Unveiling Complex Interconnections Among Companies through Learned Embeddings

Eugene Oon, Sam Shinde, and Ethan Moody

August 2023

Abstract

We explore alternative methods of company classification using learned embeddings and graph clustering techniques. We also evaluate how these methods enhance investment risk and portfolio management as a downstream task. Our approach involves fine-tuning pre-trained transformer-based models on a sector classification task using Form 10-K data, and drawing on modeled results to analyze the interconnectedness of modern businesses. Our results highlight limitations with current company classification schemes and also offer practical significance to investment diversification strategies.

1 Introduction

The rise of global supply chains and recent technological advancements have profoundly transformed business models and the way they operate. Companies are much more likely to be intricately connected and operationally complex today than they were several years ago, as many now span a wide range of sales, service, and production functions and leverage digital platforms to reach wider markets.

Company classifications help investors, financial researchers, and policy makers compare businesses and analyze specific industries relative to broader markets. A variety of classification taxonomies exist, including the North American Industry Classification System (NAICS), the Standard Industry Classification System (SIC), and the Global Industry Classification Standard (GICS). Each classification scheme aims to group like-businesses together. Some focus more heavily on production processes (NAICS) and others on core products or services (SIC and GICS). However, as both Dolphin et al. (2022) and Rizinski et al. (2023) point out, these established approaches may not adequately capture the dynamic similarities and distinctions among businesses today. They allocate each company to a specific industry and sector without allowing for intersections between these groups. They can also suffer from inconsistencies, inherent subjectivity, and outdated data.

The advent of transformers in the field of natural language processing (NLP) has provided a new opportunity for company classification systems. Studies from Slavov et al. (2019) and Ito et al. (2020) explore how these models – in conjunction with other machine learning (ML) techniques – can be applied to text classification and company categorization tasks with promising results. These methods not only have the potential to speed up the classification process, but also provide a way to gain new insights into company relationships.

1.1 Our Contributions

In this paper, we examine two related research pathways. First, we explore systematically classifying companies into GICS sectors, applying NLP techniques on business descriptions from Form 10-Ks of U.S. publicly listed companies. Our research builds on the work outlined in Rizinski et al. (2023) and Dolphin et al. (2022) by incorporating the fine-tuning of pre-trained transformers and text summarization into the classification task. Second, we propose a novel way to classify and investigate the complex relationships between companies. This objective expands on the research of Ito et al. (2020) and others by using NLP models to learn company embeddings. We extract these embeddings and use graph theory and clustering algorithms to uncover their relationships. Finally, we demonstrate how these results can be used to enhance risk and portfolio management.

2 Related Work

The present study fits within a body of research exploring (1) how pre-trained transformers perform on company classification tasks, and (2) how companies or stocks can be represented by embeddings and connected networks. Gao et al. (2020), Ito et al. (2020), and Rizinski et al. (2023) show that transformers can effectively classify companies by industries (i.e., with

a level of success comparable to established classification schemes, like GICS) using information from annual reports and business descriptions. Yang et al. (2016) propose a new classification scheme – known as the Business Text Industry Classification (BTIC) system – using Doc2Vec to extract embeddings from Form 10-Ks of S&P 500 companies. The authors apply these embeddings to group stocks using hierarchical clustering. Other research by Dolphin et al. (2022) shows that multimodal neural models can be trained to capture company embeddings from historical stock price data and financial news.

Additionally, we note the work of Du and Tanaka-Ishii (2020), which paves the way for applying stock representations to areas outside of stock price prediction, such as portfolio optimization. Lastly, Sarmah et al. (2022) shows that company embeddings can be learned by representing a stock correlation matrix as a network and applying graph ML algorithms. Here, the authors use Node2Vec to create sentence-like structures from the weighted network, apply Word2Vec to construct sets of context and target stocks, and obtain word embeddings from model training.

Our work builds on many of these findings, yet remains distinct. As in Rizinski et al. (2023), we leverage pre-trained transformer-based models and textual business descriptions to classify companies and compare results against GICS sector labels. However, we use detailed descriptions from Form 10-Ks instead of one-line descriptions from the Wharton Research Data Services (WRDS) Compustat dataset. We also go beyond this task to (1) generate embeddings of U.S. publicly listed companies from our modeling, (2) use these vector representations to develop company sector clusters, and (3) – like Sarmah et al. (2022) – examine these clusters to better identify how companies relate to one another. Lastly, like Du and Tanaka-Ishii (2020), we compute portfolio performance metrics to examine how portfolios constructed under our approach perform against the S&P 500 index.

3 Methods

In this section, we outline our key research objectives, data, and approach. We summarize our three primary objectives this way: *(1) Can we systematically classify U.S. publicly listed companies into GICS sectors using pre-trained transformer-based models? (2) Can we use the learned company embeddings from these models to examine their complex interconnections? (3) Can we enhance investment risk and portfolio management based on these connections?*

3.1 Relevant Data

For company classification, we drew from two sources of data: existing GICS sector classifications for U.S. publicly listed companies, and business descriptions from Form 10-Ks. We used the GICS sector classifications as labels for multiclass predictive modeling and the company business descriptions for fine-tuning different transformers.

We extracted current GICS sector labels from TD Ameritrade's trading platform[1] and 2022 Form 10-K data from SEC API.[2] After reviewing several Form 10-K examples, we chose to focus exclusively on the "business description" subsection of Item 1, as it offered the most comprehensive view of each company's business activities, operations, objectives, products, services, and competitors. Our final dataset contained all 500 S&P 500 companies after we manually added business descriptions for 13 missing companies, and 3,682 non-S&P 500 companies after we removed records with missing or shortened Item 1 information. We used the non-S&P 500 companies for model training – with 15% allocated for validation – and the S&P 500 companies for model testing.

For portfolio optimization and performance measurement, we downloaded adjusted closing prices from Yahoo Finance.[3] We used a time period of roughly 1.5 years – January 1, 2022, through July 29, 2023 – so as not to assume annualized figures for 2023 based only on Q1 and partial-Q2 results.

3.2 Text Summarization

We explored several pre-trained transformers for company classification. Some of these models have a maximum token limit of 512. Input texts exceeding this limit must be tokenized into smaller chunks, abbreviated, or summarized. Because of this limitation, we created summarized versions of the Form 10-K business descriptions for each company for model training, and also preserved the original descriptions.

We experimented with three summarization approaches: (1) BERT[4], (2) GPT-2[5], and (3) XLNet[6]. Manual reviews of the output from all three approaches led us to choose XLNet over BERT and GPT-2 based on overall readability and coherence.

3.3 Company Classification Modeling

Company classifications are inherently imbalanced across the 11 GICS sectors, with no sector containing more than 25% of U.S. publicly listed companies (see, e.g., Rizinski et al. (2023)). We noticed similar imbalance across sectors within our training, validation,

and testing sets as we prepared our data for modeling. Thus, for a baseline model, we chose to avoid a "predict majority class" approach, which would have been overly simplistic and only marginally accurate. Instead, we combined Multinomial Naïve Bayes (MNB) with a Bag-of-Words (BoW) model to compare our transformer-based fine-tuning approaches with established, non-SoTA NLP techniques.

We performed company classification modeling by fine-tuning pre-trained versions of BERT[7] (see Devlin et al. (2019)), RoBERTa[8] (see Liu et al. (2019)), FinBERT[9] (a domain-specific variant of BERT trained on a large financial corpus, see Araci (2019)), and Sentence Transformers[10] (see Reimers and Gurevych (2019)). For fine-tuning BERT and RoBERTa, we unfroze all layers during training and validation, added two additional hidden layers of 256 and 128 neurons, and incorporated a neural dropout rate of 0.20 in between each hidden layer to help prevent overfitting. With FinBERT, we took a similar approach, but found strong performance using only a single additional hidden layer of 512 neurons and a neural dropout rate of 0.30. We also paired a CNN with FinBERT – using filters and kernels of varying sizes[11] – but noticed this approach was more susceptible to overfitting. For these models, we used the XLNet summaries of Form 10-K business descriptions and the pooler [CLS] token for making classification predictions.

We also experimented with Sentence Transformers to encode text from Form 10-K company descriptions. We tokenized the full text into separate sentences using NLTK's "sent_tokenize" submodule, and then encoded each sentence through Sentence Transformers[12] to generate fixed-length sentence embeddings. We leveraged the model's pre-trained weights – without any fine-tuning – in our first experiment, aiming to capture the semantic relationships and contextual information within each sentence and aggregate them to create a meaningful representation for the entire document. Next, we used these embeddings as an input to a neural network (NN) and trained the classification head to predict company GICS sectors. We included two hidden layers of 512 and 256 neurons within the NN and incorporated a neural dropout rate of 0.10 in between each hidden layer to help prevent overfitting. In a second experiment, we fine-tuned both the Sentence Transformers and the classification head utilizing the SetFit framework.[13]

To address class imbalance, we trained with class weights for all models except the SetFit variation. For SetFit, class imbalance was addressed through data sampling for contrastive learning.

3.4 Alternative Company Clustering Methods

We extracted the learned company embeddings from these models to devise several alternative company clustering methods. We hypothesized these methods would better reflect the interconnectedness of modern businesses than static, unidimensional company classifications, and would also provide a more forward-looking operational outlook for these businesses.

Louvain modularity optimization: First, we created an undirected weighted network of the S&P 500 companies adapted from the methodology described by Sarmah et al. (2022) – i.e., each network node represents a stock with their cosine similarity as edge weights. Then, we applied Louvain modularity optimization to detect stock communities.

Minimum Spanning Tree (MST): MST is a widely adopted method for exploring dense networks (see Marti et al. (2020)). We applied Kruskal's algorithm to construct the MST based on computed cosine distances between stocks, as MST algorithms are designed to minimize the total distance in the tree while ensuring all nodes are connected. We considered different pruning thresholds to identify stock clusters.

Hierarchical clustering: We employed Ward's hierarchical clustering method to create clusters based on the cosine similarity between stocks. We used the resulting dendrogram to explore different thresholds to group stocks into clusters.

K-Means clustering: We applied K-Means clustering to the company embeddings and organized stocks into clusters based on their assignment. We used the Elbow Method to set the optimal number of clusters to 15. Then, we explored two different methods of organizing the stocks: (1) assigning each stock to a single cluster based on its shortest distance to the centroid of that cluster ("KMeans-1"), and (2) assigning a stock to multiple clusters if its distance to a centroid fell within the top 10th percentile of the distribution of stock distances for that particular centroid ("KMeans-2"). With KMeans-1, we wanted to achieve a straightforward allocation of stocks to their most fitting clusters. With KMeans-2, we selected the 10th percentile threshold to mitigate the impact of noise and allow stocks to be associated with multiple relevant clusters (where applicable). To achieve weighted assignments, we normalized the distances to the different cluster centroids.

Multi GICS clustering: Lastly, we developed a "multi GICS clustering" approach built on the final softmax activation and classification layer of our best performing model. We normalized the predicted prob-

abilities from this layer to derive weighted assignments of stocks to sector classes, allowing us to associate stocks with *multiple* GICS sector clusters based on their respective probabilities.

3.5 Portfolio Construction

As noted in James et al. (2022), equity sectors that have distinct risk and return characteristics can contribute to portfolio diversification, especially during periods of financial stability. We drew on this finding to (1) assess the performance of investment portfolios constructed from our alternative company clustering methods focused on S&P 500 stocks and (2) compare their performance against the S&P 500 index[14] and unmodified GICS classifications.

To conduct our portfolio analysis, we constructed risk parity portfolios for each of our alternative company clustering methods. We did this using the covariance matrices calculated from the 2022 closing price data (same time period as the Form 10-K business descriptions) and portfolio optimization methods from Riskfolio-Lib.[15] To create the risk parity portfolios, we treated each cluster as a unique risk factor. For each cluster, we allocated the constituent stocks in such a way that they had equal contributions to the overall standard deviation of the cluster portfolio. Similarly, we composed each final portfolio in such a way that each cluster portfolio contributed equally to the final portfolio's standard deviation. We hypothesized that the clustering and weights used in portfolio construction represent a forward outlook of how the constituent companies are expected to correlate with each other, and assessed how portfolio stocks performed between January 1, 2022, and July, 29 2023.

4 Results and Discussion

The results of our analysis (below) are mapped back to our three primary objectives. We first outline our classification results and note a few examples of model errors. We then share our findings on company interconnections. Finally, we discuss how our findings can enhance investment risk and portfolio management.

4.1 GICS vs. Modeled Classifications

Table 1 shows the results of our classification experiments for the S&P 500 companies. Consistent with prior research (see **Related Work**), we found that our transformer-based approaches outperformed our baseline BoW model across all evaluative metrics (accuracy, precision, recall, and F1-score). We also found mostly consistent results across the transformer models, with BERT, FinBERT, and STrans showing differences of only $0.1 - 0.2$ points across all metrics.

Model	Acc	Prec	Rec	F1
BoW MNB (*B)	0.75	0.70	0.75	0.71
BERT MLM	0.88	0.89	0.88	0.88
RoBERTa	**0.91**	**0.91**	**0.91**	**0.91**
FinBERT	0.87	0.89	0.87	0.88
FinBERT CNN	0.86	0.87	0.86	0.86
STrans	0.86	0.87	0.86	0.86
STrans SetFit	0.82	0.82	0.82	0.82

Table 1: Testing set classification results by model. Note: GICS sector labels were reduced from 11 to 10. Class weights were applied to address imbalance. *BoW: Bag of Words; MNB: Multinomial Naïve Bayes; *B: Baseline; STrans: Sentence Transformers.*

For model evaluation, we assigned highest priority to accuracy. Our aim was to fine-tune a model that was well-calibrated against the GICS sector labels – i.e., one that minimized erroneous predictions overall – so that the company embeddings generated by the model were anchored in an established classification scheme, yet still contained the underlying details necessary for examining company interconnections. With respect to this metric, our fine-tuned RoBERTa model showed the strongest performance ($Acc = 0.91$), followed by BERT MLM ($Acc = 0.88$), FinBERT ($Acc = 0.87$), and STrans ($Acc = 0.86$). We observed some overfitting with FinBERT CNN ($Acc = 0.86$) after a third epoch, as well as with STrans SetFit ($Acc = 0.82$).

A combination of network layer architecture, hyperparameter, and textual data differences explain the variation in results across our models. With RoBERTa and BERT MLM, we employed a learning rate (LR) schedule to reduce the LR if the accuracy on our validation set stagnated, and trained for a maximum of 15 epochs. For FinBERT and FinBERT CNN, we employed a constant LR and trained for only 5 epochs to avoid overfitting. There were also hidden layer and neural dropout rate differences between these models (covered under **Methods**). We chose a simpler architecture for our FinBERT model since FinBERT was already pre-trained on a financial corpus, and even after adding a CNN, we did not see significant change in performance. Unlike for RoBERTa, BERT MLM, and both FinBERT variations, we did not use summarized business descriptions for STrans or STrans SetFit. Our results are not conclusive around the performance benefits of non-summarized versus summarized versions of business descriptions within this clas-

sification task, since we noticed the performance of our STrans model was comparable to or only slightly lower than FinBERT/FinBERT CNN.

Initially, we ran each model against the 11 GICS sector labels and examined performance metrics by class (sector). We noticed that our models consistently confused the Consumer Staples sector with the Consumer Discretionary sector. For example, with BERT MLM, we saw that only 41% of Consumer Staples companies were predicted correctly, while 55% were predicted as Consumer Discretionary. Because both sectors are comprised of companies offering consumer products and services, and the lines between "essential" and "non-essential" products and services can be (at times) ambiguous in company business descriptions – e.g., various retail products can technically fit within either sector – we decided to combine these sectors into a single super-sector called "Consumer Discretionary and Staples" to simplify to 10 classes and reduce model prediction errors. This decision improved model performance without compromising our ability to leverage the company embeddings for drawing insights on company interconnections, including between consumer-oriented companies and all others.

After we re-ran each model against our 10 sector labels, we noticed that prediction errors became more dispersed across the classes. For our best performing model (RoBERTa), the most common errors were typically with the Real Estate (Acc : $0.77 - 0.84$) and Industrials (Acc : $0.85 - 0.87$) sectors, though accuracy rates varied slightly with each model iteration. Since GICS classifies such a small share of S&P 500 companies as Real Estate (6.2%), it was difficult to identify the reasons for prediction errors with this sector, as the maximum number of misclassified examples for any non-Real Estate sector was 2. Given the slightly higher share of Industrials companies (15.0%), we examined a set of summarized business descriptions that were misclassified as Information Technology (most common error). We noticed that many of these descriptions included phrases like "energy solutions" and "energy storage", and words like "power", "kilowatt", and "sustainability". We hypothesize these keywords and phrases made it challenging for the model to disentangle the Industrials and Information Technology sectors during model training, as they relate to business activities most commonly associated with electrical/industrial businesses. We suspect the integration of additional data sources describing business activities – e.g., company annual reports – or the usage of shorter business description summaries – e.g., under 30 tokens to keep descriptions very focused – are two potential ways that future experiments could eliminate some misclassifications.

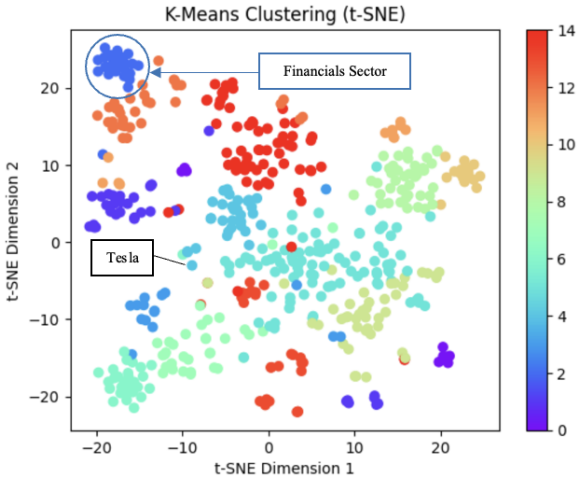

Figure 1: t-SNE plot of KMeans-1 clustering approach. *Example sector cluster and company are annotated.*

4.2 Unveiling Company Interconnectedness

As we applied Louvain modularity optimization, we consistently identified four distinct communities.[16] This finding suggests two potential factors: (1) the stocks are highly interconnected, making it challenging to delineate distinct clusters, and (2) certain equities might span multiple GICS sectors due to the diverse nature of their business activities, resulting in a more cohesive community structure with fewer clusters. For MST, we successfully obtained a sparse weighted network of stocks. However, different pruning thresholds did not yield sufficiently distinct clusters of stocks. This observation further reinforced the notion of dense connections within the network.

We focused on K-Means clustering for the remainder of our analysis, as Hierarchical clustering did not offer any distinct perspectives beyond what we gained from K-Means. Figure 1 illustrates the clustering from our KMeans-1 approach. While most clusters appeared to match the GICS sectors, we noticed some contained stocks from different GICS sectors. These stocks appeared to be densely interconnected or have similar multiple assignments to other clusters. Figure 2 shows a deeper analysis of one of these stocks, Tesla Inc. The appearance of certain stocks (like Tesla) in the intersections of different clusters suggests a degree of overlap in their characteristics. This was a key reason we analyzed a "Multi GICS" approach and assigned stocks to multiple clusters and sectors in an attempt to capture their interconnected nature.

Stock Name: Tesla Inc. (TSLA)
GICS Sector: Consumer Discretionary

The five companies most similar to Tesla based on cosine similarity of company embeddings:

1. Enphase Energy Inc (GICS: Information Technology)
2. Cummins Inc (GICS: Industrials)
3. General Motors Co (GICS: Consumer Discretionary)
4. Generac Holdings Inc (GICS: Industrials)
5. Aptiv PLC (GICS: Consumer Discretionary)

Our K-Means clustering revealed that Tesla, Enphase Energy, and Aptiv were grouped together with companies such as Intel, Advanced Micro Devices, and NVIDIA, all belonging to the GICS Information Technology sector. Despite GICS classification differences, these companies demonstrate strong interconnectedness. For instance, Enphase Energy and Generac are competitors with Tesla in the renewable solar energy and battery storage space. General Motors competes directly with Tesla in the electric vehicle market. Aptiv plays a crucial role as a key auto wire producer and supplier to both Tesla and General Motors.

The grouping of Tesla with Information Technology companies is unsurprising, considering its current endeavors (at the time of this writing) in developing hardware and software for autonomous vehicles using artificial intelligence and its recent advancements in energy storage and photovoltaic systems. These shared interests and areas of expertise contribute to the company clustering patterns we observed, highlighting the multidimensional nature of Tesla's business ventures.

Figure 2: Company clustering analysis for Tesla Inc. *Observed differences emphasized between the GICS sector classification and the sectors for most similar companies.*

Our results do not conclusively point to one clustering method being superior over others. As emphasized by Marti et al. (2020), the clusterings we obtained are susceptible to instability, presenting an ongoing challenge in quantitative finance. However, on the basis of our overall clustering results and specific company examples (like Tesla), we contend that traditional classification schemes may not fully capture the dynamics, distinctions, and interconnectedness of companies today. By extending our model predictions beyond the classification task, we gained valuable insights into the evolving nature of modern business models and how they contribute to this interconnectedness. We next cover how we conducted a quantitative evaluation of our alternative company clustering methods.

4.3 Downstream Investment Applications

We focused our quantitative evaluation on the clusterings from KMeans-1, KMeans-2 and Multi GICS. The results from our portfolio analysis (Table 2) confirmed that risk parity portfolios constructed across different equity sectors *do* offer benefits of diversification. All four portfolios outperformed the S&P 500 index over the timeframe analyzed, offering lower volatility and better return per unit of standard deviation (as represented by the Sharpe Ratio). KMeans-1 classification offered the best performing portfolio, with the highest return and highest Sharpe Ratio. KMeans-2 classification showed a lower volatility and Sharpe Ratio than GICS, while Multi GICS performed comparably to KMeans-2. Notably, for our Sharpe Ratio and return calculations, we assumed a risk-free rate of 0% and ignored transaction costs in order to simplify comparisons against each other.

To further validate these results, we would need to observe each portfolio over the remaining months in 2023 and possibly introduce quarterly rebalancing. We could also consider other techniques to derive more stable covariance matrices. Lastly, as James et al. (2022) showed, we would only need to hold a subset of representative stocks (as opposed to all the constituent stocks) to achieve the same outcomes.

Portfolio	Return	Volatility	Sharpe
SPY	(2.81)%	21.16%	(0.13)
GICS	1.58%	17.45%	0.09
KMeans-1	2.28%	17.65%	0.13
KMeans-2	1.15%	17.09%	0.07
Multi GICS	1.29%	17.70%	0.07

Table 2: Portfolio annualized returns and volatility. *Assumes risk-free rate of 0% and transaction costs of $0.*

5 Conclusion

This paper addresses limitations with current company classification systems and outlines novel methods for exploring the connections between companies. Experimental results suggest that fine-tuning pretrained transformer-based models (e.g., RoBERTa, FinBERT, and others) is an appropriate strategy for predicting sector classifications. However, forcing every company into a singular classification is overly simplistic. A comparison of performance metrics between the S&P 500 index and the risk parity portfolios developed from our company clustering methods reveals that learned company embeddings can overcome these issues and offer real value to investors – specifically when applied to reduce portfolio volatility, or to illustrate complex, nuanced connections between global businesses. Future research could build upon our findings by incorporating (a) financial news sentiment analysis to predict which stocks or clusters are likely to underperform/outperform the market or (b) company financial data complementing the textual description to provide valuable context. Additional research could pull company description information from other textual sources – including company websites and social media – to see if this data provides another layer of insight around company connections beyond what Form 10-Ks provide.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina

Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Rian Dolphin, Barry Smyth, and Ruihai Dong. 2022. A multimodal embedding-based approach to industry classification in financial markets.

Xin Du and Kumiko Tanaka-Ishii. 2020. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3353–3363, Online. Association for Computational Linguistics.

Haocheng Gao, Junjie He, and Kang'an Chen. 2020. Exploring machine learning techniques for text-based industry classification. *Big Data & Innovative Financial Technologies Research Paper Series*.

Tomoki Ito, Jose Camacho Collados, Hiroki Sakaji, and Steven Schockaert. 2020. Learning company embeddings from annual reports for fine-grained industry characterization. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 27–33, Kyoto, Japan. -.

Nick James, Max Menzies, and Georg A. Gottwald. 2022. On financial market correlation structures and diversification benefits across and within equity sectors. *Physica A: Statistical Mechanics and its Applications*, 604:127682.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, and Philippe Donnat. 2020. A review of two decades of correlations, hierarchies, networks and clustering in financial markets.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Miskovski, and Dimitar Trajanov. 2023. Company classification using zero-shot learning.

Bhaskarjit Sarmah, Nayana Nair, Dhagash Mehta, and Stefano Pasquali. 2022. Learning embedded representation of the stock correlation matrix using graph machine learning.

Stanislav Slavov, Andrey Tagarev, Nikola Tulechki, and Svetla Boytcheva. 2019. Company industry classification with neural and attention-based learning models. In *2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE)*, pages 1–7.

Hoseong Yang, Hye Jin Lee, Sungzoon Cho, and Eugene Cho. 2016. Automatic classification of securities using hierarchical clustering of the 10-ks. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3936–3943.

Notes

[1]https://www.tdameritrade.com/
[2]https://sec-api.io/
[3]https://finance.yahoo.com/
[4]https://pypi.org/project/bert-extractive-summarizer/
[5]https://huggingface.co/gpt2-medium
[6]https://huggingface.co/xlnet-base-cased
[7]https://huggingface.co/bert-base-uncased
[8]https://huggingface.co/roberta-base
[9]https://huggingface.co/ProsusAI/finbert
[10]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[11]For the CNN, we used filter sizes of 100, 100, 50, and 25 and kernel sizes of 2, 3, 4, and 5.
[12]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[13]https://github.com/huggingface/setfit/blob/main/README.md
[14]I.e., SPDR S&P 500 ETF Trust
[15]https://riskfolio-lib.readthedocs.io/en/latest/
[16]We also found that the specific stocks inside each community varied slightly with each run due to the stochastic nature of Louvain modularity optimization.

A Appendix

A.1 Additional Notes on Data

We initially explored publicly available stock screeners for the GICS sector labels and the WRDS SEC Analytics Suite for business descriptions sourced from Form 10-Ks for NYSE/NASDAQ companies.[17] We intended to parse relevant sections from the Form 10-K data – such as Items 1 (Business), 1A (Risk Factors), 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations), and 7A (Quantitative and Qualitative Disclosures About Market Risk) – using regex. However, we encountered two issues with this approach: (1) some of the GICS classifications were outdated, and (2) accurately extracting relevant sections from Form 10-Ks using regex proved to be challenging given the varied submission formats. We modified our approach to what is summarized under **Methods** for cleaner, more accurate data.

A.2 Fine-Tuning SetFit

For fine-tuning the SetFit model, we first sampled 24 companies for each of the GICS sectors as training labeled data; froze the final layer of the NN and applied contrastive learning with this sample data to tune the model body; and then froze the model body and tuned the classification head. A qualitative evaluation of the first 512 tokens of each company description (the maximum sequence length) proved sufficient for capturing the essential aspects of each company's business activities.

A.3 Additional Plots and Figures

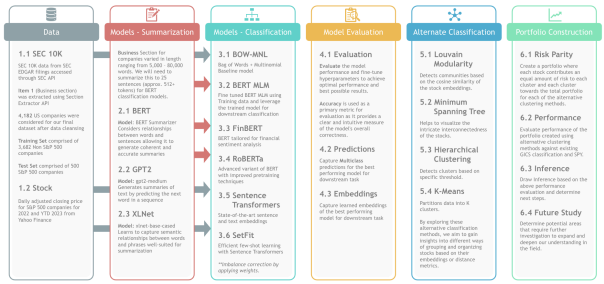

Figure 3: Visual outline (roadmap) of the present study.

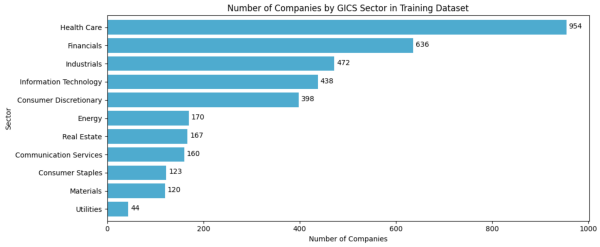

Figure 4: GICS sector distribution – Training + Validation set (Non-S&P 500 companies).

Figure 5: GICS sector distribution – Testing set (S&P 500 companies).

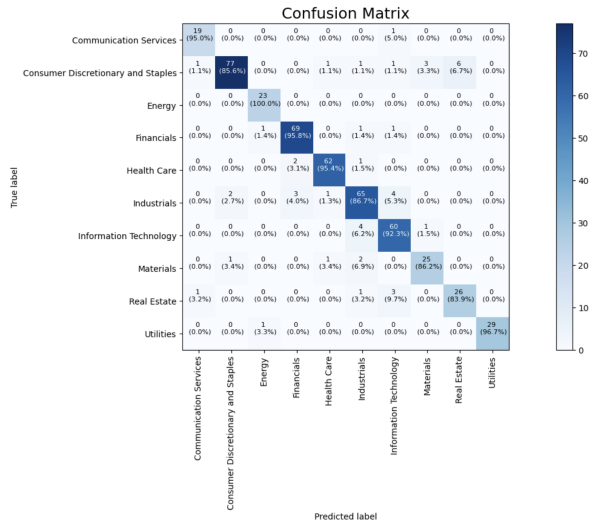

Figure 6: Testing set Confusion Matrix for RoBERTa.

	precision	recall	f1-score	support
Communication Services	0.90	0.95	0.93	20
Consumer Discretionary and Staples	0.96	0.86	0.91	90
Energy	0.92	1.00	0.96	23
Financials	0.93	0.96	0.95	72
Health Care	0.95	0.95	0.95	65
Industrials	0.87	0.87	0.87	75
Information Technology	0.86	0.92	0.89	65
Materials	0.86	0.86	0.86	29
Real Estate	0.81	0.84	0.83	31
Utilities	1.00	0.97	0.98	30
accuracy			0.91	500
macro avg	0.91	0.92	0.91	500
weighted avg	0.91	0.91	0.91	500

Figure 7: Testing set Classification Report for RoBERTa.

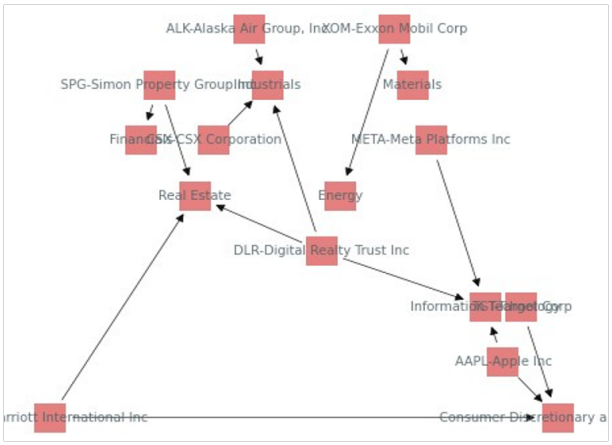

Figure 8: Illustration of MST approach.

Figure 9: Illustration of Multi GICS approach.