Old Wine in New Bottles,
Could We Learn More from the Old Data with New Techniques

Xie, Haojun

Abstract

Empirical economics had benefited a lot from the growing availability of various types of data and computing capability in recent years. Incorporating some untraditional data such as raw texts, images into economics and finance empirical research may require both the ability to extract information from the raw unstructured data and interpret the information properly. This paper would take Hoberg and Phillips (2016) as an example, using different methods on the same data to discuss the choosing of methods and the information we learned with different methods.

Because the validation was still in a very preliminary part, this version of the paper didn't include the results of all comparisons. External validation, conclusions sections, and all tables and images are omitted. The latest version of this paper should be available at here, `https://github.com/AnakinShieh/MyPapers/blob/master/OldWineinNewBottles.pdf`. All related codes will also be available in the same repository.

Introduction

Empirical economics/research has benefited a lot from the growing availability of various types and sources of data and computing capability in recent years. Incorporating some untraditional data such as texts, images into economics&finance empirical research may require both the ability to extract information from the raw unstructured data and the ability to interpret the information, connect the information with economic theory or conceptions appropriately. Usually, the latter is more challenging and important to economic theory. However, as the growing applications of new extracting information methods in economics&finance research, we may be interested about what we could gain from these new methods, both in totally new topic and data and the same data and topic compared to traditional ways, this paper aims to answer the latter part of the question by taking the classic work (Hoberg and Phillips, 2016) as an example.

The two steps mentioned above, extracting the information and interpreting the information, are not always divided clearly or connected closely. We may give a more detail definition of those two steps here, extracting information means a determined processing rule from the raw data to the output data. The determined rule here could be some machine learning algorithms that rely on the raw data to a certain extent because of the training process, a fixed rule like counting the number or ratio of certain words in all documents, and some manual work judged by human intelligence or common sense. In the case of texts and images, which could be described as high dimension data, this process usually plays a dimension-reduction role, in which classification is a special case. And the interpreting process is about constructing connections, like validations between the information extracted and the conception we are concerned with. And the conception is usually directly or indirectly connected with the model or economic theory background. The first step is similar to the codebook g in Egami et al. (2018), but not restricted in causal inference.

Because the growing methods here are usually from the development of recent machine learning, the topic has been discussed as the impacts and applications of machine learning on economics, like Athey (2017), Mullainathan and Spiess (2017). While different from the previous work, focusing on the general applications or general impacts of machine learning on economics, this paper would only focus on a small part, about the performance of different methods to calculate document similarity to construct firms similarity based on 10K filings and discuss what we have learned from the texts by using these methods. We will take several examples to illustrate different types of relations between the two steps, extracting and interpreting. Because the interpreting process is dependent on the conception we are concerned with, we could always take the determined rule of extracting information as the conception that matters, which would make the discussions meaningless. In the case of the relations between the intensity of night lights from satellite and real GDP(Henderson et al., 2012), if the conception concerned is simple the intensity of light, there wouldn't need such discussions. To avoid it, we restrict the conceptions here to be directly or indirectly related to economic conceptions, like the real GDP growth in night lights case, local economic conditions in street view images case(Gebru et al., 2017), economic policy uncertainty level in Baker et al. (2016) and product similarity between public firms Hoberg and Phillips (2016).

In most cases, the extracting information process couldn't affect the interpreting part. For example, images from the satellite were classified by machine-learning algorithms to different districts or different crops planted. Both the machine learning algorithms and the training data used here are irrelevant to the question or the key factor in the interpreting part.(Donaldson and Storeygard, 2016). In Gebru et al. (2017), they use deep learning algorithms to recognize the models of cars from street view images and found a strong correlation between the cars and socioeconomic data, political attitudes. And at the same time, the extracting process, the details of the algorithms could not be relevant to the interpretations of the relations observed.

In short, in this situation, what we learned is the well-defined tags of new data, and the tags here are not determined by the judgment from human understanding about abstract conceptions. This situation may change a little when the information we extracted is not very clear or relied on human judgment heavily. For example, in the research about economic policy uncertainty(EPU) (Baker et al., 2016), researcher have to first have a lot students decide whether certain newspaper article related to EPU or not; in the research about transparency of FOMC(Federal Open Market Committee), after creating topics using LDA model, researchers have to select topics that have good predictive power to policy opinions by LASSO while the policy opinions are collected by hand before.(Hansen et al., 2017). Both research, have some external data that are directly related to the information extracted and the conception in model/economic theory and performs as training data or criterion in the extracting information process. In the case of EPU research, the authors could use dictionary words methods to approximate the judgment about economic policy uncertainty or not by a human. This case is slighted different from the first kind relations because, in the first situation, even the information extracting process itself could be hard to interpret, but the results and what we should get from the extracting process are clear and are irrelevant with economic conceptions. In this case, the conception we're concerned with is directly involved in the extracting process. In the FOMC transparency case, LDA could perform a dimension reduction role to the transcripts, reducing the whole transcripts into a vector representing the probability distribution of topics over documents. Each topic is a collection of words and may have overlappings with some topics we are interested in. To select these topics robustly, Hansen et al. (2017) used external data, the attitudes to interest rate policy from Meade (2005) as criterions to select informative topics about policymaking, and then construct communication measure based on the selected topics.

In the latter situation, we couldn't have external validation data that is directly connected to economic conceptions and the information we extracted. The extracting process is directly

involved in the interpretations of what we get from the raw text. One example is what we will discuss in detail in this paper, Hoberg and Phillips (2016)(we will use H&B instead in the rest of this paper.), creating product similarity measures between public firms from 10K Business descriptions part.

The different relations here are relevant to the difference between supervised learning and unsupervised learning applications, which have already been emphasized in the literature like Athey (2017) and Egami et al. (2018). But not all, the key question we would like to highlight here is that, when lacking some direct criterions because the conception we're concerned with is not observable in nature which is common in economics and finance research, we could create some informative variables or factors that have robust relations with external validation data, and try to link those informative variables creation process with the economic conception, but what's the gains from different methods to the same conception and how should we employ them? And could we gain more in later research if we use some extracting information methods that were logically better or better in other evaluation tests?

H&P created a good case for us to illustrate this question, their method to create the product similarity measure is pretty straightforward and effective with good interpretability, but with a lot of small handy work, like keeping only nouns and special nouns, excluding country, state and city names, etc. Each step is reasonable but lacking a general framework at the same time, so comparing with methods with less handy work could help us understand more about what information we have learned through these methods. For example, when focusing the nouns and special nouns like brands appeared 10K files, could it overemphasize the importance of brands? And if we use some robust, more based on a semantic way to construct the measure, could it performs better? The measure created by H&B has also been widely used in recent finance and accounting literature, enabling us to validate the robustness of this method in different applications.

We hope this paper could provide a general reference about different methods applications

for researchers in economics and finance, and answer the question we mentioned above, what did we gain from these new methods and where is the performance improvement from. The rest of this paper will organize as this, in the next section we'll provide more details about the methods in Hoberg and Phillips (2016) and the details of methods we used in this paper, in section 2, we would discuss the data generation process and provide direct comparisons with Hoberg&Philips' similarity measure. Section 3 would provide the same external validation in the original paper with different generating TNIC methods. Section 4 would provide more external validation from the recent literature based on the HP classification. Section 5 is conclusions. (Section 4,5 are incomplete now)

1 Methods

In this section, we'll rephrase the methods used in origin Hoberg&Philips papers and introduce some methods we'll compare with in brief.

The work of Hoberg&Philips include two important parts, creating a product similarity measure between companies based on the nouns and special nouns appeared in the 10K Business descriptions section, which could be considered as using cosine similarity on a special type of TF-IDF, that we'll show in details in the subsection, and clustering firms into fixed industry classification. The clustering algorithm described in Appendix B in the original paper is a variant of hierarchical clustering[1]. And they use the first-year data (1997) to create the fixed industry classification and use it to infer the classification of later years. It could be considered as a train/test samples splitting procedure, which is recommended in Egami et al. (2018) and Perry (2009).To keep consistency with the HP paper, we'll use the same clustering algorithm in the paper.

Besides the two important parts roughly described above, there is a lot of handy and

[1]see more details about hierarchical learning at this Wikipedia page and Chapter 14.2 from Friedman et al. (2001)

reasonable work in H&P to create a reasonable product similarity between public firms, and we will list and explain them below.

- Keeping only nouns and special nouns from business descriptions part and excluding geographical words like country and state names, top 50 city words in the United States, and the world from the corpus vocabulary used. This procedure could be considered as a fine-tuning process to the vocabulary, excluding the possible influences from geographical words, which is often appeared in the business descriptions but irrelevant with products.

- Keeping the same level valid pairs ratio (Granularity) as SIC and NACIS. Based on the 10K texts, it's easy to create a similarity score between any two firms. Because H&B would want to create a measure that has the same fraction of membership pairs as SIC-3 Industries or NAICS-4 Industries to compare them in an unbiased fashion. They first calculate the fraction of valid membership links defined by in the same industry, for SIC-3, it was 2.05 percent of all possible firm pairs. Then set a threshold to make the similarity score pairs calculated before having the same ratio of valid links.

- Normalize by the median, before the previous step. Because there were possible heterogeneous patterns of the similarity scores between different firms, for example, there could exist such condition that certain company is more similar to the other companies than another company. So they normalize the whole similarity matrix by subtracting the median score of each firm. According to the original paper, the external validation tests would be slightly weaker if omitting this step. This step, together with the previous step would break the symmetry property of firms' product similarity, which was not discussed in the original paper. For example, one company A that is within the text-based network industry of another company B doesn't have to mean company B is within company As network industry (defined as similarity score more than 0), and

the similarity scores are usually unequal between A to B and B to A.

- Excluding possible vertical relatedness. Using the use table of the Benchmark Input-Output Accounts of the US Economy to compute input-output flow based on SIC-4 digits, and set the pair similarity score to 0 if the input fraction was more than 1 percent between two companies. The result is robust to this procedure, that is why H&P concludes that the similarity score focus on the horizontal firm product offerings but not vertical production links.

From the constructing processes, one possible concern of the use of H&P's result may be that it could overemphasize the importance of special nouns like brands when creating document vector and compute similarity. For example, iPhone, belonging to the cellphone category, would usually only appear in Apple's 10K descriptions. In the document vector of H&P's work, it would become an element in the whole vector that could distinguish Apple from other cell phone companies. And if we use the TF-IDF, the case could be worse, cause iPhone, would have a very high inverse document frequency, overemphasizing the importance much more. This issue would directly affect the Section VII. Endogenous Barriers to Entry of the original paper.

Because the logic of that section was based on that advertising and research&development costs create barriers of products for companies, weakening the competition intensity of certain companies. And the competition intensity is measured by product similarity extracted before, the higher the total similarity to the other firms, the more intense the competition the company faced. Because considering all nouns and special nouns as the same weight couldn't successfully distinguish the successful products that have created such barriers or products from products with only a special brand name, the estimation results could be weakened a lot. But at the same time, could we successfully capture the heterogeneous characteristics of products just through the 10K business descriptions or other texts remains to be unknown.

The issue we discussed above should be attributed to the ambiguity of the conception, similarity or product similarity itself, and the way to construct product similarity, which may couldn't capture the heterogeneous characteristics of brands or trademarks. In this paper, we couldn't solve the issues above properly, but we may try to use some methods that may could weaken the effects of nouns and special nouns to illustrate it, for example, word embedding methods and LDA are both robust about the special nouns cause both models are based on the concurrences contexts of words. We will use TF-IDF, Word Embedding related methods, LDA based document similarity to compare with the similarity measures from the H&P. The details of each method would be discussed in the next part of this section. And implementation details like params of each method would be provided in Section III, data part.

Running time usually is not a major concern by economists and social scientists, but because of the huge difference in running time between all methods above, we still list a rough running time table for reference. Running time should vary a lot in different computer hardware, but the relative time cost of different methods should be the same roughly.

1.1 Basic TF-IDF

TF-IDF, short for term frequency-inverse document frequency, is a classic method in information retrieval to reflect the importance of the certain word to a document in a collection or corpus [2]. It creates a statistic which is the product of term frequency weight, measuring how often a certain word appears in a certain document, and inverse document frequency weight, measuring how unique/informative the certain word to the whole corpus. There are a lot different weighting function forms for both term frequency and inverse document frequency, for example, the raw count of a word in a document creates a measure about term frequency. Inverse document frequency could be the inverse fraction of the documents in the

[2]To see more details about TF-IDF at this Wikipedia page.

whole corpus that contain a certain word, and even a constant value. In H&P's work, they choose a binary term frequency weighting scheme, that takes value 1 if the word appeared in the document, 0 if not, and a constant value 1 with a document frequency cutoff 0.25 for inverse document frequency, which takes all words whose document frequency are no more than 0.25 from the corpus equivalent. The threshold here matters as Table 5 of the original Hoberg and Phillips (2016) shows but does not vary a lot.

After creating tf-idf statistic, it's easy to transform a document to a vector, each element of the vector represents a word from the whole corpus, while the value is the tf-idf statistic representing the importance of the certain word to the document. The norms of tf-idf document vector, usually depending on the length of documents and the weighting scheme of term frequency and inverse document frequency, doesn't have exact meaning, so it's common to normalize tf-idf vector with L^2 norm, then the product of two tf-idf vectors equalizes to the cosine similarity which is widely used as direction similarity measure.

1.2 Word Embedding Related Methods

Word embedding is set of methods that aim to transform words in vocabulary to a fixed-length dense vector of real numbers. Usually, the word in vocabulary could be expressed as one dimension per word space (bags of words), word embedding performs as an embedding from the high dimension space to a much lower but dense vector space.

Word2Vec(Mikolov et al., 2013) is one of those methods, based on a shallow, two layers neural network, and could create good representations of the word. Not only direction similarity, but also the linear structures behind, like $v_{France} - v_{Paris} \approx v_{England} - v_{London}$. In the applications to social science work, like Garg et al. (2018), they use the euclidean distance between two words, names, and some positive/negative words as measures for equality between genders and ethnicity. So we may expect that, if we could include the information from word embeddings into the similarity measure construct, we may weaken the effects of

special nouns, the issues we discussed above.

Though word embedding didn't create document-level similarity directly, there is some common practice creating document vector to compare document similarity, for example merely aggregating all word vectors appeared in the document could create a document vector, or word mover's distance Kusner et al. (2015) which we will discuss in details later. Both methods are relevant with the length or norms of word embedding, so the interpretations of the length of word vectors here are essential, but lacking detail discussions in the current literature. According to Schakel and Wilson (2015), the word that appears in different contexts may be updated in various directions during training, which may affect the final norms. An extreme example is the stop words, usually with very small norms compared to other words. However, the results are all based on experiments lacking solid background theory.

Besides Word2Vec, there were also some other word embedding methods, like FastText(a variant of Word2Vec), Glove, etc., and deep contextualized word representation has been developing a lot in recent years, like ELMo(Peters et al., 2018) and BRET (Devlin et al., 2018), we plan to include results of these methods in the future. Beyond embedding on the word level, there is also some work trying on constructing a document level embedding, like doc2vec(Le and Mikolov, 2014), Seq2Seq, which has been widely used in the translation. Doc2Vec has some good properties like the word embeddings, for example, the linear structures.

Because of the good property of embedding methods, there are also some practice on the embedding with covariates, like Rudolph et al. (2017) and Tian et al. (2018), we plan to include these work into comparison in the future.

1.3 Topic Models(LDA) Related

Following the definition in Wikipedia, a topic model is a type of statistical model for discovering the abstract *topics* that occur in a collection of documents[3]. The most widely used topic model is Latent Dirichlet Allocations (LDA,Blei et al. (2003)). LDA is a generative probabilistic model, with three-level hierarchical structures, topics, words, and documents. One of the research we mentioned in the introduction part, Hansen et al. (2017) is an example of applications of LDA based topic model on economics research.

Because the basic LDA doesn't model the correlations between topics, each topic could be considered as orthogonal. To generate document similarity from topic distributions, we could use Hellinger distance, cosine similarity, Bhattacharyya coefficient, Kullback-Leibler Divergence, etc. The last three measures are also used in Hansen et al. (2017) to construct communication measures.

Similar to the embedding methods, there are also some practice about including covariates into the topic model, like correlated topic model and structural topic model. The structural topic model has been widely used in social science like politics in the past five years (Adjust confoundings texts).In the current version of this paper, we would only include the base LDA model.

2 Data

Following a similar process like other studies based on 10K fillings from SEC, we first downloaded all 10K related files from SEC EDGAR Database using a network crawler. The 10K file formats vary a lot with time. In the early periods, the 10K files were plain text without rich formats, and it gradually changed to HTML formats, which increase the difficulty of extracting texts of a certain section. We follow a standard way to preprocess the whole raw

[3]see more at this Wikipedia page

texts, clearing HTML tags, and splitting the raw texts into different sections. Because most of the processes were done with a hard-coded criterion, for example, consider a single line with format Item Num as splitters between two sections, this method could work fine with most of the 10K files, but not all, some files may be broken. Limited by time and efforts, we could not verify and fix this kind of problem to all 10K files, it may create a little difference between our comparison and H&P's origin studies.

To keep consistency with H&P, we also restricted 10K types as HP used, keeping only 10-K, 10-K405, 10KSB, 10KSB40, dropping some types like 10-KA, which are amendment version of 10-K.

Before applying the methods we discussed in Section II, we have preprocessed all texts with standard procedures in natural language processing, like POS(pos of tagging), lemmatization. Each method may vary a little, and details are as follows.

- TF-IDF on nouns and special nouns. We first use the standard POS method in NLTK to get all nouns in text and use WordNet to lemmatize all nouns, removing plurals form. The words left are our focused vocabulary because we follow a similar procedure using nouns to capture product-related issues. We limit the number words to 10000 to reduce each 10K Business Description fillings to a 10000 length vector.

- Word Embedding related. If not using pretrained word embeddings, we first use the texts after lemmatization to train Word2Vec or Doc2Vec model and then focus only on the words list nouns and special nouns created above(Word2Vec Only). We choose 100, 300 dimensions of vectors in both Word2Vec and Doc2Vec. Doc2Vec will directly reduce each document into a 100/300 length vector. Word2Vec would create a 100/300 dimension vector for each word, and we choose to use the means of the word vectors of all nouns and special nouns in each 10K fillings as vector representation of the document.

- Latent Dirichlet Allocations. We use the same sets of data as TF-IDF, and choose 100, 150 topics in this paper. LDA would directly reduce the document into the number of topics dimension vector, representing the distribution of topics on the document.

After the dimension reduction, we transform each document into a fixed length vector. In some methods like TF-IDF and LDA, the elements in the vector have a clear explainantion, the weight of word or the probability of topic, but in word embedding related methods like Word2Vec and Doc2Vec, it's hard to interpret the meaning of each element in the vector. To compare the performance of document vectors from different methods, we follow H&B's work, use the cosine similarity between document vectors as the measure of similarity.

The external validation data like operating income, sales are also from Compustat, which are also reported in the same year 10K files. All 10K files have a company identifier called cik (central index key). The cik identifier is unique but the cik of a company could have changed by time because of many reasons. In the past, researchers usually have to rely on the address and company name to get historical links between cik and gvkey, luckily there is already a cik(identifier in SEC EDGAR database) and gvkey (permanent identifier in Compustat database) links in WRDS now, which had saved a lot of efforts.

Because we didn't link the database to CRSP currently, so in this paper, we would only include the variables in Compustat as external validation. In the future, we hope we could use the similarity score created from 10K files to create predictable returns like Lee et al. (2018), which should be more reasonable to illustrate the relations between new methods and information we learned by new methods.

2.1 Basic Comparison

We first directly compare those similarity metrics after median normalization but without setting the granularity threshold, because after setting the granularity threshold, different

similarity measures would drop different pairs, so we would only be able to compare the correlations within the inner pairs set. We use Jaccard index, the size of the intersection divided by the size of the union of the sample sets, to compare the similarity of the neighbors between different methods. A formal defintion of Jaccard index is

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Table for Raw Corrleations between TF-IDF, LDA, Word2Vec and Doc2Vec

Table for Jaccard index between TF-IDF, LDA, Word2Vec and Doc2Vec

2.2

3 External Validation

4 Conclusions

5 Appendix

References

Athey, Susan, 2017, The impact of machine learning on economics, in *Economics of Artificial Intelligence* (University of Chicago Press).

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2016, Measuring economic policy uncertainty*, *The Quarterly Journal of Economics* 131, 1593–1636.

Blei, David M, Andrew Y Ng, and Michael I Jordan, 2003, Latent dirichlet allocation, *Journal of machine Learning research* 3, 993–1022.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Donaldson, Dave, and Adam Storeygard, 2016, The view from above: Applications of satellite data in economics, *Journal of Economic Perspectives* 30, 171–98.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart, 2018, How to make causal inferences using texts, *arXiv preprint arXiv:1802.02163* .

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, 2001, *The elements of statistical learning*, volume 1 (Springer series in statistics New York, NY, USA:).

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou, 2018, Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences* 115, E3635–E3644.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei, 2017, Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states, *Proceedings of the National Academy of Sciences* 201700035.

Hansen, Stephen, Michael McMahon, and Andrea Prat, 2017, Transparency and deliberation within the fomc: a computational linguistics approach, *The Quarterly Journal of Economics* 133, 801–870.

Henderson, J Vernon, Adam Storeygard, and David N Weil, 2012, Measuring economic growth from outer space, *American economic review* 102, 994–1028.

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

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger, 2015, From word embeddings to document distances, in *International Conference on Machine Learning*, 957–966.

Le, Quoc, and Tomas Mikolov, 2014, Distributed representations of sentences and documents, in *International Conference on Machine Learning*, 1188–1196.

Lee, Charles M.C., Stephen Teng Sun, Rongfei Wang, and Ran Zhang, 2018, Technological links and predictable returns, *Journal of Financial Economics* .

Meade, Ellen E, 2005, The fomc: preferences, voting, and consensus, *Federal Reserve Bank of St. Louis Review* 87.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Mullainathan, Sendhil, and Jann Spiess, 2017, Machine learning: an applied econometric approach, *Journal of Economic Perspectives* 31, 87–106.

Perry, Patrick O, 2009, Cross-validation for unsupervised learning, *arXiv preprint arXiv:0909.3052* .

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 2018, Deep contextualized word representations, in *Proc. of NAACL*.

Rudolph, Maja, Francisco Ruiz, Susan Athey, and David Blei, 2017, Structured embedding models for grouped data, in *Advances in Neural Information Processing Systems*, 251–261.

Schakel, Adriaan MJ, and Benjamin J Wilson, 2015, Measuring word significance using distributed representations of words, *arXiv preprint arXiv:1508.02297* .

Tian, Kevin, Teng Zhang, and James Zou, 2018, Cover: Learning covariate-specific vector representations with tensor decompositions, *arXiv preprint arXiv:1802.07839* .