Automatic Classification of Securities using Hierarchical Clustering of the 10-Ks

Hoseong Yang, Hye Jin Lee, and Sungzoon Cho
Department of Industrial Engineering
Seoul National University
Seoul, Korea
{hoseong, hyejinlee}@dm.snu.ac.kr, zoon@snu.ac.kr

Eugene Cho

4327 Ravensworth Rd
Annandale, VA 22003, USA
eugene.t.cho@gmail.com

***Abstract*—Industry classification has been rigorously utilized in academic research and business analytics. The existing classification schemes, however, have been constructed and maintained manually by domain experts, which require exhaustive time and human effort while vulnerable to subjectivity. Hence, the existing classification systems do not properly reflect the fast-changing trends of the firms and the capital market. As a remedy to such shortcomings, this paper proposes a new classification scheme, Business Text Industry Classification (BTIC), namely, that automatically clusters securities based on the textual information from the corporate disclosures. BTIC exploits the business section of the Form 10-Ks, in which firms provide their self-identities in a rich context. We employ doc2vec for document embedding and apply Ward's hierarchical clustering method to categorize securities into BTIC groups. Evaluation results using 12 financial ratios commonly found in financial research show that BTIC performs just as good as SIC and GICS in terms of inter- and intra-industry homogeneity, especially for the higher level of clustering. Given that, we claim that BTIC outperforms SIC and GICS in four aspects: process automation, objectivity, clustering flexibility, and result interpretability.**

***Keywords*-10-K, Industry classification, Doc2vec, Hierarchical clustering, Capital market research, SIC, GICS**

I. Introduction

The design, construction, and integration of industry classification schemes have been playing a foundational role in the field of capital market research.[1] An industry classification scheme clusters financial entities, such as companies, securities, or establishments, into bundles. Each bundle contains entities that are considered "similar" types of business given their market activities. At the same time, one bundle is (or, should be) notably different from another. These bundles, then, represent segments of the market of interest with distinct financial characteristics. For that reason, a successful industry classification may facilitate a broad range of cluster-level analysis, of which examples include: sector-wise identification of competitors, benchmarking company activities and performances, measuring economic indicators, and setting up market share [1]. A myriad of organizations and researchers has responded to the significance of the effective industry classification by developing various classification schemes. The most classic example is the Standard Industrial Classification (SIC, henceforth), established by the government of the United States in 1937. Global Industry Classification System (GICS, henceforth), among many others, are frequently utilized in the recent economic and finance literature as well.

Despite the long history and diversity, the existing classification schemes suffer from limitations in various aspects. The foremost concern involves assuring cluster quality. A good clustering scheme should ensure that clustered entities within a cluster are adequately homogenous (within-class homogeneity), while entities in different clusters are distinctly different (between-class heterogeneity). Existing classification schemes, older ones especially, such as SIC, group vastly different companies together. One conspicuous example of such a problem is the placement of Netflix, Inc. in SIC system. Currently, SIC clusters Netflix into a "Services" sector, along with Wynn Resorts. The same is true for GICS as well, where Netflix is grouped together with Wynn Resorts as the "Consumer Discretionary" sector. However, it is not so difficult to tell that these two are entirely different companies, where the former is an online streaming network, while the latter is a casino resort.

The next issue with the existing classification schemes is concerned with timely update of the classification schemes. Due to the recent surge of innovations and technological advances, the means of production have come to vary over time, with product and services provided by the firm growing evermore complex. Hence, appropriate adjustments to the classification scheme on a regular basis, reflecting the changes in the firm structure, business operation, and/or product and service provision, is essential. This issue is very challenging to tackle, since all existing classification system require human input at some point. With human input as a requirement, information updates on a regular basis will certainly be very costly and time-consuming. A good example to illustrate such a problem is the case of Amazon.com. Amazon.com Inc. has begun its business as an online bookstore in 1995, but by the beginning of 2016, it has evolved into a multi-branch retailer of all

[1]Although the term "classification" is widely used in the financial sector, it actually refers to the *clustering* of financial entities into groups. Throughout this study, we follow the naming convention and use the term "industry classification" as is.

kinds of products, as well as a cloud computing company. SIC system, however, has classified Amazon.com to be in the industry group of "Retail Stores", which hasn't been adjusted at all since 1998.

In this paper, we address issues listed above and propose a new industry classification system, Business Text Industry Clustering (BTIC, henceforth), namely, to replace the existing ones. BTIC applies text mining techniques on corporate disclosures, in order to segment securities into groups with similar "business identities". More specifically, we exploit the business section of the form 10-Ks of S&P 500 companies and cluster publicly traded equity securities, better known as "common stocks," whose content of the text appear similar. The form 10-K is a comprehensive summary of a company's business operations and market performance filed annually, as required by the U.S. Securities and Exchange Commission. Its business section, in particular, provides a finely detailed description of the company's business operations, organizational structure, risk factors, and market competitors. In other words, information presented in the business section of the 10-Ks embodies the company's self-identity, elaborately described in many different angles. For example, the form 10-K of Netflix, Inc. in 2016 states:

> Netflix, Inc. ...[omitted]...is the world's leading Internet television network with over 75 million streaming members in over 190 countries...[omitted]...we have developed an ecosystem for Internet-connected screens and have added increasing amounts of content that enable consumers to enjoy TV shows and movies directly on their Internet-connected screens. [2]

In constrast, Wynn Resorts, in their corresponding 2016 10-K, states:

> Wynn Resorts, Limited...[omitted]...is a leading developer, owner and operator of destination casino resorts (integrated resorts) that integrate hotel accommodations and a wide range of amenities, including fine dining outlets, premium retail offerings, distinctive entertainment theaters and large meeting complexes [3]

These two firms provide a very clear picture of their business operations and market activities in the excerpts, and their differences can be inferred easily. These self-identities, then, may serve as a new standard for placing firms into different groups.

The issue with timely updates may also be resolved by looking at the business section of the Form 10-K. As a supporting evidence, we present the excerpts from Amazon's Form 10-K, 1998 and 2016, which clearly shows the changes in firm structure:

Year	A part of the business section of Amazon's Form 10-K
1998	...the leading online retailer of books, [since the] opening for business as 'Earth's Biggest Bookstore' in July 1995 [4].
2016	...serve consumers through [their] retail websites and... [and] design our websites to enable millions of unique products to be sold by us and by third parties across dozens of product categories...[Amazon.com, Inc.] also manufacture[s] and sell[s] electronic devices [5].

Hence, by incorporating the textual information of corporate disclosures, the issues with Netflix and Amazon mentioned above can be easily resolved. We use the business section of the 10-K as our main data source in order to construct BTIC. BTIC scheme employs Doc2vec for document embedding, and then hierarchical clustering to group similar documents in a hierarchical manner. The result is a market segmented by a hierarchy of clusters, that group securities whose self-perceived firm identities are similar. The performance of BTIC is evaluated in comparison to SIC and GICS, the two mostly widely used industry classification systems in finance and economic literature, following the experiment design of Hrazdil et al. (2014) [6] as a benchmark. Experiment results show that BTIC's performance meets those of SIC and GICS in terms of inter- and intra-industry homogeneity. Given that it performs just as well in clustering the securities into homogeneous groups, we show that BTIC outperforms the existing classification schemes in four aspects: process-automation, objectivity, flexibility, and result interpretability.

The rest of the paper is organized as follows. In section 2, we discuss related work on the subject. Section 3 introduces and elaborates on the framework of BTIC. Section 4 presents evaluation experiment results and discusses its business application. Finally, Section 5 concludes this paper.

II. Related Works

The use of industry classification has been prevalent in both academic and business research. Its primary use has been in controlling for industry effects in various econometric analyses, a few of which the examples include: evaluations of diversification strategy [7], vertical mergers [8], or government industry development grants [9]. The purpose of industry classification is to "create discrete categories, set of classes, by maximizing the differences between industry groups and similarities of components within industry groups" [1].

In order to achieve this goal, the following industry classification schemes have mainly been utilized and discussed in the financial sector and academia. Established by the U.S. government in 1937, SIC is currently used by SEC and U.S. government agencies [10]. Full extent of SIC classification, down to the deepest division level, is publicly available at U.S. Department of Labor website.[2] However, due to its outdated standard, SIC has been replaced by NAICS in 1997. NAICS was developed on a collaborative effort among the Instituto Nacional de Estadistica y Geografia (INEGI) in Mexico, Statistics Canada and the United States Office of Management and Budget (OMB) [11]. In NAICS,

[2] https://www.osha.gov/pls/imis/sic_manual.html

companies are classified into industries defined by the means of production. GICS, developed by Morgan Stanley Capital International and Standard & Poor's, is used by the professional investment management community with a goal of portfolio diversification and asset allocation decisions [12]. GICS consisted of 4 hierarchical division levels, beginning with 10 sectors at the uppermost level. The 11-th sector, Real Estate, has just recently been added to the classification system, and it has been in use since September 1, 2016. The division then dives further into 24 industry groups, 68 industries, and finally to 157 sub-industries. The newest and full version of GICS classification data can be purchased from MCSI Inc.[3] Although it has been praised as the best industry classification scheme, GICS must still rely on yearly updates done manually, which leaves the work time-consuming and vulnerable to subjectivity.

A few studies test the effectiveness of such classification schemes. These studies rely on stock price returns and financial ratios to measure the cohesiveness of securities within industry groups. Guenther and Rosman (1994) compared correlations of intra-industry monthly stock returns and variances of intra-industry financial ratios using different SIC codes [13]. This has become a standard procedure in evalutation of industry classification. Krishnan and Press (2003) use this method to assess NAICS's effectiveness in forming industry groups, such as manufacturing, transportation, and service industries [14]. Bhojraj et al. (2003) and Hrazdil et al. (2014) also compare stock return co-movement and various key financial ratios between the SIC, NAICS, GICS system [15], [6].

III. Proposed Methods

A. Document representation

In order to calculate similarities among the business part of the Form 10-Ks of given securities, each document needs to be represented as a numeric vector. Among many, a typical method of vector representation is TF-IDF [16], one of the most common Bag-of-Words [17] techniques. TF-IDF is very simple and relatively intuitive; nonetheless, its usability quickly gets limited as number of vocabulary increases, since the vector dimension and sparcity grows the same. Efforts have been made to address this problem by a number of studies, which develop methods to represent documents as dense vectors [18], [19]. Doc2vec is one of them, developed by [20], whose framework closely resembles that of word2vec [21] yet learns words and documents simultaneously. Not only does Doc2vec significantly reduce the dimension of the learned vectors, it learns to represent words and documents on the same continuous space, hence enabling similarity calculations between them. At the same time, since Doc2vec allows unsupervised learning, it does not require syntactic information like parse trees [22]. Doc2vec has shown great performance in sentiment classification problems, and it has been studied to work well for information retrieval [20].

[3] https://www.msci.com/gics

B. Distributed representation of securities

We apply Doc2vec method on the business section of the 10-K reports and represent them as vectors. Graphical illustration of the method is presented in Figure 1. It utilizes a very simple neural network which learns words and security vectors by maximizing the probability of predicting the next word, given the context words and subject security information. S, W denote a security and word matrix, respectively, while x denotes one-hot representation for lookup. $\vec{s}_i$, $\vec{w}_i$ represents a security i and word vector, respectively.

The objective function of the model is maximize the average log probability as follows:

$$\frac{1}{N}\sum_{t=c}^{N-c}\log p(\vec{w}_t|\vec{w}_{t-c},...,\vec{w}_{t+c}) \tag{1}$$

N is the total number of words, c determines the size of context. Probability of predicting target word given context is calculated by softmax as follows:

$$p(\vec{w}_t|\vec{w}_{t-c},...,\vec{w}_{t+c}) = \frac{\exp^{y_{\vec{w}_t}}}{\sum \exp^{y_i}} \tag{2}$$

Hidden layer h is calculated by average security and word vectors.

$$h = \frac{1}{2c+1}(\vec{w}_{t-c}+...+\vec{w}_{t+c}+\vec{s}_i) \tag{3}$$

Un-normalized log-probability for each target word y is calculated as follows:

$$y = b + U \cdot h \tag{4}$$

where U, b are the softmax parameters. Finally, the weight matrix W and S are trained by stochastic gradient descent.

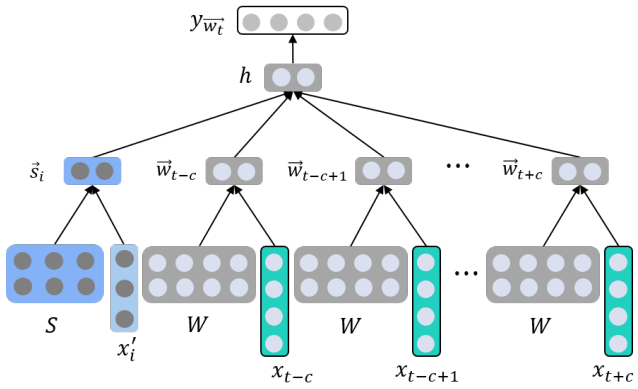

Figure 1: Visualization of the proposed model

C. Ward's Hierarchical Clustering

Industry classification schemes group entities in a hierarchical manner. BTIC employs a hierarchical clustering method to resemble the organizational structure of the existing classification schemes. There are two ways to cluster entities: an agglomerative method and a divisive method. Agglomerative method assumes that each entity is a cluster and group clusters together from bottom-up. In contrast, the divisive method is the top-down approach, which begins with all entities in one cluster and then splits them into smaller clusters recursively. However, divisive clustering methods are constrained primarily by the size of data. With n data points, there are $2^n - 1$ possible ways cluster them; consequently, time complexity increases exponentially with n.

Therefore, this study employs agglomerative hierarchical clustering. We rely on Ward's variance minimization algorithm [23], which considers information loss while clustering. At each step, Ward's method finds a pair of clusters that leads to the minimum increase in total within-cluster variance after merging the clusters. Ward's method uses Euclidean distance to cluster entities, and past studies have shown that Euclidean distances between the distributed vectors are correlated with the semantic similarity [24], [25].

IV. Experiment

We collected Form 10-Ks of 504 S&P500 companies[4], published on January 1, 2016 or later, from the U.S. Securities and Exchange Commission (SEC) website, which is publicly available. We have chosen to look at S&P500 companies specifically, since they take up the 500 largest market capitalizations. Among the 504 collected, 4 firms, Broadcam, Coca-Cola Enterprises, Cablevision Systems Corp. and ACE Limited, are discarded since their 10-Ks have not been reported. Then, we extract the business section of the 10-Ks by exploiting regular expressions. The market ratios of the corresponding firms are collected from the Center for Research in Security Prices (CRSP) database and Compustat. Information collected ranges from January 1, 2015 through December 31, 2015.

From the text corpus, we have removed stopwords[5] as well as words that appeared less than five times in the text corpus. We also discarded proper nouns, numbers and special characters by utilizing POS tagging using NLTK[6]. Upon the completion of preprocessing, we have extracted 13,012 unique tokens as a result. On average, a single business section of the Form 10-K consists of 6208.2 tokens.

We train the vectors of securities and words using `gensim` package[7] implemented by python. Input hyperparameters include: dimension of security and word vectors, the window size, and the number of training epoch, which were set at 50, 2, and 10, respectively. We present the summary statistics of BTIC, along with SIC, GICS, and TF-IDF clustering results as a comparison group, in Table I. Here, level 1 represents the broadest categories in the industry identification system, while level 2 and 3 refers to categorizations with the progressively narrow scopes.

[4]as announced on January 1, 2016

[5]http://www.lextek.com/manuals/onix/stopwords1.html

[6]http:://www.nltk.org/

[7]https://radimrehurek.com/gensim

Table I: Summary Statistics

	Level	# official categories	Mean # of firms per industry
SIC	Level 1 (broadest)	10	464
	Level 2	71	15
	Level 3	264	7
GICS	Level 1 (broadest)	10	228
	Level 2	24	103
	Level 3 (narrowest)	68	52
TF-IDF	Level 1 (broadest)	10	281
	Level 2	24	117
	Level 3 (narrowest)	68	42
BTIC	Level 1 (broadest)	10	267
	Level 2	24	118
	Level 3 (narrowest)	68	50

Notes – Mean # of firms per industry records the average number of securities per category in the corresponding division level of the subject classification scheme in our experiment data.

A. Hierarchical clustering Result

In this paper, we use Ward's variance minimization algorithm [23]. The result of hierarchical clustering on the 500 securities of our experiment is demonstrated by a dendrogram in Figure 2.

As illustrated in Figure 2, the depth of division can easily be adjusted by changing the number of thresholds. We set the depth of division to be 3 levels, so that levels 1, 2 and 3 consist of 10, 24, and 68 clusters, respectively. This is exactly the same organizational structure as GICS, and we choose this setting for the sake of consistency and better interpretability of the comparative analysis across different industry classification schemes.

Table II lists top 5 market cap firms in each cluster of Level 1 of BTIC.

It is not difficult to name Cluster 1 to be representing electric and multiutility companies; Cluster 2, energy (gas and oil); Cluster 3, real estate, and; Cluster 4, finance. Clusters 5, 6, 7, 8, 9 and 10 put together firms into fairly homogeneous groups as well, which may be compared to healthcare & pharmaceuticals, foods & leisures, retail trade, industrials, media & communications and information technology respectively.

A very interesting case is found from the result of our experiment, which is worth mentioning–the placement of Netflix. Traditionally, SIC identifies Netflix to belong in "Services", with Wynn Resorts as the same business entity group. BTIC, however, places Netflix into Cluster 10, together with IT companies such as Amazon.com and Apple

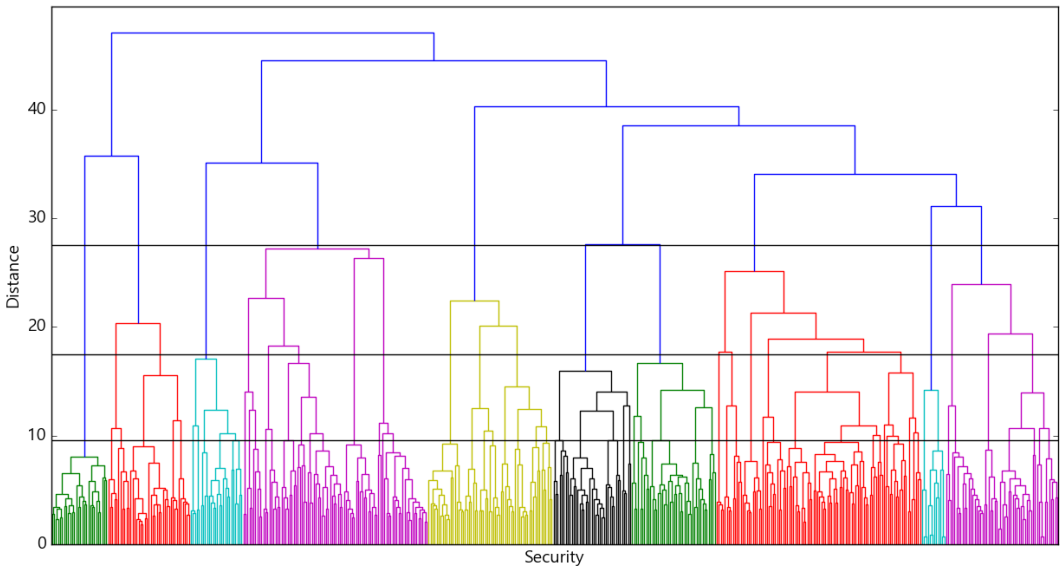

Figure 2: Visualization of BTIC result by dendrogram

Table II: Clustering result by BTIC

Cluster	Top 5 Securities
C1	Public Serv. Enterprise Inc., NiSource Inc., Southern Co. Duke Energy, SCANA Corp.
C2	Chevron Corp., Marathon Petroleum, Marathon Oil Corp. ConocoPhillips, Pioneer Natural Resources
C3	Simon Property Group Inc., Prologis, Plum Creek Timber Public Storage, American Tower Corp.
C4	Morgan Stanley, Wells Fargo, JPMorgan Chase & Co. Principal Financial Group, Bank of America Corp.
C5	Johnson & Johnson, Gilead Sciences, Allergan, Plc. Amgen Inc., Bristol-Myers Squibb
C6	PepsiCo Inc., The Coca Cola Co., Monster Beverage McDonald's Corp., Kraft Heinz Co.
C7	Wal-Mart Stores, Home Depot, Costco Co. Lowe's Cos., O'Reilly Automotive
C8	Berkshire Hathaway, Exxon Mobil Corp., General Electric Boeing Co., Raytheon Co.
C9	The Walt Disney Co., Time Warner Inc., Scripps Networks 21st Century Fox, 21st Century Fox Inc Class B
C10	Apple Inc., Amazon.com Inc., Alphabet Inc. Class C Facebook Inc., Alaphabet Inc. Class A (Netflix)

For C10, we add Netflix in a parenthesis; Netflix does not make the top 5 market cap firms in the group, but it serves as an important instance which we talk about in Section IV.

inc. Netflix is ultimately a television network streamed over internet, and BTIC captures this leading feature of the firm's characteristics very well when assigning an industry group. This is a good case that illustrates the strength of BTIC; because its classification scheme is based on the textual data provided by the firm itself, it embodies a lot more information about the firm's primary business objective, market activities, and current interests as compare to merely looking at the production and profit structures.

Another good example of BTIC detecting the changing trends of a firm's distinct characteristics is the classification of Amazon.com Inc. Amazon.com Inc. has been classified as "Retail Trade" in SIC system, and as "Consumer Discretionary" in GICS scheme, along with CostCo and WalMart. On the contrary, BTIC places Amazon.com in the same group with IT companies such as Google (Alphabet Inc.) and Facebook Inc., as illustrated in Table II. Furthermore, our experiment results report that BTIC has grouped Amazon.com together with online-transaction based firms such as eBay and PayPal on the 3rd level (the narrowest scope). While it is difficult to draw a clear line between a retail and an IT sector when it comes down to clustering Amazon.com into a specific industry, our results show that Amazon.com relies heavily on technology unlike other retail stores such as CostCo or Wal-Mart. This case certainly demonstrates that BTIC is capable of detecting the fast-changing trends of the market and reflecting them in the clustering results concurrently.

Additionally, we would like to highlight that BTIC has automatically learned to form a "real estate" group apart from other "finance and banking" entities. This is noteworthy because neither of the two major industry schemes, SIC and GICS, have not distinguished the two groups differently until very recently. GICS just announced the inclusion of a "real estate" sector to its broadest division level, a new addition to the classification structure in 17 years since its birth in 1999. BTIC, however, has automatically detected the fundamental differences and clustered real estate firms separately from other finance and banking firms using the business section of the Form 10-Ks, which adds a lot more to the strength of BTIC scheme over the existing industry classification systems.

Finally, since securities and words are trained in the same continuous space, BTIC offers interpretability by showing a list of words which are similar to each cluster. BTIC

Table III: Word vectors similar to cluster vector

Cluster 1	
word	**cos**
electric	0.7532
generating	0.7234
power	0.7093
electricity	0.6745
load	0.6646
coal-fired	0.6327
utility	0.6001
generators	0.5943
generation	0.5935
plants	0.5933

Cluster 2	
word	**cos**
oil	0.6990
onshore	0.6798
gas	0.6486
drilling	0.6398
natural	0.6269
liquids	0.6167
shale	0.6108
producing	0.6106
crude	0.5869
offshore	0.5855

Cluster 3	
word	**cos**
estate	0.6604
apartment	0.6283
buildings	0.6145
real	0.6066
properties	0.5994
mall	0.5928
space	0.5910
lease	0.5830
sublease	0.5770
rent	0.5713

Cluster 4	
word	**cos**
banking	0.6456
banks	0.6347
institutions	0.5556
deposit	0.5384
clearing	0.5235
compile	0.5213
institution	0.5086
regulators	0.5027
supervision	0.4911
bank	0.4897

Cluster 5	
word	**cos**
drugs	0.6286
medical	0.6275
drug	0.5934
indications	0.5698
pharmaceutical	0.5634
dentists	0.5563
pharmaceuticals	0.5527
patient	0.5407
physicians	0.5343
patients	0.5178

Cluster 6	
word	**cos**
flavored	0.6793
beverages	0.6571
food	0.6170
tea	0.5899
confectionery	0.5847
drinks	0.5799
drink	0.5722
beverage	0.5649
carbonated	0.5579
flavor	0.5544

Cluster 7	
word	**cos**
merchandise	0.7564
stores	0.6445
apparel	0.6279
retailer	0.6097
fashion	0.6039
showcase	0.5929
store	0.5875
assortment	0.5726
e-commerce	0.5701
accessories	0.5700

Cluster 8	
word	**cos**
aircraft	0.5281
light	0.5223
automotive	0.5151
lead-times	0.4985
aerospace	0.4963
coaxial	0.4954
fighter	0.4889
ventilation	0.4696
airfoil	0.4683
marine	0.4598

Cluster 9	
word	**cos**
television	0.8263
programming	0.7849
broadcast	0.7585
free-to-air	0.6818
movies	0.6270
radio	0.6181
audience	0.6061
sports	0.6038
viewers	0.5839
magazine	0.5780

Cluster 10	
word	**cos**
software	0.7104
computing	0.5872
hardware	0.5757
functionality	0.5649
e-services	0.5642
platforms	0.5606
cloud-based	0.5531
mobile	0.5487
desktop	0.5424
ios	0.5415

can define cluster vectors as the average of the securities vectors included in the corresponding cluster. As the distance metric, we use cosine similarity. This is one of the properties that distinguishes BTIC from other existing classification schemes. Table III reports an example of similar words for level 1 clusters. Words that are close to Cluster 10 include software, computing, hardware, and platforms, which clearly show that securities in Cluster 10 are related to information technology. For Cluster 1, similar words include electric, generating and power, which demonstrate that the clustered firms are indeed energy-related. Clusters 3 and 4 are particularly interesting to compare. The word list for Cluster 3 includes estate, apartment, buildings and properties, which leads to real estate firms. In the meantime, words like banks, institutions, deposit and regulators appear near entities in Cluster 4, which unmistakably illustrates that Cluster 4 has collected finance and bank firms.

Before concluding this section, we would like to briefly touch upon the placement of Berkshire Hathaway in Cluster 8 with other securities identified as "Industrials." Bershire Hathaway is a holding company of 62 subsidiary enterprises ranging from insurance to natural gas and fuel companies [26]. By looking at the business section of the Form 10-K of Berkshire Hathaway Inc., one can easily tell that it uses a selection of *very* general, domain-free words to describe itself[8]. Hence, it ends up being grouped together with one of the largest and most heterogeneous cluster of our ICS, the "Industrials." Exxon has been placed in Cluster 8 for similar reasons.

[8]Most frequent words in Berkshire Hathaway's Form 10-K in 2016, and its frequency, are as follows: [(products, 113), (insurance, 96), (business, 92), (businesses, 69), (markets, 64)]

B. Classification Evaluation

We follow the experiment design of Bhojraj et al. (2003) and Hrazdil et al. (2014) as the benchmark for quantitative evaluation of BTIC [15], [6]. The proposed evaluation metric is as follows: The degree of each industry's homogeneity is examined across 12 variables commonly used in capital market research. More specifically, as a measure of homogeneity for each industry cluster we look the R^2 value of the univariate regression as follows:

$$var_{i,t} = \alpha + \beta \cdot var_{k,t} + \varepsilon_{i,t} \tag{5}$$

where $var_{i,t}$ is the market ratio variables tested for security i within a particular industry group k at a monthly timestamp t, and $var_{k,t}$, the average of the market ratios of securities in industry group k at time t. Then, the R^2 would represent the portion of variations in the subject market ratio, explained by the variations in the market ratio of the corresponding industry group on average, for the given period of time. The underlying assumption here is that, if firms share similar identities in the market, then their operating characteristics will show resemblence in the long-run. Following [6], we adopt the full set of market ratio variables, of 4 different categories, to run the tests of intra- and inter-industry homogeneity. The full list of variables used for homogeneity tests is presented listed in Table IV.

We first calculate R^2 for each market ratio, and then take the average across all R^2, equally weighted. The results from the test of intra-industry homogeneity are reported below in Table V. The raw R^2 values are reported in column (A), and the incremental change in homogeneity from the higher level, in column (I). The results show that the inter-industry homogeneity of BTIC is higher than that of SIC in level 1.

Table IV: Financial Variables used in evaluation

G	Variable	Calculation
1	Calendar quarter-end returns (RET)	Returns from quarter-end to quarter-end
2	Quarter-end price-to-book (P/B)	Market capitalization / total common equity
	Enterprise value-to-sales (EVS)	(Market capitalization + debt)/ net sales
	Price-to-earnings (P/E)	Market capitalization / income before extraordinary items
3	Return-on-assets (ROA)	Income before extraordinary items / total assets
	Return-on-equity (ROE)	Income before extraordinary items / total common equity
	Profit margin (PM)	Income before extraordinary items / total common equity
	Leverage (LEV)	Total liabilities / total common equity
	Asset turnover (AT)	Total assets / net sales
	Current ratio (CR)	Total current assets / total current liabilities
4	One-quarter-ahead sales growth (SGR)	(1 quarter ahead - current net sales) / current year net sales
	R&D	Research and development expense / net sales

Notes – Group 1: Ecomonic relatedness, Group 2: Accounting measures, Group 3: Firm-level ratios, Group 4: Financial Information

Table V: Intra-Industry tests of homogeneity

	SIC		GICS		TF-IDF		BTIC	
	(A)	(I)	(A)	(I)	(A)	(I)	(A)	(I)
Level 1	0.11		0.13		0.10		0.13	
Level 2	0.49	0.38	0.20	0.07	0.16	0.06	0.19	0.06
Level 3	0.72	0.23	0.37	0.17	0.25	0.09	0.32	0.13

The high R^2 values for deeper levels of SIC is expected, since they have noticeably a lot more groups (71 categories in level 2; 264 categories in level 3, as reported in Table I), as compared to other classification schemes of the same levels. We should rather compare the results between GICS, TF-IDF and BTIC, whose organizational structures are the same, hence the corresponding results are straightforward and easier to interpret. BTIC's inter-industry homogeneity is approximately the same with that of GICS in all levels, and the incremental homogeneity across different levels are sufficiently close, too, with all differences ranging within 0.01 to 0.1. This shows that BTIC performs just as good as GICS in terms of inter-industry homogeneity. We present the case of TF-IDF as well for the baseline comparison, and it can be observed that BTIC's performance is generally better than that of the TF-IDF method.

The results from the test of inter-industry homogeneity are reported below in Table VI. The raw R^2 values are reported in column (R), next to the corresponding classification scheme and level. Differences in R^2 values across different classification systems are noted in the second panel below the raw R^2. Again, we focus on comparing the values of BTIC to GICS and TF-IDF, since they share the same organizational structure. The results show that the value of homogeneity across different levels of classification is fairly

Table VI: Inter-industry tests of homogeneity

SIC	(R)	GICS	(R)	TF-IDF	(R)	BTIC	(R)
Lv. 1	0.11	Lv. 1	0.13	Lv. 1	0.10	Lv. 1	0.13
–	–	Lv. 2	0.20	Lv. 2	0.16	Lv. 2	0.19
Lv. 2	0.49	Lv. 3	0.37	Lv. 3	0.25	Lv. 3	0.32

TF-IDF vs SIC	TF-IDF vs GICS	BTIC vs SIC	BTIC vs GICS
-0.01	-0.03	0	0
–	-0.04	–	-0.01
-0.23	-0.12	-0.17	-0.05

Notes – Lev. # denotes the #-th division level of the subject classification scheme.

close to each other, especially for the broadest level. The difference of BTIC and GICS in inter-industry homogeneity is only about -0.01 for levels 1 and 2, and about -0.05 for level 3. BTIC outperforms TF-IDF method in all three levels when compared to GICS.

Experiment results lead to a conclusion that BTIC performs just as good as GICS in terms of inter-and intra-industry homogeneity. Given that, then, BTIC outperforms GICS in four aspects: process automation, objectivity, flexibility, and interpretability. Process automation is indeed a huge advantage as compared to the past industry classification schemes. Because less human effort is required during the process, the cost of information update drops, while the clustering results are always up-to-date. Objectivity is another important feature of BTIC. Because the construction of BTIC is data-driven, results assure objectivity. By the design of the classification scheme, BTIC provides flexibility in choosing the size and depth of the levels of the clusters, by setting different thresholds. Furthermore, one can now freely choose the time-point of the industry classification when conducting market study, by simply choosing a specific year of 10-Ks and running BTIC to get industry clusters of the time of interest. Finally, BTIC's results facilitates interpretation of the result in rich context. For example, does the results of BTIC not only include hierarchical clusters, but also bags of words whose distances were close with one another for each cluster, supplying richer explanation about the characteristics and/or properties of securities grouped together.

V. Conclusion & Future Work

Effective industry classification has been the integral part of capital market research. The existing industry classification schemes, however, suffer from limitations in various aspects, such as extensive input of human labor in the development and the update of classification systems and inconsistency in the "orientation" of the classification process. This paper aims to develop a novel industry classification scheme–Business Text Industry Classification, or BTIC, namely–that addresses such issues and effectively segment the capital market into segments sharing similar characteristics. We utilize the business section of the form 10-Ks and employ Doc2vec and hierarchical clustering algorithms to group firms together with similar "self-perceived

identities". Experiment results show that BTIC performs as great as SIC and GICS at the broadest level, and GICS to the second-broadest level, the two of the most widely used industrial classification schemes in finance and business research, in terms of inter-industry heterogeneity and intra-industry homogeneity. Given the experiment results, we conclude that BTIC outperforms the existing industry classification schemes in four aspects: process automation, objectivity of the cluster results, flexibility of cluster size and depth, and interpretability of the resulting clusters. At this stage, we are advancing our algorithms further so that we can automatically cluster a greater number of securities into BTIC clusters (i.e. Russell 2000 or 3000). In addition, We plan to develop our study further by allowing multiple membership of securities in different clusters. We are also designing a trajectory analysis of the changing industry membership of securities over time.

REFERENCES

[1] R. L. Phillips and R. Ormsby, "Industry classification schemes: An analysis and review," *Journal of Business & Finance Librarianship*, vol. 21, no. 1, pp. 1–25, 2016.

[2] Netflix, Inc., "Form 10-k 2016," Retrieved from SEC EDGAR website `http://www.sec.gov/edgar.shtml`, 2016.

[3] Wynn Resorts, "Form 10-k 2016," Retrieved from SEC EDGAR website `http://www.sec.gov/edgar.shtml`, 2016.

[4] Amazon.com, Inc., "Form 10-k 2016," Retrieved from SEC EDGAR website `http://www.sec.gov/edgar.shtml`, 1998.

[5] ——, "Form 10-k 2016," Retrieved from SEC EDGAR website `http://www.sec.gov/edgar.shtml`, 2016.

[6] K. Hrazdil, K. Trottier, and R. Zhang, "An intra-and inter-industry evaluation of three classification schemes common in capital market research," *Applied Economics*, vol. 46, no. 17, pp. 2021–2033, 2014.

[7] K. Palepu, "Diversification strategy, profit performance and the entropy measure," *Strategic management journal*, vol. 6, no. 3, pp. 239–255, 1985.

[8] J. P. Fan and V. K. Goyal, "On the patterns and wealth effects of vertical mergers," *The Journal of Business*, vol. 79, no. 2, pp. 877–902, 2006.

[9] S. Wallsten, "The effects of government-industry rprograms on private r&d: the cause of the small business innovation research program," *The RAND Journal of Economics*, vol. 31, pp. 82–100, 2000.

[10] United States Department of Labor, "SIC Division Structure," Occupational Safety and Health Administration `www.osha.gov/pls/imis/sic_manual.html`, 2016.

[11] United States Census, "NAICS Update process fact sheet," 2016.

[12] MSCI Inc., "Global industry classification standard (GICS)," 2016.

[13] D. A. Guenther and A. J. Rosman, "Differences between compustat and crsp sic codes and related effects on research," *Journal of Accounting and Economics*, vol. 18, no. 1, pp. 115–128, 1994.

[14] J. Krishnan and E. Press, "The north american industry classification system and its implications for accounting research," *Contemporary Accounting Research*, vol. 20, no. 4, pp. 685–717, 2003.

[15] S. Bhojraj, C. Lee, and D. K. Oler, "What's my line? a comparison of industry classification schemes for capital market research," *Journal of Accounting Research*, vol. 41, no. 5, pp. 745–774, 2003.

[16] G. Salton, "Automatic text processing: The transformation, analysis, and retrieval of," *Reading: Addison-Wesley*, 1989.

[17] Z. S. Harris, "Distributional structure." *Word*, 1954.

[18] G. E. Hinton, "Learning distributed representations of concepts," in *Proceedings of the eighth annual conference of the cognitive science society*, vol. 1. Amherst, MA, 1986, p. 12.

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[20] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[22] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642.

[23] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[24] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[25] R. Das, M. Zaheer, and C. Dyer, "Gaussian lda for topic models with word embeddings," in *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, 2015.

[26] Berkshire Hathaway Inc., "Form 10-k 2016," Retrieved from SEC EDGAR website `http://www.sec.gov/edgar.shtml`, 2016.