation discovery, information retrieval, and downstream financial NLP applications (Wang et al., 2024; Alaparthi and Mishra, 2020). Specifically, we first propose an unsupervised fine-tuning strategy based on a dual-view similarity framework to adapt general-purpose encoders for the finance domain. Our approach leverages two key characteristics of Form 10-K filings: (1) standardized language and constrained vocabu-

[1]Nvidia: https://www.reuters.com/technology/graphic-chip-price-drop-raises-questions-whether-end-shortage-is-sight-2022-04-25/.
Wabtec: https://www.nasdaq.com/articles/whats-in-the-offing-for-wabtec-wab-this-earnings-season.

[2]Form 10-K filings are annual reports mandated by the U.S. Securities and Exchange Commission (SEC).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26302–26312
November 4-9, 2025

lary, which yield consistent lexical patterns; and (2) frequent association of risk events with date-time references, enabling temporal alignment. By modeling both lexical and chronological similarities, we construct high-quality positive training pairs that reflect semantically or temporally aligned content. This dual-view supervision guides the encoder to capture nuanced financial semantics and align similar risk disclosures across firms.

To complement this encoder, we secondly introduce a retrieval-based, interpretable scoring mechanism—*the risk relation score* (RRS)—to quantify inter-firm risk connections. RRS offers key advantages over traditional heuristics: it is symmetric, guarantees minimum similarity thresholds, and enhances interpretability by grounding each relation in explicitly retrieved mutual risk paragraphs (MRPs). This not only provides a robust measure of shared risk but also delivers textual evidence, improving the transparency and reliability of the model's output. Figure 1 gives an overview of our method for identifying inter-firm risk relations.[3]

Despite recent progress, evaluating inter-firm relations remains challenging due to the lack of standardized benchmarks and domain-specific evaluation protocols. To address this, we conduct comprehensive experiments to assess our method:

1. We show that RRS correlates strongly with the absolute values of daily stock return correlations, demonstrating real-world relevance.
2. We integrate the discovered risk relations into a graph-based model for stock price prediction, significantly improving its performance.
3. We assess the standalone retrieval capabilities of our encoder using `MultiHiertt` (Zhao et al., 2022), a financial QA benchmark built from regulatory filings, where our model consistently outperforms strong baselines.

Contributions

- **Risk Relation Scoring.** We introduce a novel metic—*risk relation score (RRS)*—to measure inter-firm risk relationships. Based on encoder-derived paragraph similarity, RRS is transparent, interpretable, and symmetric, with explicit textual evidence.
- **Domain-specific Encoder Fine-tuning.** We propose a dual-view unsupervised fine-tuning strategy using lexical and chronological similarity patterns in Form 10-K filings to adapt general NLP encoders to the financial domain.
- **Comprehensive Empirical Validation.** We demonstrate the utility of our approach through stock return correlation analysis, graph-based forecasting improvements, and strong retrieval performance on the `MultiHiertt` benchmark.

[3]A peer-reviewed system demonstration related to this work was independently developed and published as a demo paper at NAACL 2025 by Wang et al. (2025). Although both the demo paper and the present paper employ similar methodology, the former primarily emphasizes the visualization of outcomes, whereas the latter focuses on the methodological contributions and comprehensive empirical experiments across various aspects.

2 Related Works

In the field of natural language processing (NLP), pretrained encoders have been instrumental in transforming text into semantically rich representations. This section reviews advances in relation extraction, general-domain encocers, and financial-domain encoders, with an emphasis on their relevance to financial text analysis.

2.1 Relation Extraction

Early relation extraction (RE) methods relied on pattern matching and manual feature engineering, such as such as the lexico-syntactic patterns introduced by Hearst (1992). With the rise of deep learning, neural models have like CNN (Zeng et al., 2015) and PCNN (Zeng et al., 2015) improved RE through automated feature learning. The advent of pretrained language models, particularly BERT (Devlin et al., 2019), significantly advanced RE by enabling fine-tuning for contextual understanding. Approaches such as entity-aware fine-tuning (Soares et al., 2019) further boosted performance. Recent surveys (Diaz-Garcia and Lopez, 2024) highlighted the dominance of BERT-based techinques while recognizing the growing impact of large language models (LLMs) like T5 (Raffel et al., 2020).

2.2 General-domain Encoders

Transformer-based models, led by BERT (Devlin et al., 2019) and its variants such as RoBERTa (Liu et al., 2019) and SpanBERT (Joshi et al., 2020), transformed NLP by enabling transfer learning across tasks. In the context of retrieval, DPR (Karpukhin et al., 2020) introduced supervised dense retrieval with dual encoders, while Spider (Ram et al., 2021) used instruction-tuned data to enhance performance. Contriever (Izacard et al., 2021) offered an unsupervised contrastive learning alternative, effectively capturing semantic similarity. More recent models, such as Jina

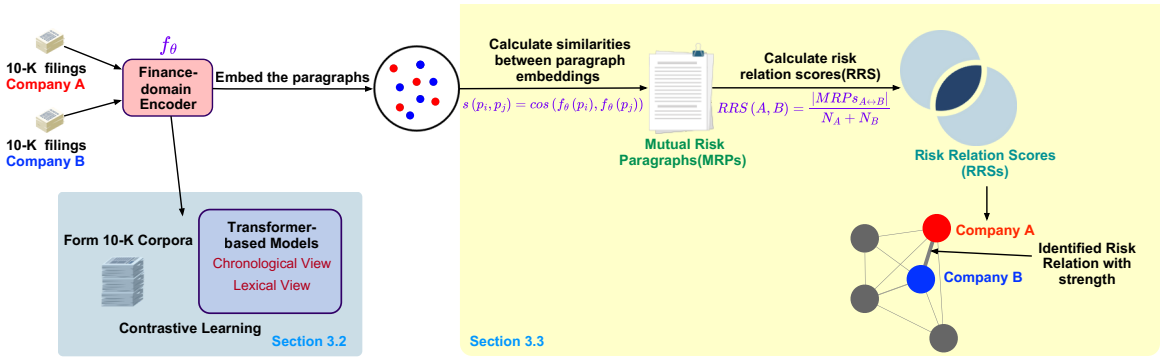

Figure 1: **Overview of risk identification pipeline.**

AI's encoder (Günther et al., 2023), incorporates novel methods like Attention with Linear Biases (ALiBi) (Press et al., 2021) to effectively process longer textual sequences. Embeddings in Xiao et al. (2024) further refine encoding strategies by considering context from multiple perspectives including functionality, granularity, and linguity.

2.3 Financial-domain Encoders

Domain-specific encoders have been introduced to better capture the nuances of financial language. FinBERT (Araci, 2019) adapted BERT by pretraining on large-scale financial corpora and has demonstrated superior performance in sentiment analysis. SEC-BERT (Loukas et al., 2022) further narrowed the focus by training exclusively on U.S. SEC filings, enhancing its applicability to regulatory documents. Another variant, FinBERT-MRC (Zhang and Zhang, 2023) reformulated financial named entity recognition as a machine reading comprehension task, improving contextual precision. Beyond BERT-based models, Fin-E5 (Tang and Yang, 2025) introduced a persona-driven synthetic data strategy to support a wider range of financial embedding tasks. In parallel, proprietary models such as BloombergGPT (Wu et al., 2023) showcased the potential of financial LLMs trained on exclusive datasets, though their closed nature spurred demand for open alternatives. In response, FinGPT (Yang et al., 2023) offers an open-source framework focused on accessible data and democratized financial LLMs.

3 Methodology

This section details our overall approach for identifying and quantifying inter-firm risk relations from financial documents. We begin by introducing the notations and terminology used throughout the paper in Section 3.1. We then present two core components of our methodology: (1) an unsupervised fine-tuning strategy that adapts a general encoder to better capture financial semantics (Section 3.2, and (2) a risk relation scoring mechanism that leverages the fine-tuned encoder to compute transparent and symmetric measures of inter-firm risk exposure (Section 3.3). Figure 1 illustrates the overall framework for identifying inter-firm risk relations.

3.1 Notation and Preliminaries

Key notations used in this paper are defined below:

- w: A token, i.e., a word or subword unit, within a paragraph.
- p: A paragraph, represented as a sequence of tokens: $p = [w_1, \ldots, w_n]$, where n is the number of tokens.
- D: A batch of paragraphs.
- f_θ: An encoder function parameterized by θ, which maps an input sequence to a dense vector representation.
- $f_\theta(p)$: The vector representation of paragraph p, obtained by mean pooling over the final-layer hidden states of the encoder.
- $s(p_i, p_j)$: The similarity score between paragraphs p_i and p_j, typically computed using cosine similarity between their embeddings.
- $\mathcal{P}_A$, $\mathcal{P}_B$: The sets of paragraphs associated with firms A and B, respectively.

3.2 Unsupervised Adaptation of a Financial Domain Encoder

We aim to train a financial-domain retriever using Form 10-K reports—authoritative and standardized corporate disclosures. The retriever is designed to retrieve semantically related paragraphs from a large corpus given a query paragraph. Unlike re-ranking methods that require pairwise comparisons, our retriever independently encodes each paragraph, enabling scalable and efficient retrieval.

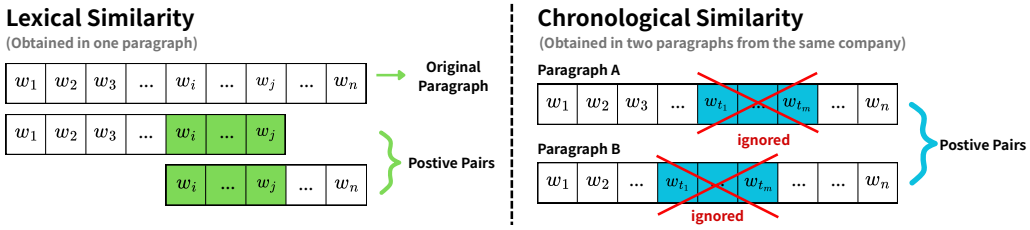

Figure 2: **Illustration of the dual-view training strategy.** Left (lexical view): Two overlapping spans within the same paragraph are sampled (green) and treated as a positive pair. Right (chronological view): Two paragraphs from the same company that share identical date-time tokens form a positive pair; date-time tokens (blue, crossed out) are excluded prior to encoding to prevent trivial matching.

Each paragraph p is encoded by a transformer-based encoder f_θ, and the final representation $f_\theta(p)$ is obtained by averaging the encoder's final-layer token embeddings. Given two paragraphs p_i and p_j, we compute their relevance score using the cosine similarity of their vector representations:

$$s(p_i, p_j) = \cos(f_\theta(p_i), f_\theta(p_j)). \quad (1)$$

3.2.1 Unsupervised Training of the Encoder

Contrastive Learning We fine-tune the encoder using contrastive learning to bring similar paragraphs closer in embedding space. Given an anchor text piece p, a positive counterpart p^+, and a set of negatives D^-, the InfoNCEloss (van den Oord et al., 2018) is defined as:

$$L(p, p^+, D^-) = -\frac{\exp^{(s(p,p^+)/\tau)}}{\exp^{(s(p,p^+)/\tau)} + \sum_{p^- \in D^-} \exp^{s(p,p^-)/\tau}},$$

where τ is a temperature hyperparameter. Minimizing this loss encourages higher scores to positives and lower scores to negatives.

Forming Positive Pairs We construct high-quality positive pairs using two complementary perspectives: chronological similarity and lexical similarity, both grounded in empirical observations from financial reports.

- *Chronological View.* Firms typically experience only one significant event per day (excluding standard accounting dates). We pair paragraphs from the same firm that share identical date-format tokens$[w_{t_1}, w_{t_2}, \ldots, w_{t_m}]$ (e.g., "July 8, 2024"). To avoid superficial matching, we remove all date tokens from both paragraphs during training and validation.
- *Lexical View.* Due to regulatory conventions, Form 10-K filings often reuse phrasing to describe similar events. We exploit this by creating overlapping spans within the sample paragraph. Given $[w_1, w_2, \ldots, w_n]$, we randomly select indices $i < j$ and form the pair $([w_1, \ldots, w_j], [w_i, \ldots, w_n])$.

An illustrative example for both view is provided in Figure 2.

Forming Negative Pairs Constructing diverse and informative negative pairs is critical for effective contrastive learning. We adopt the widely used *in-batch* negatives strategy, as implemented in retrieval models like Contriever. This approach treats all other positive samples within the same training batch as negatives, offering a scalable and memory-efficient solution without requiring additional sampling or computation.

Let $D = \{(p_i, p_{i^+})\}_{i=1}^{B}$ be a batch of B positive pairs. For each anchor text piece p_i, its positive is p_{i^+}, and the set of negatives D_i^- consists of all other positive samples in the batch except p_{i^+}:

$$D_i^- = \{p_j \in D \mid j \in \{1^+, 2^+, \ldots, B^+\},\ j \neq i^+\}.$$

This setup provides $B-1$ negative examples per anchor, enhancing the contrastive signal. While positive pairs are constructed based on the chronological and lexical similarity views described in Section 3.2.1, negative pairs are dynamically formed from the remaining batch entries. This strategy yields a robust and scalable training procedure, consistent with techniques from prior work (Chen et al., 2017, 2020).

3.3 Scoring Mechanism for Risk Relations

We utilize our trained encoder to generate paragraph embeddings, where a higher cosine similarity between embeddings indicates potential chronological or lexical alignment. Given a similarity threshold ξ (a tuned hyperparameter), we consider two paragraphs p_i and p_j to discuss similar risk content if $s(p_i, p_j) \geq \xi$ (see Eq. (1)).

Using this similarity criterion, we define the set of *mutual risk paragraphs* (MRPs) between firms A and B, denoted as $\text{MRPs}_{A \leftrightarrow B}$:

$$\text{MRPs}_A = \left\{ p_i \in \mathcal{P}_A \,\middle|\, \exists p_j \in \mathcal{P}_B : s(p_i, p_j) \geq \xi \right\},$$
$$\text{MRPs}_B = \left\{ p_j \in \mathcal{P}_B \,\middle|\, \exists p_i \in \mathcal{P}_A : s(p_i, p_j) \geq \xi \right\},$$
$$\text{MRPs}_{A \leftrightarrow B} = \text{MRPs}_A \cup \text{MRPs}_B,$$

where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the sets of all paragraphs associated with firms A and B, respectively. MRPs_A (resp. MRPs_B) consists of paragraphs from firm A (resp. B) that share high semantic similarity with at least one paragraph from firm B (resp. A). These MRPs serve as explicit, interpretable evidence of shared risk exposure.

We then define the *risk relation score* (RRS) between firms A and B as the proportion of mutual risk paragraphs relative to the total number of paragraphs from both firms:

$$\text{RRS}(A, B) = \frac{|\text{MRPs}_{A \leftrightarrow B}|}{N_A + N_B},$$

where N_A and N_B denote the total numbers of paragraphs from firm A and B, respectively. The RRS ranges from 0 (no shared risk) to 1 (complete overlap), quantifying the degree of risk-related connection between firms.

Advantages of RRS

- **Symmetry**: RRS is symmetric by construction, i.e., $\text{RRS}(A, B) = \text{RRS}(B, A)$, unlike many retrieval-based methods.
- **Minimum Similarity Guarantee**: By using a threshold-based approach rather than top-k selection, only sufficiently similar paragraph pairs contribute to the score, reducing noise.
- **Interpretability**: Each RRS is grounded by MRPs, allowing transparent inspection of the evidence behind identified risk relations.

4 Experiment

This section presents experimental setup and results. We begin by detailing our encoder training details, threshold calibration settings, and baseline models. Subsequently, we conduct extensive evaluations designed to answer the following research questions (RQs):

- **RQ1:** How well does the proposed risk relation identification align with real-world stock price co-movements?
- **RQ2:**What are the individual contributions of the chronological and lexical views to the overall performance?
- **RQ3:** Can the identified risk relations improve downstream tasks such as stock price movement prediction when used as features?
- **RQ4:** How does our dual-view fine-tuned encoder perform on financial information retrieval benchmarks compared to existing models?

4.1 Encoder Training Details

To prevent information leakage during evaluation, we restrict the training data to Form 10-K filings from 2018 to 2020. Filings from all firms and all reported sections of the 10-Ks are included to ensure broad data coverage and diversity. Each input text piece is truncated or padded to 256 tokens.

For model training, we construct 8,500 positive paragraph pairs and 1,000 validation pairs for each of the two views: chronological and lexical similarity. We fine-tune our encoder starting from the BERT-base-uncased pretrained model[4]. Training is performed using contrastive learning with a batch size of 64 and a learning rate of 2×10^{-5}, optimized via Adam optimizer (Kingma and Ba, 2014) with a linear warmup scheduler. L2 regularization is applied to improve generalization, and training proceeds for up to 50 epochs with early stopping to mitigate overfitting.[5]

4.2 RRS Calculation Details

To ensure reliable identification of shared risks, we apply a threshold-based filtering mechanism that excludes paragraph pairs with insufficient semantic similarity. The similarity threshold ξ is tuned incrementally in the range [0.5, 0.9] with a step size of 0.05. The optimal threshold for our encoder is empirically determined to be 0.75, balancing precision and coverage across downstream tasks. We provide additional hyperparameter sensitivity analysis in Appendix A. Moreover, to better focus on risk-related content, we restrict retrieval to paragraphs from Item 1A (Risk Factors) and Item 7A (Quantitative and Qualitative Disclosures about Market Risk) of Form 10-K filings, as these sections are most likely to contain discussions of companies' risk exposures.

[4] https://huggingface.co/google-bert/bert-base-uncased

[5] For computational resources, training was conducted on a single NVIDIA V100 GPU for approximately 8 hours.

4.3 Encoders for Comparison

Since our encoder is fine-tuned from the BERT-base-uncased pretrained model, we include several BERT variants for comparison.

- **BERT-base-uncased**: The original pretrained model, used as a baseline to measure the effect of our domain-specific fine-tuning.
- **Contriever** (Izacard et al., 2021): An unsupervised retrieval model trained with contrastive learning.
- **DPR** (Karpukhin et al., 2020): A supervised retrieval model trained on question-answer pairs.
- **FinBERT** (Araci, 2019) and **SEC-BERT** (Loukas et al., 2022): Two domain-specific models pretrained on financial corpora. We use the base version of SEC-BERT to ensure fairness in model size.
- **Llama-3.2** (Touvron et al., 2024): A widely recognized open-source large language model (LLM). We employ the 3B variant to provide a fair comparison in terms of model size with our encoder.

Among these, Llama-3.2 is primarily designed for general-purpose text generation, whereas Contriever and DPR are specifically designed for retrieval tasks. In contrast, FinBERT and SEC-BERT focus on domain adaptation without retrieval-specific objectives.

4.4 Risk Relation Identification Evaluation (RQ1)

This experiment assesses how well the risk relations identified by our method align with real-world stock price co-movements. The underlying intuition is that firms exposed to similar risks often experience correlated stock movements due to the simultaneous impact of common events.

4.4.1 Data Sources

We evaluate on firms consistently listed in the S&P 500 Index from 2018 to 2024, based on the 2024 constituent list. Dataset comprises stock price data from Yahoo Finance and Form 10-K reports obtained via the SEC API.[6] We retain only firms with complete filings and exclude those involved in mergers or lacking risk disclosures. After preprocessing to remove all HTML, XBRL tags, and tables, the final dataset covers 2,136 filings from 337 companies.

[6] https://www.sec.gov/search-filings/edgar-application-programming-interfaces

4.5 Compared Methods

We categorize our comparison methods into two groups: Human-based methods and Model-based methods. Below, we briefly describe each:

- **Human-based Baselines:** The Global Industry Classification Standard (GICS)[7] is a widely adopted taxonomy that assigns each company to exactly one of 11 sectors and 74 industries. As a human-labeled reference, GICS serves as a proxy for manually defined inter-firm relationships based on industry affiliation.
- **Model-based Methods**: In addition to our dual-view fine-tuned encoder, we apply the same risk relation scoring framework to all encoders described in Section 4.3 to ensure a fair and consistent comparison. All models follow the same inference procedure: paragraph embeddings are generated by averaging the final-layer token representations, followed by L2 normalization for cosine similarity computation.

4.5.1 Evaluation Metrics

We propose a metric ρ to measure alignment between risk relation scores (RRSs) and the correlation of the absolute values of daily stock returns (CAVDSR). Formally: $\rho = \mathrm{corr}(\mathrm{RRS}, \mathrm{CAVDSR})$. CAVDSR is computed using the full year of daily return data for each firm pair, and the corresponding RRS is calculated from the annual 10-K filings of that same pair. For GICS-based baselines, we assign binary RRS values: 1 if the two firms belong to the same sector or industry, and 0 otherwise. We use absolute returns to account for divergent effects from the same event (e.g., COVID-19's differing impact on the healthcare sector and the travel sector). A higher ρ indicates better alignment with real-world risk co-movement.

4.5.2 Performance Analysis

As shown in Table 1, human-based methods yield lower ρ, highlighting the limitations of their coarse granularity. Among model-based baselines, retrieval-focused models (DPR and Contriever) generally outperform general-purposed LLM models (Llama-3.2-3b) and other encoders (Bert-base-uncased, FinBERT and SEC-BERT). Specifically, domain-specific models are pretrained on financial text without retrieval-specific objectives, hence underperforming retrieval-focused models. Our

[7] https://www.msci.com/indexes/index-resources/gics

	Domain-specific	Retrieval Fine-tuning	Methods	2020	2021	2022	2023	2024	Avg.
Human-based			GICS Sector	0.1657	0.2881	0.2964	0.2971	0.2389	0.2572
			GICS Industry	0.1806	0.3336	0.2961	0.3316	0.3115	0.2907
Model-based			Bert-base-uncased	0.0804	0.2609	0.2939	0.1945	0.2471	0.2796
			Llama-3.2-3B	0.2031	0.3965	0.3233	0.4154	0.4336	0.3544
		✓	Contriever	0.2136	0.3964	0.3131	0.4112	0.4406	0.3556
		✓	DPR	0.2138	0.4008	0.3222	0.4158	0.4439	0.3583
	✓		FinBERT	0.1352	0.3013	0.2706	0.2732	0.3070	0.3058
	✓		SEC-BERT	0.1708	0.3545	0.3307	0.3460	0.3633	0.3569
	✓	✓	Ours	**0.2141**	**0.4079**	**0.3412**	**0.4233**	**0.4531**	**0.3711**

Table 1: **Correlation between RRS and CAVDSR.** Bold marks the best, underline the second-best.

	2020	2021	2022	2023	2024
Chronological	0.1659	0.3691	**0.3637**	0.3671	0.3908
Lexical	0.2137	0.4104	0.3336	0.4231	0.4510
Ours (both views)	**0.2161**	**0.4150**	0.3421	**0.4270**	**0.4553**

Table 2: **Ablation study on different views.** Bold indicates the best overall result, while underline denotes the second-best.

encoder significantly surpasses all baselines, confirming that leveraging chronological and lexical views enhances the identification of shared risk exposures.

4.6 Ablation Study on Different Views (RQ2)

To assess the individual contributions of our two training views, we train separate encoders using only the chronological or lexical similarity view. As shown in Table 2, the lexical view generally yields stronger performance, suggesting that consistent phrasing in financial reports is particularly effective for capturing risk relationships.

An interesting outcome occurs in 2022, where the chronological-view encoder performs on par with both its lexical-only and dual-view counterparts. This result can be attributed to the year's unique market conditions—marked by systemic events such as the Fed's rate hikes, the Russia–Ukraine war, and rising U.S.–China tensions—which triggered widespread, time-aligned impacts across firms. In such cases, temporal alignment is particularly useful for risk identification.

4.7 Applying Risk Relations to Stock Price Movement Prediction (RQ3)

We evaluate the practical utility of our risk relation metric by applying it to a downstream financial task: stock price movement prediction. Specifically, we integrate our risk relations into the attribute-driven graph attention networks (ADGAT) (Cheng and Li, 2021) to enhance stock price movement prediction. The original ADGAT framework notes that using human-defined relations (e.g., sector or industry links) often introduces biases and degrades performance, as such links are static, binary, and lack contextual nuance. To address this, we replace ADGAT's predefined relations with those derived from our method.

4.7.1 Experimental Setup

We closely follow the experimental setup and implementation details of ADGAT. Specifically, we use 280 days of training data, 70 for validation, and 70 for testing, covering the period from January 1, 2023, to September 4, 2024, with the final 70 days used for evaluation.

For training, we adopt ADGAT's original hyperparameters: Adam optimizer with a learning rate of 5×10^{-4}, a batch size of 15, a dropout rate of 0.2, and training up to 300 epochs, with early stopping.

Each configuration is run 15 times with different seeds, and the averages of the top 5 results are reported. To evaluate statistical significance, we apply a Mann–Whitney U test (Mann and Whitney, 1947) to compare performance distributions.

4.7.2 Performance Analysis

As shown in Table 3, replacing ADGAT's predefined relations with the ones derived from our model yields a 2.3% improvement in mean AUC. This gain is statistically significant ($p < 0.05$) and highlights the value of our method in real-world financial prediction tasks.

4.8 Retrieval Performance Evaluation (RQ4)

Beyond risk relation identification, we evaluate the retrieval effectiveness of our dual-view, financial-domain encoder using `MultiHiertt` benchmark, comparing it against several strong baselines.

Method	Mean AUC ± Std.
ADGAT (w/o our relation)	0.5807 ± 0.012
ADGAT (w/ our relation)	**0.5939 ± 0.006**
Improvement	2.27%

Table 3: **Performance comparison with/without our relations.** Bold marks the best.

4.8.1 Data Source

`MultiHiertt` is a financial question-answering (QA) benchmark designed to test multi-step numerical reasoning over hierarchical tables. Although originally proposed for QA, `MultiHiertt` also serves as a retrieval benchmark, where the task is to retrieve relevant paragraphs from a corpus of 10,475 documents given a query.

4.8.2 Experimental Setup

We extract 290 queries and their 1,331 corresponding relevant paragraphs. A two-stage retrieval pipeline is adopted. In the first stage, we retrieve the top 200 documents based on cosine similarities between the query and document embeddings. In the second stage, the candidates are re-ranked using the BAAI/bge-reranker-v2-m3 (Li et al., 2023; Chen et al., 2024). To evaluate our encoder, we replace the retriever component in this pipeline with ours and compare its performance against several baseline encoders.

4.8.3 Performance Analysis

Table 4 reports standard retrieval metrics such as normalized discounted cumulative gain (NDCG), precision, and recall at various top-k cutoffs. Our encoder significantly and consistently outperforms all the baselines. Notably, the two financial-domain encoders, FinBERT and SEC-BERT, perform poorly, likely due to the lack of retrieval-specific fine-tuning. Similarly, general-domain models (Llama-3.2-3b and BERT-base-uncased) struggle to adapt effectively to financial retrieval tasks. The key advantage of our encoder lies in its training strategy: the use of dual-view (chronological and lexical) supervision and domain-specific financial data. These design choices enable more effective modeling of semantic relevance in financial retrieval, as reflected in its superior performance.

5 Case Study

To demonstrate the practical utility of our method, we present a case study involving Enphase Energy (ENPH) and Meta Platforms (META)—two seemingly unrelated firms from the clean energy and technology sectors, respectively.

In 2023, our model ranks the risk relation between ENPH and META in the 95th percentile, uncovering a shared exposure to supply chain disruptions stemming from the COVID-19 pandemic. Although the connection is not immediately obvious, mutual risk paragraphs (MRPs) from their Form 10-K filings reveal a common theme and also provide interpretability for the risk relation between them:

- **Item1A, ENPH:** *...The global spread of COVID-19 and other actual or threatened epidemics, pandemics, outbreaks, or public health crises may adversely affect our results of operations and disrupt global supply chains...*
- **Item1A, META:** *...We rely on third parties to manufacture and manage the logistics of transporting and distributing our consumer hardware products, which subjects us to a number of risks that have been exacerbated as a result of the COVID-19 pandemic. We have experienced, and may in the future experience, supply or labor shortages or other disruptions in logistics and the supply chain...*

External news sources corroborate these findings. For instance, the *Financial Times* reported Meta's hardware delays due to supply chain shocks,[8] while *pv magazine USA* highlighted persistent pandemic-related disruptions affecting the solar industry.[9] This example highlights the encoder's ability to uncover subtle, non-obvious risk links with real-world relevance. A higher risk relation score (RRS) indicates that two companies are closely connected through shared risk exposures (e.g. supply chain disruptions), as demonstrated in the case above. These risk relations provide valuable interpretability, offering meaningful insights that can support more informed financial decision-making.

6 Conclusion

This paper proposes a novel framework for identifying inter-firm risk relations directly from unstructured financial text. By leveraging chronological and lexical similarities in Form 10-K filings, we develop an unsupervised fine-tuning strategy and introduce a transparent, symmetric risk relation core (RRS) to quantify shared exposures.

[8] https://www.ft.com/content/c7e9cfa9-3f68-47d3-92fc-7cf85bcb73b3

[9] https://pv-magazine-usa.com/2023/01/04/three-solar-industry-trends-to-watch-in-2023/

Model	NDCG@1	NDCG@5	NDCG@10	P@3	P@5	P@10	R@1	R@5	R@10
Bert-base-uncased	0.0377	0.0159	0.0140	0.0160	0.0103	0.0062	0.0052	0.0079	0.0095
Llama-3.2-3B	0.0171	0.0092	0.0089	0.0091	0.0062	0.0038	0.0037	0.0059	0.0070
Contriever	0.1712	0.0966	0.0930	0.0902	0.0616	0.0329	0.0394	0.0718	0.0764
DPR	0.1575	0.0819	0.0785	0.0696	0.0480	0.0260	0.0407	0.0581	0.0622
FinBERT	0.0034	0.0026	0.0033	0.0023	0.0021	0.0017	0.0009	0.0019	0.0039
SEC-BERT	0.0274	0.0143	0.0149	0.0137	0.0089	0.0055	0.0071	0.0097	0.0121
Ours	**0.2021**	**0.1114**	**0.1111**	**0.0993**	**0.0678**	**0.0377**	**0.0518**	**0.0859**	**0.0942**
Improvement	18.05%	15.32%	19.46%	10.09%	10.09%	14.59%	27.27%	19.63%	23.30%

Table 4: **Retrieval performance on `MultiHiertt`.** Bold marks the best, underline the second-best.

Extensive experiments validate the effectiveness of our method: (1) RRS shows strong correlation with real-world stock price co-movements, (2) its integration improves downstream stock prediction in a graph-based model (ADGAT), and (3) the encoder achieves superior performance on the `MultiHiertt` retrieval benchmark. A case study further demonstrates the method's ability to uncover subtle but meaningful risk connections.

7 Limitations

Although our method demonstrates strong performance in identifying nuanced inter-firm risk relationships, several limitations should be acknowledged. First, the framework is designed specifically to uncover relations based on shared risk exposures. As such, it may not generalize well to tasks involving other types of firm interactions, such as strategic partnerships, mergers and acquisitions, or product-market complementarities.

Second, our current approach relies solely on Form 10-K filings as the data source. Although these documents are structured and reliable, their annual frequency limits our method's responsiveness to short-term market changes or evolving risk profiles within a fiscal year. This restricts its applicability for real-time or high-frequency financial decision-making.

Lastly, while our evaluation leverages public financial and market data, it does not incorporate expert financial judgment. Practical decision-making often involves qualitative insights and domain expertise that automated models alone cannot fully capture. Future work could benefit from integrating expert input to enhance interpretability and real-world applicability.

References

Shivaji Alaparthi and Manit Mishra. 2020. Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey. *arXiv:2007.01127. Version 1.*

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv:1908.10063. Version 1.*

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv:2402.03216. Version 4.*

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Rui Cheng and Qing Li. 2021. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 55–62.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186.

Jose A. Diaz-Garcia and Julio Amador Diaz Lopez. 2024. A survey on cutting-edge relation extraction techniques based on language models. *arXiv:2411.18157. Version 1.*

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammed Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina Embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv:2310.19923. Version 4.*

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv:2112.09118. Version 4.*

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980. Version 9.*

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *arXiv:2312.15503. Version 1.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692. Version 1.*

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. *arXiv:2203.06482. Version 2.*

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv:2108.12409. Version 2.*

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.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2021. Learning to retrieve passages without supervision. *arXiv:2112.07708. Version 2.*

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Yixuan Tang and Yi Yang. 2025. Finmteb: Finance massive text embedding benchmark. *arXiv:2502.10900. Version 2.*

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Édouard Grave, and Guillaume Lample. 2024. The llama 3 herd of models. *arXiv:2407.21783. Version 3.*

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748. Version 2.*

Jiajia Wang, Jimmy X. Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.

Yu-Hsiang Wang, Wei-Ning Chiu, Yi-Tai Hsiao, Yu-Shiang Huang, Yi-Shyuan Chiang, Shuo-En Wu, and Chuan-Ju Wang. 2025. "SURF: A system to unveil explainable risk relations between firms". In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 260–267.

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:2303.17564. Version 3.*

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv:2306.06031. Version 1.*

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Yuzhe Zhang and Hong Zhang. 2023. FinBERT-MRC: Financial named entity recognition using BERT under the machine reading comprehension paradigm. *Neural Processing Letters*, 55(6):7393–7413.

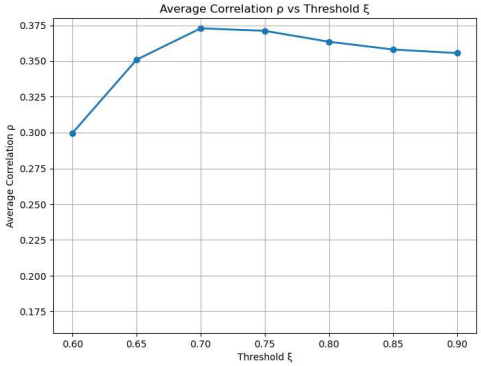

Figure 3: **Hyperparameter Sensitivity Analysis**

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi-hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6588–6600.

A Hyperparameter Sensitivity Analysis

We conduct a sensitivity analysis on the similarity threshold used by our encoder for identifying mutual risk paragraphs. Specifically, the threshold is tuned from 0.6 to 0.9, and the results of the first experiment described in Section 4.5.1 are reported in Figure 3. The results show that our encoder's performance remains relatively stable across different threshold settings, highlighting both its robustness and effectiveness in consistently identifying inter-firm risk relations.