Submitted to *Management Science*
manuscript (Revision 2)

Using Representation Learning and Web Text to Identify Competitor Networks

Gerard Hoberg, Craig Knoblock, Gordon Phillips, Jay Pujara, Zhiqiang Qiu and Louiqa Raschid

Understanding the competitive landscape of public and private companies is essential for a range of activities. Prior work has often characterized competition using inflexible industry classifications or relied on proprietary data for public companies. The almost total lack of coverage of private companies has also been a severe limitation. This paper addresses these limitations by using a vast and untapped resource for understanding the competitor network for both public and private companies: Web text. We use representation learning techniques to generate robust representations (word embeddings) of companies in a high-dimensional vector space and to accurately identify the competitor network of peers for a focal company. We evaluate the competitor network against multiple downstream applications: (1) Predicting profitability; (2) Validating the self-identified competitors of a focal company; (3) Identifying In-Sector (IS) peers that occur in the same industry sector as the focal company; (4) Determining industry classification codes (NAICS) for a focal company. Our proposed approaches match or improve on the performance of prior baselines that relied on curated corpora, despite the use of noisy Web text. In particular, embedding based models outperform prior baseline models in identifying In-Sector (IS) peers, in identifying peers that are in the same NAICS sector as the focal company, and in predicting the NAICS code for private companies. A use case identifies scenarios where companies can gain the most benefit from high quality Web text about their products and services.

Key words: Word embedding; Large Language Models; Doc2Vec; Representation learning; Peer networks; Siamese network; Competitor identification.

Acknowledgments

Hoberg, Knoblock, Pujara, and Qui are from the University of Southern California, Phillips: Dartmouth College, and Raschid: University of Maryland. We thank colleagues and seminar participants for helpful comments. We alone are responsible for any errors and omissions.

1. Introduction

The task of creating a peer network for competitor identification is of general interest to researchers, analysts, investors, and regulators. For example, an entrepreneur wishing to deploy a new product, or to expand operations, must identify the most likely competitors. Regulators may target the potential for monopolistic behavior, and block proposed mergers when there are fewer competitors. Similarly, policymakers may wish to create incentives for entry into certain sectors, e.g., green energy. In all of the above scenarios, it is important to identify both public and (emerging) private competitors.

Competitor networks must support multiple use cases (Pant and Sheng (2015)). One use case is an insider approach where there is interest in developing a detailed profile for a focal company, and the competitor network provides relevant knowledge. An alternative is where there is interest in a specific industry sector or product-related value chain, and the peer network can be used to identify relevant companies and relationships, and the peer network can be used to identify relevant companies and relationships. Yet another common use case is to develop a panel of companies to be used for further analysis, e.g., a longitudinal panel to understand year over year financial performance. These use cases can also benefit from data on private companies, which are far more numerous and poorly understood, in comparison to public companies.

The task of competitor identification has its origins in the economics and finance literature (Hotelling (1929), Chamberlin (1933)). Past approaches have used industry classifications (Pearce (1957)) and / or financial metrics (Fama and French (1997), Bhojraj and Lee (2002)). Companion approaches in the finance literature extended the comparison from financial metrics to use information from regulatory filings. For example, the Text-Based Network Industry Classification (TNIC) (Hoberg and Phillips (2010), Hoberg and Philips (2016)) used the co-occurrence of product/service-specific terms in the SEC 10-K regulatory filings of public companies. More recent activity has occurred in the Information Systems research literature as companies expanded their Web footprints, providing new datasets and opportunities to study competition. Potential datasets

range from large textual corpora to financial news/articles to Web search logs (Bao et al. (2008), Bernstein et al. (2002), Ma et al. (2011), Pant and Sheng (2015), Wei et al. (2016)),

Data-driven models of market structure and competitors are most useful when they cover the entirety of the market participants. Unfortunately, the majority of readily available data and models only cover companies that are publicly listed on stock exchanges. In this paper, we present an approach to extend the coverage of market structure and competitor modeling beyond public companies to include privately held companies. We achieve this using a publicly available source of knowledge, i.e., Web text. Companies have a vested interest in providing Web text that provides rich detail explaining products and their features. Web text is publicly available and comparatively easy to process, e.g., using representation learning approaches. Most important, there is Web text available for an extensive set of private companies over time, and this increases the potential to have a broad impact.

By employing representation learning techniques over Web text, we learn the vector space embedding of each company. We use the embedding to identify the business focus of a company and its competitors. Our key contribution is to build a spatial competitor network that will place both public and private companies within the same high-dimensional space. The distance between a pair of embeddings can be interpreted as a spatial representation of the distance between any pair of companies. In this context, the spatial distance between a pair of companies that sells the same product or service, in the same market, i.e., direct competitors, should be small. In contrast, two companies selling unrelated products should have a high spatial distance.

We begin by creating a corpus of company Web text from over 800,000 companies using the Wayback Machine of the Internet Archive project. In addition to being able to cover both public and private companies, our collection also has a temporal dimension corresponding to the time period covered by the Wayback Machine. This will eventually allow us to model how the network of competitors and product offerings change over time. We note that temporal evolution is not addressed in this paper.

While Web text is widely available, extracting the product and service relevant vocabulary from company Web text poses multiple challenges. First, Web pages can have diverse structures within and across industries. Landing pages may vary in their objectives, from pointing to specific geographic sites to describing subsidiary organizations, to linking to product catalogs, to focusing on corporate values and employees. Web text from different industry sectors, e.g., legal or manufacturing, may also use similar terms but with different semantics. For example, the term `process` indicates different semantic intent in the two sectors, e.g., a legal process versus a manufacturing protocol. In addition, the vocabulary used in industries is not static. Over time, novel and disruptive companies develop, and new terms such as `ridesharing` or `wearables` are introduced. These must seamlessly be integrated into the vocabulary. For all of these reasons, navigating Web text to identify product and service information is challenging.

A second challenge is that even when product text can be collected successfully, computing the embedding-based similarities between hundreds of thousands of companies can pose a formidable big data scalability bottleneck. For example, a pairwise network with 800,000 public and private companies (as used in this project) could require hundreds of billions of pairwise similarity calculations. We note that a significant fraction of these similarity scores will be close to zero.

Tuning the competitor network to identify the most relevant peers is the final challenge. Consider the task of identifying competitors in the primary product or service sector of a focal company. In addition to product or service-specific information, company Web pages often also contain non-product or non-service information. For example, many companies may discuss their employees, corporate leadership values, and issues of sustainability and diversity. While this information is useful for other important research questions, this content is not specific to products and services and could lead to poor prediction of, for example, company profitability. To illustrate, we find that an "untuned" competitor network incorrectly identifies large numbers of peers that are not In-Sector (IS), i.e., the peer is not associated with products and services related to the focal company. Note that we use the (reduced) five Fama / French industry portfolios (Fama and French (1997))

to label In-Sector (IS) peers. An important contribution of our research is the ability to fine-tune the learned representations so that the peer network comprises mostly IS peers.

To illustrate the effectiveness of our learned representation model and the predicted competitor network, we address several downstream tasks. A key group of tasks is predicting financial performance, e.g., company profitability. A second is predicting the self-reported competitors that public companies identify in their annual SEC 10-K filings, or identifying In-Sector (IS) peers. Another group of tasks is predicting industry classification codes (NAICS or SIC codes) (Office of Management and Budget (2017), Pearce (1957), US Department of Labor (2022)). Our final task is a use case that explores scenarios when companies can benefit the most from an investment in generating high quality Web text about their products and services.

Our research makes the following contributions:

- We demonstrate that the learned representation-based approaches are remarkably effective. A Doc2Vec document embedding (Mikolov et al. (2013)), applied to the 2016 SEC 10-K corpus for 3800 public companies, and tested on profitability prediction, can improve on the results from the original TNIC benchmark (Hoberg and Philips (2016)) benchmark. While our main contribution is the use of Web text for both public and private companies, this result in improving the TNIC benchmark for public companies is nevertheless significant.
- The Doc2Vec embedding over the 2016 SEC 10-K corpus provides one baseline for comparison for our proposed approaches using Web text (upper bar). Another baseline will make use of the six-digit or four-digit NAICS codes to construct a peer network; we note that these latter two baselines do not perform well on the downstream tasks (lower bar).
- A Doc2Vec embedding that was trained on the Web text for the 3800 public companies, and another embedding that was trained on the Web text for thirty two thousand private companies, both experience a small drop in performance in comparison to the upper bar baseline. Our proposed solution - D2V+LDA - a Doc2Vec embedding that is further tuned so that the peer network primarily consists of IS peers, similarly shows a small drop in performance. All of the embedding-based approaches performed better than the lower bar.

- Notably, our method D2V+LDA showed the best performance at identifying In-Sector (IS) peers, across each of the five Fama / French industry sectors. For some Fama / French industry sectors, the ability to identify IS peers was exceptional. Similarly, D2V+LDA showed the best performance at identifying peers in the same NAICS sector as the focal company, when considering 2-digit to 6-digit NAICS codes. Again, the performance was exceptional to very good for several of the five Fama / French sectors. Our solution (and the other approaches) provided their best results for the `Health+` sector, while the `Consumer+` sector experienced the worst results.
- A very encouraging result is that D2V+LDA can predict NAICS classification codes for private companies with very high accuracy. We emphasize that our work is the first to attempt any predictions outside the narrow realm of publicly held companies in the US.
- A final contribution comes from a use case that provides insights from scenarios where companies can benefit the most from an investment in generating high quality Web text about their products and services. We use well studied measures from the finance literature to characterize companies. We identify the scenarios where the Web text of two companies *are more likely to predict that they are IS peers*. The use case reflects a virtuous cycle of features that increase the likelihood of identifying IS peers, in precisely those scenarios where accurately identifying IS peers may have the greatest benefit. Example scenarios include companies that are operating in a competitive product market, or a market with significant venture capital funding, or a market with product fluidity, as well as companies that target significant resources toward advertising.

The paper is organized as follows: Section 2 summarizes research on peer networks and competitor identification, and representation learning. Section 3 presents our approach including (1) Representation learning; (2) Link prediction; (3) Tuning the competitor peer network. Section 4 provides details of the dataset, tasks, and evaluation protocol, and Section 5 presents the results. Finally, Section 6 presents the use case and Section 7 provides a summary and discusses future research.

2. Related Work

We summarize the literature on competitor identification. The task is of general interest to researchers, analysts, investors and regulators. This research has its origins in the economics and finance literature, with more recent activity taking place in the Information Systems research domain, as companies expanded their Web footprints, providing new opportunities and datasets to study competition. Competitor networks are most useful when they cover the entirety of the market participants. Unfortunately, the majority of models rely on often proprietary data, and the networks are typically limited to public companies that are listed on stock exchanges. This is a surprisingly small segment of the true competitive landscape. In this paper, we present an approach to extend the coverage of market structure and competitor modeling beyond public companies to include privately held companies. We briefly summarize the extensive research on representation learning. Additional details on the specific embedding and learning approaches that we utilize will be included in the next section.

2.1. Competitor Identification

Competitor networks must support multiple use cases (Pant and Sheng (2015)). One use case is an insider approach where there is interest in developing a detailed profile for a focal company, and the competitor network provides relevant knowledge. An alternative is where there is interest in a specific industry sector or product-related value chain, and the competitor network can be used to identify relevant companies and relationships. Yet another common use case is to develop a panel of companies to be used for further analysis, e.g., to understand their comparative financial performance. Lastly, potential entrants and existing firms deciding whether or not to expand need to have information about existing private firm competitors.

The problem of characterizing competition in markets has been well studied, starting with seminal work in (Hotelling (1929), Chamberlin (1933)). Companies were characterized based on product (or service) descriptions that referenced industry categories or sectors. Competitors were also identified using financial metrics.

Among the earliest approaches to qualifying competition was the use of industry classification codes, such as the Standard Industrial Classification (SIC) (Pearce (1957), US Department of Labor (2022)). SIC has been replaced by the North American Industry Classification System (NAICS) (Office of Management and Budget (2017)). Both SIC and NAICS attempt to assign a hierarchical classification to companies based on business processes and outputs. A common criticism of such classification systems is the lack of flexibility since companies are assigned just one or a small number of codes. This approach may prevent the capture of the full breadth of their business operations for large companies with many subsidiaries. Moreover, membership in an industry sector alone does not capture the strength of competition between any pair of companies. Further, as new companies innovate and join the market, precise classifications may be difficult to determine, and the process of assigning classifications can be slow and labor-intensive.

Another line of research attempts to identify competitors based on financial metrics such as asset pricing (Fama and French (1997)), asset holdings (Rauh and Sufi (2011)), valuations (Bhojraj and Lee (2002), Bhojraj et al. (2003)), or the co-movement of stock prices (Yaros and Imieliński (2015)). The underlying premise of such models is that a latent industry assignment can play the role of an explanatory factor in predicting financial outcomes. An inferential process can potentially estimate industry membership or peer relationships. A primary drawback of such approaches is that they rely on financial data. While financial data is available for public companies, it is difficult to obtain or not publicly available for private companies.

Companion approaches in the finance literature extended the comparison from financial metrics, to use information from regulatory filings. For example, the Text-Based Network Industry Classification (TNIC) (Hoberg and Phillips (2010), Hoberg and Philips (2016)) used the co-occurrence of product / service-specific terms in the SEC 10-K regulatory filings of public companies. The primary drawback of TNIC and similar approaches is the limited coverage of the competitive landscape since most are limited to large public companies. The TNIC website[1] has attracted over

[1] See https://hobergphillips.tuck.dartmouth.edu/

55,000 visits from academic and practitioner users from all over the world since its launch in 2010. The research based on this work has also received over 4200 Google citations at the time of this writing.

The explosion of information available online about companies, ranging from company Web pages, to financial news / articles (Bernstein et al. (2002), Ma et al. (2011)), to Web forums (Xu et al. (2011)), to large textual corpora including Web search logs (Bao et al. (2008), Wei et al. (2016)), provided significant opportunities to explore competition further. This research has primarily been pursued in the Information Systems research domain. These approaches generally use the co-occurrence of company names in text or the co-occurrence of associated contextual terms, often augmented with metrics such as network centrality, to identify potential competitors of a focal company. Some limitations are that news articles predominantly cover large and highly capitalized companies. Web forums primarily provide data for companies with many consumer-facing products or services. Further, data sources such as Web search logs are difficult to obtain at scale for most researchers.

An excellent exemplar of such approaches is presented in (Pant and Sheng (2015)). Building upon the concept of isomorphism of competing firms, they explore and construct a parallel phenomenon of online isomorphism based on the Web footprints of companies. They propose several metrics to capture online isomorphism for a pair of companies as follows: The similarity of in links (outlinks) to/from other websites; The similarity of Web text (extending a well-known measure of document similarity); Co-occurrence in new articles or in search engine queries. They present an extensive evaluation of using online isomorphism to predict In-Sector (IS) peers; they explore a range of prediction models for their classifier, as well as variations of training data. While they do not address the task of profit prediction or prediction of NAICS codes, this research has many parallels to our research objectives. We do note that our approach of relying on Web text and learned representations has many practical advantages over the much more labor-intensive metrics that they utilize. Further, we address the task of NAICS prediction for private companies, which is unique to our research.

2.2. Representation Learning

Data-driven machine learning often uses a methodological approach based on representation learning to extract salient features from an underlying collection. When the collection is a document collection, there is a long history of methods inspired by natural language processing. The methods can range from traditional, token-based approaches that represent words in a document, to distributional approaches that develop probabilistic models of language patterns. For example, word distributions over topics are used in the Latent Dirichlet allocation (LDA) Topic Model (Blei et al. (2003)). Extensions include supervised topic models (Blei and McAuliffe (2007)) and structured topic models (Roberts et al. (2013)). More sophisticated representation learning methods are better able to identify nuanced features from word sequences, e.g., Doc2Vec document embeddings (Mikolov et al. (2013)) and RoBERTa (Liu et al. (2019)), which can handle sentence fragments such as phrases. We provide details of both in the next section. We note that there are very sophisticated state-of-the-art Large Language Models that can handle generative tasks such as question answering. Such advanced models are typically not needed for the relatively simpler task of finding competitors based on document (Web text) similarity. We note that these more sophisticated approaches have the limitation of requiring massive data collections for training. Further, supervised approaches such as RoBERTa that are trained on high quality document collections may not be able to improve on the performance of unsupervised approaches such as Doc2Vec, over noisy Web text collections. This will be discussed further in our evaluation.

3. Methodology

Our approach to creating a competitor network introduces a multi-stage pipeline as illustrated in Figure 1. The first stage of the pipeline processes Web text and applies one of several **Representation Learning** techniques to learn a high-dimensional vector for each company. The second stage of the pipeline uses the learned vector representations for **Link Prediction** to generate a set of potential competitors using the spatial organization of the vector space. The third stage is **Tuning** to increase the proportion of In-Sector (IS) peers. The final stage of the pipeline, **Evaluation**, produces a set of tangible, economically-grounded predictions such as the profit prediction of a company, predicting its SIC or NAICS code, or predicting a set of peers identified by analysts.

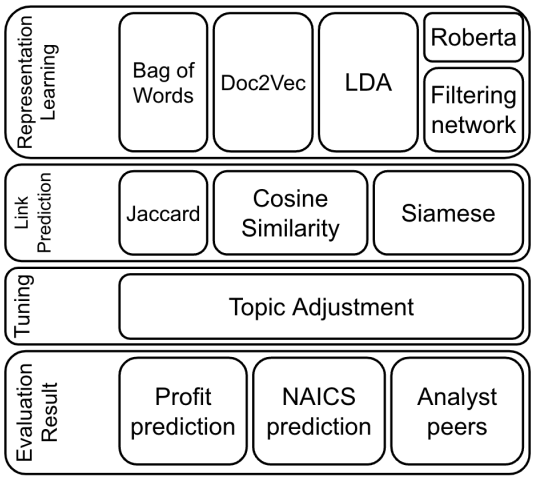

Figure 1 Architecture of our pipeline

3.1. Problem Definition

Given a set of entities, E, the goal is to predict a set of competitor links, $L(e_i, e_j), \text{s.t.} e_i, e_j \in E$. In this paper, we assume that each entity e_i, has an associated corpus of documents, D_{e_i}. Our approach is to use a representation learning algorithm, REPR, and the corpus, to learn a k-dimensional representation, $R_{e_i} \in \mathbb{R}^k$, such that $R_{e_i} = \text{REPR}(D_{e_i})$.

Next, these k-dimensional representations are used as input to a competitor link prediction algorithm, LP, to generate a set of pairwise scores between entities, $s(e_i, e_j) = LP(e_i, e_j)$. Finally, we apply a decision function, D, over the scores to determine the links to include in the competitor network: $\forall_{i,j} L(e_i, e_j)$ iff $D(s(e_i, e_j))$.

3.2. Representation Learning

In this paper, we experiment with several well-established approaches for representation learning. These include baseline models that use a bag-of-words representation, a probabilistic approach using latent Dirichlet allocation (LDA), and two deep learning approaches, Doc2Vec and RoBERTa.

3.2.1. Bag of Words: Prior work (Hoberg and Philips (2016)) represented companies by the words that occurred in the business description section of the annual SEC 10-K filings. Filters were applied to remove high-frequency terms and retain only nouns and proper nouns. We do not use any additional filtering, e.g., parts of speech. Each company will be represented by a vector of word occurrence, and these vectors will be normalized to have unit length.

3.2.2. Latent Dirichlet Allocation: Topic modeling is a popular approach to identifying the latent feature space, or topics, with a large document corpus. Topic models typically use unsupervised probabilistic models to associate individual documents with a probability distribution over latent groups, known as topics. Latent Dirichlet allocation (LDA) (Blei et al. (2003)) is a sophisticated probabilistic model that has served as the foundation for much recent work in topic modeling. The method, defined for k topics and $|D|$ documents, models topics as distributions over words (β_k), and document as mixtures of topics (θ_d). For each word $w_{d,n}$ in a document $d \in D$, a topic $z_{d,n}$ is chosen from the document's topic distribution (θ_d), and a word is chosen from the topic's distribution over words (β_k). The Dirichlet distribution of θ_d and β_k are parameterized by terms α and λ respectively. This generative approach is reflected in the joint distribution shown in Equation 1. Topic models are learned through inference on the generative model described. This inference problem is intractable at scale, and approximate techniques such as Gibbs Sampling or variational methods are used during learning (Blei et al. (2003)).

$$p(w, z, \theta, \beta|\alpha, \lambda) = \prod_k p(\beta_k|\lambda) \prod_d p(\theta_d|\alpha) \prod_n p(z_{d,n}) p(w_{d,n}|\beta_{z_{d,n}}) \tag{1}$$

To use topic models for representation learning, we use the inferred topics for a given document as the learned representation. Since each company Web site is associated with many web pages, we concatenate the content of these pages into an aggregate document and we learn a distribution over the aggregated document. Details of parameters for model training are in the evaluation section.

3.2.3. Doc2Vec: Word embeddings represent each word as a vector of a m-dimensional vector, most commonly in $\mathbb{R}^m$. Many approaches have been proposed for training word embeddings, spanning from early work in latent semantic indexing (Hofmann (2017)), to more recent tools such as Word2Vec (Mikolov et al. (2013)) and GloVe (Pennington et al. (2014)). We adopt a variant of the Word2Vec model, Doc2Vec (Le and Mikolov (2014)). Doc2Vec differs from Word2Vec in that it uses a global contextual vector associated with the document as part of its predictive model.

The training objective of Doc2Vec is to learn word vector representations that are good at predicting nearby words. In the model, $v(w) \in \mathbb{R}^m$ is the vector representation of the word $w \in W$, where W is the vocabulary and m is the embedding dimensionality. $v(e_i) \in \mathbb{R}^m$ represents a latent vector associated with each document. Given a pair of words (w_t, c), the probability that the word c is observed in the context of word w_t in a document associated with a latent vector e_i is as follows:

$$P(c = 1 | v(w_t), v(c), v(e_i)) = \frac{1}{1 + e^{-v(wt, e_i)^T v(c)}}$$

Given a training set containing the sequence of words, the word and document embeddings are learned by maximizing the log-likelihood score. For the collection of Web pages for each company, we learn a single latent document vector for the company, as well as the vectors for each of the individual words across all the pages. Through this process, each company will be associated with a vector that models the pattern of contextual word co-occurrence across the Web pages for that company.

3.2.4. RoBERTa: The final representation learning approach we apply is RoBERTa, a contextual, transformer-based pre-trained language model. In contrast to Doc2Vec, which is primarily trained to learn word representations from a fixed set of contextual words, RoBERTa is designed to learn representations based on a sequence of contextual tokens, resulting in a different representation for a word based on its surrounding context. Using RoBERTa, we map a sequence of words, $w_1, w_2, ..., w_n$ to a vector in $\mathbb{R}^m$. A limitation of RoBERTa is that the sequence of tokens for which it can generate representations is limited to 512 tokens. Most websites will yield many chunks of 512 tokens, and the diversity of chunks makes generating a combined representation difficult. Alternatives such as LongFormer (Beltagy et al. (2020)) have been proposed to overcome this limitation; however, such approaches require substantially more computational resources.

To overcome the limitation of 512 tokens, we develop a filtering technique to identify the most relevant portions of a company Web page. To do this, we train a classifier to differentiate between

text that characterizes a company's product or services versus less relevant text. As positive examples of text, we use the first paragraph of the business description section in SEC 10-K filings. As negative examples, we use documents from an NLTK standard corpora (Loper and Bird (2002)), e.g., the Brown Corpus. We introduce a filter head (Figure 2), which is trained to classify these documents into positive and negative classes. Using the RoBERTa representations from the CLS tag of each document, we train a fully connected multi-layer Perceptron to predict the class of each document.

We incorporate the trained filter head and classifier in the RoBERTa model as shown in Figure 3. Each document is first split into smaller chunks of 512 tokens. These chunks are each assigned a probability of being relevant to the company's business function. We rank these chunks and utilize the CLS embeddings of the top K chunks for each company.

Let $\vec{Z}^j$ be the output vector of the filter head for chunk j, $\vec{Z}^j = f(c_j) = [Z_+^j, Z_-^j]$. Let S_+ and S_- be the (softmax) scores to predict that the chunk is business relevant (+) or general text (-). We define the relevance score for chunk j, denoted as RS_j, as follows:

$$RS_j = \frac{e^{Z_+^j}}{e^{Z_+^j} + e^{Z_-^j}}$$

To maintain an equitable representation across documents, we use RS_j to order all token chunks associated with the document, and select the top K scoring chunks. These chunks are combined by computing the arithmetic mean over the k chunks, producing a vector in $\mathbb{R}^m$.

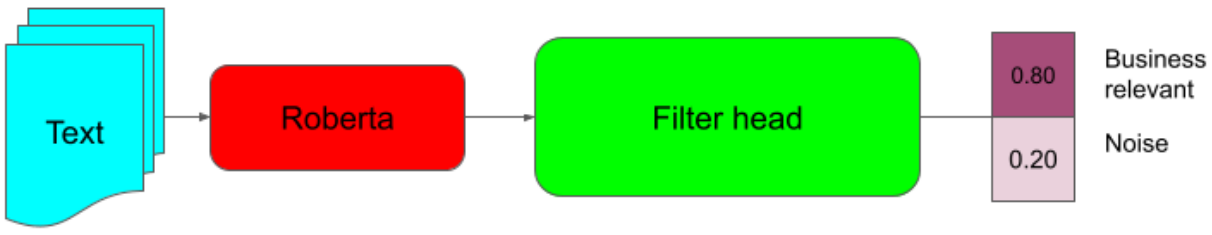

Figure 2 Filtering technique to classify relevant chunks of a long document for use in RoBERTa.

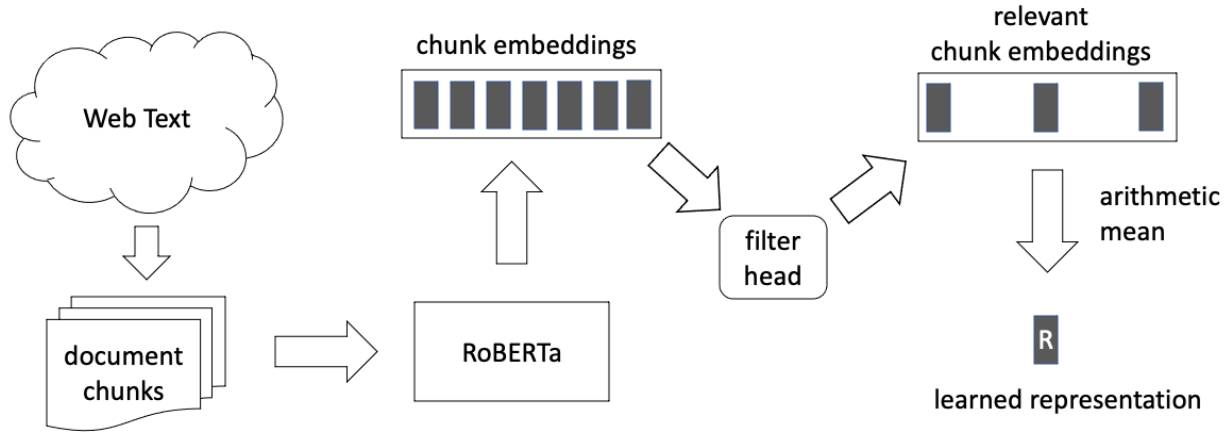

Figure 3 **Protocol to create a single representation R from the chunks of a long document using RoBERTa.**

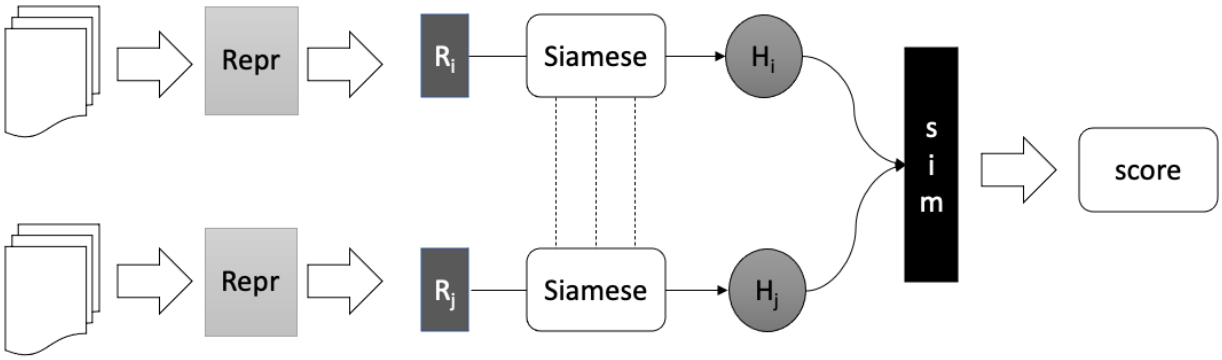

Figure 4 **Predicting the peer similarity using Siamese style neural network(the dashed line show shared weights)**

3.3. Link Prediction

The task of predicting a relationship between two entities is known as *link prediction*. We cast the problem of inferring the competitor network between companies as a link prediction task. In this setting, given a learned representation, e_i, e_j, for entities i and j, the goal is to use an algorithm, LP, that will assign a probability that the two entities have a competitive relationship, $s(e_i, e_j) = LP(e_i, e_j)$.

We explore two different strategies for link prediction. The first strategy is an unsupervised approach that uses cosine similarity to score competitor pairs. This approach relies on the similarity of the learned representations in high-dimensional space. The second strategy is a supervised approach to learning a mapping from the learned representation to a competitor score.

The advantage of unsupervised approaches is a lack of reliance on training data. Training data for understanding competition is scarce. The majority of training data is biased towards larger publicly traded firms. There are many diverse definitions of competition. Thus, predictive approaches that use training data may be biased towards a particular definition based on the training data that was used and may need to be re-trained. Conversely, representation learning captures a vast amount of information, some of which may be irrelevant for assessing competition. Thus, unsupervised approaches may be limited in the specificity of their predictions.

In both of these approaches, we use the concepts introduced in the prior section. For a given company entity, e_i, we use an associated set of documents (d_i) to learn a representation $R(d_i)$. Next, we define a competition score for entities e_i and e_j, we define $s(e_i, e_j)$.

3.3.1. Unsupervised: Cosine Similarity The guiding premise of unsupervised prediction is that documents containing similar information should have similar representations in the high-dimensional vector space. In our setting, we expect that companies that have similar products, and operate in similar markets, should have similar information and structure in their company websites. Thus, we use similarity in the learned representation as a proxy for competition. We can then expect that greater accuracy of the learned representation of documents will lead to greater accuracy of the predicted competitor network, for downstream tasks.

Cosine similarity is an unsupervised approach and is known to be a reliable metric to measure the similarity of the embeddings in a high dimensional space. We use cosine similarity to calculate the pairwise similarity between the learned vector representations. We define cosine similarity with respect to two learned representations $R(d_i)$ and $R(d_j)$ as:

$$sim_{i,j}^{cos} = \frac{R(d_i) \cdot R(d_j)}{||R(d_i)||||R(d_j)||}$$

3.3.2. Supervised: Siamese Network We also consider a supervised approach to learn the similarity of two learned representations. A common approach is based on the Siamese network, (Roy et al. (2019)). Siamese networks consist of two symmetric neural networks with tied weights.

They can transform input representations to align with a particular distance metric. Symmetry and tied weights ensure that distances are symmetric and the order in which two instances are presented does not alter their distance. Siamese networks have proven to be very useful in one shot learning problems and matching.

The training and prediction procedure for the Siamese network is depicted in Figure 4. We use the Siamese network to predict the distance between two representations $R(d_i)$ and $R(d_j)$. The network produces a hidden representation h_i, h_j for the respective document representations. It then computes a weighted pairwise distance, $\alpha \cdot h_i \cdot h_j$, over the hidden representations. In our experiments, the objective of the Siamese network is to predict the pair-wise TNIC score for pairs of public firms, computed from the SEC 10-K filings (Hoberg and Phillips (2010), Hoberg and Philips (2016)). The network is trained to minimize the RMSE loss between the predicted and actual similarity scores. We discuss the training procedure and evaluation of the model in detail in the experiment section.

3.4. Tuning the Competitor Network to Favor In-Sector (IS) Peers

The next stage of the pipeline is to tune the competitor network. Accurately identifying peers in the (primary) product or service sector of a company, i.e., In-Sector peers or IS peers, is key to the challenge of using the competitor network to make good predictions or to create cohorts of companies for economic and financial research. However, we observe that the competitor network can include peers that may be outside the primary product or service sector of a focal company. A preliminary evaluation of peers that were chosen based on the competitor network using cosine similarity over the Doc2Vec embedding revealed approximately 30% of non-product or non-sector peers. In contrast, the baseline TNIC network (Hoberg and Phillips 2010, Hoberg and Philips 2016) included approximately 4% of non-product or non-sector peers. We used the Fama-French-5 sectors (Technology, Manufacturing, Health, Consumer, Other) for this labeling of IS peers.

Including non IS peers could confound the information in our spatial model of the competitor network. Thus, a key contribution of our research is the methods used to tune the learned

Table 1 LDA selected topics that have more than 100 documents belonging to it out of 3800 public companies

Topic ID	Topic	Topic Words with high ϕ value
11	Debt	"expense" + "consolidated" + "debt" + "adjusted" + "fair"
19	Medical	"clinical" + "pipeline" "phase" + "patients" + "cancer"
52	News	"innovation" + "diversity" + "culture" + "newsroom" + "digital"
59	Measures	"kb" + "fiscal" + "diluted" + "adjusted" + "ir" + "lookup"
73	Banking	"banking" + "loans" + "checking" + "savings" + "loan"
84	Electronics	"devices" + "wireless" + "smart" + "series" + "digital"

representations and the resulting competitor network in order to ensure that peers are primarily determined using information about products and services. We use the results of the LDA-based topic models to tune the Doc2Vec competitor network. In the later evaluation section, we refer to the tuned competitor network inferred from adjusting the cosine similarity score of Doc2Vec embeddings using LDA topics as D2V+LDA.

Table 1 illustrates six topics and the keywords with the highest probability for those topics; the details of constructing the LDA model are in the next section. The Topics 19, 73 and 84 appear to be product specific, whereas topics 11, 52, and 59 appear to represent non-product topics that companies may wish to feature on Websites. Other notable non-product topics include countries of operation, eco-conscious practices, diversity in hiring, etc. We briefly discuss the following steps taken to use the LDA based topics to tune the Doc2Vec competitor network to produce D2V+LDA:

- From the 500 LDA-based topics produced from the 2016 Web text, we select the Top K significant topics, based on the topic weight distribution for the 3800 public companies across the five Fama / French sectors. We note that the Top 20 topics appeared to be a good threshold for tuning; including more topics did not have much impact.
- For each pair i and j of public companies with a Doc2Vec based cosine similarity score of $s_{i,j}$, we set the value of $f_{i,j}$ to 1 if the two companies are in different sectors; $f_{i,j}$ is 0 otherwise. We then estimate the following regression where $T_{i,k}$ and $T_{j,k}$ are the topic weights for companies i and j (of the 3800 public companies) for topic k (of the 20 topics), respectively:

$$f_{i,j} \times s_{i,j} = a_0 + \sum_{k=1}^{20} (a_k (T_{i,k} \times T_{j,k}) + \epsilon_{i,j}$$

- The predicted values from this regression, $s^{'}_{i,j}$, now represent the contributions from the LDA topics to pairs of companies that are not in the same sector. Thus, to obtain the adjusted tuned score of $s^{t}_{i,j}$, we subtract $s^{'}_{i,j}$ from the original similarity score of $s_{i,j}$. Next, we apply the same regression coefficients to the entire dataset of public and private companies to obtain the adjusted tuned scores for (public, private) and (private, private) competitor pairs.

3.5. Generating the Peer Network

The final step of our pipeline is generating a peer network. Specifically, we introduce a competitor link $L(e_i, e_j)$ between entities i and j, based on a decision function D over all entity scores: $\forall_{i,j} L(e_i, e_j)$ iff $D(s(e_i, e_j)$.

Our approach to determine which links to include uses a global scoring approach. We generate all pairwise company scores and retain the top X% scoring links. This procedure ensures that the number of peers generated is consistent across methods since it produces a fixed granularity of peers. Further, it naturally adapts to variability in peer sizes since larger firms may have many higher-scoring competitors, while smaller firms may have fewer competitors. We found that a threshold of top 2% worked well and provided a competitor network that had similar (good) properties compared to the TNIC network.

4. Evaluation Protocol

We consider the following evaluation tasks to compare the effectiveness of the learned representations and the competitor network:

1. Company profitability prediction.
2. Self-reported competitor prediction.
3. In-Sector (IS) peers.
4. Identification of the industry classification (NAICS) codes for the company.

The 10-K is a comprehensive report filed annually by a publicly-traded company about its financial performance and it is required by the U.S. Securities and Exchange Commission (SEC).

The 10-K was the primary source of ground truth for public companies. Proprietary databases from CapitalIQ and Orbis were the source of (limited) ground truth for private companies. In the rest of this section we provide details of the ground truth and the evaluation process. Then we describe the data pre-processing and experiment setup for these tasks.

4.1. Ground Truth Labels and Prediction Tasks

Profitability Prediction Task for Public Companies:

Following the approach of (Bhojraj and Lee 2002, Bhojraj et al. 2003), we evaluate the effectiveness of our competitor relationships based on the ability to predict financial outcomes using these competitors. For each company, c_i, we define financial metric(s) of interest, F. One metric is the *return on net operational assets (rnoa)* or the asset-adjusted profits. The second commonly used metric is the *profit margin (pm): net income over net sales*. We compute these metrics using the competitors R_i for company c_i; the competitors are generated as described in §3.5. We fit a regression model to estimate $\hat{F}$ based on the average metric value, $\overline{F(R_i)}$, over all competitors, and we learn values for λ and c:

$$\hat{F}(c_i) = \lambda \overline{F(R_i)} + c$$

We then evaluate the quality of the set of competitor relationships on the basis of the coefficient of determination:

$$R^2 = 1 - \frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{\sum_i (F(c_i) - \overline{F})^2}$$

A high value for R^2 suggests that the regression over the profits of the competitor network successfully explains the observed financial results for the focal company c_i.

We also consider the quality of the prediction for individual companies c_i by using the root mean square error (RMSE) as follows:

$$RMSE = \sqrt{\frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{N}}$$

Ground truth for Self-Reported Competitor Prediction for Public Companies:

Publicly traded companies typically self-report their competitors in the SEC 10-K annual filing. A company, e.g., a bank, may compete with all other companies in their sector,e.g., other banks. Typically, each company may strategically list only some of its true competitors, and it may also report some false competitors. Competitors may be identified and discussed in Item 1A (Risk Factors) or Item 7 (Management's Discussion and Analysis).

Competitor Prediction Task:

As discussed in §3.5, we generate a competitor network based on a fixed network granularity of 2% or 1%. Our method generates an initial symmetric bidirectional network; however, the 2% or 1% subset that is chosen may not be symmetric. In the real world, we expect many competitor relationships between firms to be bidirectional. However, based on the self-reported competitors, less than half of the competitor relationships in our ground truth are bidirectional. For this evaluation, we considered both a directed and bidirectional version of the ground truth relationships; we did not observe significant differences with respect to the different approaches. We report on the results from the bidirectional ground truth.

Monopoly Rate:

We define the monopoly rate as the fraction of companies with no competitors, or isolated nodes with no edges in the peer network. A typical shortcoming of embeddings is that there are dense sections of the embedding, resulting in many competitors for some subset of companies, as well as sparse sections, resulting in the failure to predict competitors for some other subset of companies. High monopoly rates may suggest shortcomings of our approach to predicting competitors using embeddings.

Ground Truth for In-Sector (IS) Peers

Portfolio analysis over equity (stock) in public companies is typically performed using the Fama / French factor model. The model has been evaluated against multiple industry sectors, e.g., up to 49 sectors; each sector is typically identified using SIC codes (Fama and French (1997)). For

simplicity, we use a reduced grouping of the following five industry sectors to determine In-Sector (IS) peers:

- (Consumer) Durables and Non-Durables; Wholesale; Retail; Some Services.
- (Manufacturing); Energy; Utilities.
- (HiTec): Business Equipment; Telecommunications; TV; Control; Computational Services.
- (Health); Healthcare; Medical Equipment; Drugs.
- (Other) Multiple sub-sectors; Finance.

Ground Truth for NAICS and SIC Codes:

NAICS is a six-digit hierarchical classification system from the US Census Bureau (Office of Management and Budget (2017)). The 2-digit prefix represents a set of 22 categories; each subsequent digit beyond the 2-digit prefix represents sub-categories. SIC is a 4-digit code (Pearce (1957)) from the US Department of Labor. Publicly traded companies include NAICS and SIC codes in their 10-K filings. Private companies do not have to self-report their industry classification. Several widely used databases such as CapitalIQ and Orbis report on NAICS and SIC codes for private companies. We will focus on private companies where the NAICS code has been validated by two sources or where it is obtained from a single high-quality source. We note that the US government wishes to adopt the NAICS code universally, but SIC codes are used more widely in industry for historical reasons. This push for widespread adoption increases the importance of identifying the NAICS sector for private companies.

NAICS Code Prediction Task: We report on the percentage of peers who share the same NAICS code. A majority vote over the NAICS codes of these peer companies will then be used for prediction. We first consider the 2-digit prefix and continue until we reach the complete six-digit NAICS code.

4.2. Data Pre-processing

The primary source for historical company Web pages and Web text is the Wayback Machine project of the Internet Archive. Our first corpus comprises the Web pages of approximately five

thousand public companies and 800,000 private companies, over a span of up to 20 years. We identified approximately 800,000 private companies and their URLs from the Orbis and CapitalIQ databases. We then crawled the Internet Archive using the URLs. The Internet Archive protocol for crawling follows a set of standards that produces high-quality results. For the limited number of cases where the crawl was not successful, we supplemented the corpus with additional pages from the company Web site.

A second corpus comprises the 10-K annual filings with the SEC, collected following the protocol in Hoberg and Philips (2016). We report on results for the two corpora from 2016.

We apply the following pre-processing steps:

- SEC 10-K filings: We included words formed with letters from the English alphabet and excluded other text or characters, including numbers, punctuation, and special characters.
- Web text: The raw HTML files are processed using the Beautiful Soup library; this step removes tags, javascript code snippets, etc. We use all pages up to a depth of three from the root page. We replace all special characters, characters not in the English alphabet, numeric digits not in 0-9, and punctuation with spaces. We replaced longer white space and tabs with a single space. We retain periods(.), commas(,), and quotes (",').

We experimented with different document-cleaning processes. This included using space-separated text without any alterations to the raw text; replacing all punctuation and special characters with spaces; replacing only selective punctuation with space; leaving some special characters undisturbed. We included the English alphabet, numbers, and a few punctuation and special characters, e.g., period(.), comma(,), and quotes (",'). These variations produce comparable results with less than 1% variation.

From the SEC 10-K filings corpus for 2016, we curated a subset of 3800 public companies and report on the results of the evaluation tasks for this subset. The subset met the following conditions:

1. The company Web text contained meaningful content.
2. The company reported at least one NAICS code.

3. The company reported profitability data.
4. At least one competitor was included.

For the private companies, we randomly sampled a subset of 32,000 companies from the corpus of approximately 800,000 companies. The sampling was performed so that the distribution of companies over the five Fama / French sectors was similar for both the public and private companies.

4.3. Parameters for Model Training

Doc2Vec: We use the Gensim Doc2Vec implementation for training the document embedding (Rehurek and Sojka (2011)). The best combination of hyperparameters was 100 epochs of training, the paragraph vector (PV-DBOW) setting, a dimensionality of 300, 1e−5 downsampling, and 10 negative sampling (Mikolov et al. (2013)). We observed that using negative sampling over hierarchical softmax (Mikolov et al. (2013)) performed better for the profit prediction task. We report on the model constructed from the 2016 dataset.

LDA: The key parameter to tune the LDA models is the vocabulary size of the corpus. We limited the vocabulary to only include words that occur in at least two documents; we also remove words that occur in more than 20% of the documents. We limit each document (company) to 2400 words. Documents are tokenized into a bag of words (BOW). The model was trained with the Gensim ldamulticore model with 100 topics and 200 passes. We also experimented with 500 topics when we used the topics to tune the Doc2Vec / Cosine similarity scores to produce the D2V+LDApeer network.

RoBERTa: We use the fairseq implementation of a pre-trained RoBERTa model (Ott et al. (2019)). For the SEC 10-K filing document corpus, we used the text from Item 1 (Business). We did not filter out any special characters and we used word chunks of 500 words.

For the Web text corpus, we applied a trained filter head over the document content, and we filtered out only the top 10 most relevant word chunks of 500 words each. The protocol to train the filter was described in §3.2. We used 3943 instances of SEC filings (positive labels) and 3670

instances of general documents from the NLTK Brown corpus (negative labels). We used the Adam optimizer and the PyTorch platform with cross-entropy loss (Paszke et al. (2019)). We performed a five-fold cross-validation on the filter head to assess its accuracy. We were able to achieve an accuracy of 99.94 ± 0.08 showing that the filter head can easily differentiate between relevant documents discussing products and services and noisy or irrelevant documents.

Siamese network:

The Siamese network was trained using the Doc2Vec embedding over the SEC 10-K filings, and the Doc2Vec embedding over the Web text, for the 3800 public companies, as input. The network was trained to predict the ground truth - the TNIC similarity score between each pair of companies, computed using the 10-K filings. The model was trained to minimize the root mean squared error (RMSE loss) between the predicted and the actual TNIC scores. Since most pairs of companies were not similar, the training data had a strong bias toward low or zero-valued TNIC scores. The model was able to achieve an R-squared (RSQ) prediction accuracy of 87.93% for predicting the TNIC values, when it was trained on the SEC 10-K filings. The RSQ value was 81.19% when the model was trained using the Doc2Vec embedding over the Web text. This is not surprising since both the ground truth TNIC similarity score and the embedding was computed over the same corpus of the SEC 10-K filings in the former case.

5. Evaluation Results

We first summarize the results and then provide detailed explanations.

5.1. Results Summary

Tables 2 and 3 report on the results *for public companies* for the tasks of (1) profit prediction and (2) predicting the self-reported competitors. The models make predictions for 3800 public companies. The top panel of Table 2 reports on using an embedding over the SEC 10-K filings for prediction. The second panel reports on an embedding over the Web text for the 3800 companies. The third panel reports on an embedding over the Web text for thirty two thousand private companies. Finally, the bottom fourth panel reports on a baseline that uses the NAICS codes to

identify the peer network. Table 3 explores these results in depth and reports on the R^2 and the $RMSE$ metrics for a subset of the approaches.

- We demonstrate that the learned representation-based approaches are remarkably effective for these tasks. A Doc2Vec document embedding (Mikolov et al. (2013)), applied to the 2016 10-K corpus for 3800 public companies, and tested on profitability prediction, improved on the results from the original TNIC benchmark (Hoberg and Philips (2016)) benchmark. This baseline serves as an upper bar for performance. The baseline that uses the (six digit and four digit) NAICS codes to create the peer network did not perform well; this is in the lower fourth panel of Table 2 and serves as a lower bar for performance.
- The Doc2Vec embeddings trained on the Web text for the 3800 public companies and on the Web text for thirty two thousand private companies experiences a small drop in performance. Similarly, our proposed solution - D2V+LDA - a Doc2Vec embedding, that is further tuned so that the peer network primarily consists of IS peers similarly, has a small drop in performance.

Tuning the peer network to identify (3) In-Sector (IS) peers is important. Tables 4 and 5 report on the performance of the different approaches for this task. Notably, our method D2V+LDA showed the best performance at identifying In-Sector (IS) peers across each of the five Fama / French industry sectors. For some Fama / French industry sectors, the ability to identify IS peers was exceptional.

Tables 6 and 7 report on the quality of the peer network with respect to the task of (5) predicting the NAICS sector. As before, our method D2V+LDA showed the best performance at identifying peers in the same NAICS sector as the focal company, when considering 2-digit to 6-digit NAICS codes. Again, the performance was exceptional to very good for several of the five Fama / French sectors. Our solution (and the other approaches) provided the best results for the `Health` sector, while the `Consumer` sector experienced the worst results. A possible explanation is that healthcare companies use a distinct specialized vocabulary for their products and services, whereas consumer offerings are diverse and may not use a specialized vocabulary.

Table 2 Profit Prediction and Competitor Prediction for Public Companies

Model	Similarity	Average Profit R-squared	Monopoly Rate(%)	Competitor Recall(%)
Embedding over the SEC 10-K filings				
TNIC	Jaccard	0.3666	17.63	31.39
BoW	Jaccard	0.3605	43.30	24.90
Doc2Vec	Cosine	**0.4750**	**0.03**	**47.94**
Doc2Vec	Siamese	0.3870	3.24	31.73
LDA	Cosine	0.3668	4.37	37.99
RoBERTa	Cosine	0.3456	19.81	33.76
Embedding over Web text for 3800 public companies				
BoW	Jaccard	0.2497	34.20	13.53
Doc2Vec	Cosine	**0.418**	**0.34**	**30.20**
Doc2Vec	Siamese	0.2823	3.81	23.26
LDA	Cosine	0.2858	0.89	19.88
RoBERTa	Cosine	0.3455	13.60	11.10
Embedding over Web text for thirty two thousand private companies				
D2V+LDA	Cosine	**0.379**	**0.684**	**32.36**
Alternate Baseline using the NAICS Code to Select Competitors				
NAICS 6 digit		0.354	6.37	36.27
NAICS 4 digit		0.343	2.18	30.4

Finally, Table 8 reports a very encouraging result: D2V+LDA can predict the NAICS sector for private companies with very high accuracy. We emphasize that our work is the first to attempt any predictions outside the narrow realm of publicly held companies in the US.

5.2. Profit, Monopoly Rate and Competitor Prediction for Public Companies

Table 2 reports on the performance of a range of models and datasets for the tasks of (i) profit prediction and (iv) predicting the self-reported competitors. All results are reported for the 3800 public companies. We report on the R^2 value, averaged over the *return on net operational assets (rnoa)* - labeled `Profit / Assets` and the *profit margin (pm): net income over net sales* - labeled `Profit / Sales`. Table 3 reports on these metrics separately and in greater detail.

The top panel reports on models constructed using an embedding over the the 2016 SEC 10-K filings. The Doc2Vec learned representation, using Cosine similarity to construct the competitor network, significantly outperforms the TNIC benchmark (Hoberg and Philips (2016)) . We note that TNIC uses a bag-of-words approach and Jaccard similarity without any sophisticated techniques for filtering and normalizing data. The R^2 value for the Average Profit prediction increases from 0.3666 (TNIC) to 0.4750 (Doc2Vec/Cosine). Further, Doc2Vec/Cosine also improves on the monopoly rate (0.03% versus 17.63% for TNIC) and recall for competitor prediction (47.94% versus 31.39% for TNIC).

A surprising result is that the other sophisticated approaches, e.g., RoBERTa and Doc2Vec combined with a Siamese network, do not appear to perform very well, and they do not outperform the simpler TNIC benchmark. A possible explanation is that the Siamese network was trained to predict the TNIC similarity score. This may have resulted in limiting the prediction performance of the Doc2Vec/Siamese peer network, to be no better compared to the TNIC peer network. The RoBERTa model used a pre-trained set of parameters that was not tuned for the specific Web text dataset and task; this may explain its mediocre performance. We discuss these limitations at the end of this section.

The second panel of Table 2 reports on models constructed using the Web text for the 3800 public companies. We observe that Doc2Vec, using Cosine similarity to construct the competitor network, continues to demonstrate the best performance. In comparison to the top panel, the Average R^2 for profit prediction reduces somewhat from 0.4750 (using the 10-K filings) to 0.418

. The recall for competitor prediction for Doc2Vec/Cosine is 30.2% compared to 31.39% of TNIC and 47.94% Doc2Vec/Cosine/10-K filings (top section). These lower evaluation results reflect some degradation of signal quality when going from the SEC 10-K fully curated text to noisy Web text. Of note, however, is that the monopoly rate using Web text continues to remain low. Overall, the results reflect that the Web text corpus captures sufficient signal, despite the noise, to be able to construct a rich competitor network comparable to the peer network obtained by using the curated SEC 10-K corpus. We can infer that the representation learning approaches are thus able to overcome the noise and variance of Web text. As before, we note that RoBERTa and Doc2Vec/ Siamese do not outperform Doc2Vec/Cosine.

The third panel of Table 2 shows the performance as we expand the training set to include thirty two thousand private companies. The companies were randomly sampled (from a collection of 800,000 private companies) to reflect a similar distribution over the five Fama / French sectors, in comparison to the 3800 public companies. Our method D2V+LDA - Doc2Vec/Cosine tuned using LDA topics - has an Average R^2 value of 0.379 for profit prediction. This Average R^2 value is a further small reduction from the performance of Doc2Vec/Cosine with the public Web text for 3800 companies. The monopoly rate remains low while the recall for competitor prediction increases to 32.36%.

The bottom fourth panel of Table 2 reports on a baseline that uses the four-digit and six-digit NAICS codes to construct the peer network. This baseline sets a lower bar for performance; the Average R^2 value is 0.354 (six-digit) and 0.343 (four-digit). This baseline also has relatively high monopoly rates, i.e., where no competitors are identified in the peer network; the monopoly rates are 6.37% (six-digit) and 2.18% (four-digit).

Table 3 drills deeper to compare the performance of the embedding over the Web text for 3800 public companies (top panel) to the performance of the embedding over the Web text for the thirty two thousand private companies (lower panel). We report on the R^2 and $RMSE$ results for the two measures of profit prediction. One is the *return on net operational assets (rnoa)* -

labeled `Profit / Assets` and the second is the *profit margin (pm): net income over net sales* - labeled `Profit / Sales`. This evaluation also includes two additional features to characterize each company, as follows: 1. `logassets:` This is the natural logarithm of the assets of the company plus one. 2. `logage:` This is the natural logarithm of the age of the company plus one, where age is measured as the number of years since the company's initial appearance in the Compustat database. We also report on the percentage of In-Sector (IS) peers.

Doc2Vec/Cosine, with an embedding over the 3800 public companies, has the highest values of R^2; the values are 0.503 and 0.447, respectively, for the two metrics. We note that, as expected, these values are higher than the R^2 reported in Table 2. The addition of the two extra features (assets and age) has a positive impact on predicting profitability. The R^2 values for D2V+LDA are slightly lower. When considering the $RMSE$, there is low variance across the approaches and datasets. Finally, D2V+LDA shows the best performance in identifying In-Sector (IS) peers in the peer network; this task is explored in detail in the next section.

The behavior of these variants is consistent with our knowledge of the strengths and weaknesses of the underlying approaches, and the characteristics of the datasets. Some intuitive observations and explanations are as follows:

- The outstanding performance of the benchmark variant that constructs a Doc2Vec embedding over the 10-K filings is consistent with expectations. While there is no strict template for these filings, public companies typically follow a similar template for these filings, and they have the incentive to provide high-quality responses backed by evidence.
- In contrast, Web text can cover a heterogeneity of topics and provide diverse content. What is notable is that the drop in prediction performance is limited, i.e., the embedding approach is robust and can overcome heterogeneity and noise.
- The benchmark variant that uses SIC and NAICS codes to select peers performs poorly. Each industry sector may include a diversity of companies with very different financial performance. Thus, it is important to use additional features from the 10-K filings or the Web text to identify peers.

Table 3 Drill Down on Profit Prediction - Assets and Sales - for Public Companies

Model	% IS peers		Profit/Assets		Profit/Sales	
	2%	1%	R-sq	RMSE	R-sq	RMSE
Prediction on public Web text based embedding						
Doc2Vec;Cos	66	73	**0.503**	0.186	**0.447**	0.290
Prediction on public and 32K private Web text based embedding						
Doc2vec;Cos	72	79	**0.492**	0.188	**0.436**	0.292
D2V+LDA;Cos	**86**	**95**	0.469	0.192	0.426	0.295
Doc2Vec;Siam	80	93	0.432	0.198	0.402	0.302

- The more sophisticated supervised models were unable to improve on performance due to a range of reasons including the lack of relevant training data.
- Our method D2V+LDA benefits from the Web text and the robust embedding approach. It is further enhanced/tuned in peer selection since it can exploit product-specific topics that are identified by an LDA topic model when selecting peers.

5.3. In-Sector (IS) Peers

Tables 4 and 5 provide insights into the task of (ii) identifying In-Sector (IS) peers in the competitor network. As discussed earlier, models for financial analysis such as the Fama / French model are typically evaluated against multiple industry sectors that are identified using SIC codes (Fama and French (1997)). For simplicity, we use a reduced grouping of the following five industry sectors to determine In-Sector (IS) peers: (1) `(Consumer)` Durables and Non-Durables; Wholesale; Retail; Some Services. (2) `(Manufacturing)` Manufacturing; Energy; Utilities. (3) `(HiTec:)` Business Equipment; Telecommunications; TV; Control; Computational Services. (4) `(Health:)` Healthcare; Medical Equipment; Drugs. (5) `(Other:)` Multiple sub-sectors including Transportation and Construction; Finance. The `(Other)` group comprises 1303 or approximately one third of the 3800 companies in our ground truth data set. The other four groups range from 536 companies (14%) for `(Consumer)` to 678 companies (17%) for `(HiTec)`.

We consider two thresholds to identify the closest peers; the first considers the Top 2% of all peers and the second considers the Top 1%. Table 4 reports on the distribution of these peers across the five sectors. We report on Doc2Vec (3800 public), Doc2Vec (3800 public and 32K private), and D2V+LDA (3800 public and 32K private). We observe the following trends across all three approaches: Sector (4) `(Health)` with 16% of the public companies dominates the peer distribution with the count of peers ranging from 23% to 49% of all peers; the high of 49% is at the 1% cutoff for D2V+LDA. The next highest count is observed for Sector (5) `(Other)`; with 34% of the public companies, in this sector, the count of peers ranges from 36 to 40% of all peers. Sector (1) `(Consumer)` has 14% of public companies; however, it has the smallest count of peers, from 9% to a low value of 2% (1% cutoff for D2V+LDA). These trends reflect that all three approaches to identifying peers perform significantly well for companies in Sectors (4) and (5). They are less successful in identifying peers for Sector (1).

Table 5 explores the percentage of In-Sector (IS) peers, among all peers reported in Table 4, for the three approaches and across the five sectors. We observe clear differences in the performance of the three approaches. Doc2Vec (public and 32K private) outperforms Doc2Vec (public), across all of the sectors in identifying IS peers. The latter has 66% IS peers (2%) and 73% IS peers (1%) overall, while the former has 72% (2%) and 79% (1%), respectively; these values are shown in the first column of the table. This reflects that there is a significant advantage of considering a much larger corpus of 32K private companies compared to the 3800 public companies in identifying IS peers.

Further, we see the significant benefit of tuning for D2V+LDA (3800 public and 32K private); it has 86% IS peers (2%) and 95% IS peers (1%) overall. The performance improvement for D2V+LDA ranges from 9% to 23%, for the different sectors and thresholds, compared to the other two approaches. As expected, the percentage of IS peers is highest for Sector (4), followed by Sector (5). We note, for example, that with a 1% cutoff for D2V+LDA, the percentage of IS peers for Sector (4) `(Health)` is at an extraordinarily high level of 98%. What is also notable is

Table 4 Distribution of All Peers across ff5 Sectors for Public Companies

Dataset	Model	ff5:1 (536 14%)		ff5:2 (665 17%)		ff5:3 (678 17%)		ff5:4 (619 16%)		ff5:5 (1303 34%)	
		2%	1%	2%	1%	2%	1%	2%	1%	2%	1%
public	Doc2Vec;Cos	11	9	14	13	14	12	23	25	38	40
pub+32K	Doc2Vec;Cos	8	6	14	12	12	9	30	34	37	38
pub+32K	D2V+LDA	5	2	8	5	10	5	41	49	36	39

Table 5 Distribution of IS Peers across ff5 Sectors for Public Companies

Dataset	Model	All		ff5:1 (536)		ff5:2 (665)		ff5:3 (678)		ff5:4 (619)		ff5:5 (1303)	
		2%	1%	2%	1%	2%	1%	2%	1%	2%	1%	2%	1%
public	D2V	66	73	41	47	49	57	52	58	80	85	74	81
pub+32K	D2V	72	79	40	43	54	63	53	59	87	91	78	84
pub+32K	D2VSiam	80	93	26	28	57	76	43	46	88	96	86	95
pub+32K	D2V+LDA	86	95	50	63	67	80	68	76	95	98	89	96

the performance in identifying IS peers for companies in Sector (1) `(Consumer)`. Recall that the distribution of all peers across the sectors is such that very few peers are identified for companies in Sector (1). Despite this, the percentage of IS peers is 63% with a 1% cutoff for D2V+LDA.

To summarize, Doc2Vec (3800 public and 32K private) outperforms Doc2Vec (3800 public), showing the advantage of considering an additional thirty two thousand private companies. We also see a significant improvement in performance from the tuning of D2V+LDA (3800 public and 32K private), resulting in the identification of up to 98% of IS peers in Sector (4) `(Health)`, with a 1% cutoff of the peer network.

5.4. NAICS Code Identification for Public Companies

The final evaluation task is (iii) the identification of the NAICS code for publicly traded and private companies. This task is of particular importance for private companies since there are no publicly available reference datasets that provide this identification for them. Further, there is significant

Table 6 Distribution of Matching NAICS Peers across ff5 Sectors for Public Companies. We consider the Top 2% Peers within each ff5 Sector.

ff5	Model	2-digit	3-digit	4-digit	5-digit	6-digit
ff5 1 Consumer	Doc2Vec; Pub	19.6	11.1	6.9	5.2	3.3
	Doc2Vec;Pub+32KPriv	18.5	10.8	6.7	4.9	3.1
	D2V+LDA;Pub+32KPriv	20.1	11.9	7.4	5.4	3.3
ff5 2 Manufacturing	Doc2Vec; Pub	39.8	23.1	17.0	15.6	14.6
	Doc2Vec;Pub+32KPriv	41.8	26.0	20.1	18.6	17.6
	D2V+LDA;Pub+32KPriv	46.3	29.2	23.0	21.4	20.4
ff5 3 HiTec	Doc2Vec; Pub	39.6	21.9	13.7	12.9	10.2
	Doc2Vec;Pub+32KPriv	39.9	23.6	14.4	13.5	10.3
	D2V+LDA;Pub+32KPriv	43.4	26.6	17.1	16.1	12.6
ff5 4 Health	Doc2Vec; Pub	75.4	72.1	71.4	71.4	43.4
	Doc2Vec;Pub+32KPriv	84.8	83.2	82.9	82.9	51.6
	D2V+LDA;Pub+32KPriv	93.3	93.1	93.1	93.1	57.7
ff5 5 Other	Doc2Vec; Pub	61.4	53.3	49.4	30.7	30.4
	Doc2Vec;Pub+32KPriv	67.2	59.3	55.6	34.5	34.3
	D2V+LDA;Pub+32KPriv	84.4	79.8	75.8	47.7	47.6

interest in the widespread adoption of the NAICS code, since there are many follow on value-added services that rely on the NAICS code being identified for any company.

For the NAICS code identification task, Tables 6 and 7 report on the percentage of peers that are in the *same NAICS sector as the focal company*, with a 2% and 1% cutoff, respectively, of the competitor network. A majority vote over the NAICS codes of these peer companies will then be used for predicting the code. The evaluation ranges from considering the 2-digit to the 6-digit NAICS code, and the results are reported for each of the five Fama / French sectors.

Table 7 Distribution of Matching NAICS Peers across ff5 Sectors for Public Companies. We consider the Top 1% Peers within each ff5 Sector.

ff5	Model	2-digit	3-digit	4-digit	5-digit	6-digit
ff5 1 Consumer	Doc2Vec; Pub	23.7	15.0	9.9	7.5	4.8
	Doc2Vec;Pub+32KPriv	21.8	14.2	9.7	7.4	4.6
	D2V+LDA;Pub+32KPriv	24.3	16.3	10.9	8.0	5.1
ff5 2 Manufacturing	Doc2Vec; Pub	47.2	31.3	24.8	22.9	21.7
	Doc2Vec;Pub+32KPriv	49.7	35.5	29.2	27.5	26.2
	D2V+LDA;Pub+32KPriv	57.8	43.0	36.5	34.9	33.6
ff5 3 HiTec	Doc2Vec; Pub	43.8	25.6	16.8	15.7	12.6
	Doc2Vec;Pub+32KPriv	44.2	27.8	18.2	17.1	13.6
	D2V+LDA;Pub+32KPriv	47.7	31.3	21.5	20.3	16.5
ff5 4 Health	Doc2Vec; Pub	80.2	77.3	76.7	76.7	49.0
	Doc2Vec;Pub+32KPriv	87.5	86.3	86.1	86.1	56.0
	D2V+LDA;Pub+32KPriv	95.4	95.4	95.4	95.4	60.8
ff5 5 Other	Doc2Vec; Pub	71.8	65.0	61.4	38.0	37.8
	Doc2Vec;Pub+32KPriv	75.9	69.4	66.1	41.2	41.0
	D2V+LDA;Pub+32KPriv	94.8	93.1	91.3	56.9	56.9

Since the distribution of peers was not uniform across the five Fama / French sectors, as reported in Table 4, we make a modification to the peer selection process, for this experiment. Instead of selecting the Top 2% (1%) peers across all of the 3801 companies and five sectors, the modified approach will consider the subset of companies in each of the five sectors separately and will select the Top 2% (1%) peers within that sector. Consequently, companies in Sectors (1) `(Consumer)`, (2) `(Manuf)` and (3)`(HiTech)`, that were initially at a disadvantage in the selection of peers, will be assigned a higher count of peers when using the modified peer selection approach.

The highest percentage of matching NAICS code peers in the peer network is reported for our method D2V+LDA (3800 public and 32K private). The matching percentages for Sector (4) `(Health)` range from 93% (2% peers) to 95% (1% peers) for 2-digit to 5-digit NAICS codes. This match drops to 57.7% (2% peers) and 60.8% (1% peers), respectively, for a 6-digit match. The matching percentages for Sector (5) `(Other)` ranges from 94.8% (2-digit; 1% peers) to 75.8% (4-digit; 2% peers). This match drops to 47.6% (2% peers) and 56.9% (1% peers), respectively, for a 6-digit match. The matching percentages for Sector (2) `(Manuf)` ranges from a high of 57.8% (2-digit; 1% peers) to a low of 20.4% (6-digit; 2% peers). The matching percentages for Sector (3) `(HiTech)` and Sector (1) `(Consumer)` are somewhat lower. The match is in the range of 47.7% (2-digit; 1% peers) to 12.6% (6-digit; 2% peers) for Sector (3) `(HiTech)` and in the range of 24.3% (2-digit; 1% peers) to 3.3% (6-digit; 2% peers) for Sector (1) `(Consumer)`.

As with the previous task of evaluating IS peers, there is an advantage of using the additional 32K private companies and the tuning for D2V+LDA, when selecting peers in the same NAICS sector. Thus, D2V+LDA (3800 public and 32K private) outperforms both Doc2Vec (public) and Doc2Vec (3800 public and 32K private). We observe the greatest performance improvement over Doc2Vec (public) for Sector (5) `(Other)`; the improvement ranges from 23% (2-digit, 1% peers) to 17% (6-digit; 2% peers). There is a similar performance improvement for Sector (4) `(Health)`. The improvement for Sector (2) `(Manuf)` ranges from 12% (6-digit; 1% peers) to 6% (6-digit; 2% peers). There is a small performance improvement for Sectors (1) `(Consumer)` and (3)`(HiTech)`.

To summarize, for the task of selecting peers in the *same NAICS sector as the focal company*, D2V+LDA (3800 public and 32K private) outperforms both Doc2Vec (public) and Doc2Vec (3800 public and 32K private), and demonstrates excellent performance for Sectors (5) `(Other)` and (4)`(Health)`.

5.5. NAICS Prediction for Private Companies

Table 8 reports the results on NAICS prediction for 32,000 private companies. This task is a key contribution since we are the only approach to use Web text and to attempt prediction for private

Table 8 NAICS Prediction for Private Companies

Model	Similarity	Filter Rate(%)	2-digit	3-digit	4-digit	5-digit	6-digit
			Accuracy (%)				
Doc2Vec	Cosine	top 2%	65.70	63.06	58.01	57.16	56.43
LDA	Cosine	top 2%	27.66	24.58	16.05	14.95	13.14
Doc2Vec	Siamese	top 2%	62.28	60.19	56.12	55.83	53.43
D2V+LDA	Cosine	top 2%	67.37	66.81	64.89	64.76	64.63
Doc2Vec	Cosine	top 1%	75.31	74.26	71.23	71.02	70.85
LDA	Cosine	top 1%	41.63	37.62	24.58	22.86	20.10
Doc2Vec	Siamese	top 1%	64.22	61.00	57.14	55.99	55.03
D2V+LDA	Cosine	top 1%	**84.85**	**84.17**	**81.62**	**81.37**	**81.03**

companies. We report on results for the embedding over the 3800 public and thirty two thousand private companies. Unlike the previous task that reported on the percentage of competitors in the same NAICS sector, for this task we simply use a majority vote over the NAICS code of the competitors to predict the NAICS sector. The top panel of Table 8 reports on an experiment where we consider the top 2% of the competitor network, while the lower panel considers the top 1% of the network.

The results for predicting the NAICS codes for private companies follow similar trends to the results for the public companies. Our method D2V+LDA, Doc2Vec with LDA tuning, outperforms Doc2Vec/Cosine and Doc2Vec/Siamese. The very encouraging result is that our method D2V+LDA, Doc2Vec with LDA tuning, performs with a very high prediction accuracy of over 80%, when predicting the 2-digit to 6-digit NAICS codes, while using the top 1% competitor network. We do note that there is a tradeoff since considering the top 1% of competitors can potentially reduce the recall for the task of predicting the competitor network. We note that for this task, Doc2Vec/Siamese has only a small performance degradation in comparison to Doc2Vec/Cosine.

This is in contrast to the task of profit prediction, where Doc2Vec/Siamese experienced a much larger performance degradation.

5.6. Summary of Contributions to Experiment Design and Best Practices

We make the following methodological contributions - relevant to experiment design and best practices - based on our extensive evaluation of representation learning approaches over a range of datasets. The first set of observations are with respect to the unsupervised approaches.

- We note the outstanding performance of the benchmark variant that constructs a Doc2Vec embedding over the SEC 10-K filings. This is consistent with expectations; the 10-K filings include high quality content and the 10-K filings often follow a common template.
- Our evaluation included a comparison of Doc2Vec embeddings over (i) SEC 10-K filings for 3800 public companies; (ii) Web text for 3800 companies; (iii) Web text for thirty two thousand private companies. Web text can cover a heterogeneity of topics with diverse content, and the content can be noisy and fragmented. What is notable is that the drop in prediction performance for identifying competitors is relatively small, as we move from the (i) 10-K filings to (ii) and (iii), i.e., the Doc2Vec embedding approach over Web text appears to be robust and can overcome heterogeneity and noise.
- Our proposed model, D2V+LDA , trained on Web text from thirty two thousand private companies, and tuned using LDA Topic Model topics, provided the best performance when identifying In-Sector (IS) peers. This model also performed remarkably well at identifying peers in the same NAICS sector as the focal company, for several Fama/French sectors. This reflects the benefits of a larger diverse training corpus of thirty two thousand private companies, in comparison to 3800 public companies, despite the potential noise from Web text.

The next two observations are with respect to supervised approaches. We used a Siamese network in an attempt to improve over the use of unsupervised cosine similarity to construct the competitor network. We used the similarity scores obtained from the TNIC approach over the SEC 10-K filings (Hoberg and Philips (2016)) to train the Siamese network. It appears that the use of the TNIC

similarity scores as training data placed an upper bound or other limitation on the performance of the Siamese network approach in creating the competitor network. Consequently, Doc2Vec/Siamese demonstrated inferior performance compared to the unsupervised Doc2Vec/Cosine for all of the evaluation tasks.

A final contribution is with respect to identifying the limitations of the more sophisticated trained models such as RoBERTa. We constructed a classifier that identified text chunks from the SEC 10-K filings, and used that classifier to filter chunks of Web text that would be more relevant to the task of competitor prediction. The addition of the filter did improve performance using RoBERTa. However, a mismatch between the parameters (from the training dataset) of the pre-trained RoBERTa and the potentially noisy Web text appeared to be a barrier. Consequently, the approach using RoBERTa was unable to improve on the performance of the unsupervised Doc2Vec models.

6. Proof of Concept Case Study

We conclude our evaluation with a proof of concept case study that explores when companies can benefit the most from generating high quality Web text about their products and services. We use well-studied measures from the finance literature (Acikalin et al. (2022), Hoberg et al. (2014)) to characterize companies. We identify the scenarios where the Web text of two companies *are more likely to predict that they are IS peers.*

For this study, for each embedding/peer network, we label the outcome variable of those pairs of peer companies that are in the same Fama / French sector, i.e., IS peers, as positive; those that are out of sector are labeled as negative. The independent variables, listed below, are computed using Compustat data (circa 2016). As our dependent variable is computed for each pair of peer companies, the corresponding independent variables are computed as the product of the values for each of the companies in the pair. For example, the `logassets` variable would test whether our model is more likely to correctly classify pairs of large companies (small companies) as IS peers.

- `logassets`: This is the natural logarithm of the reported assets (+1) of the company.

- `logage:` This is the natural logarithm of the age (+1) of the company, where age is measured from the first appearance of the company in the Compustat database.
- `mktbook:` This is the ratio of the market capitalization of the company over the book value of its assets.
- `vcf:` This is a measure of the intensity of venture capital funding activity in the corresponding industry of the firm; this measure was introduced in (Acikalin et al. (2022)).
- `fluid:` This is a measure of product market fluidity. It measures the extent to which a company is operating in a product market that is entrepreneurial, with a rapid rate of product innovation and more agile competitors. The measure was introduced in (Hoberg et al. (2014)).
- `adsales:` This is a ratio of the expenditure on advertising scaled by the reported sales.

We use a regression-based classifier to determine which features increase the likelihood of a pair of companies being identified as IS peers. Table 9 presents the results for IS peers for the 3800 public companies where the peers are identified by (i) Doc2Vec model (embedding over 3800 public companies) - in the middle column - and (ii) D2V+LDA (embedding over 3800 public and thirty two thousand private companies) - in the third column. All independent variables are standardized prior to running the regression so coefficient magnitudes can be compared. Table 9 reports on the values for the coefficients with the t-statistics in parentheses. Standard errors are clustered by company.[2] Recall that D2V+LDA had the best performance in identifying IS peers; the intercept for D2V+LDA (last column) confirms this and indicates that there are 85% IS peers, while Doc2Vec;Public has only 63.9% peers.

This analysis leads to the following novel insights regarding the scenarios when companies can benefit the most from an investment in generating high quality Web text about their products and services, and how these scenarios can influence the choice of Web text:

- The most significant feature for identifying IS peers is `fluid`, a measure of product market fluidity, product innovation and agile competitors. Our proposed method D2V+LDA has a

[2] Clustering by company accounts for spatial correlations and results in more conservative standard errors.

Table 9 Regression Results for In-Sector Peers

Variable	Doc2vec;Public	D2V+LDA ;Pub+32KPriv
Intercept	0.639 (175.03)	0.850 (245.7)
logassets	0.040 (11.310)	0.036 (9.120)
logage	0.027 (8.800)	0.009 (3.350)
mktbook	–0.024 (–7.590)	–0.012 (–3.700)
adsales	0.065 (21.540)	0.059 (18.990)
fluid	0.196 (54.620)	0.129 (28.750)
vcsimm	0.050 (13.510)	0.029 (6.940)

high overall likelihood of identifying IS peers for companies operating in such markets. These results suggest that companies in such markets produce high quality Web text that focuses on unique product offerings, as a potential tool to remain competitive. Intuitively, these earlier-stage entrepreneurial companies have the incentive to do so, since they need to quickly build market share. A product-focused Web presence serves this objective directly. Older and more mature companies, on the other hand, may choose to include additional content that is less relevant to their products, and they may favor Web text with a focus on ESG initiatives, or other indicators of corporate culture.

- The next most significant feature is `adsales`. This is a measure of the level of expenditure on advertising scaled by sales. We can expect that companies that have high advertising expenses are also going to ensure high quality Web text, and this improves the likelihood of identifying IS peers.
- We also note that `vcsimm` is significant; this is a measure that reflects high levels of venture capital funding. This suggests that companies may invest in producing high quality Web text with a product focus, and may rely on the Web text as being a proxy or signal to attract venture investment.

Overall, this use case provides interesting results. It reflects a virtuous cycle of features that increase the likelihood of identifying IS peers, in precisely those scenarios where accurately identifying IS peers may have the greatest benefit.

7. Summary and Future Research

In this paper, we present representation learning enabled approaches to use Web text to predict competitor networks. Unlike prior benchmarks, our approach provides broad coverage of both public and private companies and exploits non-proprietary public data. Our approach(es) applied state-of-the-art language modeling techniques, including different approaches for representation learning and link prediction. We performed an extensive evaluation using four downstream tasks: (i) Company profit prediction; (ii) Self-reported competitor prediction; (iii) Identification of In-Sector (IS) peers; (iv) Identification of NAICS industry codes.

We expanded the scope of the competitor network from the approximately 5,000 publicly traded companies (3800 in our experiment dataset) to include over 800,000 private companies. Using the 3800 public companies and a representative sample of thirty two thousand private companies, we demonstrated that the learned representations using noisy Web text can remarkably match or improve the accuracy of prior industry and competitor classification benchmarks. Prior benchmarks include TNIC that was trained on curated SEC 10-K regulatory filings (Hoberg and Philips (2016)) and a benchmark that used the four-digit and six-digit NAICS codes. In a setting limited to the curated SEC 10-K filings, our representation learning based models could outperform TNIC for making predictions for public companies. Finally, an encouraging result is that our proposed method D2V+LDA, a Doc2Vec based model that is tuned using LDA topics, can predict the NAICS classification codes of both public and private companies with very high accuracy. These empirical results demonstrate that representation learning approaches can extract a broader set of insights about products and services, to generate the competitor network, and can also do so for private companies where only noisy Web text is available.

Our approach does have some limitations. We identified the limitations from applying more sophisticated supervised approaches such as RoBERTa, in comparison to the unsupervised Doc2Vec embedding over Web text; a key reason for the limitations is the lack of high quality training data. Supervised approaches are also computationally very expensive, and it may be difficult to scale

these approaches to handle noisy Web text over millions of companies. An interesting area for future research is to explore distance supervision and similar extensions that may overcome the lack of training data.

Web text is only one of many sources of data about companies. Companies increasingly use social media platforms to communicate with potential customers or employees or investors. In future research, we plan to supplement Web text with content from social media platforms such as LinkedIn, Facebook, and Yelp. The strength of our initial results suggests that the same models that were used to construct the competitor networks could also be applied to identifying the structure of supply chain networks, or to trace innovation through industries. Another promising direction is to use representation learning to characterize how products, services, and the competitor network, evolves over time.

Appendix

Reproducible Research You can find our pipeline code at `https://osf.io/wrtda/?view_only=2814e0e5eeb04e4194da4a15d9d74f2c`. We will make a fully documented system available upon publication.

References

Acikalin U, Caskurlu T, Hoberg G, Phillips G (2022) ntellectual property protection lost and competition: An examination using large language models. *Tuck School of Business Working Paper No. 4023622* URL `http://dx.doi.org/10.2139/ssrn.4023622`.

Bao S, Li R, Yu Y, Cao Y (2008) Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering* 20(10):1297–1310, URL `http://dx.doi.org/10.1109/TKDE.2008.98`.

Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. *arXiv:2004.05150* .

Bernstein A, Clearwater S, Hill S, Perlich C, Provost F (2002) Discovering knowledge from relational data extracted from business news. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on MultiRelational Data Mining.*

Bhojraj S, Lee CMC (2002) Who is my peer? a valuation-based approach to the selection of comparable firms. *Journal of Accounting Research* 40(2):407–439.

Bhojraj S, Lee CMC, Oler DK (2003) What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41(5):745–774.

Blei D, McAuliffe J (2007) Supervised topic models. *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 121–128, NIPS'07, ISBN 9781605603520.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Chamberlin E (1933) *A Theory of Monopolistic Competition* (Harvard University Press).

Fama EF, French KR (1997) Industry costs of equity. *Journal of financial economics* 43(2):153–193.

Hoberg G, Philips GM (2016) Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy* 124(5).

Hoberg G, Phillips G, Prabhala N (2014) Product market threats, payouts, and financial flexibility. *Journal of Finance* 69(1):293–324.

Hoberg G, Phillips GM (2010) Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies* URL `https://doi.org/10.1093/rfs/hhq053`.

Hofmann T (2017) Probabilistic latent semantic indexing. *ACM SIGIR Forum* 51(2):211–218.

Hotelling H (1929) Stability in competition. *Economic Journal* 39(153):41–57.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *International Conference onmMachine Learning*, 1188–1196.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach.

Loper E, Bird S (2002) Nltk: The natural language toolkit. *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.*

Ma Z, Pant G, Sheng OR (2011) Mining competitor relationships from online news: A network-based approach. *Electronic Commerce Research and Applications* 10(4):418–427.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Office of Management and Budget (2017) *North American Industry Classification System* (United States Census).

Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M (2019) fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*.

Pant G, Sheng O (2015) Web footprints of firms: Using online isomorphism for competitor identification. *Information Systems Research* 26(1):188–209, URL `http://dx.doi.org/10.1287/isre.2014.0563`.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds., *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc.), URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Pearce E (1957) *History of the Standard Industrial Classification* (Bureau of the Budget, Office of Statistical Standards).

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Rauh JD, Sufi A (2011) Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance* 16(1):115–155.

Rehurek R, Sojka P (2011) Gensim–python framework for vector space modelling. *Technical Report, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3(2).

Roberts M, Stewart B, Tingley D, Airoldi E (2013) The structural topic model and applied social science. *International Conference on Neural Information Processing*.

Roy SK, Harandi M, Nock R, Hartley R (2019) Siamese networks: The tale of two manifolds. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

US Department of Labor (2022) *Standard Industrial Classification (SIC) Manual*. URL https://www.osha.gov/data/sic-manual.

Wei Q, Qiao D, Zhang J, Chen G, Guo X (2016) A novel bipartite graph based competitiveness degree analysis from query logs. *ACM Trans. Knowl. Discov. Data* 11(2), ISSN 1556-4681, URL http://dx.doi.org/10.1145/2996196.

Xu K, Liao SS, Li J, Song Y (2011) Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems* 50(4):743–754, ISSN 0167-9236, URL http://dx.doi.org/https://doi.org/10.1016/j.dss.2010.08.021, enterprise Risk and Security Management: Data, Text and Web Mining.

Yaros JR, Imieliński T (2015) Data-driven methods for equity similarity prediction. *Quantitative Finance* 15(10):1657–1681, URL http://dx.doi.org/10.1080/14697688.2015.1071079.