Text-based industry classification by Autoencoder

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

Hanyang University[1]

Ulsan National Institute of Science and Technology[2]

Abstract

Industry classification has been one of the crucial issues in financial analysis. However, classical industry classification systems have several limitations. Several studies have been progressed to overcome the limitations by using the text information that firms use to describe their business process and products. In this paper, we propose an industry classification methodology based on their business descriptions by reducing high dimensions using autoencoder to avoid a high dimensionality problem in vector space. The main contribution of this paper is first, we overcome the limitation of cosine similarity measure where the word vector is large and highly sparse by reducing the dimension of word vector utilizing the autoencoder. Second, we are able to visualize the relative industry relation of the firms based on the lower dimensional information extracted from the business description text. The relative location 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.

1. Introduction

Industry classifications have been an important issue for practitioners in financial industry and scholars in the field of finance and economics. Researchers in academia have adopted a variety of approaches to group homogenous firms to analyze consequence of corporate reorganization or changes in financial and investment policy. Most of the methods for grouping economically similar firms is based on the industry affiliations.

Traditionally, there are three types of method that classify industries in the financial sector. Standard Industrial Classification (SIC) code, which aggregate firms selling end products and using similar production process, have been used for this purpose (Chan, Lakonishok, & Swaminathan, 2007). Despite the wide application of SIC codes in empirical studies, several types of research have questioned the usage of SIC code as a method of selecting samples. Walker and Murphy (2001) mentioned that the classification system, however, does not sufficiently reflect a shifting the main product, business process, and emerging markets because the system mainly emphasizes on manufacturing operations relative to service processes. The Fama and French (1997) proposed 49 industry grouping system based on the firms' four-digit SIC codes. The system aggregates many similar subgroups into one group. Some firms coded 22, 23,30, and 35, which are related to "transportation equipment" from two-digit SIC major group are grouped as Fama-French classification code "automobiles and trucks" of 37. The Fama-French classification has been essentially influential in the academic area of asset pricing, corporate finance, accounting, and investment. Fama and French, however, did not provide sufficient evidence on validation and verification of performance their classification system even though its broad usage in financial fields. How well their classification system produces groups of economically similar firms still debate as one of the key questions. Another approach to classifying industry is Global Industry Classification System(GICS), which is widely used among investment analysts and portfolio management. The system categorizes firms based not only on their operational characteristics also on investors' perceptions of what constitutes the firm's mainstream of their business (Kile & Phillips, 2009). Because of the characteristics, there are some cases on different industry assignment result between GICS and SIC codes. For example, GATX Corporation, which leases and operates railroad equipment and ships, is classified in the financial sector by GICS but is assigned in the transportation equipment sector as SIC codes. These conflicting assignments arise whenever firms are not only involved in one industry area, products or

business process because the classification systems mentioned above 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) utilized the product description of firms reported to the Securities and Exchange Commission (SEC). They formed new industry classifications based on the strong tendency of product market vocabulary to cluster among firms operating in the same market. The method is based on the products that firms supply to the market rather than production processes. By analyzing the word set from the documents each firm reported, the approach allows us to generate a new set of industries and be able to capture the changes in firms' business, namely diversification and pivoting. They measure the pairwise cosine similarity of word vectors converted from the reported document of firms, which distinguish peer firm as a rival and competitor. The measure can identify how similar a firm is compared to the other firms. 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, which refers that we cannot infer the overall map of the industry relationship. In addition, recent some research results show that the concept of proximity, the distance 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 word vectors of their research have larger than 60000 dimensions, which implies that the similarity measure 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 cosine similarity comparison used in previous research. We reduce the number of features which can sufficiently describe the relationship between firms as an industry rather than directly use the whole dimensions of word vector. The classification is processed by using the reduced features from the coded layer output of the autoencoder. We then compare the classified result of the method proposed in this research with the Fama-French 12 classification scheme.

The main contribution of this paper is first, we overcome the limitation of cosine similarity measure when the vector is large and highly sparse by reducing the vector dimension

utilizing the autoencoder. Second, we are able to visualize the relative location of the firms based on the lower dimensional information extracted from the business description text. The relative location can also describe the industry-level relationship as well as individual firms which provide a variety of products and processes, and thus, originally were involved in conflicting assignment problem as an aspect of the classical classification scheme.

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

We gathered 10-K annual reports filed with the Securities and Exchange Commission (SEC) using web crawling algorithm to extract business description part of each firm from 2013 to 2016 where the total number of 10-K reports is 21631. The 10-K business descriptions are legally required followed by item 101 of Regulation S-K which requires that firms describe the core products they offer to the market. We merged the SIC codes using historical segment data of Compustat from Wharton Research Data Services (WRDS).

The reporting period and financial activity period are able to 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 highest 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. We only keep the information of 2013 in the case that the reporting year is 2015. We use primary SIC code among the segment information of each data. We remove rows if firms have missing primary code at a certain year. The CIK code and reporting year are used to merge the Compustat data and 10-K annual reports. After merging the SIC code with individual firms, the number of report data merged with the industrial classification code is 14560. The documents of the business descriptions are processed to build(construct) a bag of words in the next step.

3. Methodology

3.1. Bag of words representation

The bag of word model is widely used in information retrieval and text mining (Singhal, 2001). It can be represented as a bag of words, where words are assumed to appear independently, and the order is immaterial. Words are counted in the bag, which differs from the mathematical definition of *set*. Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension.

The words from the 10-K business description 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 the 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 word “memory” 67 times to describe their business and products in a single document. It is essential to extract appropriate words which are able to reflect and distinguish industry or market segment that the firm belongs to from the product description texts. In this manner, we only focus on nouns and proper nouns that use no more than 20 percent of all 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 indicated that the minor modifications of the threshold do not affect the similarity 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.

Insert Table 1 in here

In this study, a conventional text mining process is applied to structure a set of unique words from the preprocessed documents. Suppose that there are W unique words used in all documents reported by all firms in training samples. We only use 2000-unique-word in order of frequent appearance among all the W unique words. The selection of 2000-unique-word is

called as a bag of words in this research. The bag of words is utilized to convert the individual documents of business description to the corresponding word vectors. A given firm i's business description can be represented as a binary (coded) vector Vi, which indicates that each element of the vector is allocated by the number 1 if the given bag of words contains the element corresponding the word and otherwise 0.

$$W_{BoWs} = [\, w_1, w_2, \dots, w_{1999}, w_{2000}] \quad (1)$$

$$V_i = [\, v_1, v_2, \dots, v_{1999}, v_{2000}] \text{ where } v_i = 1 \; or \; 0 \quad (2)$$

3.2. Dimensionality reduction by using the autoencoder

Hoberg and Phillips (2016) utilize a whole set of unique words 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, which implies that the remaining 59500 elements of word vector are coded as the number 0. The high dimensional vector space with sparsity 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 some notion of spread of the distribution of all pairwise distances within a data set to converge to 0 as dimensionality increases, which indicate that expectation of pairwise cosine similarity measure becomes constant and its variance shrinks with increasing dimensionality.

Insert Figure 1 in here

Autoencoder is a dimensionality reduction technique based on the 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 of the output node is the same as the number of the input node. The first research was to implement PCA-like dimensionality reduction method using backpropagation method in the shallow neural network. The model consists of two parts which are encoding layers and decoding layers. The purpose of the model is to reduce the number of hidden (encoded) nodes while minimizing the difference between input vector and (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. Based on Hinton and Salakhutdinov, 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 implemented the regime of the autoencoder as a dimensionality reduction method to represent specifications containing information of industry that each firm belongs to.

The dimension of the 2000-word-vector reduced to hidden nodes of 10 which can refer as features for the classification and clustering step. 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 encoding layer (linear function) and last of the decoded layer(sigmoid). We minimize the value of the binary cross-entropy as a loss function used in this learning stage. To measure the value of the loss function, the sigmoid function is selected when computing the loss function at the last decoding (output) layer. This structure sufficiently discriminates and explain information on the 2000-length word vectors even we modified the number of a hidden node of the last encoding layer.

Insert Figure 2 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 is Gaussian. For high-dimensional data such as text documents, cosine similarity has been shown to be a superior measure compared to Euclidean distance (Strehl, Ghosh, & Mooney, 2000). The implication is that the direction of a document vector is more important than the magnitude. The spherical k-means algorithm to maximize 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.

The SIC codes were collected from historical segment data of Compustat(WRDS). The codes are merged with the individual 10-K filings based on the CIK code and corresponding year. The labeled SIC codes are converted to the Fama-French's 12 industry classification code. We then classify the industry of firms by clustering the reduced features of outputs from the coded layer. We use a spherical K-means clustering algorithm which is a suitable clustering algorithm for the vector space model of the corpus. The number of clusters is 12 for the consistency of the Fama-French's classification. We compare the clustering result with the Fama-French classification code of each firm.

4. Result

For a visual representation, we modified the number of hidden nodes at the last encoding layer which used for features in classification. 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. 3 shows the 2D scatter plots of values of the coded layer from after training the autoencoder. The figure shows that firms labeled as same Fama-French industry code are clustered with the spike-like shape, which indicates that firms in the same industry have a similar degree with different vector norm as an aspect of the polar coordinate system. For example, the SIC code range of the cluster colored by yellow-green (Figure. 3. A) is from 6000 to 6999 which refers the financial firms in terms of Fama-French industry classification. This result is consistent with the clustering result of the Hoberg and Phillips' because they used the cosine similarity to measure the degree (distance) between two text reports. Figure 4 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 3 in here

There are some cases that firms are far from their group(spike) on the 2D graph, which indicates that the firms select and use quite different words when they described their business process and products compared to other firms belongs to the same industry according to the SIC and Fama-French industry classification code. These firms might improperly report their business description or SIC code of the firm is wrongly assigned because the firm recently changed their major products and business process through diversification and pivoting. The method proposed in this paper is able to solve the latter problem while determining the industry by their words used in the annual reports.

In detail, a SIC code of a firm "TEAM HEALTH HOLDINGS INC" is 7363, which the Fama-French classification code is 12 (Others) as well as listed in Table 2. Another firm named as "WELLCARE HEALTH PLANS INC" is coded as 6324 with Fama-French classification code of 11 (Money). The two firms are clustered in the same industry as we use the business description text for industry clustering. The other firms clustered with the two firms are in the SIC code range of 8000-8099. The second case is related to the energy and mining including oil, gas, and materials. The first two firms of the "Case 2" below the table belong to the "Shops" in terms of the Fama-French classification and SIC code. The points 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 they provide in the markets. Figure 4 illustrates the situation of the firms mentioned in Table 2. The firms related to healthcare products and process are in the purple ellipse along the same direction (angle). Many firms involved in the industry of oil, gas, and coal extraction and products are grouped with the red marks.

Insert Table 2 in here

In contrast, there is a case that some firms are able to be grouped as the same industry in terms of Fama-French 12 Classification code. Three or four spike-groups are seen within the industry colored as a green circle as shown in Figure 4 (Case 3). Each spike indicates different

clusters in terms of descriptions of their business process and product even though the firms are all related to money or financial industry. We are able to clarify the location of the firms in the plot and divide this industry group into subgroups (sub-industries). In addition to subgroup division, one can recognize the firms in subgroups of the financial industry, which is located at the dashed green circle (ellipse), are closer to the industry group in the purple circle (Case 1 group) than other subgroups within the financial industry.

Insert Figure 4 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 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 a suitable for the vector space model of the corpus. The number of clusters is 12 for the consistency of 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 polar coordinate system. The classification result also shows that 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.* Paper presented at the International conference on database theory.

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.* Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

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.* Paper presented at the Workshop on artificial intelligence for web search (AAAI 2000).

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

Table 1. Most used words in the business description of sample firms with different industries

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. Sample firms having different SIC code ranges but allocated to the same label of the cluster.

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. Word lists contained in the business descriptions text of the sample firms

Unique words out of 2000 words in the bag-of-words	Redundancy
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

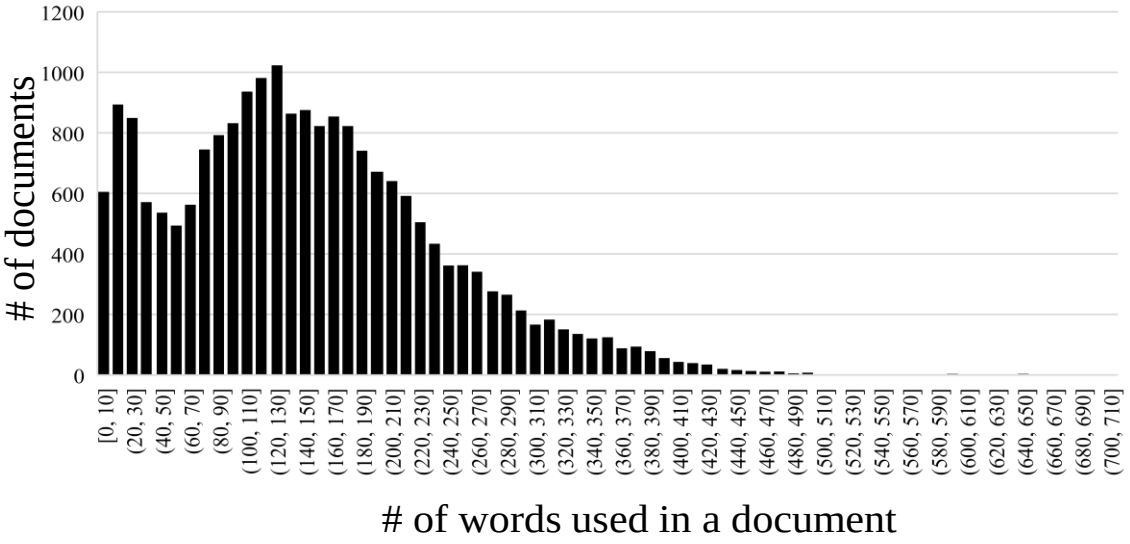

Figure 1. A histogram of the number of words used in business descriptions of each firm

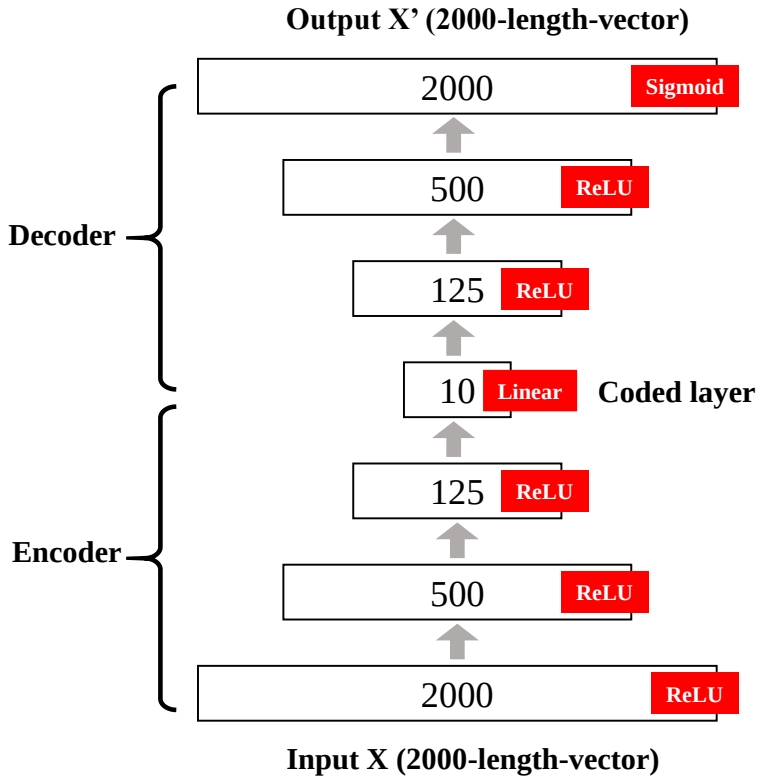

Figure 2. A schematic structure of the autoencoder with 5 fully-connected hidden layers used in the paper.

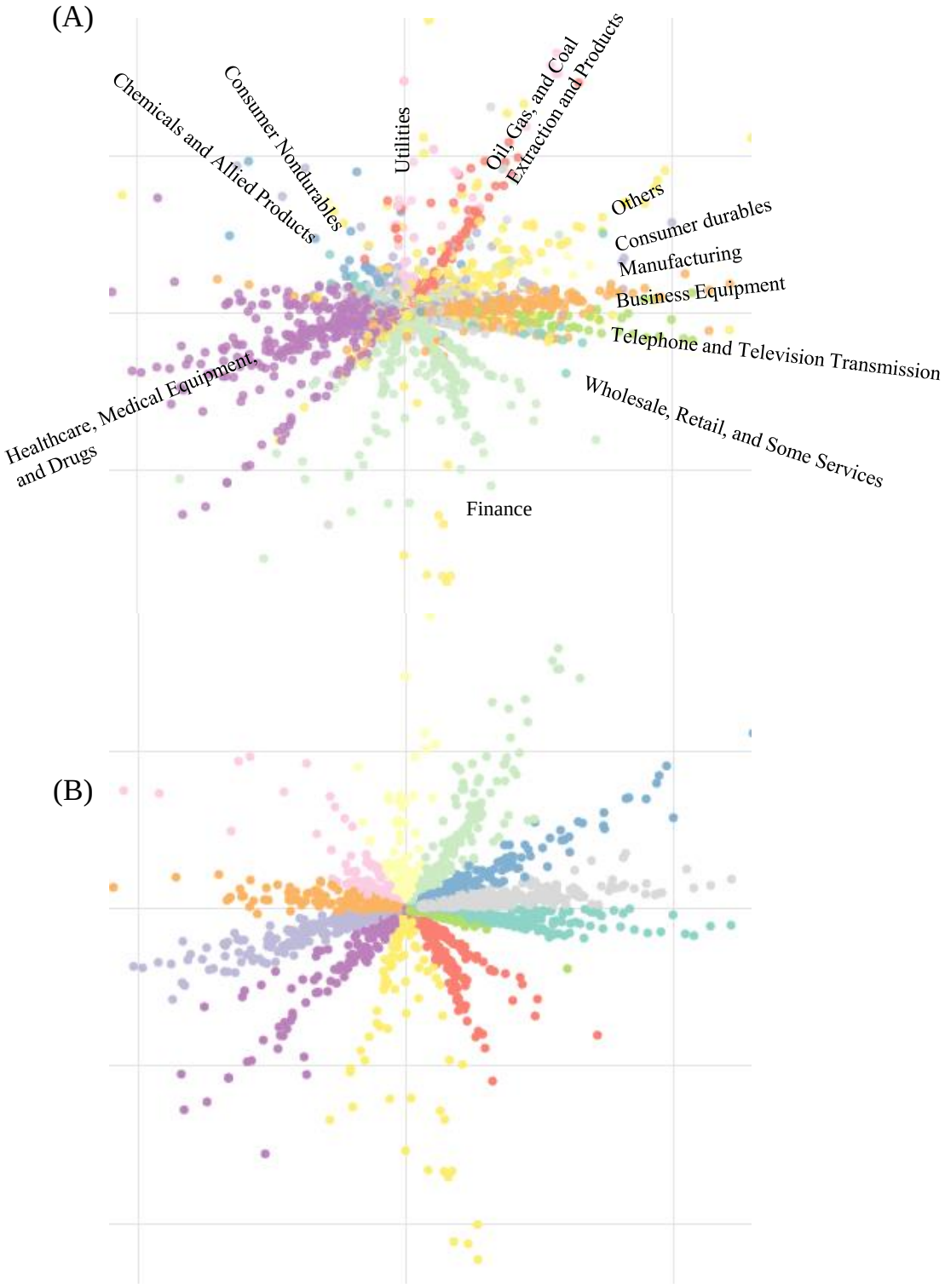

Figure. 3. The scatter plot of the two-dimensional codes from the coded layer of autoencoder in terms of the Fama-French's 12 industry classification code (A), and spherical clustering result using the dimensionality-reduced features of the coded layer (B).

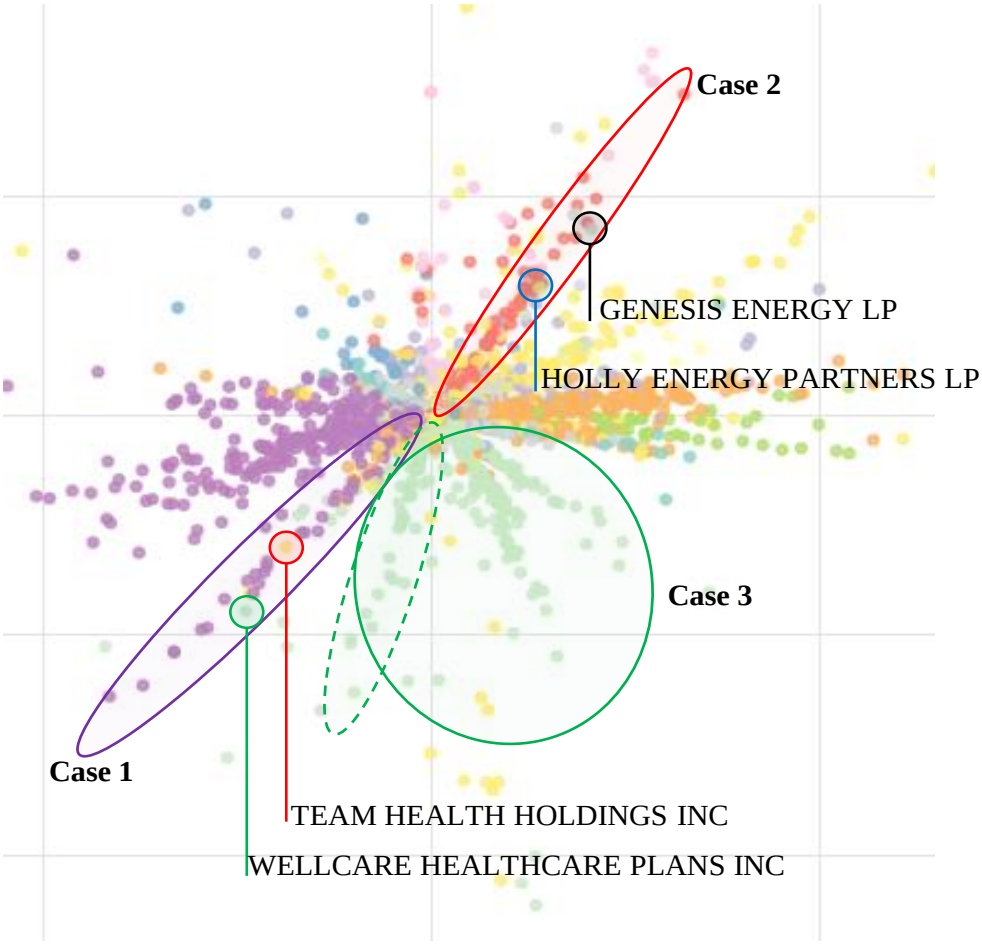

Figure 4. Sample firms having different SIC code ranges but allocated to the same label of the cluster.