applied sciences

MDPI

Article

Industry-Discriminative Sentence Reconstruction for Inter-Firm Similarity Analysis Using 10-K Reports

Sangmin Jo [1], Taehyeon Kim [1], Jeongho Park [2], Jihyeon Yu [1], Hyerim Bae [1] and Jihoon Chung [3,*]

[1] Department of Industrial Engineering, Pusan National University, Busan 46241, Republic of Korea; 27461a@pusan.ac.kr (S.J.); tae6709@pusan.ac.kr (T.K.); dorcus0043@pusan.ac.kr (J.Y.); hrbae@pusan.ac.kr (H.B.)
[2] Major in Industrial Data Science, Graduate School of Convergence Science and Technology, Pohang University of Science and Technology, Pohang 37673, Republic of Korea; wjdgh1354@postech.ac.kr
[3] Department of Industrial Engineering, Hanyang University, Seoul 04763, Republic of Korea
* Correspondence: chungjh@hanyang.ac.kr

Abstract

Peer firm identification provides a basis for financial decision-making. Text-based approaches increasingly utilize 10-K annual reports for peer selection because they describe firms' business activities and strategic positioning. However, most existing methods construct firm representations using full-document embeddings that uniformly utilize the entire textual content. This study proposes a sentence-level industry-discriminative reconstruction framework for peer selection using 10-K reports. The framework trains a sentence-level classification model to predict each firm's GICS Sector label and uses the predicted probabilities to prioritize sentences reflecting its industry identity and business characteristics. The classifier was trained using 2021 Item 1. Business sections from 4150 non-S&P 500 firms. Reconstruction ratios were validated using 2021 reports from 466 S&P 500 firms and their corresponding 2022 GICS, stock-return, and financial data. The determined ratios were then applied to the same firms' 2022 reports, with final performance evaluated using 2023 outcomes. Based on the selected sentences, 10-K business descriptions were reconstructed before generating document embeddings for peer selection. Experiments across five embedding models and three peer group settings demonstrate that the proposed framework generally achieves higher GICS alignment, stock-return correlation, and financial homogeneity than full-text, length-matched random-deletion, alternative sentence-selection baselines and benchmark methods.

Keywords: peer group selection; text-based analysis; 10-K annual reports

Check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 14 July 2026
Revised: 13 August 2026
Accepted: 20 August 2026
Published: 23 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Identifying similar firms is a widely used analytical task in finance that supports portfolio construction, relative valuation, and risk analysis [1]. Accurately identifying peer firms is essential not only for selecting comparable firms but also for understanding industry structure and supporting investment decisions. Therefore, firm similarity has been extensively studied in both academia and practice.

Due to its importance, industry classification systems such as the Global Industry Classification Standard (GICS) have been widely used to define firm similarity [2]. GICS organizes firms hierarchically based on their primary business activities and provides a structured view of industry organization. However, such classification systems are inherently categorical and have fundamental limitations [1]. In practice, firms often operate

across multiple business segments but are assigned to a single category, which limits the ability to represent complex business structures. In addition, categorical labels cannot capture within-industry heterogeneity or cross-industry similarity, and do not allow ranking firms by degree of similarity. As a result, firm similarity is reduced to coarse group membership, highlighting the need for more refined and continuous representations.

To address these shortcomings, researchers have proposed alternative similarity systems based on structured data, including fundamentals and accounting information [3], patent citations [4], co-search behavior [5], and analyst choice [6]. However, each approach has limitations. Accounting data are largely historical and often miss forward-looking aspects such as strategic shifts or market expansion. Patent data are irrelevant for firms with few or no patents, particularly in the service sector, and suffer from long approval lags. Co-search data are rarely accessible [5], while analyst-based measures introduce subjectivity and potential bias [6].

To construct more continuous and fine-grained firm representations for peer selection, recent studies have increasingly explored text-based approaches that directly analyze firms' business descriptions. In particular, many studies have attempted to construct continuous peer-selection systems by embedding diverse business information contained in text [7,8]. Among textual sources, U.S. Securities and Exchange Commission (SEC) 10-K annual reports have become widely used for firm similarity analysis [1,2,9]. These reports are publicly available and provide detailed descriptions of firms' business activities, products, and strategic positioning. However, most existing text-based approaches embed 10-K business descriptions in their original form without distinguishing informative sentences for peer selection. Since peer selection aims to identify firms with similar operations and business activities [10], sentences describing firms' core products, markets, and business activities are particularly informative for identifying similar firms [8]. Since such information also may reflect characteristics that differentiate firms across industries, these sentences are likely to contain industry-discriminative information. As a result, these important industry-discriminative signals may be diluted by generic or less industry-relevant content in 10-K reports in the existing studies [11]. Accordingly, this study is guided by the following main research question: Can selectively prioritizing industry-discriminative sentences in 10-K business descriptions produce firm representations that identify more meaningful peer firms than representations constructed from unfiltered business descriptions?

To address this limitation, this study proposes a sentence-level industry-discriminative framework for peer selection that constructs firm representations by selectively prioritizing informative sentences from 10-K business descriptions. Specifically, the framework trains a sentence-level classification model to predict each firm's GICS Sector label from individual sentences in its 10-K business description. Because GICS categorizes firms according to their primary business activities into sectors (i.e., industries) such as Financials, Health Care, and Energy [8], the predicted sector probabilities provide a useful signal for identifying industry-discriminative information. Therefore, sentences that strongly predict a firm's Sector can be interpreted as containing stronger evidence of industry identity and business similarity. Based on these probabilities, the framework assigns an industry-discriminative score to each sentence and selectively prioritizes sentences with stronger industry-related signals when constructing firm representations for peer selection. Lower scores do not imply that a sentence is meaningless, but rather less informative for identifying similarities relevant to peer selection. The resulting firm representations are subsequently transformed into firm embeddings, and peer firms are identified based on cosine similarity within the embedding space.

To the best of our knowledge, no prior text-based peer-selection study has explicitly investigated sentence-level selection of 10-K business descriptions for peer identification.

The effectiveness of peer selection based on the proposed framework is evaluated using three representative measures [5,8,12]. First, GICS alignment between target firms and their selected peers is measured [11]. Because the framework is trained using GICS Sector labels, finer-grained GICS categories, including Industry Group, Industry, and Sub-Industry, are additionally evaluated to examine whether the learned representations capture detailed industry identity. Second, stock-return correlations between target firms and their selected peers are examined as a market-based validation metric, motivated by the tendency of firms with similar business activities to exhibit correlated stock price movements [13]. Third, financial similarity is evaluated using adjusted R^2 values for financial variables associated with industry identity, following [14]. This evaluation builds on prior evidence suggesting that firms with similar business models tend to exhibit comparable financial reporting characteristics [15]. The main contributions of this study are as follows.

- This study proposes a sentence-level industry-discriminative framework for peer selection that uses GICS Sector prediction probabilities to prioritize sentences reflecting firms' industry identity and business characteristics when constructing firm representations.
- This study demonstrates that the proposed framework achieves higher peer-selection performance in terms of GICS alignment, stock-return correlation, and financial similarity using actual 10-K reports compared to benchmark methods.

2. Review of Related Work

The related existing studies of inter-industry similarity analysis are provided in this section. Specifically, Section 2.1 introduces research on industry classification, which groups and clusters firms based on similarities in industrial structure or activities, while Section 2.2 reviews studies on peer group selection, which identify sets of firms most comparable to a target firm. Afterward, the research gaps in the current work are identified in Section 2.3.

2.1. Industry Classification Research

For the industry classification research indicating the general process of clustering and categorizing firms into industries based on their similarities [16], traditional industry classification systems such as Standard Industry Classification (SIC), North American Industry Classification System (NAICS), and GICS are widely used [5]. However, the systems rely on manual code assignments by experts, making it difficult to reflect rapidly changing industry relations and structures in real time [9]. Furthermore, these systems only provide the categorical classification of an industry and do not provide any information about the similarity or distance between industries [1]. To overcome these limitations, various efforts have been made to develop new industry classification frameworks. For example, Mantegna [17] defined a distance matrix using correlation coefficients for daily logarithmic stock price differences and constructed the Minimal Spanning Tree and the associated subdominant ultrametric hierarchical tree, which provides an economically meaningful hierarchical organization of the market. Dalziel [18] proposed a systems-based approach to industry classification that defines sectors based on the similarity of needs to which firms collectively respond and disaggregates them into subsectors and industries through recursive hierarchical dependency. Gay and Karger [4] constructed a system based on patent citation relationships between firms as a measure of similarity, and applied the Walktrap clustering algorithm to build a dynamic industry classification system. Kakushadze and Yu [19] clustered variance-normalized returns using the K-means algorithm and improved the stability and performance of their statistical industry classification system by aggregating classifications from multiple samplings. In addition, Fodor et al. [20] applied K-means

clustering to classify firms using common-size financial statement data from balance sheets and income statements.

Recently, relevant studies have been increasingly turning to utilize text-based methods that analyze the actual business descriptions of firms, aiming to uncover industry structures through the semantic content embedded in the firms themselves. Table 1 classifies the relevant prior studies according to five dimensions, including the primary output, the text source used, the text-representation and similarity method, whether industry information is used as supervision, and whether task-relevant sentences are explicitly selected before constructing the final text representation. For example, Yang et al. [9] proposed the method called Business Text Industry Classification, which generated firm-level embeddings using Doc2Vec on 10-K reports and applied Ward's hierarchical clustering. Gao et al. [16] explored various embedding and clustering techniques and constructed a text-based industry classification from U.S. and Chinese firm descriptions, showing that Latent Semantic Indexing with K-means yields results comparable to GICS. In addition, Kim et al. [2] embedded 10-K reports using a bag-of-words, applied a deep autoencoder for dimensionality reduction, and employed spherical K-means clustering to construct a new industry classification method. Vamvourellis et al. [8] utilized business descriptions from SEC 10-K reports as input to a large language model, which was pretrained and fine-tuned to predict GICS industry labels. The resulting firm embeddings were then reduced in dimensionality using Uniform Manifold Approximation & Projection (UMAP) and clustered via spectral clustering to construct a data-driven industry classification system. However, these text-based industry-classification studies primarily focus on constructing firm- or industry-level representations for categorical classification and do not explicitly select relevant original sentences before constructing the final representation.

Table 1. Comparison of text-based industry classification and peer-selection studies.

Study	Output	Text Source	Method	Industry Supervision	Relevant Sentence Selection
Yang et al. [9]	Industry Classification	10-K	Doc2Vec + Ward hierarchical clustering	X	X
Gao et al. [16]	Industry Classification	Company descriptions (Yahoo/CSRC)	Multiple text representations + multiple clustering	X	X
Kim et al. [2]	Industry Classification	10-K	Binary BoW + autoencoder + spherical k-means	X	X
Hoberg and Phillips [1]	Both	10-K	Binary BoW + cosine similarity	X	X
Kee [12]	Peer Selection	Financial news (Reuters)	Word2Vec + cosine similarity	X	X
Ito et al. [11]	Peer Selection	10-K (US)/Japanese annual reports	Multi-task BERT + cosine similarity	O	X
Bergmann et al. [7]	Peer Selection	Corporate websites	Multiple text representations + cosine similarity	X	X
Vamvourellis et al. [8]	Both	10-K	Multiple text representations + cosine similarity/UMAP + spectral clustering	O	X
Proposed	Peer Selection	10-K	Multiple text representations + cosine similarity	O	O

2.2. Peer Group Selection Research

While classification studies have focused on establishing criteria for clustering industries, peer group selection research centers on the problem of quantitatively identifying the most similar set of competitor firms for a given individual firm. For example, Cheng and McNamara [21] used Industrial Compustat data to compare the price-earnings, the price-

book, and combined price-earnings and price-book valuation methods. They identified which peer-selection criteria, including industry membership, firm size, return on equity, and their combinations, yielded the highest accuracy for each valuation method. Bhojraj and Lee [3] conducted peer selection using warranted multiples, which are derived through regression analysis and large-sample estimation based on firms' related variables such as profitability and growth. Lee et al. [5] applied a co-search algorithm to internet traffic from the SEC's Electronic Data Gathering, Analysis and Retrieval (EDGAR) website. The study selected Search-Based Peers (SBPs) based on annual search fractions. Compared to GICS peers, SBPs identified through frequent co-searches by multiple users explain a substantially greater proportion of cross-sectional variations in firms' out-of-sample monthly returns, valuation multiples, financial ratios, and other fundamentals. Noels et al. [22] proposed a graph-based distance metric for financial statements using Earth Mover's distance. This method demonstrated strong performance in terms of both accuracy and computational efficiency when applied to practical tasks such as peer identification.

Recently, there has been growing interest in combining large-scale unstructured text with language models to generate continuous, real-time similarity scores as shown in Table 1. For example, Hoberg and Phillips [1] embedded industry descriptions from 10-K reports using a bag-of-words approach and measured industry-level similarity via cosine similarity. Kee [12] trained Word2Vec on a 10-year corpus of 21 million Reuters financial news articles and used cosine similarity to produce peer firms. The study showed that the Word2Vec-based peers provided greater financial homogeneity than existing classification systems and enabled tracking of peer group changes over time. Ito et al. [11] fine-tuned BERT on annual reports from publicly listed firms in the U.S. and Japan using multi-task learning and generated firm embeddings. The study illustrated that peer groups could be effectively selected using cosine similarity between these embeddings. Bergmann et al. [7] collected 'About ' pages from firm websites, which hold the key information about a firm, and used BERT model to obtain textual representations. Then, they calculated cosine similarity between these embeddings to identify technologically similar peer firms. Finally, Vamvourellis et al. [8] compared pretrained and fine-tuned language models to predict GICS industry labels, in order to examine whether fine-tuning leads to more effective embeddings for peer-selection tasks. Although Vamvourellis et al. [8] incorporate industry supervision into representation learning, they use the available textual content as input when constructing firm-level representations rather than explicitly selecting which original sentences are provided to the representation model. In contrast, explicitly selecting relevant sentences and reconstructing the text before document embedding offers several practical advantages. The reconstructed text can be used consistently across different pretrained embedding models without modifying their architectures. It also provides explicit textual evidence of which original information is used to construct firm-level representations, while removing less relevant sentences reduces the amount of text processed by downstream embedding models and may thereby reduce the associated computational burden.

2.3. Research Gap

Despite the progress in both industry classification and peer group selection research, several important limitations still remain. Industry classification methods discussed in Section 2.1 primarily focus on grouping firms into discrete industrial categories, but they generally do not provide continuous similarity representations between individual firms, limiting their ability to capture fine-grained business relationships and relative similarity structures among firms. In contrast, peer group selection methods reviewed in Section 2.2 aim to directly quantify inter-firm similarity using continuous representations derived from structured or textual information. Recently, text-based approaches using 10-K reports and

language models have demonstrated promising performance by embedding firms' business descriptions into semantic vector spaces. However, most existing methods construct firm representations using full-document embeddings that uniformly utilize the entire textual content of 10-K reports. As discussed in Section 1, sentences describing firms' core products, markets, and business activities are particularly informative for identifying firms with similar industry identity and business characteristics, while generic administrative or less industry-relevant content may weaken these important industry-discriminative signals. Therefore, there is a need for an approach that explicitly identifies and selects industry-relevant sentences from the source text before constructing firm-level representations, so that informative business content can be retained while less relevant information is excluded. Although prior studies have used supervised signals to automatically identify and select sentences relevant to their respective downstream tasks [23,24], research that uses such relevance signals to explicitly reconstruct source documents for inter-firm peer selection remains limited. To address this research gap, the present study selects and reconstructs the textual input based on sentence-level industry-discriminative relevance and then provides the reconstructed text to downstream embedding models for firm representation and peer identification.

3. Proposed Methodology

This section presents a novel framework for quantifying inter-firm similarity using 10-K annual reports. The framework supports peer group selection, validated against GICS alignment [11], stock-return correlations [8], and financial similarity measured using adjusted R^2 across key accounting variables [5].

Figure 1 illustrates the overall architecture. Each firm's 10-K report is first collected as input data and undergoes basic preprocessing (Section 3.1). A sentence-level model is then applied to estimate the industry alignment of each sentence using GICS labels. Based on this, the reports are reconstructed by selecting sentences with higher industry-discriminative scores (Section 3.2). The reconstructed reports are embedded using a language model to obtain vector representations for peer group selection (Section 3.3).

3.1. Data Preprocessing

The 10-K annual reports of U.S. companies, collected from the SEC's EDGAR system, serve as the primary dataset for this framework [25]. As shown in Figure 2a,b, the 10-K report is a standardized document containing descriptions of a firm's business overview, risk factors, and financial analysis. Because it systematically records a firm's characteristics and strategic information, it is highly suitable for peer group selection [1]. Among the multiple Items of the 10-K report illustrated in Table A1 in Appendix A, we focus on 'Item 1. Business'. This Item has been widely used across various financial tasks and applications [1,2,26,27].

As mentioned in Section 1, this Item provides essential content on industry characteristics and corporate strategies, making it particularly relevant for industry-level semantic analysis. Item 1 of each firm is segmented into sentences using the sentence segmentation functionality of spaCy [28]. In addition to textual data from 10-K reports, GICS industry classification labels are obtained from the Compustat database through WRDS [29], while firm-level financial statement variables are collected via the Financial Modeling Prep API [30].

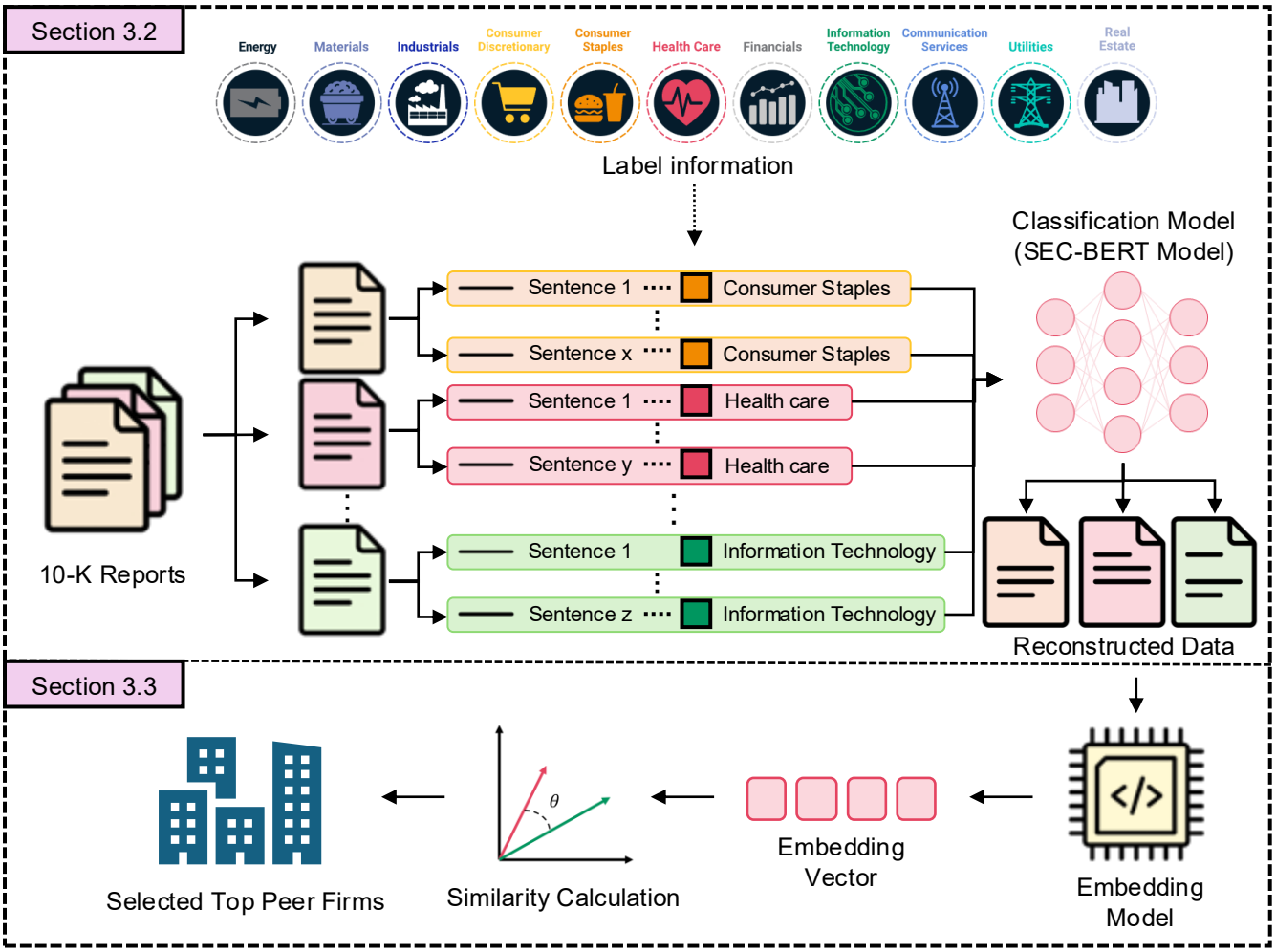
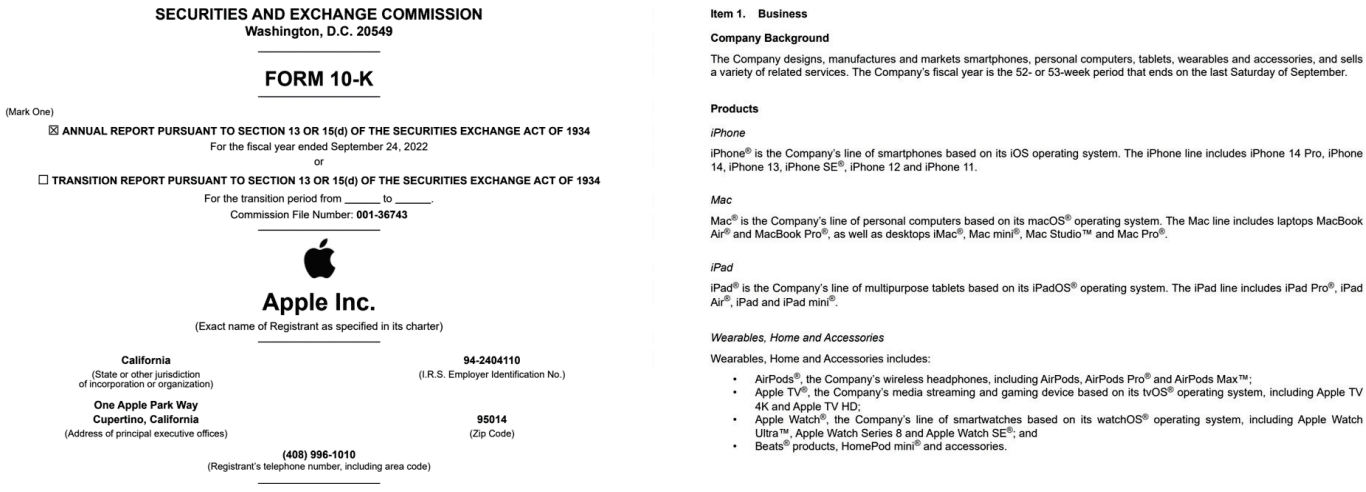


Figure 1. Overall framework of the proposed method.



SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended September 24, 2022
or
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from _____ to _____.
Commission File Number: 001-36743

Apple Inc.
(Exact name of Registrant as specified in its charter)

California	94-2404110
(State or other jurisdiction of incorporation or organization)	(I.R.S. Employer Identification No.)
One Apple Park Way Cupertino, California	95014
(Address of principal executive offices)	(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

(**a**)

Item 1. Business

Company Background

The Company designs, manufactures and markets smartphones, personal computers, tablets, wearables and accessories, and sells a variety of related services. The Company's fiscal year is the 52- or 53-week period that ends on the last Saturday of September.

Products

iPhone

iPhone® is the Company's line of smartphones based on its iOS operating system. The iPhone line includes iPhone 14 Pro, iPhone 14, iPhone 13, iPhone SE®, iPhone 12 and iPhone 11.

Mac

Mac® is the Company's line of personal computers based on its macOS® operating system. The Mac line includes laptops MacBook Air® and MacBook Pro®, as well as desktops iMac®, Mac mini®, Mac Studio™ and Mac Pro®.

iPad

iPad® is the Company's line of multipurpose tablets based on its iPadOS® operating system. The iPad line includes iPad Pro®, iPad Air®, iPad and iPad mini®.

Wearables, Home and Accessories

Wearables, Home and Accessories includes:

- AirPods®, the Company's wireless headphones, including AirPods, AirPods Pro® and AirPods Max™;
- Apple TV®, the Company's media streaming and gaming device based on its tvOS® operating system, including Apple TV 4K and Apple TV HD;
- Apple Watch®, the Company's line of smartwatches based on its watchOS® operating system, including Apple Watch Ultra™, Apple Watch Series 8 and Apple Watch SE®; and
- Beats® products, HomePod mini® and accessories.

(**b**)

Figure 2. Example structure of the Form 10-K report. (**a**) Cover page of the Form 10-K report; (**b**) Item 1. Business section of the Form 10-K report.

3.2. Text Reconstruction

Following the preprocessing stage, we apply the proposed framework to reconstruct the input text for similarity analysis. As illustrated in Figure 2, 10-K reports contain a wide

range of information, including business descriptions as well as additional details such as executive listings, headquarters addresses, and organizational structures. Since sentences differ in their ability to distinguish between firms, treating all content equally may hinder effective text-based similarity analysis.

To address this issue, the framework reconstructs each report by selecting sentences with higher industry-discriminative power, resulting in more focused and task-relevant documents (Figure 3). Specifically, GICS label information is used as a supervisory signal to estimate the industry alignment of each sentence using a pretrained SEC-BERT model. The reconstructed report is then formed by prioritizing sentences that are more useful for distinguishing between firms. In addition, the reconstructed Item 1 preserves the selected original sentences, allowing the textual evidence used for firm representation to be directly inspected. The detailed reconstruction procedure consists of the following steps.

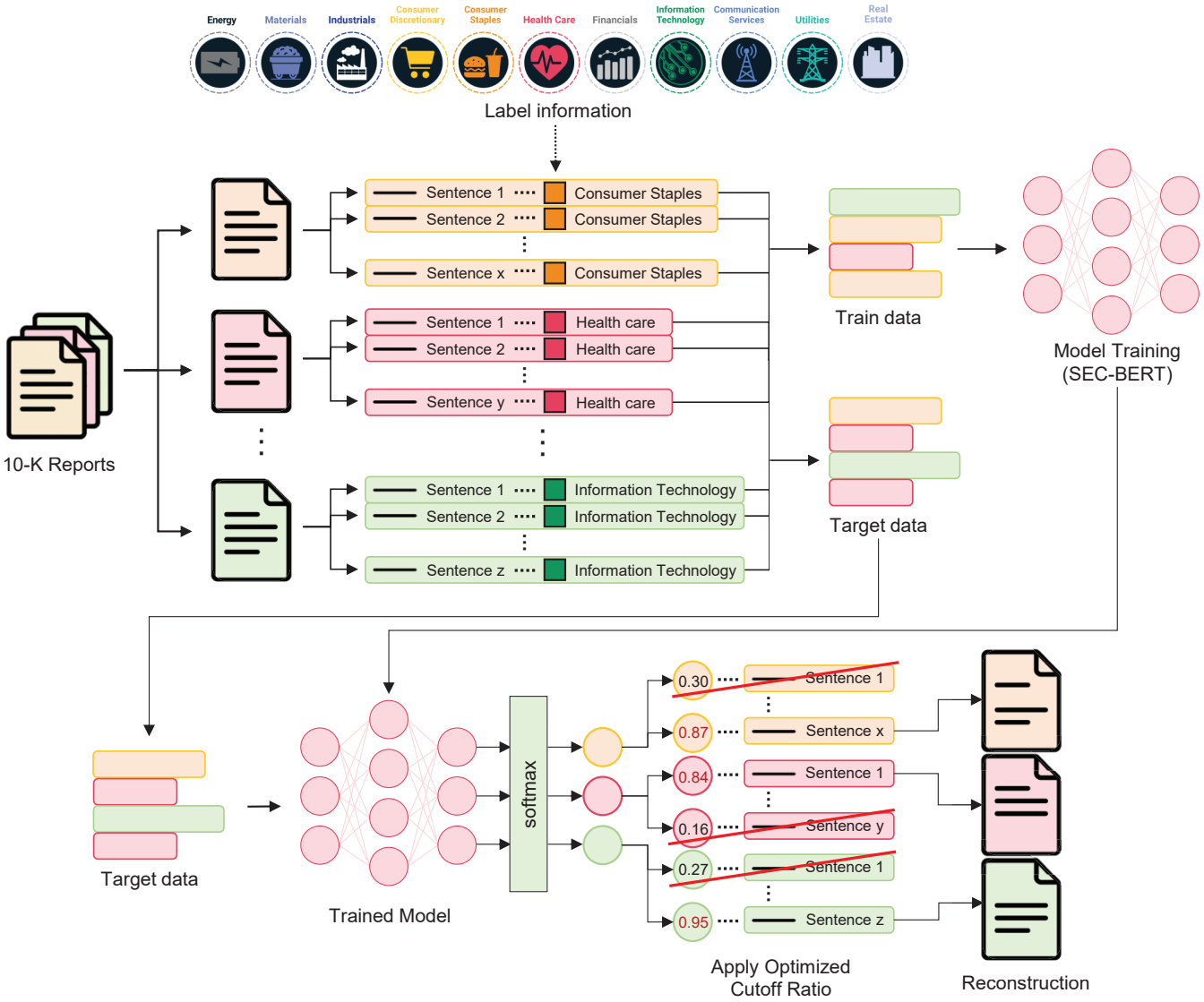


Figure 3. Procedures for the reconstruction of 10-K reports.

(1) Segmentation: First, each firm's annual report is segmented into individual sentences. Each sentence is assigned a label corresponding to the GICS sector of the firm it originated from. Since GICS has 11 sectors, as shown in Figure 3, each sentence is labeled within 11 sectors.

(2) Classification: Second, using this labeled and segmented dataset, a multi-class text classification model is trained to predict the GICS label of each sentence. Specifically,

a classifier is designed by attaching a multi-layer perceptron (MLP) to a pretrained SEC-BERT model [31]. SEC-BERT is adopted because it is pretrained on approximately 260,000 10-K filings from the SEC EDGAR database (1993–2019). Since the proposed framework is also trained on 10-K reports, the pretrained model is well aligned with the domain-specific vocabulary and linguistic patterns of financial disclosures encountered during fine-tuning. The labeled sentences were split into training and internal development subsets at an 8:2 ratio, with the internal development subset used exclusively for early stopping.

Let N denote the number of training samples (i.e., sentences) and G the number of GICS labels. For the j-th sample that is j-th sentence (i.e., $sent^{(j)}$), $y^{(j)}$ represents its true label index, and $y_i^{(j)}$ is the one-hot encoded label that equals one if the sample belongs to label i and zero otherwise. The term $\hat{p}_i^{(j)}$ denotes the predicted probability that the j-th sample belongs to label i. The model is trained by minimizing the cross-entropy loss, as defined in Equation (1).

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{G} y_i^{(j)} \log \hat{p}_i^{(j)}. \tag{1}$$

(3) Probability extraction: Third, the annual report for the evaluation year is passed through the trained model. For the sentence j in the annual report, let $z_i^{(j)}$ denote the logit score assigned to label i. The predicted probability vector is denoted as $\hat{p}^{(j)} = (\hat{p}_1^{(j)}, \hat{p}_2^{(j)}, \ldots, \hat{p}_G^{(j)})$, where each element $\hat{p}_i^{(j)}$ is obtained by applying the softmax function, as defined in Equation (2).

$$\hat{p}_i^{(j)} = \frac{\exp(z_i^{(j)})}{\sum_{k=1}^{G} \exp(z_k^{(j)})}. \tag{2}$$

Since the true label index of sentence j is denoted by $y^{(j)}$, the component corresponding to $y^{(j)}$ from the predicted probability vector is extracted and used as the industry-discriminative score $s^{(j)}$, as defined in Equation (3).

$$s^{(j)} = \hat{p}_{y^{(j)}}^{(j)}. \tag{3}$$

(4) Report reconstruction: Finally, based on the industry-discriminative score of all sentences, only those with the highest discriminative scores are utilized to generate the reconstructed reports. This procedure aims to improve the performance of similarity analysis by reconstructing the annual reports with high discriminative scores. Importantly, this score is not the maximum softmax probability across GICS Sectors, but the probability assigned to the source firm's actual GICS Sector, so sentence ranking reflects sector-specific relevance rather than general prediction confidence.

Specifically, for firm p, let X_p denote the set of its sentences, where j-th sentence $sent^{(j)}$ has an industry-discriminative score $s^{(j)}$ defined in Equation (3). Given a cutoff ratio $r \in (0, 1]$, a score threshold $\tau_p(r)$ is set to the $(1-r)$-quantile of the score set of firm p that $\{\, s^{(j)} : sent^{(j)} \in X_p \,\}$. Then, the reconstructed report $\tilde{X}_p(r)$ retains the sentences whose score is at least this threshold, as defined in Equation (4).

$$\tilde{X}_p(r) = \left\{\, sent^{(j)} \in X_p \;:\; s^{(j)} \geq \tau_p(r) \,\right\}, \qquad \tau_p(r) = Q_p(1-r), \tag{4}$$

where $Q_p(\cdot)$ denotes the quantile function of the score set, so that $\tau_p(r)$ is the score exceeded by the top fraction r of the sentences. A smaller ratio r raises the threshold, yielding a more aggressively reconstructed report that keeps fewer but more industry-discriminative sentences.

Therefore, the cutoff ratio determines the proportion of top-ranked sentences retained for report reconstruction and is determined as follows. The sentence classifier is first trained on the training set, which excludes all target firms used in the subsequent validation and test stages. Reconstruction-ratio selection is then performed on the validation set. For each candidate ratio ranging from 0.1 to 0.9, peer firms are identified using the reconstructed reports and evaluated against the corresponding evaluation measures. The ratio achieving the highest performance for each predefined evaluation setting is selected using only the validation set and subsequently fixed for evaluation on the temporally separated test set.

Specifically, candidate reconstruction ratios $\mathcal{R}$ ranges from 0.1 to 0.9 in increments of 0.1, covering a broad range from highly selective to near-full-text retention. Ratios 0 and 1 are omitted because they retain no text and the full text, respectively. For embedding model m, evaluation measure q, and peer size K, the ratio is selected as

$$r^*_{m,q,K} = \arg\max_{r \in \mathcal{R}} \ \mathcal{V}_{m,q,K}(r), \tag{5}$$

where $\mathcal{V}_{m,q,K}(r)$ is validation performance and q denotes the evaluation measure. One ratio is selected for each (m, q, K) setting using all target firms in the validation set. Then, the corresponding ratio is used to evaluate the performance of the testing data.

Figure 4 illustrates the reconstruction process of an actual 'item 1. Business' of the 10-K annual report from the firm 'PulteGroup, Inc.'. The original report undergoes a preprocessing stage, and then the proposed reconstruction method is applied with a determined rate of 0.5 (i.e., the top 50% of sentences by relevance score are retained). Sentences highlighted in red indicate those removed by the filtering process, while the remaining black text constitutes the final reconstructed report.

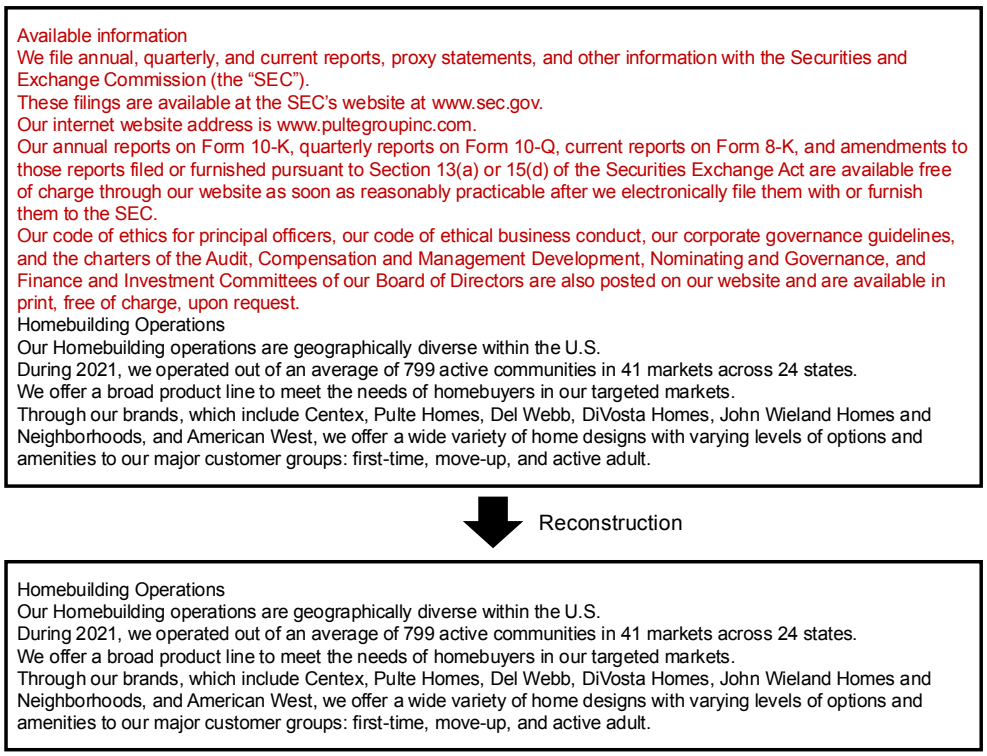


Figure 4. Illustration of the reconstructed 10-K report of 'PulteGroup' (Atlanta, GA, USA).

The result illustrates a clear distinction between the retained and removed content. The retained sentences primarily describe the firm's homebuilding operations, geographic markets, product offerings, brands, and target customer groups, which reflect its industry

identity and business characteristics. In contrast, the removed sentences mainly contain administrative and disclosure-related information, such as SEC filing availability, website information, and corporate governance documents. This example illustrates that the proposed reconstruction method prioritizes sentences containing more industry-discriminative information while excluding content that is relatively less relevant to industry identity and business characteristics.

3.3. Peer Group Selection and Analysis

After reconstructing the reports, document embeddings are generated to represent each report as a vector for further analysis. To this end, a chunking-based embedding approach with LLM is applied [32] (Figure 5). Specifically, each firm's reconstructed text is segmented into chunks according to the model's maximum token length, and each chunk is embedded with LLM to produce chunk-level vectors. These vectors are then aggregated through a weighted average, where the weight corresponds to the token size of each chunk, to form a single document-level embedding for each firm.

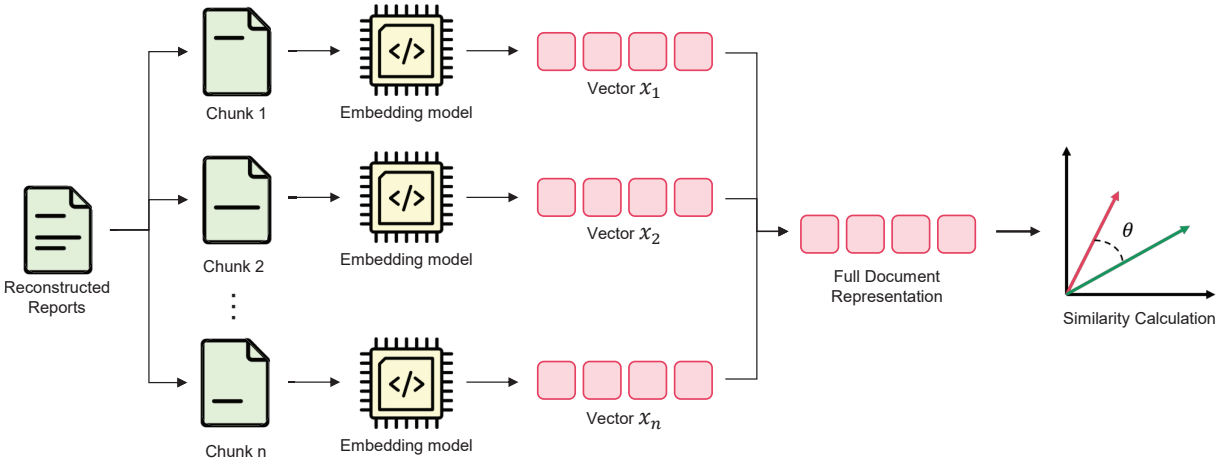


Figure 5. Procedures of peer group selection.

Because the reconstructed 10-K reports may exceed the maximum input length of the embedding models, each document is divided into consecutive, non-overlapping token-based chunks. Each chunk contains at most the model-specific maximum number of tokens and is independently encoded using the same embedding model. Because chunks may differ in length, token-count weighting is used to avoid disproportionately weighting shorter chunks when averaging chunk embeddings [33].

Formally, let the reconstructed document of firm i be divided into C_i chunks, and let $v_{i,c}$ and $n_{i,c}$ denote the embedding vector and the token count of the c-th chunk, respectively. Before aggregation, each chunk embedding is L2-normalized [34]. The document-level embedding u_i of firm i is obtained as the token-weighted average of its chunk embeddings, as defined in Equation (6).

$$u_i = \frac{\sum_{c=1}^{C_i} n_{i,c} \left(\frac{v_{i,c}}{\|v_{i,c}\|} \right)}{\sum_{c=1}^{C_i} n_{i,c}}. \tag{6}$$

where larger chunks contribute proportionally more to the aggregated representation. The resulting embedding u_i is used to measure inter-firm similarity in Equation (7).

For the peer group selection task, a pairwise similarity matrix among firms is calculated based on the similarity of vectors. To construct a pairwise similarity matrix, cosine similarity is used, defined as follows:

$$\text{cos_sim}(u_i, u_j) = \frac{u_i \cdot u_j}{\|u_i\| \, \|u_j\|}, \tag{7}$$

where u_i and u_j denote the document-level embedding vectors of firms i and j, respectively, obtained through the token-weighted chunk aggregation described above. Equation (7) ($\text{cos_sim}(u_i, u_j)$) forms the (i, j)-th entry of the pairwise similarity matrix, which is utilized for peer group selection. Algorithm 1 summarizes the proposed framework.

Algorithm 1: SEC-BERT-based text reconstruction framework for peer selection.

Input: X_{train}, X_{val}, $X_{\text{test}} = \{X_i\}_{i=1}^{n}$: training, validation, testing 10-K reports; candidate ratios $\mathcal{R}$; chunk length T; embedding model M.
Output: Similarity matrix S
Data preprocessing (Section 3.1)
Construct labeled dataset $\mathcal{D} = \{(sent^{(j)}, y^{(j)}) \mid \forall j \in X_{\text{train}}\}$.
Text refinement & reconstruction (Section 3.2)
Train SEC-BERT classifier on $\mathcal{D}$ via Equation (1).
Cutoff ratio selection
foreach $r \in \mathcal{R}$ **do**
  Reconstruct X_{val} by retaining top-r sentences ranked by relevance.
  Embed documents, select peers via cosine similarity, and record validation scores.
end
Select optimal cutoff ratio r^*.
for $i = 1$ **to** n **do**
  Segment X_i into sentences $\{sent^{(j)} \mid \forall j \in X_i\}$ and compute softmax probabilities $\{\hat{p}^{(j)} \mid \forall j \in X_i\}$ via Equation (2).
  Compute relevance scores $\{s^{(j)} = \hat{p}^{(j)}_{y^{(j)}} \mid \forall j \in X_i\}$ via Equation (3) and reconstruct $\tilde{X}_i$ by retaining top-r^* sentences.
end
Peer group selection (Section 3.3)
for $i = 1$ **to** n **do**
  Split $\tilde{X}_i$ into chunks of length $\leq T$ and embed each with M.
  Obtain u_i by token-weighted averaging of chunk embeddings.
end
Form $U = [u_1, \ldots, u_n]^\top$ and compute S via Equation (7).

4. Real-World Case Studies

This section presents real-world case studies on peer group selection to demonstrate the effectiveness of the proposed framework. Specifically, Section 4.1 describes the experimental setup, including data descriptions, hyperparameters, and evaluation metrics. Section 4.2 shows the performance of the sentence-level classifier. Section 4.3 reports validation studies that show that the effectiveness of the proposed framework is not affected by the embedding model, demonstrating its generalizability in text-analysis applications. Finally, Section 4.4 provides comparative analyses against state-of-the-art benchmark methods, supplemented by statistical hypothesis testing to confirm the significance of the results.

4.1. Experiment Setup

The classifier in this study is built upon a pretrained SEC-BERT model for sentence classification in 10-K reports. During training, both the pretrained SEC-BERT parameters and the classifier layers are updated. The model is trained on sentences from 4150 companies that submitted 10-K reports in 2021. Firms with missing Item 1. Business sections or fewer than 20 extracted sentences are excluded. S&P 500 firms, which constitute a widely used large-cap benchmark of major U.S. firms [35], are excluded from classifier training to prevent data leakage and are subsequently used for validation and testing.

The validation and testing data consists of 466 firms that are included in the S&P 500 in both 2021 and 2022, have usable Item 1. Business sections in both reporting years, and have the required subsequent-year GICS, stock-return, and financial data for the evaluation measures. Firm records across the 10-K reports, GICS data, stock-price data, and financial data are matched using both CIK and ticker identifiers.

Specifically, no GICS Sector changes are observed among these firms between 2021 and 2022. Stock returns are calculated using available trading-day observations from 1 January to 31 December 2022 for cutoff-ratio validation and from 1 January to 31 December 2023 for final evaluation. Financial variables are obtained from FMP annual data according to FMP fiscal-year classifications, with fiscal year 2022 used for cutoff-ratio validation and fiscal year 2023 used for final evaluation.

During cutoff-ratio validation, the 2021 Item 1. Business sections of the 466 firms are used for peer selection, and the resulting peer relationships are evaluated using the corresponding evaluation measures based on 2022 GICS, stock-return, and financial data. The selected ratios are then fixed and applied to the 2022 Item 1. Business sections of the same firms, and final peer-selection performance is evaluated using the corresponding 2023 data. Using subsequent-year evaluation data separates the reports used for peer selection from the outcomes used for evaluation.

Table 2 summarizes the sample composition, textual inputs, supervision and evaluation data, and the temporal roles of the data used at each stage of the study.

Table 2. Data structure and temporal design of the study.

Stage	Sample	10-K Text	Supervision/Evaluation Data	Purpose	Data Source
Sentence-classifier training	4150 non-S&P 500 firms	2021 Item 1	2021 GICS Sector	Train the SEC-BERT sentence classifier	SEC EDGAR, WRDS
Reconstruction-ratio validation	466 S&P 500 firms	2021 Item 1	2022 GICS, stock returns, and financial data	Determine reconstruction ratios	SEC EDGAR, WRDS, FMP
Final evaluation	Same 466 S&P 500 firms	2022 Item 1	2023 GICS, stock returns, and financial data	Final peer-selection evaluation	SEC EDGAR, WRDS, FMP

Table 3 presents the distribution of firms, sentences, and tokens across the 11 GICS Sectors in the classifier-training dataset.

4.1.1. Hyperparameters and Experimental Settings

Table 4 summarizes the classifier architecture and key hyperparameters used for analysis. The labeled sentences were split into training and internal development subsets at an 8:2 ratio, with the internal development subset used exclusively for early stopping based on the development loss. All experiments are conducted using an NVIDIA L40S GPU with PyTorch 2.2.0 and CUDA 12.1. The code for the case studies is available at github.com/smj1331/industry-discriminative-reconstruction.

Table 3. Distribution of firms, sentences, and tokens across GICS Sectors.

GICS Sector	No. of Firms	No. of Sentences	No. of Tokens
10 Energy	216	72,786	2,519,589
15 Materials	172	41,383	1,349,819
20 Industrials	514	123,138	4,112,522
25 Consumer Discretionary	403	102,276	3,435,995
30 Consumer Staples	160	36,787	1,218,521
35 Health Care	1008	547,709	20,511,790
40 Financials	685	250,752	8,982,556
45 Information Technology	492	117,204	3,868,439
50 Communication Services	168	45,304	1,577,300
55 Utilities	131	54,092	1,762,194
60 Real Estate	201	40,894	1,496,948
Total	4150	1,432,325	50,835,673

Table 4. Hyperparameters of the classifier.

Parameters	Value
Model	SEC-BERT (`nlpaueb/sec-bert-base`)
Optimizer	AdamW
Learning rate	3×10^{-5}
Number of epochs	100
Patience	5
Batch size	16
Loss function	CrossEntropyLoss
Tokenizer	AutoTokenizer (`nlpaueb/sec-bert-base`)
Max sequence length	512
Number of labels	11
Classifier head	Sequence classification head (`num_labels=11`)
Weight decay	0.01

4.1.2. Performance Metric

All evaluation metrics are computed under three peer set sizes: Top-1, Top-5, and Top-10. This multi-scale evaluation allows us to examine whether the performance trends hold consistently as the number of selected peers increases, as well as to investigate how performance systematically changes with the size of the peer set, thereby capturing both best-match precision and broader group-level tendencies.

(1) GICS alignment: Prior related-company extraction studies have evaluated whether nearest-neighbor firms retrieved from the embedding space are consistent with existing industry classification systems [11,36]. Building on this evaluation perspective, we use GICS as the reference taxonomy for assessing industry-classification alignment among selected peers. The GICS is organized as a hierarchical taxonomy consisting of four levels: Sector (11 categories), Industry Group (25 categories), Industry (74 categories), and Sub-Industry (163 categories) as shown in Figure 6.

Let $g_\ell(i)$ denote the GICS label of firm i at level ℓ, and let $\mathcal{N}_{Kp}(i)$ denote the set of top K peer firms selected for target firm i based on cosine similarity of embedding vectors. For each selected peer $j \in \mathcal{N}_{Kp}(i)$, the alignment indicator is defined as follows:

$$I_\ell(i,j) = \begin{cases} 1, & \text{if } g_\ell(i) = g_\ell(j), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

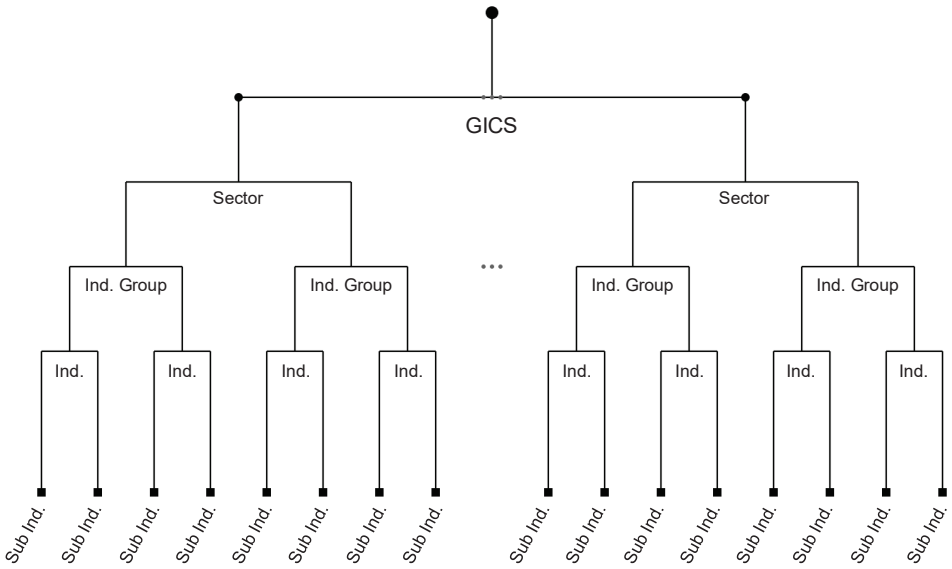


Figure 6. GICS structure.

Using this indicator, the GICS alignment accuracy at level ℓ is defined as follows:

$$\mathrm{Acc}_{\ell,\mathrm{peer@K}} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_{Kp}(i)} I_{\ell}(i,j), \tag{9}$$

where N denotes the total number of firms and K is the number of selected peers for each target firm. This metric measures the proportion of selected peer firms whose GICS labels match that of the target firm at the specified classification level ℓ. A higher value of $\mathrm{Acc}_{\ell,\mathrm{peer@K}}$ indicates that the selected peers are more closely aligned with the existing GICS classification system. Because GICS Sector labels are used to supervise sentence selection, GICS alignment is interpreted as a diagnostic measure of the industry-related information retained by the reconstructed representation rather than as independent external validation.

(2) Stock price return correlation coefficients: In addition to GICS alignment, stock returns are used as a market-based performance metric [5,8]. Specifically, the Pearson correlation coefficient between the time series of returns of two firms, i and j, is defined as follows:

$$\rho(i,j) = \frac{\sum_{t=1}^{T}(r_{i,t}-\bar{r}_i)(r_{j,t}-\bar{r}_j)}{\sqrt{\sum_{t=1}^{T}(r_{i,t}-\bar{r}_i)^2}\cdot\sqrt{\sum_{t=1}^{T}(r_{j,t}-\bar{r}_j)^2}}, \tag{10}$$

where $r_{i,t}$ represents the return of firm i at time t and $\bar{r}_i$ indicates the average return of firm i during the given period.

To quantitatively evaluate the peer-selection results, the average Pearson correlation is computed. For each firm i, the top K peers with the highest cosine similarity scores are selected, and the average Pearson correlation of return time series between the target firm and its selected peers is computed to define the following evaluation metric:

$$\mathrm{AvgCorr}_{\mathrm{peer@K}} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_{Kp}(i)} \rho(i,j). \tag{11}$$

Here, $\mathcal{N}_{Kp}(i)$ denotes the set of top K peers for target firm i selected based on cosine similarity, and N is the total number of firms. The validity of the peer selection is assessed by averaging the Pearson correlation coefficients across all $N \times K$ target-peer pairs as illustrated in Equation (11). The correlation analysis is conducted separately using daily

and weekly stock returns to examine whether the peer-selection results are consistent across different return frequencies.

For both GICS alignment and stock-return correlation, one-sided paired t-tests were conducted at the firm level to determine whether the proposed method achieves higher performance than each comparison method, with observations paired by target firm because both methods were evaluated on the same firms. In addition, to account for the increased probability of obtaining small p-values by chance when conducting repeated tests across multiple hypothesis tests, we applied the Benjamini–Hochberg false-discovery-rate correction to adjust p-value [37]. All hypothesis tests reported within the same result table were treated as one multiple-testing family because each table contains a set of related tests addressing the same comparison method and evaluation domain across different experimental conditions. The BH correction was applied separately to each table, and results with BH-adjusted p-values below 0.05 were considered statistically significant.

Furthermore, Cohen's d_z is calculated to show magnitude of the effectiveness of the proposed method [38]. Cohen's d_z values of approximately 0.2, 0.5, and 0.8 can be interpreted as small, medium, and large effects, respectively [39]. Finally, we additionally report a GICS industry-level cluster-bootstrap 95% confidence interval to account for potential dependence among firms within the same industry, as firms in the same industry may share common characteristics and be exposed to common economic shocks [40]. Industry clusters are resampled with replacement while all firm-level observations within each sampled industry are retained together. The paired performance difference is recomputed for each bootstrap sample, and the cluster-bootstrap 95% confidence interval is obtained from the 2.5th and 97.5th percentiles of the resulting bootstrap distribution. A cluster-bootstrap CI that excludes zero indicates that the performance difference remains statistically supported after accounting for potential within-industry dependence.

(3) Financial homogeneity: To assess how financially homogeneous the selected peer firms are relative to the target firm, we evaluate peer group quality using the adjusted coefficient of determination (R^2) from cross-sectional regressions, following the regression-based peer-evaluation approach of [5,12,14]. A well-constructed peer group should explain a larger share of cross-sectional variation in firms' financial characteristics. Consequently, more homogeneous peer selection provides higher cross-sectional explanatory power when peer group information is used to predict a target firm's characteristic.

For each target firm i and financial variable k, let $y_{i,k}$ denote the value of variable k for firm i. Let $\bar{y}_{\mathcal{P}_i,k}$ denote the mean value of variable k across the selected peer set $\mathcal{P}_i$ (excluding the target firm) as follows:

$$\bar{y}_{\mathcal{P}_i,k} = \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} y_{j,k}.$$

Then, we estimate the following cross-sectional regression across all N firms as follows:

$$y_{i,k} = \alpha_k + \beta_k \bar{y}_{\mathcal{P}_i,k} + \varepsilon_{i,k}. \quad (12)$$

We report the adjusted R_k^2 separately for each financial variable from Equation (12). A higher R_k^2 indicates that peer group averages more precisely capture the target firm's financial characteristics and is interpreted as evidence of greater financial homogeneity among the selected peers. Prior peer-firm studies have evaluated financial comparability using multiple dimensions of firm characteristics, including valuation multiples, operating efficiency, profitability, and growth [5,12,14]. Accordingly, we use Enterprise Value-to-Sales

(EVS), Asset Turnover (AT), Return on Assets (ROA), and Revenue Growth to represent these four dimensions, respectively.

The financial variables are obtained from Financial Modeling Prep (FMP), and their calculations and corresponding data fields follow the definitions provided by FMP [30].

The EVS ratio measures how much investors are willing to pay per dollar of sales [41]. In this study, EVS is based on the FMP definition and is calculated as follows:

$$\text{EVS} = \frac{\text{Enterprise Value}}{\text{Revenue}}. \tag{13}$$

AT captures a firm's ability to convert its assets into revenue, reflecting its asset utilization [42]. Based on the FMP definition, AT is calculated as follows:

$$\text{AT} = \frac{\text{Revenue}}{\text{Total Assets}}. \tag{14}$$

ROA is a profitability measure that represents profit relative to a firm's assets [43]. Based on the FMP definition, ROA is calculated as follows:

$$\text{ROA} = \frac{\text{Net Income}}{\text{Total Assets}}. \tag{15}$$

Revenue Growth represents the percentage change in a firm's revenue from the previous year to the current year [44]. The corresponding FMP fields used for EVS, AT, ROA, and Revenue Growth are `evToSales`, `assetTurnover`, `returnOnAssets`, and `growthRevenue`, respectively [30]. To mitigate the influence of outliers, we remove the top and bottom 1% of observations for each financial variable [14].

4.1.3. Table Notation

The notation used throughout the experimental tables in the following sections is summarized in Table 5. Here, s_P and s_C denote the performance scores of the proposed and comparison methods, respectively. The paired-t and bootstrap 95% CIs are calculated for the performance difference between the proposed and comparison methods. All hypothesis tests were evaluated at the 0.05 significance level for both the unadjusted and BH-adjusted p-values.

Table 5. Notation used in experimental tables.

Notation	Description
r^*	Selected retained ratio for the proposed method
s	Performance score of a method under the corresponding evaluation metric
s_P	Performance score of the proposed method
s_C	Performance score of the comparison method
Paired-t 95% CI	95% confidence interval of the mean paired performance difference between the proposed and comparison methods
p	One-sided paired t-test p-value
p_{BH}	Benjamini–Hochberg-adjusted p-value
Cohen's d_z	Paired-sample effect size
Bootstrap 95% CI	Industry-level cluster-bootstrap 95% confidence interval of the performance difference between the proposed and comparison methods
RD SD	Standard deviation across five random-deletion repetitions
RD 95% CI	95% confidence interval across five random-deletion repetitions

4.2. Classification Results

To evaluate the sentence classifier on firms not used for classifier training, we conducted a held-out evaluation using the 2021 10-K reports of 466 S&P 500 firms. The evaluation set contains 141,614 sentences, with each sentence assigned the GICS Sector label of its source firm. Table 6 reports the overall sentence-level classification performance.

The classifier achieved an accuracy of 0.6074, a macro precision of 0.6251, a macro recall of 0.5633, a macro F1 score of 0.5814, and a weighted F1 score of 0.6040. Table 7 further reports class-specific performance across the 11 GICS Sectors. Performance varies across sectors, with F1 scores ranging from 0.3989 for Consumer Staples to 0.8407 for Utilities. These results provide a sector-level view of the classifier's performance and show that classification performance is not uniform across GICS categories.

Table 6. Classification performance on the 2021 S&P 500 sample.

Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
0.6074	0.6251	0.5633	0.5814	0.6040

Table 7. Class-specific classification performance on the 2021 S&P 500 sample.

Sector	Sector Name	Firms	Support	Precision	Recall	F1
10	Energy	21	6619	0.6255	0.6842	0.6535
15	Materials	29	7894	0.6634	0.3894	0.4907
20	Industrials	70	17,109	0.3943	0.5214	0.4490
25	Consumer Discretionary	50	13,660	0.5301	0.5072	0.5184
30	Consumer Staples	35	6776	0.5810	0.3037	0.3989
35	Health Care	57	25,029	0.6476	0.7341	0.6882
40	Financials	64	23,502	0.6901	0.7221	0.7057
45	Information Technology	62	16,620	0.5571	0.5755	0.5661
50	Communication Services	21	6210	0.6457	0.4931	0.5592
55	Utilities	27	11,492	0.8345	0.8470	0.8407
60	Real Estate	30	6703	0.7067	0.4180	0.5253

4.3. Validation Studies

The validation study evaluates the effectiveness of the proposed text refinement framework by comparing it with the following baseline setting. All controls use the same firms, 2022 Item 1. Business sections, embedding model, and cosine-similarity-based Top-K peer-selection procedure.

- Full text: No sentence filtering. The complete Item 1. Business section is embedded with the same embedding model as the proposed method.
- Length-matched random deletion: Sentences are randomly retained, and the same retained ratio ($r^*_{m,q,K}$) is used. Five repetitions are summarized by their mean, SD, and CI.
- Longest-sentence ranking: Sentences are ranked by sentence length, and the same retained ratio ($r^*_{m,q,K}$) is used.
- TF-IDF ranking: Sentences are ranked by sentence-level TF-IDF score [45], and the same retained ratio ($r^*_{m,q,K}$) is used.
- Low-entropy ranking: Sentences are ranked by predictive entropy computed from the same SEC-BERT 11-class probability vector, with lower-entropy sentences prioritized, and the same retained ratio ($r^*_{m,q,K}$) is used.

By demonstrating consistent performance improvements across multiple embedding models and evaluation metrics, the results validate the generalizability of the proposed method. The embedding models are described in Section 4.3.1, and detailed results are presented in Sections 4.3.2–4.3.4.

4.3.1. Embedding Models

In this section, to evaluate the generalizability of the proposed framework, we select five representative embedding models: Fin-E5 [46], E5-Mistral [47], NV-Embed-v2 [48],

OpenAI text-embedding-3-large [49], and Voyage-3-large [50]. Consistent performance improvements across these models suggest that the effectiveness of the proposed reconstruction framework is not dependent on a specific embedding model but is broadly applicable across different representation settings.

- Fin-E5 [46]: A finance-adapted embedding model built by fine-tuning e5-mistral-7b-instruct on a synthetic financial corpus, designed to capture domain-specific semantics in financial text more effectively. We use `abacinlp-text-v1` in this study.
- E5-Mistral [47]: An instruction-tuned embedding model based on the Mistral-7B architecture, designed for high-quality semantic retrieval and text representation. We use e5-mistral-7b-instruct.
- NV-Embed-v2 [48]: A large-scale general-purpose embedding model proposed to enhance retrieval and downstream embedding quality through improved model design and training methodology. We use `NV-Embed-v2`.
- OpenAI [49]: A commercial embedding model designed for semantic retrieval, similarity measurement, clustering, and recommendation tasks. We use `text-embedding-3-large`.
- Voyage-3-Large [50]: A large general-purpose embedding model optimized for high-quality multilingual and retrieval-focused representations. We use `voyage-3-large`.

4.3.2. Validation Results for GICS Alignment

Tables 8–10 report GICS alignment against matched random deletion, the full-text baseline, and alternative sentence-selection criteria. Higher alignment means that a larger share of the Top-K peers has the same GICS label as the target firm. The evaluation covers Sector, Industry Group, Industry, and Sub-Industry even though sentence ranking is learned from Sector-level supervision, allowing us to examine whether the reconstructed reports also retain finer industry structure.

Compared with matched random deletion (Table 8), the proposed method has higher GICS alignment in all 60 comparisons, and 58 remain significant after BH correction. Because random deletion retains the same proportion of text, this comparison controls for the reduction in document length. The higher alignment is therefore associated with sentence selection rather than merely with shortening the report itself. The paired-difference 95% CIs are generally narrow and lie above zero, indicating that the estimated firm-level performance gains are positive and relatively precise in most settings. The corresponding Cohen's d_z values range from small to large and generally increase with Top-K, indicating that the relative benefit of the proposed method becomes more pronounced when a broader set of peer firms is considered. In addition, the industry-level cluster-bootstrap 95% CIs exclude zero in 57 of the 60 comparisons, showing that the performance advantage generally remains supported after accounting for potential dependence among firms within the same industry.

Compared with the full-text baseline (Table 9), the proposed method has higher alignment in 58 of the 60 comparisons, with 50 remaining significant after BH correction. Because the full-text baseline retains every sentence in Item 1, these results indicate that using all available text is not necessarily optimal for industry-based peer identification. The performance differences and corresponding Cohen's d_z values are generally more pronounced at larger Top-K values, suggesting that selective filtering becomes particularly useful when a broader peer set is considered. Moreover, the industry-level cluster-bootstrap 95% CIs are above zero for most comparisons, indicating that the principal findings are generally maintained after accounting for potential within-industry dependence.

For the alternative sentence-selection analysis, E5-Mistral is used as the representative embedding backbone because it shows consistently strong performance across the main evaluation metrics. In Table 10, the proposed method achieves better alignments than

entropy, sentence-length, and TF-IDF ranking in all 36 comparisons, with 32 remaining significant after BH correction. The paired-difference 95% CIs and Cohen's d_z values show that the advantage of the proposed ranking is generally more pronounced at larger Top-K values, particularly relative to sentence-length and TF-IDF ranking. The industry-level cluster-bootstrap 95% CIs also support most of these differences after accounting for potential dependence among firms within the same industry, although intervals containing zero indicate that some comparisons are less robust to industry-level variation. Sentence length and TF-IDF therefore do not recover the same industry information simply by favoring longer or lexically salient sentences. Entropy performs more competitively, showing that classifier confidence is informative, but conditioning the score on the firm's actual Sector in the proposed method provides a more task-specific signal for industry-based peer selection.

4.3.3. Validation Results for Stock-Return Correlation

Tables 11–13 report stock-return correlations using daily and weekly returns. A higher correlation means that the target firm and its selected peers move more similarly in the market. Reporting both frequencies checks whether the result depends on the return aggregation used in evaluation.

Compared with matched random deletion (Table 11), the proposed method records higher return correlations in 29 of the 30 comparisons, and 28 remain significant after BH correction. Because both methods retain the same amount of text, the gain cannot be attributed solely to document shortening. These results suggest that the sentence-ranking criterion selects content that is more useful for identifying firms with similar return behavior than an equally sized random subset. The advantage generally becomes more pronounced as Top-K increases. In particular, Cohen's d_z is largest at Top-10 for both daily and weekly returns across all five embedding models, indicating that the relative benefit of the proposed method is more evident when a broader set of peer firms is considered. Although the absolute return correlations tend to decrease as Top-K increases because progressively less similar firms are included, the performance gap relative to matched random deletion generally widens. Moreover, the industry-level cluster-bootstrap 95% CIs are above zero in 28 of the 30 comparisons, indicating that the performance advantage is generally maintained after accounting for potential dependence among firms within the same industry.

Compared with the full-text baseline (Table 12), the proposed method records higher return correlations in 28 of the 30 comparisons, with 22 remaining significant after BH correction. Thus, retaining every sentence is not required to preserve market-related similarity, and selective removal can improve the resulting peer representation. The advantage is particularly consistent at Top-10, where all ten daily and weekly comparisons remain significant after BH correction and their industry-level cluster-bootstrap 95% CIs are above zero. The corresponding Cohen's d_z values are also generally larger than those at Top-1, suggesting that selective filtering becomes more beneficial relative to the full-text representation when similarity is evaluated over a broader peer set.

Table 13 compares the sentence-ranking criteria with E5-Mistral fixed as the embedding model. The proposed ranking records higher return correlations in 16 of the 18 pairwise comparisons, with one tie, and 12 differences remain significant after BH correction. Its advantage over sentence-length and TF-IDF ranking indicates that longer sentences or greater lexical salience alone do not account for the return-correlation gains. These differences and their corresponding Cohen's d_z values are generally more pronounced at Top-10, suggesting that task-specific sentence selection becomes more useful as the peer set expands. The industry-level cluster-bootstrap 95% CIs support the majority of these differences after accounting for potential within-industry dependence, although several intervals include zero.

Table 8. Comparison between the proposed method and matched random deletion for GICS alignment.

GICS Level	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Sector	1	[46]	0.1	0.897	0.770	0.010	[0.758, 0.783]	[0.097, 0.156]	<0.001	<0.001	0.389	[0.086, 0.173]
		[47]	0.2	0.933	0.867	0.010	[0.855, 0.879]	[0.044, 0.089]	<0.001	<0.001	0.272	[0.040, 0.098]
		[48]	0.1	0.933	0.806	0.023	[0.777, 0.834]	[0.101, 0.155]	<0.001	<0.001	0.431	[0.093, 0.170]
		[49]	0.1	0.921	0.837	0.033	[0.795, 0.878]	[0.060, 0.107]	<0.001	<0.001	0.324	[0.060, 0.111]
		[50]	0.1	0.923	0.823	0.009	[0.811, 0.834]	[0.075, 0.125]	<0.001	<0.001	0.363	[0.068, 0.136]
	5	[46]	0.1	0.814	0.671	0.005	[0.664, 0.677]	[0.125, 0.161]	<0.001	<0.001	0.716	[0.114, 0.173]
		[47]	0.2	0.867	0.777	0.005	[0.771, 0.784]	[0.075, 0.104]	<0.001	<0.001	0.553	[0.064, 0.122]
		[48]	0.1	0.838	0.707	0.008	[0.697, 0.717]	[0.112, 0.151]	<0.001	<0.001	0.612	[0.099, 0.168]
		[49]	0.1	0.828	0.741	0.012	[0.726, 0.757]	[0.070, 0.105]	<0.001	<0.001	0.458	[0.064, 0.112]
		[50]	0.1	0.848	0.739	0.007	[0.730, 0.747]	[0.093, 0.127]	<0.001	<0.001	0.601	[0.086, 0.137]
	10	[46]	0.1	0.743	0.595	0.004	[0.589, 0.600]	[0.132, 0.164]	<0.001	<0.001	0.840	[0.120, 0.175]
		[47]	0.1	0.822	0.690	0.006	[0.683, 0.698]	[0.117, 0.146]	<0.001	<0.001	0.807	[0.104, 0.163]
		[48]	0.1	0.771	0.633	0.006	[0.625, 0.641]	[0.121, 0.156]	<0.001	<0.001	0.723	[0.105, 0.173]
		[49]	0.1	0.766	0.669	0.009	[0.657, 0.680]	[0.082, 0.112]	<0.001	<0.001	0.599	[0.076, 0.119]
		[50]	0.1	0.801	0.675	0.006	[0.667, 0.683]	[0.112, 0.141]	<0.001	<0.001	0.798	[0.103, 0.151]
Industry Group	1	[46]	0.1	0.792	0.671	0.006	[0.664, 0.679]	[0.085, 0.156]	<0.001	<0.001	0.309	[0.069, 0.172]
		[47]	0.2	0.871	0.789	0.012	[0.774, 0.805]	[0.053, 0.111]	<0.001	<0.001	0.257	[0.048, 0.120]
		[48]	0.1	0.824	0.705	0.028	[0.671, 0.739]	[0.085, 0.154]	<0.001	<0.001	0.313	[0.077, 0.165]
		[49]	0.1	0.820	0.748	0.030	[0.711, 0.786]	[0.040, 0.102]	<0.001	<0.001	0.208	[0.022, 0.114]
		[50]	0.4	0.863	0.795	0.004	[0.790, 0.799]	[0.042, 0.093]	<0.001	<0.001	0.241	[0.034, 0.104]
	5	[46]	0.1	0.670	0.544	0.007	[0.536, 0.553]	[0.106, 0.145]	<0.001	<0.001	0.585	[0.091, 0.158]
		[47]	0.2	0.752	0.663	0.007	[0.654, 0.672]	[0.072, 0.106]	<0.001	<0.001	0.481	[0.060, 0.123]
		[48]	0.1	0.692	0.575	0.017	[0.554, 0.596]	[0.097, 0.138]	<0.001	<0.001	0.522	[0.079, 0.156]
		[49]	0.1	0.696	0.616	0.012	[0.601, 0.631]	[0.061, 0.099]	<0.001	<0.001	0.384	[0.042, 0.115]
		[50]	0.1	0.725	0.619	0.008	[0.610, 0.629]	[0.088, 0.124]	<0.001	<0.001	0.534	[0.074, 0.137]
	10	[46]	0.3	0.568	0.496	0.004	[0.491, 0.502]	[0.057, 0.085]	<0.001	<0.001	0.467	[0.051, 0.091]
		[47]	0.1	0.674	0.550	0.006	[0.542, 0.557]	[0.109, 0.140]	<0.001	<0.001	0.716	[0.092, 0.159]
		[48]	0.1	0.599	0.485	0.011	[0.471, 0.498]	[0.098, 0.131]	<0.001	<0.001	0.627	[0.082, 0.148]
		[49]	0.1	0.615	0.522	0.008	[0.512, 0.532]	[0.078, 0.107]	<0.001	<0.001	0.586	[0.066, 0.118]
		[50]	0.1	0.653	0.541	0.005	[0.535, 0.547]	[0.097, 0.126]	<0.001	<0.001	0.686	[0.085, 0.140]
Industry	1	[46]	0.6	0.635	0.615	0.019	[0.592, 0.639]	[−0.010, 0.050]	0.098	0.100	0.060	[−0.017, 0.056]
		[47]	0.6	0.745	0.720	0.006	[0.713, 0.727]	[0.000, 0.049]	0.024	0.025	0.092	[0.002, 0.046]
		[48]	0.1	0.682	0.576	0.013	[0.559, 0.592]	[0.070, 0.144]	<0.001	<0.001	0.261	[0.062, 0.152]
		[49]	0.1	0.685	0.626	0.022	[0.599, 0.654]	[0.024, 0.093]	<0.001	<0.001	0.153	[0.009, 0.102]
		[50]	0.6	0.738	0.703	0.014	[0.685, 0.720]	[0.007, 0.064]	0.007	0.008	0.115	[0.005, 0.065]

Table 8. *Cont.*

GICS Level	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Industry	5	[46]	0.1	0.505	0.398	0.004	[0.393, 0.403]	[0.088, 0.125]	<0.001	<0.001	0.525	[0.074, 0.136]
		[47]	0.4	0.562	0.525	0.003	[0.522, 0.529]	[0.023, 0.049]	<0.001	<0.001	0.256	[0.017, 0.057]
		[48]	0.1	0.505	0.430	0.011	[0.416, 0.443]	[0.058, 0.094]	<0.001	<0.001	0.380	[0.044, 0.106]
		[49]	0.1	0.512	0.456	0.008	[0.447, 0.466]	[0.038, 0.074]	<0.001	<0.001	0.283	[0.015, 0.091]
		[50]	0.1	0.563	0.468	0.007	[0.459, 0.476]	[0.078, 0.113]	<0.001	<0.001	0.499	[0.067, 0.122]
	10	[46]	0.4	0.414	0.349	0.004	[0.344, 0.354]	[0.053, 0.077]	<0.001	<0.001	0.505	[0.040, 0.088]
		[47]	0.2	0.469	0.414	0.005	[0.408, 0.420]	[0.043, 0.067]	<0.001	<0.001	0.412	[0.032, 0.077]
		[48]	0.1	0.402	0.336	0.006	[0.328, 0.344]	[0.054, 0.079]	<0.001	<0.001	0.477	[0.039, 0.090]
		[49]	0.1	0.420	0.361	0.005	[0.355, 0.367]	[0.046, 0.071]	<0.001	<0.001	0.432	[0.029, 0.087]
		[50]	0.1	0.458	0.381	0.002	[0.379, 0.384]	[0.064, 0.089]	<0.001	<0.001	0.549	[0.053, 0.098]
Sub-Industry	1	[46]	0.6	0.528	0.518	0.015	[0.499, 0.537]	[−0.021, 0.041]	0.256	0.256	0.030	[−0.026, 0.043]
		[47]	0.7	0.655	0.624	0.006	[0.617, 0.631]	[0.004, 0.057]	0.011	0.012	0.106	[0.007, 0.055]
		[48]	0.1	0.562	0.462	0.018	[0.439, 0.485]	[0.062, 0.139]	<0.001	<0.001	0.236	[0.058, 0.144]
		[49]	0.1	0.586	0.501	0.025	[0.470, 0.532]	[0.049, 0.121]	<0.001	<0.001	0.212	[0.038, 0.128]
		[50]	0.7	0.650	0.626	0.012	[0.611, 0.641]	[−0.004, 0.053]	0.044	0.046	0.079	[−0.007, 0.056]
	5	[46]	0.1	0.366	0.284	0.004	[0.279, 0.289]	[0.066, 0.098]	<0.001	<0.001	0.470	[0.057, 0.105]
		[47]	0.3	0.427	0.390	0.007	[0.381, 0.399]	[0.024, 0.050]	<0.001	<0.001	0.257	[0.016, 0.059]
		[48]	0.1	0.364	0.312	0.006	[0.304, 0.320]	[0.037, 0.068]	<0.001	<0.001	0.306	[0.029, 0.074]
		[49]	0.1	0.379	0.334	0.009	[0.323, 0.345]	[0.030, 0.061]	<0.001	<0.001	0.265	[0.015, 0.072]
		[50]	0.5	0.417	0.392	0.002	[0.390, 0.394]	[0.014, 0.036]	<0.001	<0.001	0.206	[0.011, 0.040]
	10	[46]	0.3	0.266	0.232	0.003	[0.228, 0.236]	[0.024, 0.042]	<0.001	<0.001	0.338	[0.019, 0.048]
		[47]	0.3	0.315	0.286	0.004	[0.281, 0.290]	[0.021, 0.037]	<0.001	<0.001	0.339	[0.014, 0.043]
		[48]	0.1	0.264	0.227	0.003	[0.223, 0.230]	[0.027, 0.047]	<0.001	<0.001	0.341	[0.018, 0.054]
		[49]	0.1	0.283	0.246	0.006	[0.239, 0.253]	[0.027, 0.047]	<0.001	<0.001	0.347	[0.017, 0.056]
		[50]	0.2	0.308	0.273	0.005	[0.267, 0.279]	[0.025, 0.045]	<0.001	<0.001	0.330	[0.015, 0.056]

Table 9. Comparison between the proposed method and the full-text baseline for GICS alignment.

GICS Level	Top-K	Embedding Model	r^*	s_P	s_C	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Sector	1	[46]	0.1	0.897	0.843	[0.019, 0.088]	0.001	0.002	0.141	[0.018, 0.091]
		[47]	0.2	0.933	0.876	[0.031, 0.084]	<0.001	<0.001	0.199	[0.031, 0.089]
		[48]	0.1	0.933	0.609	[0.279, 0.370]	<0.001	<0.001	0.649	[0.260, 0.393]
		[49]	0.1	0.921	0.712	[0.166, 0.250]	<0.001	<0.001	0.452	[0.160, 0.260]
		[50]	0.1	0.923	0.903	[−0.009, 0.048]	0.090	0.104	0.062	[−0.018, 0.060]
	5	[46]	0.1	0.814	0.744	[0.049, 0.090]	<0.001	<0.001	0.304	[0.045, 0.096]
		[47]	0.2	0.867	0.791	[0.060, 0.092]	<0.001	<0.001	0.420	[0.050, 0.106]
		[48]	0.1	0.838	0.508	[0.299, 0.362]	<0.001	<0.001	0.957	[0.276, 0.388]
		[49]	0.1	0.828	0.595	[0.204, 0.262]	<0.001	<0.001	0.742	[0.197, 0.272]
		[50]	0.1	0.848	0.785	[0.045, 0.081]	<0.001	<0.001	0.318	[0.036, 0.092]
	10	[46]	0.1	0.743	0.676	[0.050, 0.085]	<0.001	<0.001	0.352	[0.044, 0.093]
		[47]	0.1	0.822	0.734	[0.073, 0.103]	<0.001	<0.001	0.531	[0.061, 0.118]
		[48]	0.1	0.771	0.459	[0.284, 0.340]	<0.001	<0.001	1.009	[0.258, 0.369]
		[49]	0.1	0.766	0.523	[0.217, 0.269]	<0.001	<0.001	0.864	[0.208, 0.279]
		[50]	0.1	0.801	0.723	[0.063, 0.093]	<0.001	<0.001	0.461	[0.055, 0.103]
Industry Group	1	[46]	0.1	0.792	0.753	[−0.003, 0.080]	0.034	0.041	0.084	[−0.010, 0.085]
		[47]	0.2	0.871	0.811	[0.027, 0.093]	<0.001	<0.001	0.164	[0.024, 0.100]
		[48]	0.1	0.824	0.489	[0.282, 0.387]	<0.001	<0.001	0.582	[0.260, 0.408]
		[49]	0.1	0.820	0.609	[0.162, 0.259]	<0.001	<0.001	0.396	[0.138, 0.279]
		[50]	0.4	0.863	0.850	[−0.020, 0.046]	0.220	0.235	0.036	[−0.027, 0.053]
	5	[46]	0.1	0.670	0.625	[0.023, 0.066]	<0.001	<0.001	0.189	[0.018, 0.071]
		[47]	0.2	0.752	0.685	[0.049, 0.086]	<0.001	<0.001	0.326	[0.037, 0.100]
		[48]	0.1	0.692	0.377	[0.282, 0.347]	<0.001	<0.001	0.885	[0.254, 0.376]
		[49]	0.1	0.696	0.467	[0.200, 0.259]	<0.001	<0.001	0.708	[0.183, 0.272]
		[50]	0.1	0.725	0.679	[0.027, 0.066]	<0.001	<0.001	0.217	[0.013, 0.079]
	10	[46]	0.3	0.568	0.541	[0.010, 0.042]	<0.001	<0.001	0.149	[0.004, 0.047]
		[47]	0.1	0.674	0.607	[0.051, 0.084]	<0.001	<0.001	0.373	[0.036, 0.100]
		[48]	0.1	0.599	0.318	[0.254, 0.308]	<0.001	<0.001	0.941	[0.223, 0.340]
		[49]	0.1	0.615	0.386	[0.205, 0.254]	<0.001	<0.001	0.855	[0.190, 0.267]
		[50]	0.1	0.653	0.597	[0.041, 0.071]	<0.001	<0.001	0.343	[0.031, 0.082]
Industry	1	[46]	0.6	0.635	0.665	[−0.067, 0.007]	0.943	0.959	−0.074	[−0.082, 0.017]
		[47]	0.6	0.745	0.721	[−0.004, 0.051]	0.047	0.055	0.078	[−0.003, 0.050]
		[48]	0.1	0.682	0.363	[0.265, 0.374]	<0.001	<0.001	0.536	[0.241, 0.397]
		[49]	0.1	0.685	0.470	[0.162, 0.267]	<0.001	<0.001	0.372	[0.140, 0.284]
		[50]	0.6	0.738	0.734	[−0.030, 0.039]	0.403	0.417	0.011	[−0.036, 0.041]
	5	[46]	0.1	0.505	0.476	[0.009, 0.050]	0.003	0.004	0.129	[−0.001, 0.058]
		[47]	0.4	0.562	0.526	[0.021, 0.050]	<0.001	<0.001	0.226	[0.016, 0.056]
		[48]	0.1	0.505	0.250	[0.227, 0.283]	<0.001	<0.001	0.828	[0.202, 0.307]
		[49]	0.1	0.512	0.329	[0.157, 0.210]	<0.001	<0.001	0.628	[0.140, 0.225]
		[50]	0.1	0.563	0.529	[0.015, 0.054]	<0.001	<0.001	0.161	[0.000, 0.068]
	10	[46]	0.4	0.414	0.380	[0.022, 0.047]	<0.001	<0.001	0.250	[0.010, 0.056]
		[47]	0.2	0.469	0.437	[0.019, 0.046]	<0.001	<0.001	0.218	[0.006, 0.059]
		[48]	0.1	0.402	0.198	[0.183, 0.226]	<0.001	<0.001	0.854	[0.153, 0.255]
		[49]	0.1	0.420	0.255	[0.144, 0.184]	<0.001	<0.001	0.749	[0.128, 0.199]
		[50]	0.1	0.458	0.431	[0.013, 0.040]	<0.001	<0.001	0.180	[0.003, 0.049]
Sub-Industry	1	[46]	0.6	0.528	0.588	[−0.097, −0.023]	0.999	0.999	−0.148	[−0.100, −0.025]
		[47]	0.7	0.655	0.618	[0.006, 0.067]	0.010	0.012	0.109	[0.006, 0.066]
		[48]	0.1	0.562	0.260	[0.248, 0.357]	<0.001	<0.001	0.503	[0.233, 0.371]
		[49]	0.1	0.586	0.354	[0.180, 0.283]	<0.001	<0.001	0.409	[0.167, 0.295]
		[50]	0.7	0.650	0.644	[−0.028, 0.041]	0.357	0.376	0.017	[−0.037, 0.047]
	5	[46]	0.1	0.366	0.355	[−0.006, 0.029]	0.107	0.119	0.058	[−0.013, 0.034]
		[47]	0.3	0.427	0.399	[0.014, 0.042]	<0.001	<0.001	0.177	[0.006, 0.050]
		[48]	0.1	0.364	0.164	[0.175, 0.224]	<0.001	<0.001	0.739	[0.156, 0.244]
		[49]	0.1	0.379	0.224	[0.132, 0.179]	<0.001	<0.001	0.610	[0.116, 0.195]
		[50]	0.5	0.417	0.408	[−0.005, 0.022]	0.099	0.113	0.060	[−0.009, 0.027]
	10	[46]	0.3	0.266	0.259	[−0.004, 0.017]	0.114	0.124	0.056	[−0.013, 0.025]
		[47]	0.3	0.315	0.298	[0.008, 0.024]	<0.001	<0.001	0.181	[0.001, 0.032]
		[48]	0.1	0.264	0.124	[0.123, 0.157]	<0.001	<0.001	0.752	[0.104, 0.177]
		[49]	0.1	0.283	0.162	[0.105, 0.136]	<0.001	<0.001	0.690	[0.090, 0.153]
		[50]	0.2	0.308	0.295	[0.003, 0.023]	0.007	0.009	0.114	[−0.008, 0.034]

Table 10. Comparison between the proposed method and alternative sentence-selection methods for GICS alignment.

GICS Level	Top-K	Method	r^*	s	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Sector	1	Proposed	0.2	0.933	–	–	–	–	–
		Entropy	–	0.891	[0.019, 0.067]	<0.001	<0.001	0.161	[0.013, 0.076]
		Sentence Length	–	0.873	[0.033, 0.088]	<0.001	<0.001	0.199	[0.029, 0.093]
		[45]	–	0.863	[0.041, 0.100]	<0.001	<0.001	0.219	[0.035, 0.113]
	5	Proposed	0.2	0.867	–	–	–	–	–
		Entropy	–	0.814	[0.038, 0.068]	<0.001	<0.001	0.317	[0.030, 0.079]
		Sentence Length	–	0.786	[0.063, 0.099]	<0.001	<0.001	0.410	[0.050, 0.117]
		[45]	–	0.764	[0.084, 0.122]	<0.001	<0.001	0.507	[0.072, 0.138]
	10	Proposed	0.1	0.822	–	–	–	–	–
		Entropy	–	0.748	[0.060, 0.088]	<0.001	<0.001	0.485	[0.051, 0.100]
		Sentence Length	–	0.683	[0.120, 0.158]	<0.001	<0.001	0.665	[0.101, 0.181]
		[45]	–	0.689	[0.115, 0.151]	<0.001	<0.001	0.666	[0.101, 0.167]
Industry Group	1	Proposed	0.2	0.871	–	–	–	–	–
		Entropy	–	0.835	[0.011, 0.062]	0.003	0.003	0.130	[0.007, 0.069]
		Sentence Length	–	0.796	[0.042, 0.108]	<0.001	<0.001	0.205	[0.035, 0.117]
		[45]	–	0.796	[0.040, 0.111]	<0.001	<0.001	0.193	[0.032, 0.121]
	5	Proposed	0.2	0.752	–	–	–	–	–
		Entropy	–	0.710	[0.028, 0.057]	<0.001	<0.001	0.260	[0.022, 0.065]
		Sentence Length	–	0.677	[0.055, 0.096]	<0.001	<0.001	0.336	[0.039, 0.113]
		[45]	–	0.664	[0.068, 0.110]	<0.001	<0.001	0.391	[0.053, 0.125]
	10	Proposed	0.1	0.674	–	–	–	–	–
		Entropy	–	0.618	[0.044, 0.070]	<0.001	<0.001	0.396	[0.036, 0.081]
		Sentence Length	–	0.546	[0.109, 0.147]	<0.001	<0.001	0.602	[0.089, 0.170]
		[45]	–	0.560	[0.095, 0.133]	<0.001	<0.001	0.557	[0.076, 0.151]
Industry	1	Proposed	0.6	0.745	–	–	–	–	–
		Entropy	–	0.727	[−0.007, 0.041]	0.079	0.081	0.066	[−0.004, 0.038]
		Sentence Length	–	0.727	[−0.012, 0.046]	0.124	0.124	0.054	[−0.007, 0.040]
		[45]	–	0.710	[0.007, 0.062]	0.007	0.008	0.115	[0.011, 0.058]
	5	Proposed	0.4	0.562	–	–	–	–	–
		Entropy	–	0.545	[0.006, 0.026]	<0.001	0.001	0.146	[0.003, 0.030]
		Sentence Length	–	0.524	[0.022, 0.053]	<0.001	<0.001	0.222	[0.013, 0.064]
		[45]	–	0.526	[0.019, 0.052]	<0.001	<0.001	0.198	[0.009, 0.064]
	10	Proposed	0.2	0.469	–	–	–	–	–
		Entropy	–	0.448	[0.013, 0.030]	<0.001	<0.001	0.224	[0.008, 0.038]
		Sentence Length	–	0.416	[0.039, 0.067]	<0.001	<0.001	0.344	[0.024, 0.082]
		[45]	–	0.412	[0.042, 0.072]	<0.001	<0.001	0.349	[0.025, 0.088]
Sub-Industry	1	Proposed	0.7	0.655	–	–	–	–	–
		Entropy	–	0.631	[−0.001, 0.048]	0.028	0.031	0.089	[0.000, 0.048]
		Sentence Length	–	0.618	[0.006, 0.067]	0.010	0.012	0.109	[0.002, 0.071]
		[45]	–	0.633	[−0.007, 0.050]	0.070	0.074	0.068	[−0.008, 0.050]
	5	Proposed	0.3	0.427	–	–	–	–	–
		Entropy	–	0.419	[−0.001, 0.017]	0.049	0.053	0.077	[−0.006, 0.023]
		Sentence Length	–	0.393	[0.018, 0.049]	<0.001	<0.001	0.199	[0.010, 0.059]
		[45]	–	0.398	[0.014, 0.044]	<0.001	<0.001	0.174	[0.002, 0.057]
	10	Proposed	0.3	0.315	–	–	–	–	–
		Entropy	–	0.307	[0.001, 0.013]	0.010	0.012	0.108	[−0.002, 0.017]
		Sentence Length	–	0.286	[0.020, 0.038]	<0.001	<0.001	0.289	[0.012, 0.047]
		[45]	–	0.285	[0.021, 0.040]	<0.001	<0.001	0.289	[0.013, 0.046]

Table 11. Comparison between the proposed method and matched random deletion for stock-return correlation.

Freq.	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Daily	1	[46]	0.7	0.566	0.553	0.002	[0.550, 0.555]	[0.004, 0.023]	0.003	0.003	0.128	[0.003, 0.024]
		[47]	0.6	0.582	0.571	0.002	[0.569, 0.573]	[0.003, 0.018]	0.003	0.003	0.130	[0.003, 0.019]
		[48]	0.1	0.550	0.523	0.005	[0.517, 0.529]	[0.015, 0.039]	<0.001	<0.001	0.204	[0.015, 0.040]
		[49]	0.1	0.555	0.535	0.012	[0.521, 0.550]	[0.008, 0.031]	<0.001	<0.001	0.154	[0.004, 0.032]
		[50]	0.9	0.587	0.579	0.002	[0.577, 0.581]	[0.002, 0.013]	0.003	0.003	0.130	[0.002, 0.014]

Table 11. *Cont.*

Freq.	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Daily	5	[46]	0.6	0.487	0.476	0.003	[0.472, 0.480]	[0.006, 0.015]	<0.001	<0.001	0.222	[0.006, 0.016]
		[47]	0.7	0.508	0.502	0.001	[0.500, 0.503]	[0.003, 0.009]	<0.001	<0.001	0.187	[0.002, 0.009]
		[48]	0.1	0.489	0.463	0.002	[0.460, 0.465]	[0.020, 0.033]	<0.001	<0.001	0.397	[0.018, 0.035]
		[49]	0.1	0.492	0.471	0.004	[0.466, 0.476]	[0.015, 0.026]	<0.001	<0.001	0.332	[0.010, 0.030]
		[50]	0.5	0.507	0.498	0.002	[0.495, 0.500]	[0.006, 0.013]	<0.001	<0.001	0.250	[0.005, 0.015]
	10	[46]	0.5	0.452	0.434	0.002	[0.431, 0.436]	[0.015, 0.022]	<0.001	<0.001	0.451	[0.012, 0.024]
		[47]	0.4	0.476	0.463	0.001	[0.462, 0.465]	[0.010, 0.015]	<0.001	<0.001	0.395	[0.008, 0.017]
		[48]	0.1	0.454	0.428	0.004	[0.424, 0.433]	[0.021, 0.031]	<0.001	<0.001	0.482	[0.019, 0.033]
		[49]	0.1	0.458	0.435	0.002	[0.432, 0.437]	[0.019, 0.028]	<0.001	<0.001	0.495	[0.016, 0.031]
		[50]	0.4	0.472	0.460	0.002	[0.457, 0.462]	[0.009, 0.015]	<0.001	<0.001	0.362	[0.007, 0.017]
Weekly	1	[46]	0.7	0.595	0.578	0.003	[0.574, 0.583]	[0.005, 0.028]	0.002	0.003	0.133	[0.004, 0.030]
		[47]	0.6	0.608	0.597	0.002	[0.594, 0.600]	[0.002, 0.019]	0.008	0.009	0.111	[0.002, 0.018]
		[48]	0.1	0.582	0.554	0.004	[0.549, 0.559]	[0.015, 0.042]	<0.001	<0.001	0.193	[0.014, 0.043]
		[49]	0.1	0.591	0.560	0.016	[0.540, 0.579]	[0.019, 0.043]	<0.001	<0.001	0.228	[0.016, 0.045]
		[50]	0.4	0.607	0.598	0.007	[0.589, 0.607]	[−0.002, 0.020]	0.050	0.052	0.076	[−0.002, 0.019]
	5	[46]	0.6	0.518	0.506	0.003	[0.502, 0.509]	[0.007, 0.017]	<0.001	<0.001	0.209	[0.006, 0.018]
		[47]	0.5	0.537	0.532	0.003	[0.528, 0.536]	[0.001, 0.009]	0.003	0.003	0.127	[0.001, 0.010]
		[48]	0.1	0.522	0.495	0.003	[0.491, 0.499]	[0.020, 0.034]	<0.001	<0.001	0.355	[0.019, 0.035]
		[49]	0.1	0.523	0.502	0.004	[0.497, 0.508]	[0.014, 0.026]	<0.001	<0.001	0.297	[0.010, 0.030]
		[50]	0.9	0.531	0.532	0.002	[0.529, 0.535]	[−0.003, 0.002]	0.632	0.632	−0.016	[−0.004, 0.003]
	10	[46]	0.4	0.489	0.464	0.002	[0.462, 0.466]	[0.020, 0.030]	<0.001	<0.001	0.485	[0.016, 0.034]
		[47]	0.4	0.506	0.496	0.002	[0.494, 0.498]	[0.007, 0.013]	<0.001	<0.001	0.281	[0.005, 0.015]
		[48]	0.1	0.486	0.463	0.004	[0.457, 0.468]	[0.018, 0.029]	<0.001	<0.001	0.396	[0.017, 0.030]
		[49]	0.1	0.491	0.467	0.003	[0.463, 0.471]	[0.019, 0.029]	<0.001	<0.001	0.438	[0.016, 0.032]
		[50]	0.4	0.505	0.494	0.002	[0.492, 0.496]	[0.007, 0.014]	<0.001	<0.001	0.271	[0.006, 0.015]

Table 12. Comparison between the proposed method and the full-text baseline for stock-return correlation.

Freq.	Top-K	Embedding Model	r^*	s_P	s_C	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Daily	1	[46]	0.7	0.566	0.561	[−0.006, 0.017]	0.178	0.205	0.043	[−0.006, 0.017]
		[47]	0.6	0.582	0.572	[0.001, 0.019]	0.012	0.017	0.105	[0.002, 0.018]
		[48]	0.1	0.550	0.433	[0.097, 0.135]	<0.001	<0.001	0.557	[0.094, 0.139]
		[49]	0.1	0.555	0.480	[0.057, 0.093]	<0.001	<0.001	0.370	[0.051, 0.099]
		[50]	0.9	0.587	0.584	[−0.003, 0.009]	0.190	0.211	0.041	[−0.003, 0.009]
	5	[46]	0.6	0.487	0.482	[−0.001, 0.010]	0.040	0.052	0.081	[−0.002, 0.011]
		[47]	0.7	0.508	0.504	[0.001, 0.007]	0.010	0.016	0.107	[0.000, 0.008]
		[48]	0.1	0.489	0.389	[0.088, 0.112]	<0.001	<0.001	0.785	[0.080, 0.120]
		[49]	0.1	0.492	0.426	[0.056, 0.076]	<0.001	<0.001	0.589	[0.051, 0.080]
		[50]	0.5	0.507	0.504	[−0.001, 0.008]	0.055	0.068	0.074	[−0.002, 0.009]
	10	[46]	0.5	0.452	0.445	[0.002, 0.012]	0.002	0.003	0.137	[0.001, 0.012]
		[47]	0.4	0.476	0.468	[0.005, 0.011]	<0.001	<0.001	0.251	[0.004, 0.013]
		[48]	0.1	0.454	0.367	[0.077, 0.097]	<0.001	<0.001	0.788	[0.067, 0.109]
		[49]	0.1	0.458	0.395	[0.054, 0.072]	<0.001	<0.001	0.647	[0.049, 0.077]
		[50]	0.4	0.472	0.466	[0.002, 0.009]	<0.001	<0.001	0.157	[0.001, 0.010]
Weekly	1	[46]	0.7	0.595	0.589	[−0.008, 0.020]	0.198	0.212	0.039	[−0.007, 0.019]
		[47]	0.6	0.608	0.595	[0.002, 0.023]	0.007	0.012	0.113	[0.004, 0.021]
		[48]	0.1	0.582	0.472	[0.089, 0.132]	<0.001	<0.001	0.463	[0.087, 0.135]
		[49]	0.1	0.591	0.512	[0.058, 0.099]	<0.001	<0.001	0.349	[0.054, 0.104]
		[50]	0.4	0.607	0.612	[−0.017, 0.007]	0.800	0.800	−0.039	[−0.016, 0.006]
	5	[46]	0.6	0.518	0.514	[−0.003, 0.010]	0.128	0.153	0.053	[−0.004, 0.011]
		[47]	0.5	0.537	0.533	[0.000, 0.009]	0.025	0.034	0.091	[−0.001, 0.010]
		[48]	0.1	0.522	0.430	[0.080, 0.104]	<0.001	<0.001	0.671	[0.073, 0.111]
		[49]	0.1	0.523	0.464	[0.048, 0.070]	<0.001	<0.001	0.488	[0.044, 0.074]
		[50]	0.9	0.531	0.532	[−0.004, 0.002]	0.771	0.798	−0.034	[−0.005, 0.002]
	10	[46]	0.4	0.489	0.480	[0.004, 0.014]	<0.001	<0.001	0.162	[0.001, 0.017]
		[47]	0.4	0.506	0.499	[0.003, 0.010]	<0.001	<0.001	0.182	[0.002, 0.012]
		[48]	0.1	0.486	0.406	[0.070, 0.091]	<0.001	<0.001	0.687	[0.061, 0.101]
		[49]	0.1	0.491	0.433	[0.049, 0.068]	<0.001	<0.001	0.573	[0.045, 0.072]
		[50]	0.4	0.505	0.497	[0.004, 0.011]	<0.001	<0.001	0.181	[0.003, 0.011]

Table 13. Comparison between the proposed method and alternative sentence-selection methods for stock-return correlation.

Freq.	Top-K	Method	r^*	s	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Daily	1	Proposed	0.6	0.582	–	–	–	–	–
		Entropy	–	0.582	[−0.007, 0.007]	0.551	0.582	−0.006	[−0.007, 0.006]
		Sentence Length	–	0.570	[0.003, 0.022]	0.004	0.009	0.122	[0.003, 0.022]
		[45]	–	0.573	[0.000, 0.019]	0.021	0.034	0.095	[−0.001, 0.019]
	5	Proposed	0.7	0.508	–	–	–	–	–
		Entropy	–	0.507	[−0.002, 0.003]	0.402	0.482	0.012	[−0.003, 0.003]
		Sentence Length	–	0.502	[0.002, 0.009]	0.001	0.003	0.143	[0.001, 0.010]
		[45]	–	0.503	[0.001, 0.008]	0.004	0.009	0.122	[0.001, 0.008]
	10	Proposed	0.4	0.476	–	–	–	–	–
		Entropy	–	0.472	[0.001, 0.007]	0.002	0.004	0.138	[0.001, 0.007]
		Sentence Length	–	0.463	[0.010, 0.017]	<0.001	<0.001	0.350	[0.007, 0.019]
		[45]	–	0.466	[0.007, 0.014]	<0.001	<0.001	0.282	[0.005, 0.016]
Weekly	1	Proposed	0.6	0.608	–	–	–	–	–
		Entropy	–	0.607	[−0.007, 0.008]	0.432	0.486	0.008	[−0.006, 0.008]
		Sentence Length	–	0.596	[0.001, 0.022]	0.017	0.031	0.098	[0.001, 0.022]
		[45]	–	0.597	[0.000, 0.021]	0.022	0.034	0.093	[−0.001, 0.021]
	5	Proposed	0.5	0.537	–	–	–	–	–
		Entropy	–	0.538	[−0.004, 0.003]	0.582	0.582	−0.010	[−0.005, 0.004]
		Sentence Length	–	0.528	[0.005, 0.014]	<0.001	<0.001	0.184	[0.004, 0.014]
		[45]	–	0.533	[−0.001, 0.009]	0.050	0.065	0.076	[−0.002, 0.011]
	10	Proposed	0.4	0.506	–	–	–	–	–
		Entropy	–	0.503	[0.000, 0.006]	0.043	0.059	0.080	[−0.001, 0.006]
		Sentence Length	–	0.494	[0.008, 0.015]	<0.001	<0.001	0.285	[0.006, 0.017]
		[45]	–	0.498	[0.005, 0.012]	<0.001	<0.001	0.207	[0.003, 0.014]

4.3.4. Validation Results for Financial Homogeneity

Tables 14–16 report financial homogeneity using adjusted R^2 for Enterprise Value-to-Sales (EVS), Asset Turnover (AT), Return on Assets (ROA), and Revenue Growth. These variables represent valuation, operating efficiency, profitability, and growth. A higher adjusted R^2 means that the selected peers more closely match the target firm's financial characteristics.

Table 14. Comparison between the proposed method and matched random deletion for financial homogeneity.

Financial Variable	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI
EVS	1	[46]	0.4	0.305	0.226	0.049	[0.164, 0.287]
		[47]	0.7	0.320	0.294	0.011	[0.280, 0.308]
		[48]	0.1	0.280	0.218	0.039	[0.169, 0.267]
		[49]	0.1	0.325	0.236	0.013	[0.220, 0.251]
		[50]	0.4	0.310	0.258	0.018	[0.235, 0.280]
	5	[46]	0.4	0.373	0.327	0.029	[0.290, 0.363]
		[47]	0.9	0.393	0.383	0.007	[0.374, 0.392]
		[48]	0.1	0.376	0.295	0.026	[0.264, 0.327]
		[49]	0.1	0.382	0.308	0.009	[0.297, 0.319]
		[50]	0.2	0.434	0.373	0.016	[0.353, 0.392]
	10	[46]	0.9	0.385	0.388	0.012	[0.373, 0.404]
		[47]	0.2	0.408	0.360	0.014	[0.343, 0.377]
		[48]	0.1	0.357	0.299	0.016	[0.279, 0.319]
		[49]	0.1	0.371	0.329	0.016	[0.309, 0.349]
		[50]	0.5	0.442	0.426	0.024	[0.397, 0.456]

Table 14. *Cont.*

Financial Variable	Top-K	Embedding Model	r^*	s_P	s_C	RD SD	RD 95% CI
AT	1	[46]	0.8	0.424	0.378	0.014	[0.360, 0.396]
		[47]	0.3	0.445	0.391	0.035	[0.347, 0.434]
		[48]	0.1	0.403	0.317	0.018	[0.294, 0.339]
		[49]	0.1	0.396	0.321	0.036	[0.277, 0.366]
		[50]	0.7	0.481	0.428	0.022	[0.400, 0.456]
	5	[46]	0.7	0.551	0.486	0.009	[0.475, 0.496]
		[47]	0.8	0.527	0.515	0.012	[0.500, 0.530]
		[48]	0.1	0.564	0.448	0.026	[0.416, 0.480]
		[49]	0.1	0.553	0.428	0.024	[0.398, 0.458]
		[50]	0.4	0.537	0.532	0.020	[0.508, 0.557]
	10	[46]	0.9	0.503	0.496	0.017	[0.475, 0.517]
		[47]	0.6	0.507	0.495	0.016	[0.475, 0.514]
		[48]	0.1	0.535	0.438	0.012	[0.422, 0.453]
		[49]	0.1	0.530	0.422	0.013	[0.406, 0.438]
		[50]	0.1	0.517	0.491	0.008	[0.481, 0.501]
ROA	1	[46]	0.9	0.105	0.096	0.028	[0.061, 0.131]
		[47]	0.7	0.139	0.093	0.018	[0.071, 0.115]
		[48]	0.1	0.130	0.098	0.040	[0.049, 0.148]
		[49]	0.3	0.092	0.074	0.020	[0.049, 0.099]
		[50]	0.9	0.179	0.155	0.007	[0.146, 0.163]
	5	[46]	0.6	0.215	0.164	0.009	[0.153, 0.176]
		[47]	0.6	0.228	0.198	0.008	[0.188, 0.209]
		[48]	0.2	0.165	0.097	0.025	[0.066, 0.127]
		[49]	0.1	0.198	0.171	0.027	[0.138, 0.205]
		[50]	0.7	0.164	0.187	0.023	[0.159, 0.216]
	10	[46]	0.4	0.187	0.175	0.023	[0.147, 0.204]
		[47]	0.2	0.222	0.185	0.005	[0.179, 0.191]
		[48]	0.1	0.191	0.155	0.023	[0.126, 0.183]
		[49]	0.1	0.195	0.181	0.013	[0.164, 0.197]
		[50]	0.3	0.205	0.183	0.014	[0.166, 0.200]
Revenue Growth	1	[46]	0.2	0.275	0.227	0.038	[0.179, 0.274]
		[47]	0.3	0.295	0.309	0.020	[0.283, 0.334]
		[48]	0.1	0.256	0.187	0.051	[0.123, 0.251]
		[49]	0.1	0.239	0.251	0.021	[0.224, 0.277]
		[50]	0.7	0.403	0.372	0.023	[0.343, 0.400]
	5	[46]	0.2	0.405	0.358	0.020	[0.334, 0.383]
		[47]	0.7	0.435	0.424	0.005	[0.418, 0.429]
		[48]	0.1	0.368	0.334	0.025	[0.303, 0.366]
		[49]	0.4	0.319	0.219	0.025	[0.188, 0.250]
		[50]	0.1	0.438	0.409	0.017	[0.389, 0.430]
	10	[46]	0.7	0.419	0.408	0.021	[0.382, 0.435]
		[47]	0.7	0.431	0.422	0.007	[0.413, 0.430]
		[48]	0.1	0.379	0.335	0.032	[0.295, 0.375]
		[49]	0.1	0.366	0.366	0.010	[0.354, 0.379]
		[50]	0.8	0.427	0.419	0.008	[0.409, 0.429]

Table 15. Comparison between the proposed method and the full-text baseline for financial homogeneity.

Financial Variable	Top-K	Embedding Model	r^*	s_P	s_C
EVS	1	[46]	0.4	0.305	0.303
		[47]	0.7	0.320	0.288
		[48]	0.1	0.280	0.137
		[49]	0.1	0.325	0.151
		[50]	0.4	0.310	0.254

Table 15. *Cont.*

Financial Variable	Top-*K*	Embedding Model	r^*	s_P	s_C
EVS	5	[46]	0.4	0.373	0.383
		[47]	0.9	0.393	0.389
		[48]	0.1	0.376	0.111
		[49]	0.1	0.382	0.261
		[50]	0.2	0.434	0.420
	10	[46]	0.9	0.385	0.381
		[47]	0.2	0.408	0.378
		[48]	0.1	0.357	0.083
		[49]	0.1	0.371	0.252
		[50]	0.5	0.442	0.436
AT	1	[46]	0.8	0.424	0.378
		[47]	0.3	0.445	0.427
		[48]	0.1	0.403	0.158
		[49]	0.1	0.396	0.247
		[50]	0.7	0.481	0.466
	5	[46]	0.7	0.551	0.521
		[47]	0.8	0.527	0.497
		[48]	0.1	0.564	0.195
		[49]	0.1	0.553	0.338
		[50]	0.4	0.537	0.564
	10	[46]	0.9	0.503	0.516
		[47]	0.6	0.507	0.501
		[48]	0.1	0.535	0.232
		[49]	0.1	0.530	0.321
		[50]	0.1	0.517	0.543
ROA	1	[46]	0.9	0.105	0.108
		[47]	0.7	0.139	0.093
		[48]	0.1	0.130	0.081
		[49]	0.3	0.092	0.121
		[50]	0.9	0.179	0.141
	5	[46]	0.6	0.215	0.166
		[47]	0.6	0.228	0.208
		[48]	0.2	0.165	0.074
		[49]	0.1	0.198	0.176
		[50]	0.7	0.164	0.205
	10	[46]	0.4	0.187	0.189
		[47]	0.2	0.222	0.184
		[48]	0.1	0.191	0.113
		[49]	0.1	0.195	0.179
		[50]	0.3	0.205	0.177
Revenue Growth	1	[46]	0.2	0.275	0.373
		[47]	0.3	0.295	0.321
		[48]	0.1	0.256	0.140
		[49]	0.1	0.239	0.209
		[50]	0.7	0.403	0.365
	5	[46]	0.2	0.405	0.420
		[47]	0.7	0.435	0.420
		[48]	0.1	0.368	0.261
		[49]	0.4	0.319	0.312
		[50]	0.1	0.438	0.447
	10	[46]	0.7	0.419	0.397
		[47]	0.7	0.431	0.421
		[48]	0.1	0.379	0.284
		[49]	0.1	0.366	0.311
		[50]	0.8	0.427	0.424

Compared with matched random deletion (Table 14), the proposed method has higher adjusted R^2 in 55 of 60 comparisons. Table 15 shows the comparison with full text. The

proposed method has higher adjusted R^2 in 48 of 60 comparisons. The proposed method therefore shows higher financial homogeneity in a majority of settings, although the pattern is less uniform than for GICS alignment or return correlation.

Table 16. Comparison between the proposed method and alternative sentence-selection methods for financial homogeneity.

Financial Variable	Top-*K*	Method	*r**	*s*
EVS	1	Proposed	0.7	0.320
		Entropy	–	0.315
		Sentence Length	–	0.296
		[45]	–	0.290
	5	Proposed	0.9	0.393
		Entropy	–	0.385
		Sentence Length	–	0.395
		[45]	–	0.386
	10	Proposed	0.2	0.408
		Entropy	–	0.375
		Sentence Length	–	0.344
		[45]	–	0.346
AT	1	Proposed	0.3	0.445
		Entropy	–	0.442
		Sentence Length	–	0.414
		[45]	–	0.364
	5	Proposed	0.8	0.527
		Entropy	–	0.540
		Sentence Length	–	0.511
		[45]	–	0.536
	10	Proposed	0.6	0.507
		Entropy	–	0.514
		Sentence Length	–	0.496
		[45]	–	0.485
ROA	1	Proposed	0.7	0.139
		Entropy	–	0.119
		Sentence Length	–	0.106
		[45]	–	0.116
	5	Proposed	0.6	0.228
		Entropy	–	0.212
		Sentence Length	–	0.173
		[45]	–	0.206
	10	Proposed	0.2	0.222
		Entropy	–	0.196
		Sentence Length	–	0.176
		[45]	–	0.173
Revenue Growth	1	Proposed	0.3	0.295
		Entropy	–	0.282
		Sentence Length	–	0.327
		[45]	–	0.357
	5	Proposed	0.7	0.435
		Entropy	–	0.425
		Sentence Length	–	0.428
		[45]	–	0.421
	10	Proposed	0.7	0.431
		Entropy	–	0.421
		Sentence Length	–	0.409
		[45]	–	0.416

In Table 16, the proposed E5-Mistral ranking has the highest adjusted R^2 in 8 of the 12 financial-variable–Top-*K* settings. The simpler criteria remain competitive for several financial dimensions, unlike the much clearer separation observed for GICS alignment. This

suggests that lexical salience, sentence length, and classifier confidence can each preserve some financially relevant information, while the actual-Sector-probability criterion provides the strongest result more often across the four dimensions.

4.4. Comparative Studies

This section compares the performance of the proposed framework with existing text-based peer-selection methods. In the comparative analysis, the proposed framework is implemented using E5-Mistral as the embedding backbone, based on its consistently strong performance in the validation studies. The benchmark approaches for the peer group selection are described in Section 4.4.1. The corresponding comparison results are presented in Section 4.4.2, Section 4.4.3 and Section 4.4.4 respectively.

4.4.1. Benchmark Methods

For the peer-selection task, several methods mentioned in Section 2.2 are utilized. The detailed explanations of the benchmark methods are provided as follows.

- [7]: Peers are selected by embedding texts using BERT and using cosine similarity.
- [12]: Peers are identified by embedding textual data with Word2Vec and utilizing cosine similarity.
- [11]: Peers are determined by embedding texts using a fine-tuned BERT model and computing cosine similarity.
- [1]: Peers are selected by embedding texts using bag-of-words and computing cosine similarity.
- [8]: Peers are determined by embedding texts using SBERT and performing cosine similarity.

The proposed and comparative methods are evaluated using the same firms, 2022 Item 1. Business sections, and cosine-similarity-based Top-K peer-selection procedure.

4.4.2. Comparative Results for GICS Alignment

Table 17 compares the E5-Mistral-based proposed framework with five existing text-based peer-selection methods. The proposed method has the highest GICS alignment in all 12 hierarchy-level–Top-*K* settings.

Table 17. Comparison between the proposed method and existing peer-selection methods for GICS alignment.

GICS Level	Top-*K*	Method	r^*	s	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Sector	1	Proposed	0.2	0.933	–	–	–	–	–
		[1]	–	0.871	[0.033, 0.091]	<0.001	<0.001	0.195	[0.030, 0.098]
		[12]	–	0.833	[0.069, 0.133]	<0.001	<0.001	0.290	[0.065, 0.142]
		[11]	–	0.818	[0.081, 0.151]	<0.001	<0.001	0.299	[0.068, 0.170]
		[7]	–	0.468	[0.416, 0.515]	<0.001	<0.001	0.855	[0.415, 0.520]
		[8]	–	0.871	[0.031, 0.093]	<0.001	<0.001	0.184	[0.030, 0.097]
	5	Proposed	0.2	0.867	–	–	–	–	–
		[1]	–	0.777	[0.070, 0.109]	<0.001	<0.001	0.421	[0.060, 0.124]
		[12]	–	0.742	[0.104, 0.146]	<0.001	<0.001	0.542	[0.091, 0.163]
		[11]	–	0.779	[0.064, 0.113]	<0.001	<0.001	0.328	[0.056, 0.124]
		[7]	–	0.372	[0.468, 0.522]	<0.001	<0.001	1.670	[0.456, 0.536]
		[8]	–	0.784	[0.063, 0.103]	<0.001	<0.001	0.372	[0.054, 0.115]
	10	Proposed	0.1	0.822	–	–	–	–	–
		[1]	–	0.712	[0.091, 0.128]	<0.001	<0.001	0.538	[0.079, 0.145]
		[12]	–	0.695	[0.108, 0.146]	<0.001	<0.001	0.607	[0.092, 0.163]
		[11]	–	0.737	[0.062, 0.109]	<0.001	<0.001	0.329	[0.046, 0.124]
		[7]	–	0.324	[0.474, 0.522]	<0.001	<0.001	1.916	[0.456, 0.539]
		[8]	–	0.730	[0.074, 0.111]	<0.001	<0.001	0.454	[0.064, 0.123]

Table 17. *Cont.*

GICS Level	Top-K	Method	r^*	s	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Industry Group	1	Proposed	0.2	0.871	–	–	–	–	–
		[1]	–	0.796	[0.039, 0.112]	<0.001	<0.001	0.188	[0.034, 0.122]
		[12]	–	0.742	[0.089, 0.169]	<0.001	<0.001	0.292	[0.086, 0.176]
		[11]	–	0.727	[0.100, 0.188]	<0.001	<0.001	0.296	[0.086, 0.208]
		[7]	–	0.322	[0.500, 0.599]	<0.001	<0.001	1.003	[0.494, 0.605]
		[8]	–	0.800	[0.032, 0.109]	<0.001	<0.001	0.168	[0.022, 0.121]
	5	Proposed	0.2	0.752	–	–	–	–	–
		[1]	–	0.669	[0.062, 0.105]	<0.001	<0.001	0.355	[0.048, 0.122]
		[12]	–	0.620	[0.108, 0.156]	<0.001	<0.001	0.501	[0.094, 0.171]
		[11]	–	0.646	[0.078, 0.134]	<0.001	<0.001	0.345	[0.066, 0.148]
		[7]	–	0.250	[0.473, 0.531]	<0.001	<0.001	1.578	[0.452, 0.550]
		[8]	–	0.678	[0.052, 0.097]	<0.001	<0.001	0.299	[0.037, 0.112]
	10	Proposed	0.1	0.674	–	–	–	–	–
		[1]	–	0.577	[0.078, 0.117]	<0.001	<0.001	0.461	[0.061, 0.137]
		[12]	–	0.551	[0.104, 0.144]	<0.001	<0.001	0.562	[0.085, 0.162]
		[11]	–	0.589	[0.061, 0.111]	<0.001	<0.001	0.311	[0.042, 0.131]
		[7]	–	0.211	[0.439, 0.488]	<0.001	<0.001	1.712	[0.413, 0.512]
		[8]	–	0.601	[0.054, 0.093]	<0.001	<0.001	0.347	[0.037, 0.111]
Industry	1	Proposed	0.6	0.745	–	–	–	–	–
		[1]	–	0.685	[0.021, 0.099]	0.001	0.001	0.141	[0.013, 0.108]
		[12]	–	0.582	[0.121, 0.205]	<0.001	<0.001	0.352	[0.126, 0.206]
		[11]	–	0.571	[0.127, 0.221]	<0.001	<0.001	0.338	[0.124, 0.226]
		[7]	–	0.208	[0.486, 0.587]	<0.001	<0.001	0.964	[0.464, 0.599]
		[8]	–	0.665	[0.039, 0.119]	<0.001	<0.001	0.180	[0.034, 0.123]
	5	Proposed	0.4	0.562	–	–	–	–	–
		[1]	–	0.513	[0.029, 0.068]	<0.001	<0.001	0.229	[0.019, 0.080]
		[12]	–	0.443	[0.100, 0.138]	<0.001	<0.001	0.562	[0.087, 0.150]
		[11]	–	0.476	[0.061, 0.111]	<0.001	<0.001	0.316	[0.049, 0.123]
		[7]	–	0.151	[0.383, 0.439]	<0.001	<0.001	1.323	[0.350, 0.465]
		[8]	–	0.514	[0.027, 0.069]	<0.001	<0.001	0.211	[0.014, 0.079]
	10	Proposed	0.2	0.469	–	–	–	–	–
		[1]	–	0.409	[0.046, 0.075]	<0.001	<0.001	0.386	[0.035, 0.086]
		[12]	–	0.369	[0.084, 0.117]	<0.001	<0.001	0.556	[0.067, 0.131]
		[11]	–	0.405	[0.042, 0.086]	<0.001	<0.001	0.270	[0.025, 0.103]
		[7]	–	0.120	[0.326, 0.372]	<0.001	<0.001	1.376	[0.289, 0.402]
		[8]	–	0.419	[0.034, 0.067]	<0.001	<0.001	0.280	[0.014, 0.086]
Sub-Industry	1	Proposed	0.7	0.655	–	–	–	–	–
		[1]	–	0.558	[0.053, 0.140]	<0.001	<0.001	0.204	[0.045, 0.150]
		[12]	–	0.494	[0.116, 0.206]	<0.001	<0.001	0.327	[0.123, 0.199]
		[11]	–	0.442	[0.161, 0.263]	<0.001	<0.001	0.379	[0.155, 0.271]
		[7]	–	0.150	[0.455, 0.554]	<0.001	<0.001	0.924	[0.440, 0.560]
		[8]	–	0.554	[0.057, 0.145]	<0.001	<0.001	0.209	[0.061, 0.140]
	5	Proposed	0.3	0.427	–	–	–	–	–
		[1]	–	0.379	[0.030, 0.067]	<0.001	<0.001	0.237	[0.019, 0.076]
		[12]	–	0.333	[0.077, 0.112]	<0.001	<0.001	0.486	[0.065, 0.123]
		[11]	–	0.329	[0.076, 0.120]	<0.001	<0.001	0.407	[0.060, 0.136]
		[7]	–	0.097	[0.303, 0.357]	<0.001	<0.001	1.112	[0.276, 0.378]
		[8]	–	0.380	[0.028, 0.067]	<0.001	<0.001	0.222	[0.015, 0.078]
	10	Proposed	0.3	0.315	–	–	–	–	–
		[1]	–	0.273	[0.030, 0.052]	<0.001	<0.001	0.343	[0.020, 0.063]
		[12]	–	0.249	[0.055, 0.077]	<0.001	<0.001	0.546	[0.045, 0.088]
		[11]	–	0.259	[0.041, 0.071]	<0.001	<0.001	0.346	[0.028, 0.084]
		[7]	–	0.076	[0.219, 0.258]	<0.001	<0.001	1.124	[0.197, 0.279]
		[8]	–	0.282	[0.020, 0.046]	<0.001	<0.001	0.238	[0.008, 0.058]

Across the five benchmark methods, the proposed method achieves higher GICS alignment in all 60 pairwise comparisons, and every difference remains significant after BH correction ($p_{BH} \leq 0.001$). The paired-difference 95% CIs and industry-level cluster-bootstrap 95% CIs are above zero in all comparisons. Cohen's d_z values range from approximately 0.14 to 1.92, representing effect sizes from small to very large depending on the benchmark method and experimental setting. The effect sizes generally become more pronounced at larger Top-K values. The advantage is observed not only at the Sector level used for sentence supervision but also at the Industry Group, Industry, and Sub-Industry levels. This finding suggests that the reconstructed reports retain information useful for distinguishing

firms within progressively finer GICS categories, even though these finer-grained labels are not directly used for sentence supervision.

4.4.3. Comparative Results for Stock-Return Correlation

Table 18 compares the proposed method with the five benchmark peer-selection methods using daily and weekly returns. The proposed method has the highest correlation in all six return-frequency–Top-K settings.

Table 18. Comparison between the proposed method and existing peer-selection methods for stock-return correlation.

Freq.	Top-K	Method	r^*	s	Paired-t 95% CI	p	p_{BH}	Cohen's d_z	Bootstrap 95% CI
Daily	1	Proposed	0.6	0.582	–	–	–	–	–
		[1]	–	0.562	[0.008, 0.033]	<0.001	<0.001	0.155	[0.007, 0.033]
		[12]	–	0.537	[0.032, 0.057]	<0.001	<0.001	0.322	[0.028, 0.061]
		[11]	–	0.508	[0.058, 0.090]	<0.001	<0.001	0.424	[0.054, 0.095]
		[7]	–	0.376	[0.185, 0.227]	<0.001	<0.001	0.889	[0.175, 0.239]
		[8]	–	0.552	[0.016, 0.044]	<0.001	<0.001	0.193	[0.013, 0.048]
	5	Proposed	0.7	0.508	–	–	–	–	–
		[1]	–	0.488	[0.014, 0.025]	<0.001	<0.001	0.303	[0.011, 0.028]
		[12]	–	0.479	[0.023, 0.034]	<0.001	<0.001	0.460	[0.021, 0.036]
		[11]	–	0.470	[0.030, 0.046]	<0.001	<0.001	0.443	[0.027, 0.049]
		[7]	–	0.346	[0.150, 0.173]	<0.001	<0.001	1.274	[0.141, 0.181]
		[8]	–	0.489	[0.012, 0.025]	<0.001	<0.001	0.264	[0.011, 0.026]
	10	Proposed	0.4	0.476	–	–	–	–	–
		[1]	–	0.449	[0.022, 0.032]	<0.001	<0.001	0.505	[0.020, 0.035]
		[12]	–	0.447	[0.024, 0.034]	<0.001	<0.001	0.505	[0.020, 0.037]
		[11]	–	0.443	[0.027, 0.040]	<0.001	<0.001	0.482	[0.023, 0.043]
		[7]	–	0.331	[0.135, 0.155]	<0.001	<0.001	1.346	[0.123, 0.167]
		[8]	–	0.456	[0.015, 0.024]	<0.001	<0.001	0.365	[0.013, 0.026]
Weekly	1	Proposed	0.6	0.608	–	–	–	–	–
		[1]	–	0.596	[−0.003, 0.025]	0.059	0.059	0.073	[−0.004, 0.027]
		[12]	–	0.563	[0.030, 0.060]	<0.001	<0.001	0.272	[0.028, 0.063]
		[11]	–	0.549	[0.040, 0.076]	<0.001	<0.001	0.295	[0.038, 0.079]
		[7]	–	0.415	[0.169, 0.215]	<0.001	<0.001	0.751	[0.162, 0.225]
		[8]	–	0.574	[0.017, 0.050]	<0.001	<0.001	0.182	[0.015, 0.053]
	5	Proposed	0.5	0.537	–	–	–	–	–
		[1]	–	0.519	[0.012, 0.025]	<0.001	<0.001	0.248	[0.009, 0.029]
		[12]	–	0.510	[0.020, 0.034]	<0.001	<0.001	0.359	[0.019, 0.036]
		[11]	–	0.502	[0.027, 0.044]	<0.001	<0.001	0.370	[0.021, 0.051]
		[7]	–	0.384	[0.141, 0.166]	<0.001	<0.001	1.136	[0.135, 0.172]
		[8]	–	0.513	[0.018, 0.032]	<0.001	<0.001	0.317	[0.016, 0.033]
	10	Proposed	0.4	0.506	–	–	–	–	–
		[1]	–	0.481	[0.019, 0.030]	<0.001	<0.001	0.400	[0.017, 0.033]
		[12]	–	0.481	[0.019, 0.031]	<0.001	<0.001	0.401	[0.017, 0.034]
		[11]	–	0.475	[0.023, 0.038]	<0.001	<0.001	0.394	[0.019, 0.042]
		[7]	–	0.368	[0.128, 0.149]	<0.001	<0.001	1.192	[0.117, 0.160]
		[8]	–	0.485	[0.016, 0.027]	<0.001	<0.001	0.343	[0.013, 0.030]

The proposed method achieves higher return correlations than the benchmark methods in all 30 daily and weekly pairwise comparisons, and 29 remain significant after BH correction. The paired-difference and industry-level cluster-bootstrap 95% CIs are above zero in the same 29 comparisons. Cohen's d_z values range from 0.073 to 1.346, indicating effects ranging from small to very large, and increase with Top-K for every benchmark method under both return frequencies. This pattern indicates that the relative advantage of the proposed method becomes more pronounced when a broader peer set is considered. The same general result is obtained with daily and weekly returns, indicating that the comparative finding is not driven by a particular return frequency. Because these benchmarks

use alternative text representations and peer-selection procedures, the higher correlations further suggest that the reconstructed reports identify peers whose similarity is reflected in observed market co-movement, rather than only in GICS membership.

4.4.4. Comparative Results for Financial Homogeneity

Table 19 compares financial homogeneity between the proposed method and the five benchmark peer-selection approaches. The proposed method has the highest adjusted R^2 in 10 of the 12 financial-variable–Top-K settings.

Table 19. Comparison between the proposed method and existing peer-selection methods for financial homogeneity.

Financial Variable	Top-K	Method	r^*	s
EVS	1	Proposed	0.7	0.320
		[1]	–	0.251
		[12]	–	0.289
		[11]	–	0.129
		[7]	–	0.003
		[8]	–	0.219
	5	Proposed	0.9	0.393
		[1]	–	0.346
		[12]	–	0.337
		[11]	–	0.261
		[7]	–	0.041
		[8]	–	0.313
	10	Proposed	0.2	0.408
		[1]	–	0.342
		[12]	–	0.333
		[11]	–	0.284
		[7]	–	0.095
		[8]	–	0.322
AT	1	Proposed	0.3	0.445
		[1]	–	0.327
		[12]	–	0.389
		[11]	–	0.339
		[7]	–	0.072
		[8]	–	0.350
	5	Proposed	0.8	0.527
		[1]	–	0.499
		[12]	–	0.493
		[11]	–	0.462
		[7]	–	0.117
		[8]	–	0.447
	10	Proposed	0.6	0.507
		[1]	–	0.514
		[12]	–	0.440
		[11]	–	0.443
		[7]	–	0.153
		[8]	–	0.465
ROA	1	Proposed	0.7	0.139
		[1]	–	0.100
		[12]	–	0.090
		[11]	–	0.021
		[7]	–	0.013
		[8]	–	0.118

Table 19. *Cont.*

Financial Variable	Top-*K*	Method	r^*	*s*
ROA	5	Proposed	0.6	0.228
		[1]	–	0.192
		[12]	–	0.181
		[11]	–	0.101
		[7]	–	0.075
		[8]	–	0.167
	10	Proposed	0.2	0.222
		[1]	–	0.177
		[12]	–	0.191
		[11]	–	0.122
		[7]	–	0.097
		[8]	–	0.184
Revenue Growth	1	Proposed	0.3	0.295
		[1]	–	0.313
		[12]	–	0.259
		[11]	–	0.221
		[7]	–	0.010
		[8]	–	0.238
	5	Proposed	0.7	0.435
		[1]	–	0.381
		[12]	–	0.405
		[11]	–	0.301
		[7]	–	0.048
		[8]	–	0.365
	10	Proposed	0.7	0.431
		[1]	–	0.369
		[12]	–	0.401
		[11]	–	0.325
		[7]	–	0.066
		[8]	–	0.342

The proposed method ranks first in all EVS and ROA settings and in most AT and Revenue Growth settings. Although the advantage is less consistent than that observed for GICS alignment and return correlation, the proposed method achieves the highest performance in 10 of the 12 settings. These results indicate that peer firms identified from the reconstructed reports exhibit relatively greater financial homogeneity in terms of valuation, operating efficiency, profitability, and growth characteristics.

5. Conclusions and Future Work

This study proposed a sentence-level industry-discriminative reconstruction framework for peer selection using 10-K annual reports. The proposed framework prioritizes sentences related to firms' industry and business characteristics based on sentence-level industry-discriminative scores derived from GICS Sector supervision, reconstructs the selected sentences in their original order, and subsequently generates firm-level document representations. Because sentence selection and text reconstruction are performed before document embedding, the same reconstructed text can be applied across different pretrained embedding models.

Across five embedding models and multiple peer group settings, the proposed framework generally achieved higher GICS alignment and stock-return correlation than the full-text baseline, matched random deletion, alternative sentence-selection criteria, and existing peer-selection methods, while also showing greater financial homogeneity in most settings. In particular, comparisons with matched random deletion, which retains the same amount of text, show that the observed performance differences cannot be ex-

plained solely by document-length reduction. However, because GICS Sector labels are used as the supervision signal for sentence selection, GICS alignment should not be interpreted as independent external validation of the proposed method. Rather, it serves as a diagnostic evaluation of the extent to which the reconstructed representations retain the industry-related information targeted by the framework. Stock-return similarity and financial homogeneity provide complementary evidence based on outcomes that are not directly used for sentence scoring, although these observational relationships should not be interpreted as causal effects of sentence selection.

The findings should also be interpreted in light of several limitations. The final evaluation is based on U.S. S&P 500 firms over a limited set of reporting and outcome years, and the results therefore do not establish generalizability to firms of different sizes, firms in other countries, or substantially different market environments. In addition, assigning a firm-level GICS Sector label to each sentence provides weak sentence-level supervision and may not fully capture secondary business segments of diversified firms. Reconstruction ratios are also selected from validation data for predefined task-specific evaluation objectives, and a ratio selected for one objective may not generalize equally well to other analytical objectives. Some of the evaluated embedding models are commercial models, and the observed results may therefore depend in part on the characteristics of the embedding models used.

Future research may evaluate the generalizability of the proposed framework across firms with a broader range of sizes and characteristics, across different countries, and over longer time periods. Future work may also assess the practical utility of the framework in downstream financial applications such as pairs trading and portfolio construction and incorporate multimodal information beyond textual disclosures into firm representations. For explanation generation, future research may move beyond the direct use of selected report content by developing a more systematic and structured RAG framework that explicitly defines the retrieval corpus, retrieval strategy, and context-selection mechanism.

Author Contributions: Conceptualization, S.J., T.K. and J.C.; methodology, S.J., T.K., J.P., J.Y. and J.C.; formal analysis, S.J. and T.K.; investigation, J.P. and J.Y.; resources, J.C.; data curation, H.B. and J.C.; writing—original draft preparation, S.J. and J.C.; writing—review and editing, S.J. and J.C.; visualization, S.J. and T.K.; supervision, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the research fund of Hanyang University (HY-202500000001271).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data underlying the results presented in this study were obtained from the following existing data sources: Form 10-K filings from the U.S. Securities and Exchange Commission's EDGAR database, available at https://www.sec.gov/; financial statement data and stock price data from Financial Modeling Prep (FMP), available at https://site.financialmodelingprep.com/ (accessed on 16 December 2025); and GICS classification data from Wharton Research Data Services (WRDS), available at https://wrds-www.wharton.upenn.edu/ (accessed on 10 September 2024). Restrictions may apply to the redistribution of third-party data obtained from FMP and WRDS.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Structure of the 10-K Report

Table A1 presents the standard structure of Form 10-K, which organizes a company's annual report into four main parts. Each part contains specific disclosure items, such as business overview, risk factors, financial statements, and executive compensation.

Table A1. Standard structure of Form 10-K.

Part	Item
Part I	Item 1. Business Item 1A. Risk Factors Item 1B. Unresolved Staff Comments Item 1C. Cybersecurity Item 2. Properties Item 3. Legal Proceedings Item 4. Mine Safety Disclosures
Part II	Item 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities Item 6. [Reserved] Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations Item 7A. Quantitative and Qualitative Disclosures About Market Risk Item 8. Financial Statements and Supplementary Data Item 9. Changes in and Disagreements With Accountants on Accounting and Financial Disclosure Item 9A. Controls and Procedures Item 9B. Other Information Item 9C. Disclosure Regarding Foreign Jurisdictions that Prevent Inspections
Part III	Item 10. Directors, Executive Officers and Corporate Governance Item 11. Executive Compensation Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters Item 13. Certain Relationships and Related Transactions, and Director Independence Item 14. Principal Accountant Fees and Services
Part IV	Item 15. Exhibits and Financial Statement Schedules Item 16. Form 10-K Summary

Appendix B. Illustrative Examples of Reconstruction

Table A2 shows illustrative reconstruction examples from Item 1 of Form 10-K. Red text indicates removed sentences.

Table A2. Examples of 10-K text before and after sentence reconstruction.

Company	Before Reconstruction	After Reconstruction
Cadence Design Systems, Inc. (San Jose, CA, USA)	We urge you to consider these factors carefully in evaluating the forward-looking statements contained in this Annual Report on Form 10-K. All subsequent written or oral forward-looking statements attributable to our company or persons acting on our behalf are expressly qualified in their entirety by these cautionary statements. Cadence is a leader in electronic system design, building upon more than 30 years of computational software expertise. We apply our underlying Intelligent System Design™ strategy to deliver computational software, hardware and IP that turn design concepts into reality. Our products and services are designed to give our customers a competitive edge in their development of integrated circuits ("ICs"), systems-on-chip ("SoCs"), and increasingly sophisticated electronic devices and systems.	Cadence is a leader in electronic system design, building upon more than 30 years of computational software expertise. We apply our underlying Intelligent System Design™ strategy to deliver computational software, hardware and IP that turn design concepts into reality. Our products and services are designed to give our customers a competitive edge in their development of integrated circuits ("ICs"), systems-on-chip ("SoCs"), and increasingly sophisticated electronic devices and systems.

Table A2. *Cont.*

Company	Before Reconstruction	After Reconstruction
Cboe Global Markets, Inc. (Chicago, IL, USA)	The following description of the business should be read in conjunction with the information included elsewhere in this Annual Report on Form 10-K for the year ended 31 December 2021. This description contains forward-looking statements that involve risks and uncertainties. Actual results could differ significantly from the results discussed in the forward-looking statements due to the factors set forth in "Risk Factors" and elsewhere in this Annual Report on Form 10-K. Overview Cboe Global Markets, Inc., a leading provider of market infrastructure and tradable products, delivers cutting-edge trading, clearing and investment solutions to market participants around the world. Cboe provides trading solutions and products in multiple asset classes, including equities, derivatives and FX, across North America, Europe, and Asia Pacific. Cboe's subsidiaries include the largest options exchange and the third largest stock exchange operator in the U.S.	Overview Cboe Global Markets, Inc., a leading provider of market infrastructure and tradable products, delivers cutting-edge trading, clearing and investment solutions to market participants around the world. Cboe provides trading solutions and products in multiple asset classes, including equities, derivatives and FX, across North America, Europe, and Asia Pacific. Cboe's subsidiaries include the largest options exchange and the third largest stock exchange operator in the U.S.
Mid-America Apartment Communities, Inc. (Germantown, MD, USA)	To continue to qualify as a REIT, MAA must continue to meet certain tests which, among other things, generally require that our assets consist primarily of real estate assets, our income be derived primarily from real estate assets and that we distribute at least 90% of our REIT taxable income (other than our net capital gains) to our shareholders annually. If MAA maintains its qualification as a REIT, MAA generally will not be subject to U.S. federal income taxes at the corporate level on its net income to the extent it distributes such net income to its shareholders annually. Even if MAA continues to qualify as a REIT, it will continue to be subject to certain federal, state and local taxes on its income and its property. Our Annual Reports on Form 10-K, along with our Quarterly Reports on Form 10-Q, Current Reports on Form 8-K and any amendments to those reports, are available on our website at https://www.maac.com as soon as reasonably practicable after such material is electronically filed with, or furnished to, the SEC. Reference to our website does not constitute incorporation by reference of the information contained on the site and should not be considered part of this Annual Report on Form 10-K.	To continue to qualify as a REIT, MAA must continue to meet certain tests which, among other things, generally require that our assets consist primarily of real estate assets, our income be derived primarily from real estate assets and that we distribute at least 90% of our REIT taxable income (other than our net capital gains) to our shareholders annually. If MAA maintains its qualification as a REIT, MAA generally will not be subject to U.S. federal income taxes at the corporate level on its net income to the extent it distributes such net income to its shareholders annually. Even if MAA continues to qualify as a REIT, it will continue to be subject to certain federal, state and local taxes on its income and its property.

References

1. Hoberg, G.; Phillips, G. Text-based network industries and endogenous product differentiation. *J. Political Econ.* **2016**, *124*, 1423–1465. [CrossRef]
2. Kim, D.; Kang, H.G.; Bae, K.; Jeon, S. An artificial intelligence-enabled industry classification and its interpretation. *Internet Res.* **2022**, *32*, 406–424. [CrossRef]
3. Bhojraj, S.; Lee, C.M.C. Who is my peer? A valuation-based approach to the selection of comparable firms. *J. Account. Res.* **2002**, *40*, 407–439. [CrossRef]
4. Gay, S.; Karger, E. Patent citations and stock performance: Constructing a dynamic industry classification. *SSRN* **2014**, SSRN:2496414. [CrossRef]
5. Lee, C.M.C.; Ma, P.; Wang, C.C.Y. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *J. Financ. Econ.* **2015**, *116*, 410–431. [CrossRef]
6. De Franco, G.; Hope, O.K.; Larocque, S. Analysts' choice of peer companies. *Rev. Account. Stud.* **2015**, *20*, 82–109. [CrossRef]
7. Bergmann, J.P.; Amin, M.; Campbell, Y.; Trela, K. How to find similar companies using websites? *World Pat. Inf.* **2023**, *73*, 102172. [CrossRef]
8. Vamvourellis, D.; Tóth, M.; Bhagat, S.; Desai, D.; Mehta, D.; Pasquali, S. Company similarity using large language models. In Proceedings of the 2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr), Hoboken, NJ, USA, 22–23 October 2024; pp. 1–9. [CrossRef]

9. Yang, H.; Lee, H.J.; Cho, S.; Cho, E. Automatic classification of securities using hierarchical clustering of the 10-Ks. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 3936–3943. [CrossRef]
10. Olbert, L. Identifying gaps between research results and education. *J. Account. Educ.* **2024**, *66*, 100884. [CrossRef]
11. Ito, T.; Camacho-Collados, J.; Sakaji, H.; Schockaert, S. Learning company embeddings from annual reports for fine-grained industry characterization. In Proceedings of the Second Workshop on Financial Technology and Natural Language Processing, Kyoto, Japan, 5 January 2020; pp. 27–33.
12. Kee, T. Peer firm identification using word embeddings. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 5536–5543. [CrossRef]
13. Gopikrishnan, P.; Rosenow, B.; Plerou, V.; Stanley, H.E. Identifying business sectors from stock price fluctuations. *arXiv* **2000**, arXiv:cond-mat/0011145. [CrossRef]
14. Bhojraj, S.; Lee, C.M.C.; Oler, D.K. What's my line? A comparison of industry classification schemes for capital market research. *J. Account. Res.* **2003**, *41*, 745–774. [CrossRef]
15. Brown, S.V.; Ma, G.; Tucker, J.W. Financial statement similarity. *Contemp. Account. Res.* **2023**, *40*, 2577–2615. [CrossRef]
16. Gao, H.; He, J.; Chen, K. Exploring machine learning techniques for text-based industry classification. *SSRN* **2020**, SSRN:3640205. [CrossRef]
17. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B* **1999**, *11*, 193–197. [CrossRef]
18. Dalziel, M. A systems-based approach to industry classification. *Res. Policy* **2007**, *36*, 1559–1574. [CrossRef]
19. Kakushadze, Z.; Yu, W. Statistical industry classification. *arXiv* **2016**, arXiv:1607.04883. [CrossRef]
20. Fodor, A.; Jorgensen, R.D.; Stowe, J.D. Financial clusters, industry groups, and stock return correlations. *J. Financ. Res.* **2021**, *44*, 121–144. [CrossRef]
21. Cheng, C.S.A.; McNamara, R. The valuation accuracy of the price-earnings and price-book benchmark valuation methods. *Rev. Quant. Financ. Account.* **2000**, *15*, 349–370. [CrossRef]
22. Noels, S.; De Ridder, S.; Viaene, S.; De Bie, T. An efficient graph-based peer selection method for financial statements. *Intell. Syst. Account. Financ. Manag.* **2023**, *30*, 120–136. [CrossRef]
23. Patra, B.; Suryanarayanan, V.; Fufa, C.; Bhattacharya, P.; Lee, C. ScopeIt: Scoping Task Relevant Sentences in Documents. In Proceedings of the 28th International Conference on Computational Linguistics: Industry Track, Online, 12–13 December 2020; pp. 214–227. [CrossRef]
24. Hsu, C.-C.; Tan, C. Decision-Focused Summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 117–132. [CrossRef]
25. U.S. Securities and Exchange Commission. Available online: https://www.sec.gov/ (accessed on 14 February 2025).
26. Song, S. The informational value of segment data disaggregated by underlying industry: Evidence from the textual features of business descriptions. *Account. Rev.* **2021**, *96*, 361–396. [CrossRef]
27. Parshakov, P.; Shakina, E. Do companies disclose intellectual capital in their annual reports? New evidence from explorative content analysis. *J. Intellect. Cap.* **2020**, *21*, 853–871. [CrossRef]
28. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. *spaCy: Industrial-Strength Natural Language Processing in Python*, Version v3.7.2 ; Zenodo: Geneva, Switzerland , 2020. [CrossRef]
29. Wharton Research Data Services. Available online: https://wrds.wharton.upenn.edu/ (accessed on 10 September 2024).
30. Financial Modeling Prep. Financial Modeling Prep API Documentation. Available online: https://site.financialmodelingprep.com/ (accessed on 16 December 2025).
31. Loukas, L.; Fergadiotis, M.; Chalkidis, I.; Spyropoulou, E.; Malakasiotis, P.; Androutsopoulos, I.; Paliouras, G. FiNER: Financial numeric entity recognition for XBRL tagging. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland, 22–27 May 2022. [CrossRef]
32. Phanse, R.; Zhou, Y.; Shi, K.; Zhang, W.; Liu, Y.; Zhao, Y.; Cohan, A. MSRS: Evaluating multi-source retrieval-augmented generation. *arXiv* **2025**, arXiv:2508.20867. [CrossRef]
33. Csányi, G.M.; Lakatos, D.; Üveges, I.; Megyeri, A.; Vadász, J.P.; Nagy, D.; Vági, R. From fact drafts to operational systems: Semantic search in legal decisions using fact drafts. *Big Data Cogn. Comput.* **2024**, *8*, 185. [CrossRef]
34. Ma, X.; Wang, L.; Yang, N.; Wei, F.; Lin, J. Fine-tuning LLaMA for multi-stage text retrieval. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 2421–2425. [CrossRef]
35. Cremers, M.; Petajisto, A.; Zitzewitz, E. Should benchmark indices have alpha? Revisiting performance evaluation. *Crit. Financ. Rev.* **2013**, *2*, 1–48. [CrossRef]
36. Takayanagi, T.; Sakaji, H.; Izumi, K. SETN: Stock embedding enhanced with textual and network information. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 2377–2382. [CrossRef]

37. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
38. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Front. Psychol.* **2013**, *4*, 863. [CrossRef] [PubMed]
39. Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **2017**, *8*, 355–362. [CrossRef] [PubMed]
40. Cameron, A.C.; Gelbach, J.B.; Miller, D.L. Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* **2008**, *90*, 414–427. [CrossRef]
41. Lyons, B. Valuation multiples: A tool for fundamental & firm analysis. *J. High. Educ. Theory Pract.* **2015**, *15*, 19–27.
42. Fairfield, P.M.; Yohn, T.L. Using asset turnover and profit margin to forecast changes in profitability. *Rev. Account. Stud.* **2001**, *6*, 371–385. [CrossRef]
43. Chen, M.-H.; Lin, C.-P. The impact of corporate charitable giving on hospitality firm performance: Doing well by doing good? *Int. J. Hosp. Manag.* **2015**, *47*, 25–34. [CrossRef] [PubMed]
44. Benkraiem, R.; Gaaya, S.; Lakhal, F. Tax avoidance, investor protection, and investment inefficiency: An international evidence. *Res. Int. Bus. Financ.* **2024**, *69*, 102258. [CrossRef]
45. El-Haj, M.; Ogden, A. Financial narrative summarisation using a hybrid TF-IDF and clustering summariser: AO-Lancs system at FNS 2022. In Proceedings of the 4th Financial Narrative Processing Workshop @ LREC2022, Marseille, France, 24 June 2022; pp. 79–82.
46. Tang, Y.; Yang, Y. FinMTEB: Finance massive text embedding benchmark. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 3620–3638. [CrossRef]
47. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Improving text embeddings with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 11897–11916. [CrossRef]
48. Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; Ping, W. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In Proceedings of the International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025; pp. 79310–79333.
49. OpenAI. Text-Embedding-3-Large. Available online: https://developers.openai.com/api/docs/models/text-embedding-3-large (accessed on 12 March 2026).
50. Voyage AI. Available online: https://docs.voyageai.com/docs/embeddings (accessed on 12 March 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.