A Semantic Approach to Financial Fundamentals

Jiafeng Chen[*1,2] , **Suproteem K. Sarkar**[*†1]
[1]Harvard University
[2]Harvard Business School
{jiafengchen, suproteemsarkar}@g.harvard.edu,

Abstract

The structure and evolution of firms' operations are essential components of modern financial analyses. Traditional text-based approaches have often used standard statistical learning methods to analyze news and other text relating to firm characteristics, which may shroud key semantic information about firm activity. In this paper, we present the Semantically-Informed Financial Index (SIFI), an approach to modeling firm characteristics and dynamics using embeddings from transformer models. As opposed to previous work that uses similar techniques on news sentiment, our methods directly study the business operations that firms report in filings, which are legally required to be accurate. We develop text-based firm classifications that are more informative about fundamentals per level of granularity than established metrics, and use them to study the interactions between firms and industries. We also characterize a basic model of business operation evolution. Our work aims to contribute to the broader study of how text can provide insight into economic behavior.

1 Introduction

Firm operations are key components of financial analyses, but fine-grained data about business characteristics are often stored in unstructured corpora. Text analysis has been used over the past decade in the study of firms and their characteristics, though most of these studies have used manual labels or simple word-based metrics [Tetlock, 2007; Loughran and McDonald, 2016]. Recently, [Hoberg and Phillips, 2016] introduced a method for clustering firms by taking the cosine similarity of bag-of-words of 10-Ks, annual reports filed by firms to the SEC, and [Ke et al., 2019] introduced a topic model-based approach to modeling sentiment and returns.

With the recent growth of large transformer models [e.g. Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020], semantically-informed approaches to domain-specific language tasks have become more tractable. While there is a preliminary literature on using pre-trained language models to measure sentiment [Hiew et al., 2019], a more consistent measure of fundamentals may instead lie in descriptions of actual firm operations. We use pre-trained language models to study business descriptions on firm filings, which are legally required to be accurate. This paper presents the following contributions:

- We introduce the Semantically-Informed Financial Index (SIFI), a quantitative representation of business operations that can be interpreted alongside existing industrial classifications of firms.
- We compare our index to existing business classifications, including those based on text, and find our measure can be more informative and queryable.

Our index allows us to measure the structure and evolution of firms and industries over time. With this semantically-informed approach, we seek to develop rich information sources for broader financial applications.

1.1 Paper Organization

In Section 2, we discuss the methods used to develop the SIFI index. In Section 3, we probe the interpretability of our index, relating it to existing classification schemes and analyzing how it can track industry-level trends. Finally, in Section 4, we quantitatively measure the expresiveness of our index compared to existing benchmarks.

2 Developing SIFI

Recent advances in natural language processing have contributed to a growing literature of transfer learning-based applications using large pre-trained transformer models [Vaswani et al., 2017], with models based on BERT [Devlin et al., 2018] achieving state-of-the-art results in a variety of domains and tasks. Transformers in particular [Vaswani et al., 2017] allow for modeling of long-range dependencies, since their architecture allows for massively parallelizable operations on GPUs—in contrast to RNN models, where evaluation is iterative along the length of the text. BERT is a transformer pre-trained on a series of self-supervised tasks on a large dataset—and can be used to learn a representation

[*]Equal Contribution.

[†]Sarkar gratefully acknowledges the support of a National Science Foundation Graduate Research Fellowship.

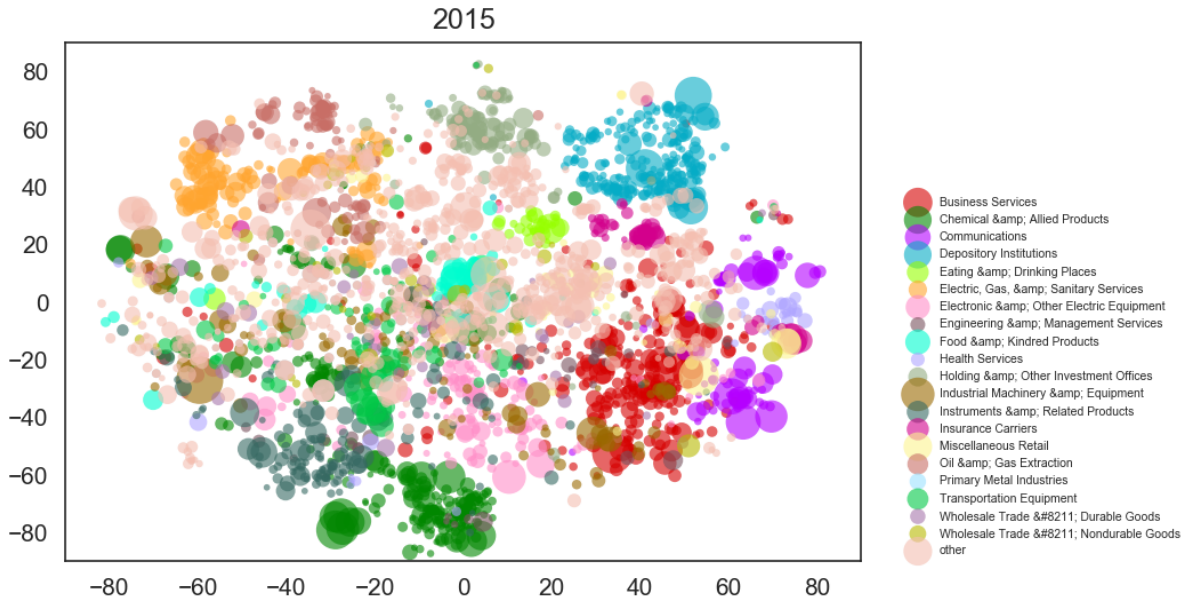

Figure 1: t-SNE decomposition of business section embeddings in 2015, labeled by SIC industry code. The size of each bubble reflects the market capitalization of each firm.

of language which allows for downstream fine-tuning [Liu, 2019].

In this paper, we build on these advances to generate concise document embeddings that capture important features of financial documents. Compared to traditional approaches using word embeddings [Mikolov et al., 2013], the semantic embeddings generated by BERT are context dependent, and can capture long-range dependencies bidirectionally [Jawahar et al., 2019]. Therefore, we believe transformer-based embeddings hold particular promise for processing financial documents, which are often long, complex, and context-dependent.

Our data encompasses the universe of Form 10-K filings to the Securities and Exchange Commission (SEC) from 2006 to 2018 collected by [Loughran and McDonald, 2016],[1] which consists of 93,480 firm-years, with an average of 7,191 firms represented per year. 10-Ks are annual filings to the SEC in which a firm's management discusses business operations and financial performance in the recent fiscal year. These documents are legally required to be correct, and contain rich information about each firm's operations, as shown in Figure 2. We use regular expressions to collect only the *Business Section* of the filing, which provides a description of firms' operations and products.

We use pre-trained uncased BERT-large (24-layer, 340 million parameters) implemented by Huggingface in PyTorch in order to compute our embeddings.[2] We break each document into 512-word segments and pass into the pre-trained transformer. We then mean-pool over the resulting word-level embeddings over the entire document, and save the resulting

```
...Across the company, machine learning and
artificial intelligence (AI) are increasingly
driving many of our latest innovations.
Within Google, our investments in machine
learning over a decade have enabled us
to build products that are smarter and
more useful -- it's what allows you to use
your voice to ask the Google Assistant for
information, to translate the web from one
language to another, to see better YouTube
recommendations, and to search for people and
events in Google Photos...
```

Figure 2: An excerpt from the business section of Alphabet's 2019 10-K.

vector as the document embedding. Our approach is entirely zero-shot, since the model is not fine-tuned on the specific dataset, though we still find interpretable and meaningful results. In this vein, we believe these findings can be interpreted as a lower bound of the power of using transformers for financial applications. Our approach can be extended in future work by fine-tuning the model on meaningful financial targets, including fundamentals and trading metrics, which we anticipate may lead to even greater separation across the semantic components that are most relevant for the financial metrics we aim to identify.

3 Querying the Index

In this section, we present a series of experiments describing the validity and interpretability of our index.

[1] https://sraf.nd.edu/data/stage-one-10-x-parse-data/

[2] https://github.com/huggingface/transformers

Table 1: Industries most similar to Heavy Construction (16)

Industry	Distance (L2)
Engineering & Management Services (87)	0.8578
Rubber & Miscellaneous Plastics Products (30)	0.8720
Industrial Machinery & Equipment (35)	0.8870
Instruments & Related Products (38)	0.9055
Wholesale Trade & Durable Goods (50)	0.9158
Electronic & Other Electric Equipment (36)	0.9237
Business Services (73)	0.9254
Printing & Publishing (27)	0.9406
Transportation Equipment (37)	0.9432
Oil & Gas Extraction (13)	0.9517

Table 2: Industries most similar to Depository Institutions (60)

Industry	Distance (L2)
Holding & Other Investment Offices (67)	0.7101
Insurance Carriers (63)	0.7898
Instruments & Related Products (38)	0.8808
Wholesale Trade & Durable Goods (50)	0.8965
Real Estate (65)	0.9121
Security & Commodity Brokers (62)	0.9491
Electronic & Other Electric Equipment (36)	0.9677
Oil & Gas Extraction (13)	1.0147
Engineering & Management Services (87)	1.0262
Business Services (73)	1.0327

3.1 Industry Characteristics

After computing embeddings for all firms in our sample period, we can plot relationships between firm-years for our sample. In Figure 1, we plot a t-SNE [Maaten and Hinton, 2008] decomposition of all firms in the 2015 sample, first transforming each firm embedding into 50-component PCA space. We label each firm by its 2-digit SIC classification code, which is an established classification scheme for businesses and has been used to evaluate industry-level operations [Fertuck, 1975].

Our results mirror those encoded by SIC classifications, separating into clusters that map onto 2-digit codes. Note that we do not introduce additional information about SIC codes to the model, so this clustering is based on the pre-trained corpus. Therefore, Figure 1 suggests our vectors separate in a similar manner as existing SIC classifications without relying on the metric during the development of the index—this leads us to believe that our embedding-based approach may mirror actual operational differences at the industry-wide scale.

Furthermore, we report the minimum-distance cluster centroids to the Heavy Construction (Table 1) and Depository Institution (Table 2) SIC classification centroids. We note that many of these SIC-2 clusters employ similar operations as their respective targets, including "Engineering & Management Services" and "Rubber & Miscellaneous Plastics Products" for "Heavy Construction," as well as "Holding & Other Investment Offices" and "Insurance Carriers" for "Depository Institutions." The cluster similarities mirror what one might expect for similar industries, but do not exclusively reside in the same 1-digit classification as the SIC index, which may reflect the evolving nature of business operations since the introduction of SIC in the 1930s. The results suggest that our embeddings may also capture semantic information about operations at a broad, industry level.

3.2 Firm Characteristics

We also query our firm space by projecting firm embeddings onto components reflecting the direction between two given firms. For example, projections onto axes from CVS Health to Amgen and CVS Health to Amazon components are reported in Figure 3. This decomposition allows us to query how the semantic embeddings of firms' operations change as

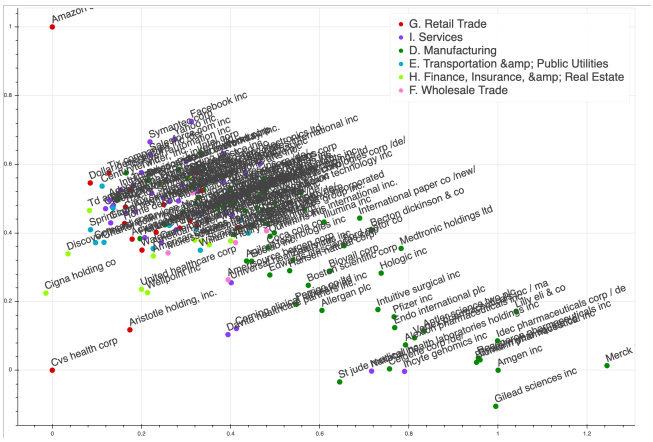

Figure 3: Projections of firm embeddings onto CVS Health-Amgen and CVS Health-Amazon components.

they move across the latent space. Note that as firms move toward Amazon they lean toward tech (e.g. Facebook, Yahoo), and as they move toward Amgen they lean toward biotech (e.g. Celgene and Regeneron). Furthermore, although firms along each axis do not neatly fall in the same 2 broad industrial classifications, they do exhibit similar focuses (e.g. St. Jude Medical and Incyte Genomics, which both have health-related operations, fall across the CVS Health-Amgen axis, even if they are not in the same broad SIC code), which suggests that the rigid boundaries of the SIC classification system do not necessarily account for all the ways that firm operations may coincide. These analyses suggest that meaningful semantic information about operations is also being captured at the firm level in our index.

3.3 Industry Dynamics

Since our sample contains filing information across several years, we can also use our document embeddings to explore the evolution of firms and industries over time. In Figure 4, we plot the within cluster mean sum of squares of the Depository Institutions and Motion Picture Industry SIC codes over time in the SIFI embeddings. The motivation behind this particular analysis is to broadly study how operations are different from one another over time. The trends in the dispersion

Table 3: Across-industry variation for each index, holding granularity fixed across learned clusters. OI denotes operating income and MV denotes market value. A higher standard deviation indicates greater informativeness for that particular fundamental. The higher variation of SIFI compared to SIC-3 and TNIC suggests that it is capturing important semantic information, which allows for greater separation of firms across business operations that are associated with fundamentals.

Across-Industry Std. Dev.	OI / Assets	OI / Sales	MV / Assets	MV / Sales	OI / MV
SIC-3	0.429	19.104	4.191	85.963	13.251
TNIC	0.971	20.736	10.055	83.692	14.457
SIFI	**1.271**	**41.088**	**11.350**	**147.763**	**23.310**

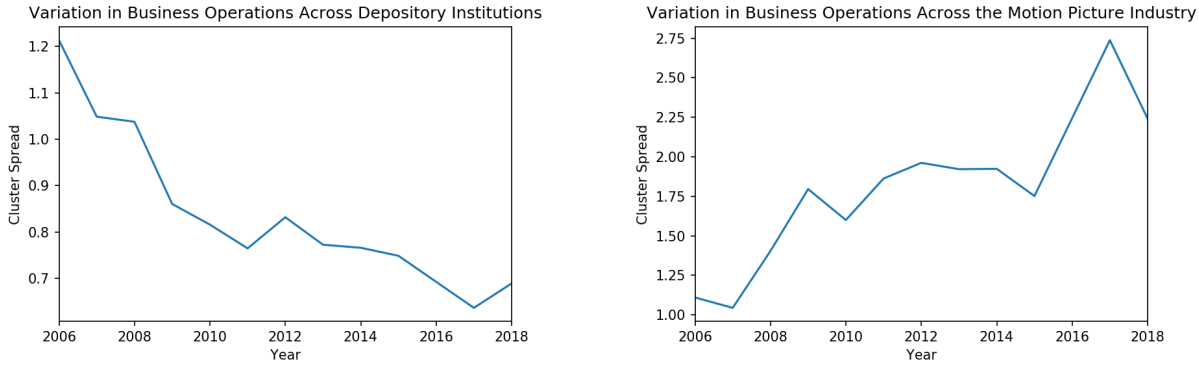

Figure 4: Business operation dispersion across selected industries over time, measured using within-cluster mean sum of squares.

of operations mirror what we might expect from the evolution of product mixes in these industries, particularly given consolidation in banks, which supports the claim that semantic embeddings may also be useful for documenting changes in operations over time.

4 Evaluating SIFI

In this section, we compare the informativeness of our index with related classifications, using metrics explored elsewhere in the literature.

4.1 Index Informativeness

An important metric for evaluating SIFI is its relative informativeness compared to existing indices. In their development of Text-Based Network Industries (TNIC), [Hoberg and Phillips, 2016] use across-industry variation in fundamentals to measure the informativeness of their index. Specifically, they cluster firms that have small pairwise distances between their respective latent space vectors. Next, they take the standard deviation of selected financial fundamentals across their mean values in each of these clusters' centers.

In Table 3, we report this measure for SIFI, TNIC, and 3-digit SIC codes across a series of fundamentals that are scale-invariant. We hold the granularity of SIFI and TNIC fixed to have the same number of clusters as the SIC-3 classification, which is also done in the informativeness calculations in [Hoberg and Phillips, 2016], in order to put our index on the same playing field as the less granular SIC index. We find that SIFI maintains the highest level of across-industry variation across different types of fundamentals, for example in predicting the ratio of operating income to sales, which is also queried by [Hoberg and Phillips, 2016].

The higher measured across-industry variation suggest that SIFI is meaningfully capturing differences in fundamentals across firms, even when it is limited to a coarser scale. This implies that using semantically-informed measures of firm operations may contribute additional information about the ways that fundamentals of related firms behave.

5 Discussion

In this paper, we develop a semantically-informed index of firm operations sourced from the business section of firm filings, which are legally required to be accurate. We demonstrate how the index can be used to understand the characteristics and dynamics of firms and industries, both at a large, industry-wide level and a smaller, firm-specific level. We further evaluate how our index might be useful for evaluating operational trends within industries, as well as for analyzing fundamentals. This paper is a step in our efforts to develop more targeted, semantically-informed measures of operations to augment the financial data toolbox. In this paper, we employed a raw pre-trained model to form embeddings, and future research would benefit from fine-tuning on operationally meaningful targets, including financial fundamentals and market shares, which may produce even better separation across fundamental types.

The universe of semi-structured firm filings provides an ideal application for transformer models, which have allowed

us to dig deeper into the structure of firm operations than previous methods based on text. Through our approach, we have sought to introduce more robust and principled approaches to knowledge discovery from semi-structured financial data, and to contribute useful metrics to the broader research community in finance and computer science.

References

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Fertuck, 1975] Fertuck, L. (1975). A test of industry indices based on sic codes. *Journal of Financial and Quantitative Analysis*, 10(5):837–848.

[Hiew et al., 2019] Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., and Xu, Y. (2019). Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

[Hoberg and Phillips, 2016] Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

[Jawahar et al., 2019] Jawahar, G., Sagot, B., and Seddah, D. (2019). What does bert learn about the structure of language?

[Ke et al., 2019] Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

[Liu, 2019] Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

[Loughran and McDonald, 2016] Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance:: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Tetlock, 2007] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.