Text-based industry classification by Autoencoder

Kyounghun Bae[1], Daejin Kim[2], Rocku Oh[3]

Abstract

Industry classification is one of the crucial issues in financial analysis. Classical industry classification systems have several limitations as an aspect of outdated and multiple assignments. In this paper, we propose an improved industry classification methodology based on firms' business descriptions by reducing high dimensional vector space to lower dimension using autoencoder to avoid a curse of dimensionality problem. The main contribution of this paper is first, we effectively reduce the dimension of word vector to cluster firms into distinctive industries by utilizing the autoencoder. The reduced dimensions overcome the limitation of Hoberg and Phillips clustering method in which the word vector is large and highly sparse. Second, we are able to visualize and clarify the boundary of industries based on the lower dimensional information extracted from the business description text. The graphical closeness between industries is able to describe the industry-level relationship as well as the closeness between individual firms which are originally involved in conflicting assignment problem in terms of the classical classification scheme.

Keywords: Industry classification, Dimensionality reduction, Autoencoder, Textual analysis

1 Hanyang University, Seoul, Korea, Republic of; Tel: +82 2-2220-1044; E-mail: khbae@hanyang.ac.kr
2 Ulsan National Institute of Science and Technology, Ulsan, Korea, Republic of; Tel: +82 52-217-3169; E-mail: daejin@unist.ac.kr
3 Ulsan National Institute of Science and Technology, Ulsan, Korea, Republic of; Tel: +82 52-217-3169; E-mail: org817@unist.ac.kr

1. Introduction

Industry classifications have been an important issue for practitioners as well as scholars in the field of finance and economics. Researchers in academia have adopted a variety of approaches to group homogenous firms to analyze industry-based studies such as consequence of corporate reorganization or changes in financial and investment policy. Traditionally, there are three types of method to classify similar firms in the financial sector. Standard Industrial Classification (SIC) code aggregates firms by using information on selling end products and similar production process (Chan, Lakonishok, & Swaminathan, 2007). Despite the wide application of SIC codes in previous empirical studies, many researchers have questioned the usefulness of SIC code as a standard for the industry classification. Walker and Murphy (2001) mention that this SIC classification system does not sufficiently reflect a shifting of the main product, business process, and emerging markets because the system mainly emphasizes on manufacturing operations relative to service processes. Fama and French (1997) propose 49 industry grouping system by merging several ranges of four-digit SIC codes. The system aggregates similar subgroups into one group. Some firms coded 22, 23,30, and 35, which are related to "transportation equipment" from major groups of two-digit SIC codes are grouped as Fama-French classification code "automobiles and trucks" of 37. The Fama-French classification is yet most influential in the academic area especially on asset pricing, corporate finance, accounting, and investment. How well their classification system produces groups of economically similar firms is still under debate. Fama and French, however, do not provide sufficient evidence of the performance of their classification system despite its broad usage in financial fields. The third approach to classifying the firms is the Global Industry Classification System(GICS). The method is widely used by the investment analysts and portfolio managers. The system categorizes firms based not only on operational characteristics of each firm but also on the investors' perceptions of what constitutes the firm's mainstream of their business (Kile & Phillips, 2009). Because of the characteristics, certain firms may be categorized into different industries based on the GICS and SIC system respectively. For example, the main business of the GATX Corporation is to lease and operate railroad equipment and ships. The firm is classified as a financial sector by GICS but is assigned in the transportation equipment sector as SIC codes simultaneously. These conflicting assignments arise whenever firms are not only involved only in one specific industry sector, product or business process. The

classification systems mentioned above also cannot reflect sufficiently diversified information on their business products and processes as well as new technologies and emerging markets.

To overcome the drawbacks of the previous industrial classification system, Hoberg and Phillips (2016) utilize the text information on product description of firms reported to the Securities and Exchange Commission (SEC). They propose a new industry classification system based on a strong tendency of vocabulary usage among firms operating in the same market. The method is based on a text description of products and process that firms supply to the market. By analyzing the word vocabulary from the documents each firm report, the approach allows us to generate new text-based industry groups and be able to capture the changes in firms' business, namely diversification and pivoting as well. They measure the pairwise cosine similarity of the word vectors extracted from the reported document of firms. The measure can identify how similar a business description of the firm is compared to all the other firms. They cluster 300 industries based on the pairwise comparison result to distinguish peer firm as a rival and competitor. The pairwise comparison approach, however, is limited due to providing only one-dimensional information, which means that the measure can only represent firm-to-firm information, not firm-to-industry and industry-to-industry. we cannot infer the overall map of the industry closeness and relationship although they validate the across industry variation of their final clusters. In addition, a research shows that the distance measure, the concept of proximity, may not even be qualitatively meaningful in high dimensional space (Aggarwal, Hinneburg, & Keim, 2001). The cosine similarity measure is mathematically identical to the L2-normalized Euclidean distance. The dimension of word vectors used in their research is larger than 60000, which implies that the similarity measures by the vectors cannot escape from the curse of dimensionality problem because the space of the cosine similarity is still high-dimensional (Skillicorn, 2012).

We utilize the deep autoencoder to reduce the dimensionality of word vectors to be used in industry classification to mitigate the high dimensionality problem of the clustering based on the cosine similarity comparison in the previous research. Our method reduces the number of features by using the machine learning algorithm rather than directly use the whole word vector extracted from the business description of 10-K annual reports. We show that the dimensionally reduced features can sufficiently describe the relationship between firms by visualization. We utilize a spherical k-means algorithm to cluster the industries. The clustering process is processed by using the reduced features from the coded layer output of the

autoencoder. We then compare the clustering result of the proposed method with the Fama-French 12 industry classification scheme.

The main contribution of this paper is first, we improve the text-based industry classification method from Hoberg and Phillips (2016) by effectively reducing the dimensions of word vector utilizing machine learning technique. The reduced dimension overcomes the limitation of Hoberg and Phillips clustering method of which the word vector is large and highly sparse. Second, we are able to visualize and specify the boundary of industries. The graphical relativeness can describe the relationship between industries as well as individual firms which are originally involved in conflicting assignment problem as an aspect of the classical classification scheme such as SIC and GICS code.

2. Data (Business descriptions of the 10-K annual report)

We collect 10-K annual reports filed by the Securities and Exchange Commission (SEC) using web crawling algorithm to extract business description parts of each document from 2013 to 2016 which results in the total number of 21,631 10-K reports. The 10-K business description section is legally required for all firms followed by item 101 of Regulation S-K which requires that firms describe the core products they offer to the market. After we clean the unwanted parts, we merge the corresponding SIC codes using Compustat historical segment data from Wharton Research Data Services (WRDS). The reporting period and financial activity period of the Compustat data could be different and redundant because each firm should report the information for the past three years at every reporting period. We, thus, only keep the latest reporting year for a given year at which financial activity occurred. For example, a certain firm reports their segment information for 2013, 2014, and 2015 in 2015. For this reason, we only keep the information of 2013 in the case that the reporting year is 2015. We use primary SIC code for the segment information of each firm and year. We remove observations when a firm has missing primary code at a certain year. The CIK code and reporting year are used to merge the Compustat data and 10-K annual report document that we crawled. After merging the SIC code by individual firms and year, the total number of samples merged with the SIC code is 18072. The documents of the business descriptions are processed to represent a bag of words and construct word vectors in the next step.

3. Methodology

3.1. Bag of words representation

The words from the 10-K business description can describe certain business process and products that each firm offers in the markets. The underlying hypothesis is based on the notion that firms classified in the same industry use more similar words to describe and offer their business and products than the firms classified in the different industries. Table 1 illustrates the most used 20 words extracted from the business description of sample firms which belong to different industries. For example, SANDISK CORP uses the word “memory” 67 times to describe their business and products in the business description part of the 10-K annual report in 2013. For a generalized form, we utilize a bag of word representation to convert the business description to a vector form. The bag of word representation is widely used in information retrieval and text mining (Singhal, 2001). The bag of word is a vector where each component of the vector is matched to the word aggregated set of documents. The representation is assumed that words appear independently, and the order of the words is immaterial. The number of words, thus, corresponds to the dimensions in vector space and each document then becomes an individual vector containing non-negative values on each dimension.

Insert Table 1 in here

In this study, a conventional text mining process is applied to structure a bag of words which is a set of unique words from the preprocessed documents. We only focus on nouns and proper nouns that use no more than 20 percent of the individual documents (business descriptions) in order to remove commonly used words. The threshold of 20 percent is selected based the Hoberg and Phillips (2016). The research indicates that the minor modifications of the threshold do not affect the main result of firms significantly. We remove geographical words such as country names as well as the name of the popular cities in the world. The business description text of firms which contains fewer than 20 unique words were excluded due to a lack of unique information of its firm. Suppose that there are unique words from all the

documents reported by all firms in training samples. We construct vector $\boldsymbol{W}$ which uses 2000-unique-word in order of frequent appearance among all the unique words after excluding the common words as we mentioned before. The vector $\boldsymbol{W}$ contains 2000-unique-word is called a bag of words in this research. The bag of words is utilized to convert the individual business description of firms to the corresponding word vectors respectively. A business description of given firm i can be represented as a 2000-dimensional binary (coded) vector $\boldsymbol{V_i}$. Each element $\boldsymbol{v_{ij}}$ of the vector $\boldsymbol{V_i}$ is allocated by the value of 1 if the given bag of words $\boldsymbol{W}$ contains the word element from the word list of individual document and otherwise 0. Figure 1 depicts a schematic illustration for the codification of word vectors. For Suppose there is a word "memory" in Firm $\boldsymbol{i}$'s business description. The system checks whether the word "memory" exists in the bag of words $\boldsymbol{W}$. In this example, the element $\boldsymbol{v_{i1}}$ is coded as a value of 1 because the bag of word has the word "memory" at the location of $\boldsymbol{w_1}$.

Insert Figure 1 in here

3.2. Dimensionality reduction by using the autoencoder

Hoberg and Phillips (2016) utilize a whole set of unique words, 60000-dimensional word vector, to compute the pairwise similarity of two words vectors using a cosine similarity measure. The number of nouns including proper nouns extracted from the product descriptions of each firm is about 500 on average based on the Hoberg and Phillips (2016), which implies that the remaining 59500 elements of word vector are coded as a value of 0. The sparsity of our samples is depicted in Figure **2**. This high dimensional and sparse vector space arises an issue of the curse of dimensionality when computing the cosine similarity and applying the clustering method as we mentioned in the introduction part. The cosine similarity formula can be directly represented in terms of the normalized Euclidean distance, which indicate that the cosine similarity measure cannot mitigate the problem of clustering based on the Euclidean distance when the vectors space is very high dimensions. In addition, Radovanović, et al. (2010) refers that the tendency of the ratio between standard deviation and mean of the distribution of all pairwise distances within a data set to converge to 0 as dimensionality increases. The result implies that the expectation of pairwise cosine similarity measure becomes a constant and its variance shrinks as increasing dimensionality.

Insert Figure 2 in here

To mitigate the high dimensionality problem of Hoberg and Phillips' approach, we utilize the dimensionality reduction method by using one of the machine learning techniques. The autoencoder is a dimensionality reduction technique based on the deep neural network, which is first introduced by Baldi and Hornik (1989) and improved by Hinton and Salakhutdinov (2006) toward a generative model by applying the greedy layer-wise pre-training technique. It is one of the unsupervised learning technique in machine learning field where the number and shape of the output node are the same as the number and shape of the input node. The first research was to implement PCA-like dimensionality reduction result using backpropagation method in the shallow neural network. The model consists of two parts which are encoding layers and decoding layers. The purpose of this model is to reduce the number of hidden (encoded) nodes while minimizing the difference between the input vector and the reconstructed output vector at the end of the model. The essential advantage of the autoencoder is to utilize the notion of the sparsity of the hidden layers when the model is trained and can be applied to a non-linear problem beyond the PCA approach. According to Hinton and Salakhutdinov (2006), when the cosine of the angle between two vectors is used to measure similarity, the autoencoder clearly outperformed latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), a well-known document retrieval method based on PCA. Autoencoders also outperform local linear embedding, a recent nonlinear dimensionality reduction algorithm (Roweis & Saul, 2000). We implement this regime of the autoencoder as a dimensionality reduction technique to represent a low-dimensional vector containing information of industry that each firm belongs to.

The dimension of the 2000-word-vector reduced to hidden nodes of 10 which is used as features for the clustering step. A whole system to extract the latent values is depicted in Figure 3. A layer structure of the learning model is 2000-500-125-10-125-500-2000 with the ReLU as an activation function of each layer except for the last of the encoding layer and last of the decoding layer. The last encoding layer uses a linear function which does not bound the lower limit of the values of the reduced vector. We minimize the value of the binary cross-entropy as a loss function used in the training stage. To measure a value of the loss function,

the sigmoid function is selected when computing the loss function at the end of the decoding (output) layer. This structure sufficiently discriminates and explains information on the 2000-length word vectors even we modify the number of a hidden node of the last encoding layer.

Insert Figure 3 in here

3.3. Industry classification by the spherical K-means clustering

The objective of standard K-means clustering is to minimize the mean-squared error of Euclidean distance within the clusters. It is well-known that the underlying probability distribution for the standard k-means algorithm follows the Gaussian. We rather use a spherical K-means clustering algorithm which is a suitable clustering algorithm for the vector space model of the corpus because the direction of a word vector of the document is more important than the magnitude itself. For high-dimensional data such as text documents, cosine similarity is a superior measure compared to the Euclidean distance (Strehl, Ghosh, & Mooney, 2000). The spherical k-means algorithm maximizes the average cosine similarity within the clusters. The main difference from standard k-means is that the re-estimated mean vectors need to be normalized to unit-length and the underlying probabilistic models are not Gaussian.

We cluster the reduced features into several industries by applying the spherical k-means algorithm. The result of the spherical k-means clusters is directly compared to the SIC, Fama-French 49 industry classification code, GICS code, and Hoberg and Phillips' fixed code. We compute the cosine similarity between firms based on the reduced features as well. We validate the proposed method by comparing within and across variations of the operating income divided by total asset, operating income divided by total sales, and estimated market beta.

4. Result

4.1. Qualitative approach

For a visual representation, we modify the number of hidden nodes at the last of the encoding

layer. The modified layer structure is 2000-500-125-2-125-500-2000. The structure only differs from the previous model with the number of the node at the coded layer which is two. Figure 4 (A) shows the 2D scatter plots of latent values from the last encoding layer after training the autoencoder. The figure shows that firms labeled as a same Fama-French industry code are clustered on a single spike from the origin, which indicates that firms in the same industry have similar directions with different vector norm. For example, the SIC code range of the cluster colored by yellow-green (Figure 4 (A)) is labeled as 6000 to 6999 which is financial firms based on the Fama-French industry classification. This result is consistent with the Hoberg and Phillips' clustering result because they use the cosine similarity to measure the degree (distance) between two business descriptions. Figure 4 (B) shows the result of the spherical K-means clustering. The clustering result indicates that financial firms and medical-related firms are divided into several sub-industries.

Insert Figure 4 in here

There is a case that some firms are far from their industry groups(spike) in terms of the Fama-French classification code. It indicates that the firms select and use different words when they describe their business process and products compared to other firms belongs to their industry on the basis of the SIC and Fama-French industry classification code. These firms may improperly report their business description or SIC code of the firms are wrongly assigned because the firms recently change their major products or business process through diversification and pivoting. The method proposed in this research is able to solve the latter problem while determining more proper industries by their words used in the annual reports.

Text-based network similarities by Hoberg and Phillips reveal that firms in newspaper publishing and printing industry (SIC-3 is 271) are highly similar to firms in radio and broadcasting stations industry (SIC-3 is 483). This example illustrates the fact that both industries are likely to cater for the same customers who want advertising and exhibit at least some substitution. The example shows that traditional classifications treat these industries as entirely unrelated. Because our classifications are based on the actual product description of firms reported, we are able to detect potential peer firms that offer related products even if they are not currently connected. We investigate the similar cases as they discuss. As shown in

Figure 5 (Case 1), a SIC code of a firm “TEAM HEALTH HOLDINGS INC” is 7363, which the Fama-French classification code is 12 (Others) as listed in Table 2. Another firm named as “WELLCARE HEALTH PLANS INC” is coded as the SIC code of 6324 and the Fama-French classification code of 11 (Money). The former firm provides outsourced healthcare professional staffing and administrative services to the hospital. The latter firm provides managed care services government-sponsored healthcare program. The two firms are clustered in the same industry based on the proposed method when the number of clusters is 12. The other firms clustered with the two firms are in the SIC code range of 8000-8099. The firms related to healthcare products and services are in the purple ellipse along the similar direction (angle).

Insert Table 2 in here

The second case is related to the energy and mining industry including oil, gas, and materials. The first two firms of the “Case 2” in Table 2 belong to industry related with “Shops” in terms of the Fama-French classification and SIC code. However, the scatter point of the two firms are closer to the “Energy” firms coded as Fama-French classification code of 4 in terms of the product and business process that they provide in the markets. Figure 5 (Case 2) illustrates the situation of the firms mentioned in Table 2. Many firms involved in the industry of oil, gas, and coal extraction and products are grouped with the red marks. Table 3 is the example of words overlapped by documents reported by the firms in “Case 1” and “Case 2”.

Insert Table 3 in here

The third case is related to the sub-industry issue. Three or four spike-groups are seen within a financial industry depicted as a green dashed-circle as shown in Figure 5 (Case 3). The Case 3 illustrates that the financial industry can be divided into different sub-industries compared to the Fama-French 12 Classification code. Each spike indicates different clusters in terms of the word presented in their business descriptions even though the firms are all related to money or financial industry. We are able to clarify the closeness and relationship of these firms as a visualization and divide this industry into subgroups(sub-industries). In addition to

subgroup division, most of the firms in subgroups of the financial industry located at the green circle (ellipse) are insurance firms such as EMPLOYEE HOLDINGS and ALLIED WORLD ASSURANCE CO HOLDINGS. The financial firms in the blue circle are security brokers and dealers such as GOLDMAN SACHS, MORGAN STANLEY, CBOE HOLDINGS and CME GROUP. The insurance firms in the green circle are much closer to the industry of healthcare services in the purple circle (Case 1 group) than the security brokers and dealers in the blue circle because the security brokers and dealers change their core business to IT services rather than traditional financial services nowadays.

Insert Figure 5 in here

4.2. Within and across industry variations

We validate the informativeness of our proposed method by comparing within and across industry variations in profitability and market risk (market beta) because the financial factors are correlated with the industry the firms belong to. We compute the degree of industry variation using firm-size-weighted approach. We first compute the value of a given characteristic as the firm-size-weighted means among its industry peers. We then calculate across-industry variation as the standard deviation of these industry characteristics across all firm-year samples. In the case of the within-industry variation, we compute the standard deviation of characteristics among the industry peers first. We compute the industry-size-weighted average of the standard deviation. We expect that the proposed method to have higher across-industry variation and lower within-industry variation that the other classification systems including SIC-3digit industries, GICS sub-industries, and TNIC fixed code.

Panel A of Table 4 shows the results of the across industry variation. The table shows that text-based industry classification systems have more across industry variation than SIC 3-digit and GICS sub-industries. We define operating income divided by total sales(OI/sales) as the profitability of the firm. The across-industry variation is 0.390, 3.455, and 4.493 for SIC 3-digit, GICS sub-industry, and TNIC 300 fixed code, respectively. Spherical k-means applied by dimensionality reduction technique, Autoencoder, result in the across variation as 10.819, which is 2.4 times larger value than the variation of the original TNIC 300 fixed code. The variation of operating income divided by total asset also increases to 0.150 from 0.139

compared to original TNIC 300 fixed code. We observe similar improvement regarding market beta. We validate that it is more informative to consider the business description of firms' report than both SIC 3-digit and GICS sub-industry classification. It is also validated that reducing the dimension of the word vector by applying autoencoder improves the informativeness of the text-based industry classification approach as well. Panel A also shows that applying the autoencoder clearly improves the performance of intransitive TNIC in terms of OI/assets and market beta. The across-industry variation of 19.081 for original intransitive TNIC increases by 5.36 percent to 20.103 regarding OI/assets. The variations of 0.678 increases by 3.69 percent to 0.703 for the market beta as well.

Panel B of Table 4 shows the results of the within industry variation. The table shows that text-based industry classification systems are an effective way to group homogenous firms in terms of firm's characteristics we consider comparing with the SIC 3-digit and GICS sub-industry system. The within-industry variation is 18.296, 13.823, and 1.243 for SIC 3-digit, GICS sub-industry, and TNIC 300 fixed code regarding OI/sales. The Spherical k-means algorithm with the autoencoder results in the within variation as 5.857, which is 57.2 percent smaller value than the variation of the original TNIC 300 fixed code. Regarding OI/asset also decreases to 0.113 from 0.130 compared to original TNIC 300 fixed code. We also validate that reducing the dimension of word vector by applying autoencoder improves the informativeness of the text-based industry classification approach as well. Panel B also shows that applying the autoencoder clearly improves the performance of intransitive TNIC in terms of OI/sales and market beta. The within-industry variation of 8.655 for original intransitive TNIC decreases by 50.9 percent to 4.250 regarding OI/sales. The variations of market beta also slight decreases when we use dimensionally reduced word vectors. We conclude that reducing the dimension of vector space increases the informativeness of the text-based industry classification system in overall.

Insert Table 4 in here

4.3. Curse of high dimensionality

As we mentioned above, Radovanović, et al. (2010) refers that the tendency of the ratio between standard deviation and mean of the distribution of all pairwise cosine similarity to

converge to 0 as dimensionality increases. The result implies that the expectation of pairwise cosine similarity measure becomes a constant and its variance shrinks as increasing dimensionality. In addition, Aggarwal, et al (2011) shows that the distance measure, the concept of proximity, may not even be qualitatively meaningful in high dimensional space. Figure **6** depicts the curse of dimensionality issue intuitively. The cosine similarity measure of all firm-pairs from the TNIC database they provide have values lower than 0.2. The descriptive statistics of the two approach is described in Table **5**. The average similarity of the original TNIC approach is 0.073 which is extremely skewed. The small average value of the cosine similarity indicates that most of the firm pairs are extremely different from each other. The standard deviation of the original TNIC approach is 0.063 as well. The result indicates that most of the firm pairs are not separated to discriminate industry peers. The simple statistics show the curse of dimensionality as the vector space increases. Figure **6** (right) depicts the cosine similarity values when dimensionality reduction technique, autoencoder, is applied. The mean and the standard deviation of the cosine similarity measures is 0.521 and 0.173, respectively. We conclude that Applying the autoencoder to the word vector effectively mitigate the curse of dimensionality problem that original TNIC algorithm has.

InsertTable **5** ***Figure 6 in here***

Insert Table 5 in here

5. Conclusion

In this paper, we gather 10-K annual reports filed with the Securities and Exchange Commission (SEC) using web crawling algorithm to extract business description part of each firm. A previous work done by Hoberg and Phillips (2016) utilize high-dimensional word vectors to cluster the firms. However, clustering using cosine similarity measure when the vectors are sparse and high dimensions arises a curse of dimensionality problem. We then use a deep learning technique which is called autoencoder as a dimensionality reduction method to reduce the dimension of the original word vector to mitigate a high dimensionality problem in

vector space. The reduced features containing the information of business description are grouped by using spherical K-means clustering algorithm which is suitable for the vector space model of the corpus. The number of clusters is 12 for a consistency to the Fama-French's classification. A visual representation of the clustered firms shows that that firms labeled as similar SIC codes are well-clustered, which indicates that firms in the same industry based on the SIC code have a similar radius with different vector norm as an aspect of the cosine similarity system. The classification result also shows the similarity and closeness between industries. The relative similarity can also describe the industry-level relationship as well as the position of individual firms which were originally involved in conflicting assignment problem in terms of the classical classification scheme.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *In International conference on database theory*, 420-434

Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks, 2*(1), 53-58.

Chan, L. K., Lakonishok, J., & Swaminathan, B. (2007). Industry classifications and return comovement. *Financial Analysts Journal, 63*(6), 56-70.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391.

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of financial economics, 43*(2), 153-193.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science, 313*(5786), 504-507.

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

Kile, C. O., & Phillips, M. E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance, 24*(1), 35-58.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). On the existence of obstinate results in vector space models. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 186-193.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science, 290*(5500), 2323-2326.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull., 24*(4), 35-43.

Skillicorn, D. B. (2012). Understanding high-dimensional spaces: *Springer Science & Business Media.*

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on web-page clustering. *In Workshop on artificial intelligence for web search (AAAI 2000), 58,* 64.

Walker, J. A., & Murphy, J. B. (2001). Implementing the North American Industry Classification System at BLS. *Monthly Labor Review*, 15.

Table 1. The 20 most frequent words presented in the business description of sample firms with different industries in 2013. The 20 most frequently used words are described. The number of the occurrence frequency of each word is the value in parentheses after the word. For example, SANDISK CORP uses the word "memory" as 67 times when describing their business and products in the business description part of the 10-K annual report in 2013.

Firm 1: SANDISK CORP (SIC code: 3572)
Business: Flash Memory Storage
Core words: memory(67), product(52), technology(44), storage(36), market(31), device(31), solution(28), NAND(26), flash(24), drive(20), manufacturer(19), design(19), corporation(18), venture(18), card(18), president(17), data(16), wafer(16), cost(15), year(15)
Firm 2: SCHEIN (HENRY) INC (SIC code: 5047)
Business: Healthcare Distribution
Core words: health(102), product(89), care(65), service(62), state(47), customer(47), law(44), practice(41), president(40), business(40), distribution(37), sale(35), drug(33), act(30), vice(30), practitioner(26), officer(25), technology(24), order(23), management(23)
Firm 3: IMPAC MORTGAGE HOLDINGS INC (SIC code: 6162)
Business: Long-term Portfolio
Core words: mortgage(174), loan(159), origination(53), portfolio(49), service(45), estate(34), operation(33), channel(33), Mae(30), interest(30), correspondent(29), lending(27), rate(27), credit(21), security(21), broker(20), sale(20), borrower(19), seller(18), act(17)
Firm 4: TENGASCO INC (SIC code: 1311)
Business: Oil & Gas
Core words: company(200), gas(96), methane(44), production(43), well(37), oil(36), agreement(33), hoactzin(30), pipeline(28), management(27), program(26), interest(25), sale(25), property(24), operation(23), swan(23), report(22), project(21), price(21), field(21)
Firm 5: ACCELRYS INC (SIC code: 7372)
Business: Software Development
Core words: product(82), software(54), customer(46), platform(34), development(32), service(29), data(27), solution(26), process(24), market(24), acquisition(24), research(24), enterprise(23), industry(23), organization(22), quality(19), informatics(19), system(18), management(18), information(17)

Table 2. The clustering result of the sample firms allocated in different SIC and Fama-French classification code ranges. The first case (Case 1) is in an industry related to healthcare, medical equipment, and drugs. TEAM HEALTH HOLDINGS INC and WELLCARE HEALTH PLANS INC are classified as a "Money" and "Others" sectors respectively in terms of the Fama-French classification system. The proposed method clusters the two firms into the same industry or sector. The second case (Case2) is in an industry related to oil, gas, and coal extraction and products. GENESIS ENERGY LP and CROSSTEX ENERGY LP are classified as a sector of "Shops" in terms of the Fama-French classification system. The proposed method clusters the two firms in the same industry with energy-related firms.

Firm name	SIC code	Fama-French 12 Classification code		Clustered code
Case 1 - Healthcare, Medical Equipment, and Drugs				
TEAM HEALTH HOLDINGS INC	7363	12	Others	11
WELLCARE HEALTH PLANS INC	6324	11	Money	11
SELECT MEDICAL HOLDINGS CORP	8069	10	Hlth	11
SYMBION INC TN	8011	10	Hlth	11
LHC GROUP INC	8082	10	Hlth	11
LIFEPOINT HOSPITALS INC	8062	10	Hlth	11
TENET HEALTHCARE CORP	8062	10	Hlth	11
AMN HEALTHCARE SERVICES INC	8090	10	Hlth	11
HCA HOLDINGS INC	8062	10	Hlth	11
Case 2 - Oil, Gas, and Coal Extraction and Products				
GENESIS ENERGY LP	5171	9	Shops	4
CROSSTEX ENERGY LP	5172	9	Shops	4
HOLLY ENERGY PARTNERS LP	4613	12	Others	4
GULFPORT ENERGY CORP	1311	4	Energy	4
CONTINENTAL RESOURCES INC	1311	4	Energy	4
UNIT CORP	1311	4	Energy	4
MID CON ENERGY PARTNERS LP	1311	4	Energy	4
CHEVRON CORP	2911	4	Energy	4

Table 3. The most overlapped words used in a set of business descriptions in the same industry. The 9 firms related to the healthcare, medical equipment, and drugs (Case 1 in Table 2) uses words “hospital”, “billing”, “Medicare” and etc. as 9 times to describe their business and product in the 10-K annual report. The word “abuse”, “therapy”, “HIPPA” is occurred 8 times out of the 9 firms.

Unique words out of 2000 words in the bag-of-words	The number of occurrences
Case 1 - Healthcare, Medical Equipment, and Drugs	
hospital, billing, physician, Medicaid, productivity, patient, submission, reimbursement, Medicare, referral	9 times out of 9 documents
beneficiary, methodology, recruitment, length, abuse, accountability, authorization, accreditation, CM, associate, prohibition, utilization, therapy, transition, employer, sanction, eligibility, safeguard, notification, fraud, worker, HIPPA(Health Insurance Portability and Accountability Act), spending, portability, admission, antikickback, update	8 times out of 9 documents
Case 2 - Oil, Gas, and Coal Extraction and Products	
crude, commodity, pipeline, hydrocarbon, petroleum, transport	8 times out of 8 documents
proximity, carrier, cleanup, liquid, pollution, barrel, index, discharge, mile, tank, basin, emergency, exploration, drilling, commerce, injection, FERC(Federal Energy Regulatory Commission), shale, formation, greenhouse, dioxide, emission, gathering, fuel	7 times out of 8 documents

Table 4. Within and Across industry variations by different industry classification systems.

Industry classification systems	# of industries	Weighed OI/asset	Weighed OI/sales	Weighed OI/asset
	A. Across-Industry Standard Deviations: Firm-size weighted Result			
1. SIC 3-digit industries	245	0.390	0.066	0.741
2. GICS sub-industries	157	3.455	0.136	0.619
3. TNIC 300 fixed industries	300	4.493	0.139	0.809
4. Autoencoder + SKmenas	300	10.819	0.150	0.924
5. Original Intransitive TNIC	300	0.125	19.081	0.678
6. Autoencoder + Intransitive TNIC	300	0.115	20.103	0.703
	B. Within-Industry Standard Deviations: Industry-size weighted Result			
1. SIC 3-digit industries	245	0.126	18.296	0.884
2. GICS sub-industries	157	0.143	13.823	0.803
3. TNIC 300 fixed industries	300	0.130	10.243	0.980
4. Autoencoder + SKmenas	300	0.113	5.857	0.856
5. Original Intransitive TNIC	300	0.124	8.655	1.055
6. Autoencoder + Intransitive TNIC	300	0.132	4.250	0.996

Table 5. The summary statistics of cosine similarity measures in terms of the original TNIC and the applied autoencoder of the TNIC system.

Industry classification systems	Mean	Standard Deviation	Min	Max
Original intransitive TNIC	0.073	0.063	0.000	0.904
Autoencoder + Intransitive TNIC	0.521	0.173	0.011	0.988

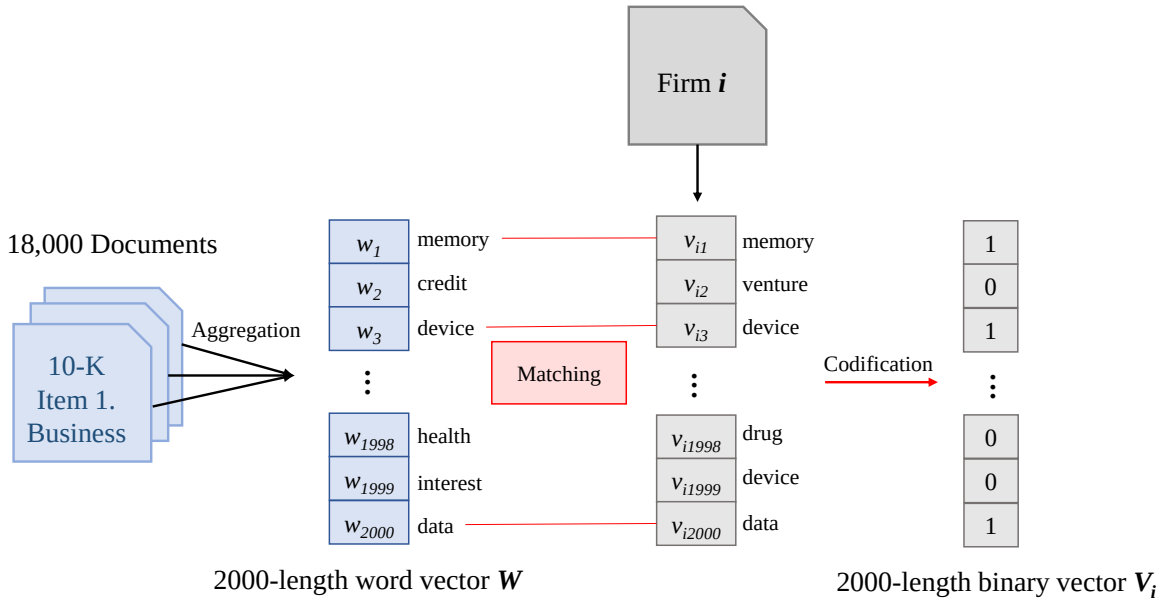

Figure 1. An example of word vector codification using the bag of word representation. The Vector W denotes

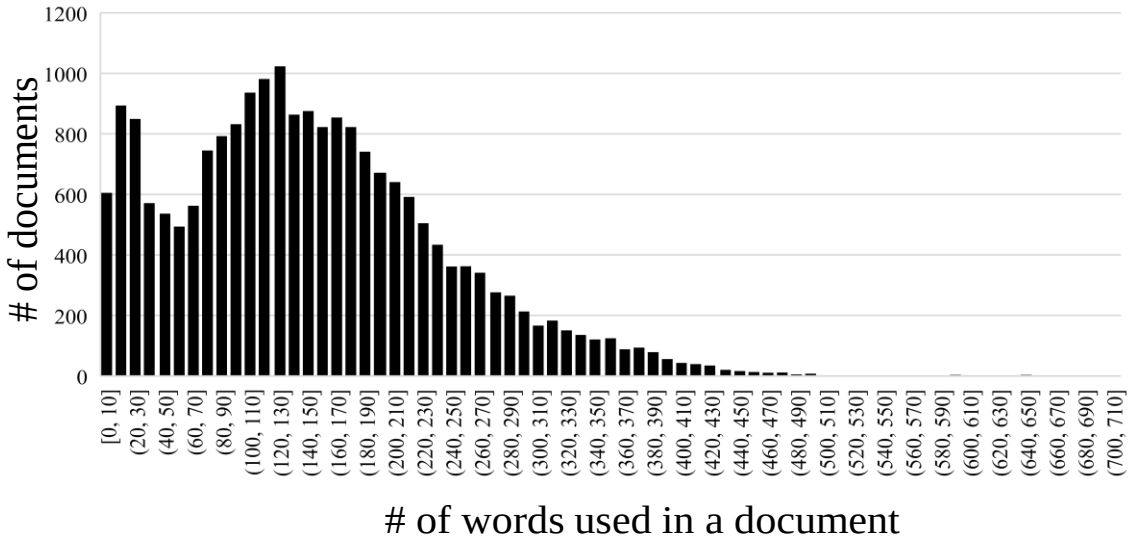

Figure 2. A histogram of the number of words used in business descriptions of each firm. The average number of unique words used in the business description is about 143. The distribution of the unique words by documents are positively skewed. The distribution describes that the word vector by a bag of words is highly sparse.

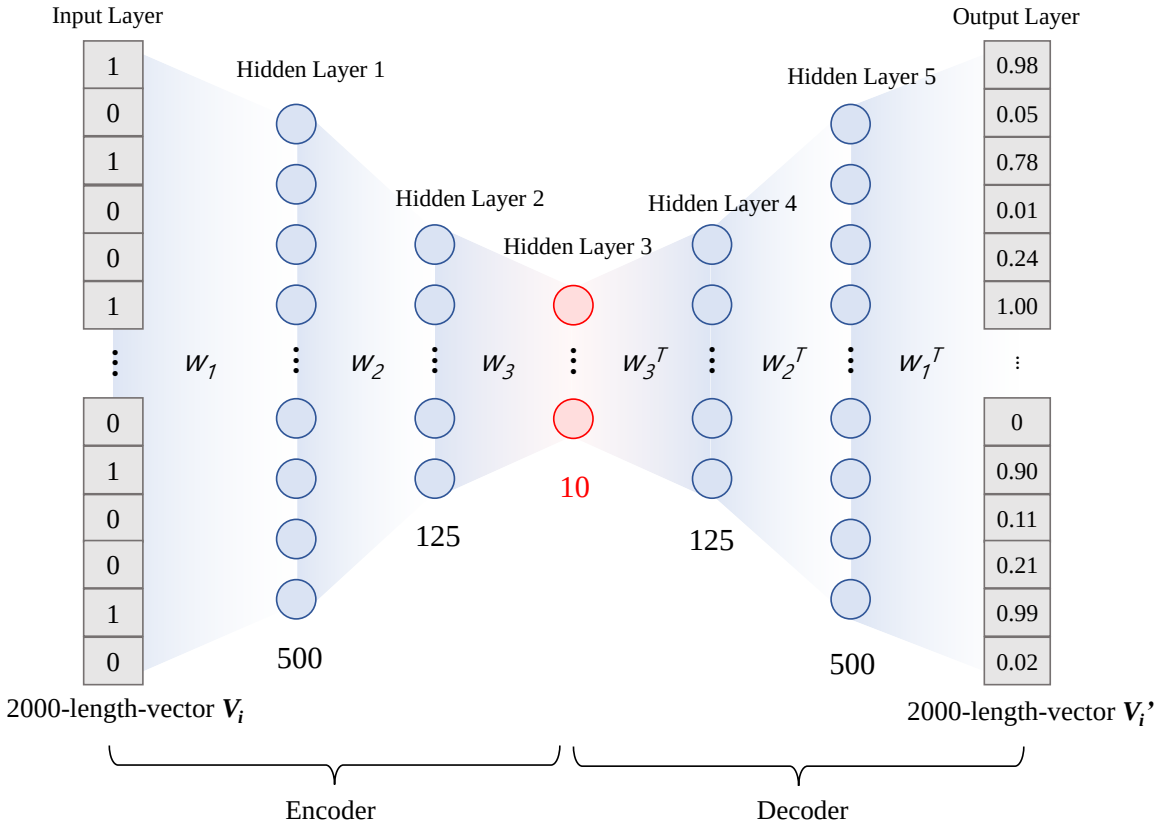

Figure 3. A schematic structure of the autoencoder with 5 fully-connected hidden layers used in the paper. The number of nodes of encoding layers is gradually decreased to 10 to reduce the dimension of information of the input vector (2000-word vector). The activation function of the last encoding layer is a linear function. The activation function of the last decoding (output) layer is a sigmoid function which bounded between 0 to 1 to match the vector representation of the input layer.

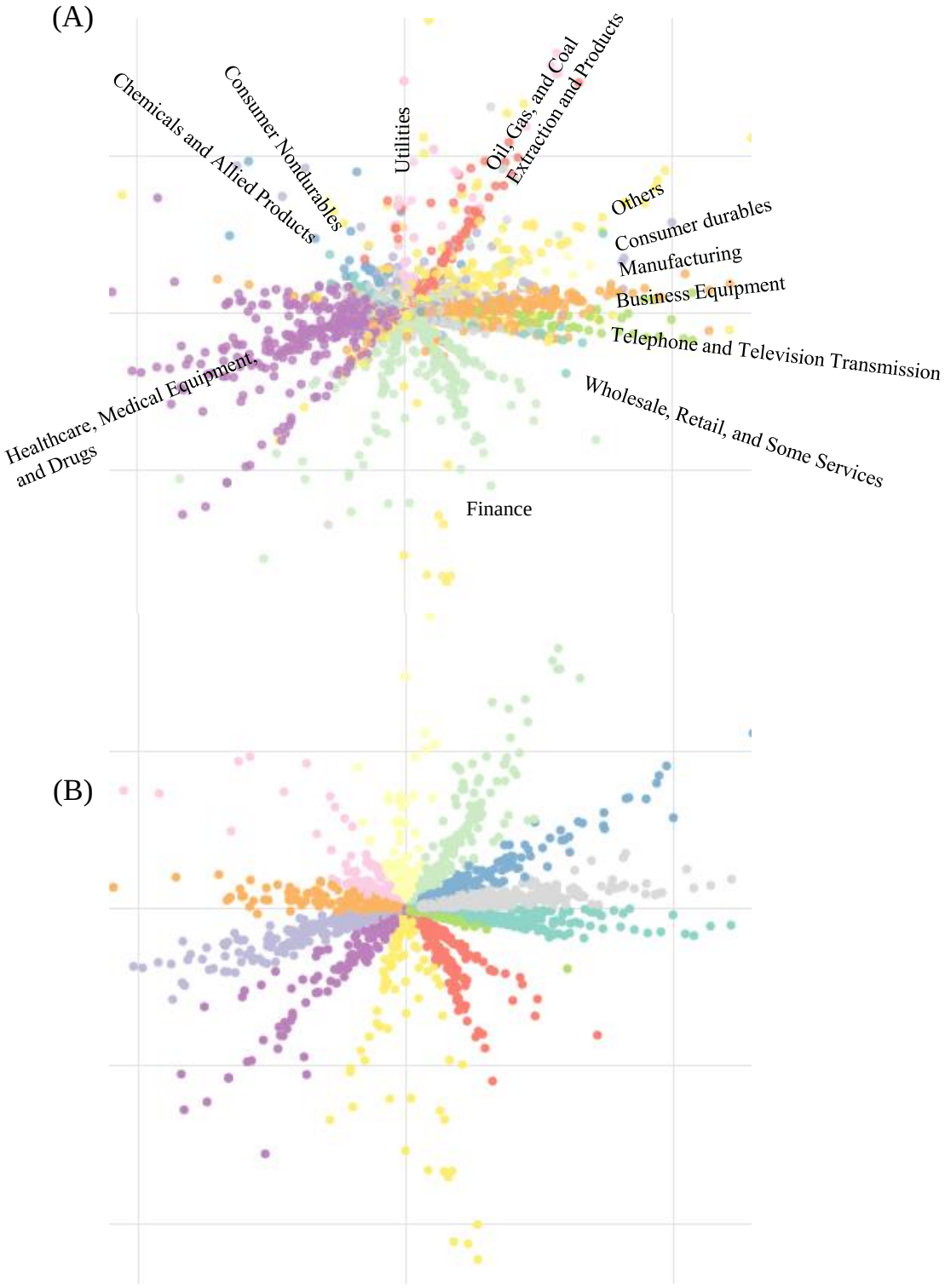

Figure 4. The scatter plot of the two-dimensional codes from the coded layer of autoencoder in terms of the Fama-French's 12 industry classification code (A), and spherical clustering result using the dimensionality-reduced features of the coded layer (B).

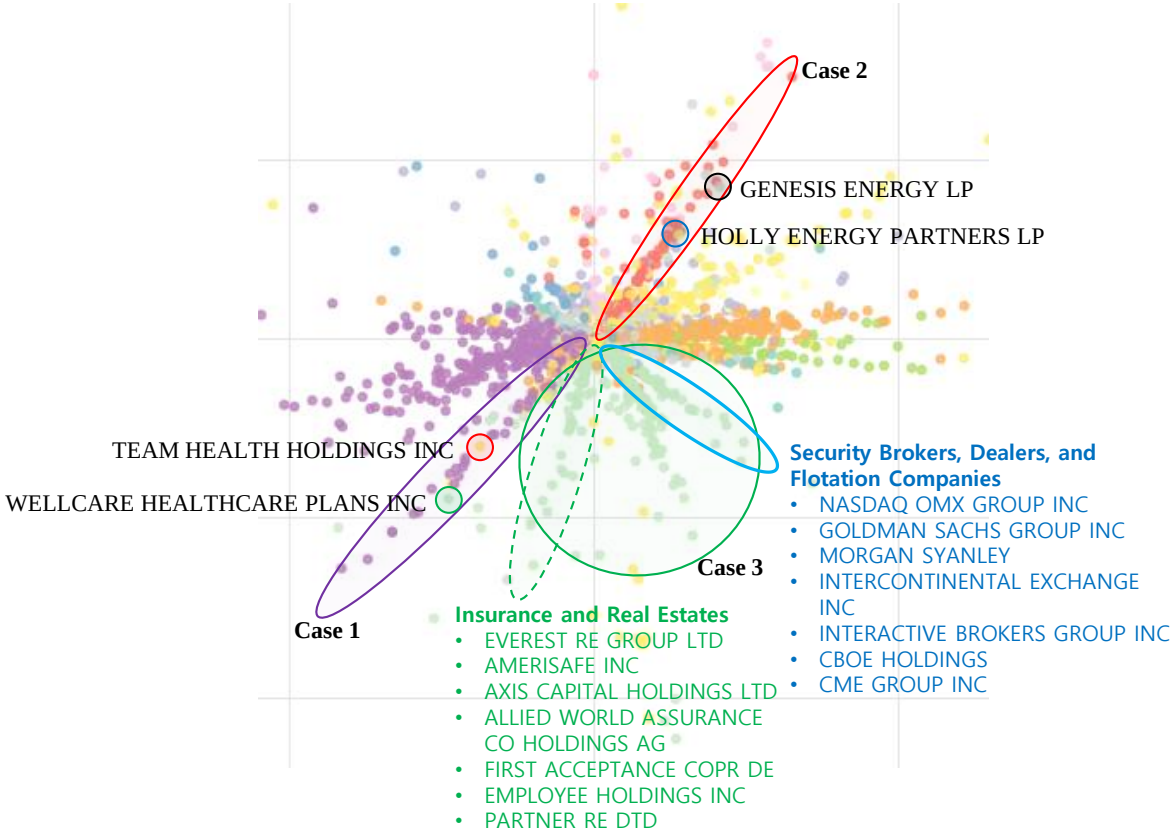

Figure 5. Sample firms having different SIC code ranges but allocated to the same label of the cluster. The firms in Case 1 (purple) circle are related to the healthcare, medical equipment, and drugs. The firms in Case 2 (red) circle are related to the Oil, Gas, and Coal Extraction and Products. Case 3 (green circle) illustrates the financial firms with different sub-industries. The firms in dashed green ellipse are closer to the firms in a purple ellipse which is Case 1 than the other financial sub-industries.

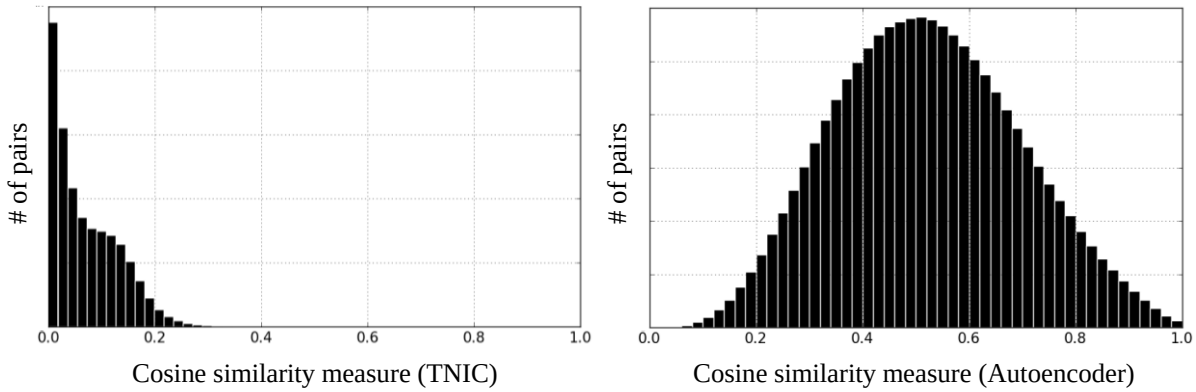

Figure 6. A Histogram of cosine similarity measures by original TNIC database (left) and a histogram of cosine similarity measures on the basis of the dimension reduction output of autoencoder (right).