An Analysis of Machine Learning Methods in Industrial Classification

John Doncy

March 14, 2021

Abstract

This paper aims to provide a critical discussion of the machine learning methods followed by Hoberg and Phillips (2016) in constructing their text-based classification system. They follow a novel approach of classifying firms using textual data in financial reports. The text-based method offers a completely new perspective in classifying industries as compared to the traditional SIC or NAICS schemes. This discussion also includes alternative methods for classification and suggestions for future research.

1 Introduction

Ever-changing industries, market structure and their competitive landscape are widely discussed topics in the field of industrial organisation. Industry structure plays a crucial role in determining the behaviour of firms and their performance in the marketplace (Bain, 1968). Although firms within an industry may be similar in certain characteristics, they often differ from one another in many ways. These differences are a result of various competitive strategies adopted by firms (Porter, 1979, p. 215). It is important to address the relationship between behaviour of firms and industry structure to frame appropriate antitrust policies and regulatory mechanisms (Einav and Levin, 2010, p. 146). But, in order to provide more clarity about firms and their competitive mechanisms, we need to have efficient industry classification systems in place. Existing industrial classification systems like SIC or NAICS form industries based on the production processes of firms. A major flaw with this approach is that their

classifications are infrequently updated and hence fails to capture the evolving product market. Firms that focus on different product markets may end up being grouped into the same broad industry. Moreover, the transitivity assumption of existing classifications does not provide a measure of similarity between firms within an industry or across industries. This can be quite misleading because two firms A and B having a common rival firm C, need not necessarily imply that A and B are also rivals. Hence, a measure of similarity between firms is important to capture the competitive environment of the industry. Also, these systems are based on pre-defined industry categories, which makes it difficult to accommodate innovations. Hence, SIC or NAICS classification systems may fail to accurately define industries and industry boundaries. These issues of time-fixed product markets and transitivity are accounted for in the novel text-based network industry classification or TNIC introduced by Hoberg and Phillips (2016). The objective of TNIC is to classify firms based on their product offerings instead of their production processes. The authors believe that product similarity is core to classifying industries and that firms in an industry use the same words in describing their products. This tendency of product market vocabulary to cluster among firms is captured using text-analysis methods. It is mandatory that publicly traded firms in the U.S submit a 10-K annual report to the SEC, providing details about the firm's financial performance. The authors analyse the text of product details within these reports to form new industry classifications. Based on the words used to describe their products, each firm is given a spatial location in this new network of industries. Text analysis also makes it easier to calculate word similarity scores for a pair of firms in a given year and these scores are used to form industries. Cosine similarity of word vectors is used to measure 'product similarity' and 'product differentiation' between firm pairs. By using this approach, the authors are able to identify a distinct set of competitors for each firm. Besides identifying close rivals, the text-based system also finds that managers discuss about potential future threats from distantly located firms in the industry network. The TNIC also helps to identify the different impacts of exogenous industry shocks. A positive shock leads to increases in competition and product similarity as firms relocate to areas of high demand. On the other hand, a negative shock leads to

firm exits and movement into differentiated product markets. Such an analysis of industry behaviour is possible since the TNIC data gets updated annually. With the help of firm similarity scores, the authors also find that firms investing in R&D and advertising activities are able to create endogenous barriers to entry. We require classification systems to account for the heterogeneity between firms and industries. These groupings provide a better context for researchers to conduct economic and financial analysis (Bhojraj, Lee, and Oler, 2003, p. 746). While few other studies have tried to come up with new industry classification systems, the text-based classification introduced by Hoberg and Phillips seems promising. The aim of this paper is to provide a critical discussion of the methods used to create the TNIC model and to look into the potential avenues for future research.

The paper is organised as follows: Section 2 provides a brief description of the data, Section 3 discusses the methodology, Section 4 includes a discussion of the importance of TNIC, its external vailidity and future extensions and Section 5 concludes.

2 Data

Electronic Data Gathering, Analysis and Retrieval system or EDGAR is the primary data source used for obtaining the 10-K annual filings of firms. It is an electronic filing system developed by the SEC to improve accessibility and transparency of financial information in the U.S. Firm 10-K report sourced from the EDGAR database contains detailed descriptions of company history, financial statements, products and services, business operations and the company's markets. The TNIC is based on the product details obtained from the business description sections of such 10-K forms. An advantage of using 10-K reports is that the information is standardized and unbiased. It is required by law that firms provide accurate descriptions of the products and services they offer. Any changes to their business is updated annually through these forms at the end of each fiscal year. All these factors make this database ideal for studying chang-

ing industry dynamics. The authors use a web crawling algorithm to collect the entire text of each 10-K report and an APL text reading algorithm extracts product description from each document. They also link the 10-K reports to Compustat data and identify that 10-K covers 97.9 percent of the listed firms in Compustat.

3 Methodology

The authors have leveraged the idea of cosine similarity in creating their text-based industry classification. Cosine similarity method measures the cosine of the angle between two vectors in a multi-dimensional space. Here, the objective is to measure the cosine similarity of word vectors for each firm pair. Higher the cosine similarity measure the greater is the similarity between firms. In order to develop the TNIC, the authors map words in the product descriptions of each firm to words in the overall set of product descriptions. This produces a binary word vector for each firm, where an element equals a value one if the word associated with it is present in the overall corpus of product words. It equals zero if the word used by the firm is not present in the overall corpus. To account for variations in the length of each firm's product description, the word vectors are normalised to unit length. Finally, they obtain a matrix wherein each row contains the normalised word vector for each firm in a given year. The next step is to calculate the dot product of word vectors for each firm pairs. This results in a pairwise similarity matrix (square matrix) where each entry is a score for product similarity between two firms. The square matrix is in fact the network representation of all firms and this forms the foundation of the text-based classification model. In general, cosine similarity is a powerful metric and its application for measuring product similarity seems efficient. The similarity measures are calculated based on the orientation of word vectors in the product space. Moreover, after normalization, the scores are bounded in the interval (0,1). Cosine similarity is used to obtain measures of product similarity and differentiation in the following way: for two firms i, j with normalised word vectors V_i and V_j,

$$Product\ Similarity = (V_i \,.\, V_j)$$

$$Product\ Differentiation = 1 - (V_i \,.\, V_j)$$

The pairwise similarity matrix is first used to form a fixed industry classification that is consistent with the properties of existing schemes like SIC and NAICS. Each firm in 1997, the earliest year of the sample, is assumed to be a separate industry i.e. if there are N firms in 1997, they form N industries. The idea is that the initial set of industries formed by firms in 1997 is held constant for all years in the time series to form a fixed classification. Then two industries with highest similarity are combined. Once two industries are combined, a new similarity score is calculated by taking the average of their firm pairwise similarity. This process of combining single firm industries continues until the desired number of industries is reached. A given firm in rest of the years is grouped into the industry with which it has the highest similarity. The cosine similarity method measures this by taking the dot product of the firm's normalized vector and the industry's normalized word vector.

The difference with the TNIC model is that the fixed location and transitivity assumption is no longer considered. In each firm i's industry, all other firm j's are included if their pairwise cosine similarity is higher than a prespecified minimum threshold. Hence, a higher threshold results in fewer rivals within an industry and a lower threshold groups larger number of firms within an industry. But, exclusion of the transitivity assumption makes it difficult to calculate industry weighted estimates of across-industry variation in profitability, sales growth and stock market risk. An area of concern would be the approach taken to form the unique set of words in the overall corpus. Only nouns and proper nouns that appear in less than 25 percent of all business descriptions are included. While this step puts more focus on the unique product words, one may also lose out on some subtler details of the products. The classification gets more efficient if it helps to uncover these finer details of the product markets served by firms. Hence, including more word classes can add to the uniqueness of a product and result in clearer industry groupings, making the model more stable. Also, firms having fewer than 20 unique words in their product

description are not included while forming the classification. While this makes the classification more meaningful and robust, it leaves a gap by omitting such firms. But, existing systems like SIC and NAICS ensure that every business is given a code, indicating its primary line of business. Hence, the text-based classification may not be fully efficient to replace existing industry classifications.

An alternative approach in classifying industries could be implemented using the Latent Dirichlet Allocation (LDA) method. LDA is a generative probabilistic model which assumes that our observed data arose from a generative process that includes hidden variables. The intuition behind LDA is that we assume each document to be a collection of a given set of topics and each of these topics is represented by a distribution of words. Here, the observed variables are the words in our documents and the hidden variables are the topic structure of these documents (Blei, 2012, p. 79-80). It is an unsupervised learning algorithm which means that the model is not trained beforehand. Fig 1 provides a graphical representation of the LDA model. For each topic k, a distribution β_k over

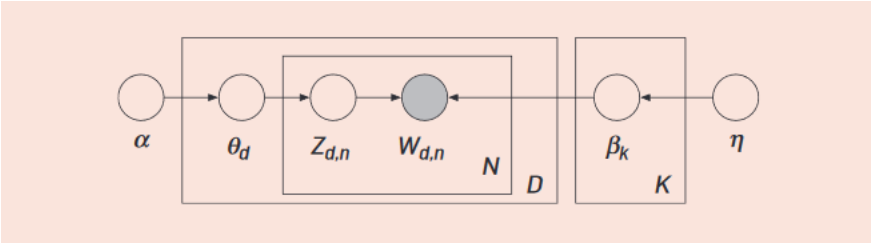

Figure 1: Graphical representation of LDA
Source: Blei (2012, p. 81)

words is drawn from a Dirichlet distribution with parameter η. Then, for each document d, a distribution over topics θ_d is drawn from a dirichlet distribution with parameter α. $Z_{d,n}$ is the topic assignment for the n^{th} word in document d which is drawn from a multinomial distribution with parameter θ_d. Finally, we draw a word from that assigned topic's distribution which is again a multinomial distribution and these multinomial distributions come from Dirichlet prior distributions. By focusing only on the product offerings of firms, such as in TNIC, we may not be making the most of the available firm specific data.

Products and services only form a part of the business model of a firm while the strategies adopted by firms help in explaining how they differentiate themselves and deal with competition. Hence, LDA can be used to extract different business model topics to classify similar firms.

4 Discussion

This section focuses on why the text-based classification is important in today's world and how we can improve it. In the recent years there has been an increased focus towards digitization of several aspects of business. Before the Internet, business operated primarily in a more tangible 'brick and mortar' setup. Today, many firms are shifting to a more intangible digital platform to expand their businesses across industries (Weill and Woerner, 2013, p. 72). Hence it is questionable whether it makes sense to rely on earlier classification systems to measure industry dynamics today. Firms are able to easily diversify and increase their presence in new markets with the help of internet. Existing systems place such firms with multiple business activities into a category reflecting their principal business. Here, TNIC seems to provide a more accurate representation of such firms and their distinct rivals as the model maps firms across industries. So, in today's fast paced and highly connected world, TNIC is far superior in comparison to earlier systems. However, the TNIC model focuses only on publicly traded firms in the U.S. Hence, the idea of using product descriptions from firm 10-K reports might provide only a partial insight into aspects of industrial organisation. Another study using business data between 1976 to 2001, show that publicly traded firms only account for a trivial share of all the firms in the U.S. (Davis et al., 2006, p. 11). Also, foreign direct investments are not taken into consideration while forming the text-based model.[1] Therefore, it raises the question whether the TNIC model is useful in providing a bigger picture covering all types of firms and industries.

1. This gap can be accounted for by including financial data of foreign firms using SEC form 20-F, which is similar to the 10-K reports.

4.1 External Validity and Future Extensions

While the TNIC model breaks new ground with its text-based classification framework, it is important that we question its generalizability. Finding a source of standardised financial data such as EDGAR is crucial for constructing classification systems like TNIC. What most countries lack is also a central financial statement database like EDGAR. Hence, it would be difficult to extend this idea of classification to other settings. Applying the methodology to other settings like Europe could help us understand the differences in firm behaviour and industry dynamics across a diverse sample. Comparing the output of different TNIC models can help identify variations in industrial organisation across countries in Europe. However, the biggest hurdle in extending the idea is to find reliable sources of firm centric data. For example, accessing information disclosed by listed companies in Europe is quite complicated. The European Securities and Markets Authority is still working on improving accessibility of financial data to the general public.[2] Another useful source for firm level information is the SEDAR[3] database used by Canadian authorities which bears close resemblance to EDGAR. This could be a good starting point for extending the TNIC model to public firms in Canada.

For capturing industry change to exogenous shocks, the study analyses two events that occurred in years 2000 and 2001. But the sample of 10-K data only begins in the year 1997 and hence there is not enough data to examine trends in behaviour of firms before these shocks. As the data gets updated annually, we could focus on more recent events like the 2007-2009 financial crisis to obtain more convincing results. Also, the TNIC only captures horizontal relatedness between firms and not vertical relatedness. Including vertical links or vertical production processes could add more stability to the text-based classification. For example, Fan and Goyal (2006) show that vertical relatedness plays a crucial role in explaining mergers between firms. This limitation exhibited by earlier systems like SIC is not accounted for in the new TNIC model. Future research could also include business risks shared by firms while calculating sim-

2. https://www.esma.europa.eu/regulation/corporate-disclosure/transparency-directive
3. https://www.sedar.com/homepage_en.htm

ilarities.[4] This could not only uncover more areas of similarity between firms but also help identify the ways in which similar firms mitigate their risks. Likewise, by including other firm specific variables one can also create an average similarity score spanning a wide range of firm characteristics. Another important application is to combine the text-based classification with price data for all the products to obtain firm-specific price levels. This could help us understand pricing strategies followed by different firms. Moreover, patent filings by firms can be used to create a measure of similarity between patents held by different firms. As 10-K reports are updated annually, it can be used to study how bigger firms react to the entry of smaller firms in their market, especially when startups make IPO's. It may be the case that bigger firms anticipating such a threat would engage in product differentiation, advertising and limit pricing behaviour. For instance, Goolsbee and Syverson (2005) find that incumbent airline firms cut their prices upon facing threat of entry from competitors. The TNIC model can be used to study such behaviour across industries by analysing changes in product similarity scores upon entry of new firms. While the TNIC is useful for research, it offers few benefits to policymakers since firms could manipulate their behaviour if they got to know that TNIC is used to analyse their activities. Setting aside this issue of manipulation by firms, the TNIC is efficient in identifying single firm industries as compared to existing classifications. This aspect can be leveraged by policymakers to regulate monopolies and protect consumer interests. We could also use TNIC to study changes in consumer behaviour to external shocks as product offerings of firms are based on underlying consumer preferences and demand.

5 Conclusion

The aim of this study was to provide a critical discussion of the idea and methodology followed by Hoberg and Phillips in constructing their text-based network industry classification. We observe how the text-based classification is far superior to existing systems like SIC and NAICS in uncovering finer details of firm

4. Risk factors section of 10-K contains details of risks the firm faces or may face in the future.

behaviour and industrial organisation. This in fact reveals that it is time we update our traditional approach of classification based on production processes and focus on more suitable aspects like product offerings of firms. By capturing similarities in textual data of firm reports, the authors are able to construct a flexible network of industries in a product market space. They apply machine learning techniques like the cosine similarity method to identify similar firms by looking at their product market vocabulary. While this method results in efficient classification, it is also important to include other aspects of similarity to form an overall measure of relatedness between firms. It could be that firms share similarities in other features and hence we need not restrict the focus to product offerings only. Moreover, the diversity of machine learning techniques can help us replicate the text-based classification using other methods. A possible alternative for grouping firms could be the LDA method. Even if firms do not explicitly describe their product offerings, we can use LDA to extract different topics from financial reports and use these topics as a basis to classify firms. Also, given the differences in availability of standardized financial data, it is important that we learn to model classifications based on different techniques. The concept of classification based on text looks promising, eventhough it is difficult to extend it to other settings with existing data. However, one could delve deeper into studying industrial organisation by using TNIC in conjunction with other available data on firms. Industry groupings are important for scientific and financial research but very few studies have focused on preparing efficient classification systems in industrial economics. The text-based network industry classification not only offers a new perspective on the way we think about industrial classification but also presents new methods for future research in this field.

References

Bain, Joe Staten. 1968. *Industrial organization.* New York: Wiley.

Bhojraj, Sanjeev, Charles MC Lee, and Derek K Oler. 2003. "What's my line? A comparison of industry classification schemes for capital market research." *Journal of Accounting Research* 41 (5): 745–774.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77–84.

Davis, Steven J, John Haltiwanger, Ron Jarmin, Javier Miranda, Christopher Foote, and Eva Nagypal. 2006. "Volatility and dispersion in business growth rates: Publicly traded versus privately held firms [with comments and discussion]." *NBER macroeconomics annual* 21:107–179.

Einav, Liran, and Jonathan Levin. 2010. "Empirical industrial organization: A progress report." *Journal of Economic Perspectives* 24 (2): 145–62.

Fan, Joseph PH, and Vidhan K Goyal. 2006. "On the patterns and wealth effects of vertical mergers." *The Journal of Business* 79 (2): 877–902.

Goolsbee, Austan, and Chad Syverson. 2005. "How Do Incumbents Respond to the Threat of Entry? Evidence from Major Airlines", NBER Working Paper 11072."

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

Porter, Michael E. 1979. "The structure within industries and companies' performance." *The review of economics and statistics,* 214–227.

Weill, Peter, and Stephanie L Woerner. 2013. "Optimizing your digital business model." *MIT Sloan Management Review* 54 (3): 71.