Information Technology and Management
https://doi.org/10.1007/s10799-023-00411-0

Analysis of the relationship between technological diversification and enterprise value using patent data

Yusuke Matsumoto[1] · Aiko Suge[1] · Hiroshi Takahashi[1]

Accepted: 17 August 2023

Abstract
As natural language processing technology advances, its application to finance is growing. We focus on technological diversification resulting from R&D activities, one of the concerns for firms requiring efficient business management. We analyze the relationship between technological diversification and enterprise value using text information of patents and examine text information's usefulness in corporate finance. Specifically, we created a firm's technological diversification index from the text information in the patents and the firm's excess value from its financial data and analyzed the relationship between them. We analyzed Japanese firms that have been listed on the First Section of the Tokyo Stock Exchange and have applied for patents during the period 2002–2015. As a result, we found that (1) the spread of technology in multiple ways could impair enterprise value, and (2) the use of text information can be more valuable than the use of specific standards indexes such as the International Patent Classification (IPC). This study shows interesting results on the relationship between technological diversification and enterprise value and implies the applicability of unstructured data to finance research. A detailed analysis is for further study.

Keywords Corporate finance · Diversification · Excess value · Patent · Natural language processing

JEL Classification G10 · G32

1 Introduction

Thanks to the progress of information technology, we can now analyze unstructured data such as images, voices, and texts, which used to be challenging to deal with and gain new insights. For example, patents, one of the outcomes of a firm's R&D activities, contain indicators such as the IPC and citations, as well as text and figures. In the past, when we analyzed patent data, we usually used IPC and other indicators. Recent progress in information technologies, such as the natural language process, has allowed us the possibility to analyze text and figures in patents.

Finance is one of the fields where structured data has been mainly used while text, images, and other data have existed. In recent years, the use of unstructured data, such as text in finance, has been expanding with the progress of information technology, especially natural language processing. For example, the impact of words and sentence structure used in news articles on the stock market has been discussed in the market finance field. In corporate finance, natural language is utilized in research on the relationship of M&A probability using sentences and words in investor relations materials such as Form 10-K.

In recent years, there has been a growing trend, like ESG investment, to evaluate non-financial aspects, such as human resources, technologies, and environmental contributions, in addition to the financial aspects of the enterprise. As a measure of a firm's technology, patents are well-used. Patents consist of unstructured data, such as text and figures, and structured data, such as IPC (International Patent Classification). In corporate finance, structured data such as IPC

✉ Yusuke Matsumoto
yusukeM-aichi@keio.jp

Aiko Suge
aikosuge@keio.jp

Hiroshi Takahashi
htaka@kbs.keio.ac.jp

[1] Graduate School of Business Administration, Keio University, 4-1-1 Hiyoshi, Kohoku-ku, Yokohama-City, Kanagawa-Pref. 223-8526, Japan

Published online: 11 October 2023

is well-used in research to evaluate enterprise value as an indication of the technological capabilities possessed by a firm. The development of natural language processing technology enables efficient analysis of the text contained in patent documents. It can bring new insights into the study of corporate valuation. Thus, in this study, we try to extract new insights into the finance field using patent text data and indices.

Innovation is a way for firms to increase their enterprise value. The firm's R&D activity is the engine for realizing innovation. Their proactive R&D activities in other domains and their firm's primary product and service domains will lead to technological integration and innovation. In other words, firms expand their R&D activities and expand and/or possess technologies in a wide variety of domains. It will lead to the expansion of their technology platforms and create new business opportunities.

R&D activities, however, tend to diverge in any direction, which may reduce the speed of business innovation compared to investment and lead to a conglomerate discount. The conglomerate discount refers to the tendency of diversified firms to be undervalued by the stock market in comparison to not-diversified firms [11, 4, 17, 24, 47]. Therefore, firms must adjust and/or manage the degree and/or direction of divergence of R&D activities and resulting technology. Since the 1950s, firms in Japan have enhanced their presence globally with the strength of their technology. Recently, however, it has been pointed out that Japanese firms have declined and stagnated in technological capabilities and/or enterprise value compared to their global counterparts. Japanese firms are being questioned whether they can enhance their enterprise value and improve their presence worldwide through efficient management of R&D activities and technological diversification. In this paper, we analyze the relationship between enterprise value and the management of technological diversification, using indices and/or text data in patents, and try to extract new insights for efficient firm management.

The structure of this paper is as follows. First, Chapter 2 describes previous research related to the theme of this paper, and Chapter 3 describes the data used for analysis. Chapter 4 describes how to create the variables used in the analysis. Chapter 5 explains the method of empirical analysis and discusses the results. Finally, Chapter 6 describes the summary and issues of this paper.

2 Related works and our motivation

2.1 Technological diversification

A firm's technological resources and capabilities are central to its competitive success, and technological innovation is one of the ways for a firm to enhance its enterprise value [1]. Therefore, many firms have expanded their technological fields, that is, technological diversification [9, 22]. However, there are opinions that technological diversification positively impacts enterprise value while it has a negative impact and A unanimous consensus view has not been reached. At first, Miller [35] and Lin and Chang [30] report that technological diversification positively impacts enterprise value. Pugliese et al. [40] show that firms with more consistent technological diversification have higher performance and labor productivity. As mentioned above, many reports claim that technological diversification improves corporate profits and enterprise value.

On the other hand, there are reports that technology diversification has a negative impact on firm performance or enterprise value. Granstrand and Oskarson [12] reported that technological diversification does not necessarily lead to increased revenues. Yamaguchi [50] analyzed the relationship between R&D investment diversification and profitability of Japanese firms during the period 2000–2004 and reported that the higher the degree of R&D investment diversification, the lower the profitability.

In real cases, some firms attempt technological diversification. For example, Japanese electronics manufacturer Canon had diversified its technology from its core technology of fine optical technology and precision machinery technology, which have been refined since the firm's founding, to semiconductor equipment technology used in steppers and/or microelectronics technology [22]. Thanks to its technological diversification, while many Japanese electronics manufacturers struggled to adapt to the transition to the information society in the 1990s, Canon was able to continue to increase profits [48]. On the other hand, in the power generation area, ABB in Swiss, which expanded its technology portfolio, performed lower than General Electric in the United States, which focused on its technology portfolio. As a result, ABB's financial condition deteriorated and it chose to sell off its power generation business to Alstom in France [3].

Studies that claim a positive or negative relationship between technological diversification and enterprise value suggest that the management of technological diversification, knowledge management, is essential. Silverman [43] claims that firms tend to diversify into areas related to each firm's existing technologies. Oikawa and Takahashi [38] use an agent-based model to model a firm's behavior in the technological space and analyze the relationship between a firm's behavior and enterprise value. They claim if there is a less accumulation of technology around the technological space to which a firm belongs, the firm will generate a large number of quality-adjusted patents. Iwaki [18] also reports that firms that select the technology domain rather than rashly expand their technology domain have higher enterprise value since Japan's bubble economy collapsed. It states that the situation may change depending on how they manage technological diversification. Miller [34] reports that a lack of management of the firm's internal technical knowledge leads to a decline in the enterprise value.

The degree of technological diversification of a firm can be obtained by calculating the degree of similarity between technologies within a firm, that is, the degrees to which the firms' technologies differ, using patents. Jaffe [21] examines which technology domain a firm submits as a patent and expresses the share of the technology domain within a firm using a vector. Specifically, Jaffe [21] expresses the share in the technological domain in the firm q as a vector of $F_q = \left(F_{q1}, F_{q2}, \dots, F_{qp}\right)$ when the technological domain can be represented by $p = 1, 2, 3, \dots, P$. Based on this method, two different methods have been used to calculate the similarity between technologies within a firm. The first method is to obtain a firm's financial indicators, such as R&D expenses for each firm's business segment, and calculate their Entropy [10, 20, 28, 36, 39, 50]. The second method counts the kinds of IPC described in the patent document. The IPC is a hierarchical system of the technological domain using codes. Lerner [27], Iwaki [18], Iwaki and Okada [19], and Zabala-Iturriagagoitia et al. [52] use IPC to express the breadth of the technological domain in the form of the Herfindahl-Hershman index (HHI) and Entropy.

2.2 Utilization of text information in finance

Thanks to the rapid development of information technology, it has become possible for us to handle unstructured data, especially text data. In studies in finance, the use of text data is expanding. As for studies that deal with the text data in market finance, we can see Loughran and Mcdonald [31] and Nishi et al. [37]. Loughran and Mcdonald [31] created a financial dictionary based on the words used in Form 10-K and analyzed their association with stock prices. Nishi et al. [37] evaluate news articles' sentences distributed to financial markets in terms of volatility of stock prices and construct a model to predict and evaluate stock price fluctuations. As in these studies, research using words and sentences has been conducted in market finance. As for studies that deal with text data in corporate finance, we can see Hoberg and Phillips [15] and Bellstam et al. [2]. Hoberg and Phillips [15] use words in each firm's product description from its Form 10-K to measure product similarity between firms and analyze its association with mergers and acquisitions. Specifically, they calculate product similarity between firms using the distance between words in the product description and report that the probability of M&A increases when firms have products that are more broadly similar to all other firms. Bellstam et al. [2] construct an innovation index based on texts within analyst reports through Latent Dirichlet Allocation (LDA) and analyze its relationship with enterprise value and profit margins. They find that the text-based innovation measure effectively evaluates a firm's innovation.

We have the potential to find new insights through text data analysis. Studies of market finance by Tetlock et al. [45] and corporate finance by Bellstam et al. [2] suggest that news and/or analyst report texts may hold firm's fundamental information that is difficult to quantify. As the advantages of the analysis through the process of the natural language, Bellstam et al. [2] claim that it can be computed for firms that do not patent and do not use R&D, which meaningfully expands the scope of innovation that can be studied, by using the text in analyst reports. If we use natural language processing, we could deepen and/or widen our research. The usefulness of natural language processing will likely lead to more research in finance.

When using text data in corporate finance research, we often focused on words used in the text, such as word kinds and word counts. Now we can efficiently obtain an effective distributed representation of the entire text through Doc2Vec [26], SCDV [33], and other methods, thanks to the advancement of deep learning technique and computer power. Therefore, this study will not focus on words but the text and attempt to apply it to the corporate finance field. Specifically, we analyze the relationship between technological diversification and enterprise value using patent indexes and/or text data and attempt to extract insights for corporate management, as well as examine the usefulness of text data in corporate finance.

2.3 Our motivation

When we measure technological diversification in the two ways described in Sect. 2.1, we often use indexes assigned based on specific criteria such as IPC, R&D investment, and firm segment information. However, several issues have been pointed out in dealing with these commonly used data. For example, the accuracy of the assignment of indicators based on specific criteria, such as IPC, is challenged by the fact that the assignment of classification indicators differs from person to person or by the revision of the criteria.

Regarding firms' R&D investment and segment information, changes in accounting standards require careful data utilization over a long time. It isn't easy to obtain detailed information on R&D fields. Segment information depends on the arbitrariness of the firms concerned. These challenges make it difficult to create indices that reflect the firm's actual conditions in a precise manner. Due to these data issues before analysis, we may have yet to reach a consistent view of the relationship between a firm's technological diversification and enterprise value.

Patent documents are a candidate for available data that confirms the firm's R&D activities. Patent documents have a possibility to solve these problems of classification accuracy, data availability, and firms' arbitrariness as much as possible. For example, we do not need to consider the assignment to a specific classification because the patent document is text data. Patent text data is publicly available in many countries, and we can obtain the data without effort. Unlike indicators, there is no revision or conversion of texts. A third party checks the patent texts for the grant of rights. We could consider that arbitrariness of firms is eliminated as much as possible from the patent text. Therefore, using patent documents may lead to a unified view of the relationship between technological diversification and enterprise value.

3 Data

We use the firm's patent data and financial data. We obtained the patent data from the DWPI (Derwent World Patent Index) from 2002 to 2015. We have obtained patent data for each firm by name collation using the firm's name, including the notations such as “Corp.”, “Inc.”, and “KK.”, which indicate a company limited.[1]

Patent data consists of the title, publication date, the International Patent Classification (IPC) code, and abstracts that DWPI experts summarize each patent. A unique point of our patent data used in this analysis is that DWPI experts write the abstract of each patent. The text's writing style and/or wording in each patent differs from author to author. By having a DWPI expert summarize each patent according to a specific standard, the habit of the text of each patent caused by different authors can be reduced, making it easier to use for analysis, such as natural language processing. The abstract written in English by DWPI experts is text data and consists of the following four parts, and we use the four parts; (1) the novelty of the invention, (2) a detailed description, (3) application, and (4) superiority. We make use of the four parts of the analysis.

This study focuses on sentences, not words. It is generally better to utilize the text because it expresses what the author wants to say. Even if we focus on words in dealing with sentences such as Form-10K and/or patent documents, it is generally necessary for us to reconstruct from words to sentences to analyze them at the firm level [15, 16, 29].

It has been pointed out that the method of reconstructing words into sentences by combining them does not consider that words have different meanings in different contexts. Patents are challenging because they require the expression of inventions related to more sophisticated, complex technology and because they contain technical terms that show the higher-level conceptualization of their invention. When we focus on words and obtain numerical expressions with word2vec or other methods, the same word will result in the same numerical expression. However, it may be difficult to express the firm’s technology in detail by restoring the numerical representation of a word to a sentence since each firm may have different methods of applying and composing technology.

Natural language processing technology has been developed to enable numerical expressions considering the meaning and/or context of entire sentences. SCDV, which we used to obtain the numerical representation of the patent document in this study, shows superior results in several tasks compared to reconstructing the numerical representation of words to sentences. Therefore, numerical representations acquired directly from sentences may provide a more detailed representation of the technology represented by each patent compared to the word-focused approach.

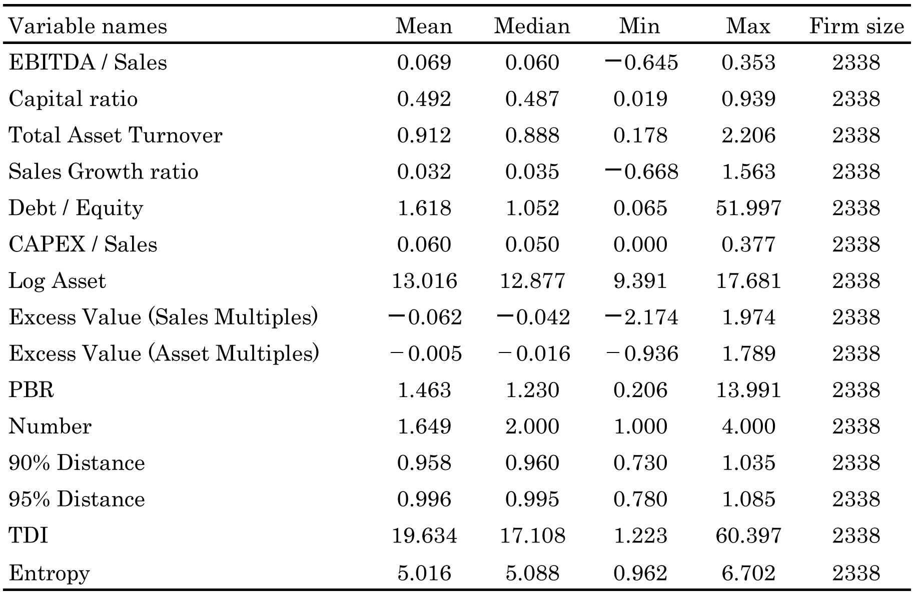
Next, we obtained the financial data since 2002 of firms listed on the First Section of the Tokyo Stock Exchange from Nikkei NEEDS. We mainly use four types of data about the firm's financial data; operating margin, equity ratio, total asset turnover, and sales growth rate. The descriptive statistics for each financial indicator are shown in rows 2–5 of Fig. 1. We select the firms listed on the Tokyo Stock Exchange because the listed firms are a collection of relatively large firms.

The analysis period was 14 years, from 2002 to 2015, when we obtained patents and financial data from each database. The number of firms in this analysis is 167. These firms have applied for patents yearly and have been listed for 14 years. In other words, our data has all the patent and financial data for 167 firms for 14 years.

4 Method

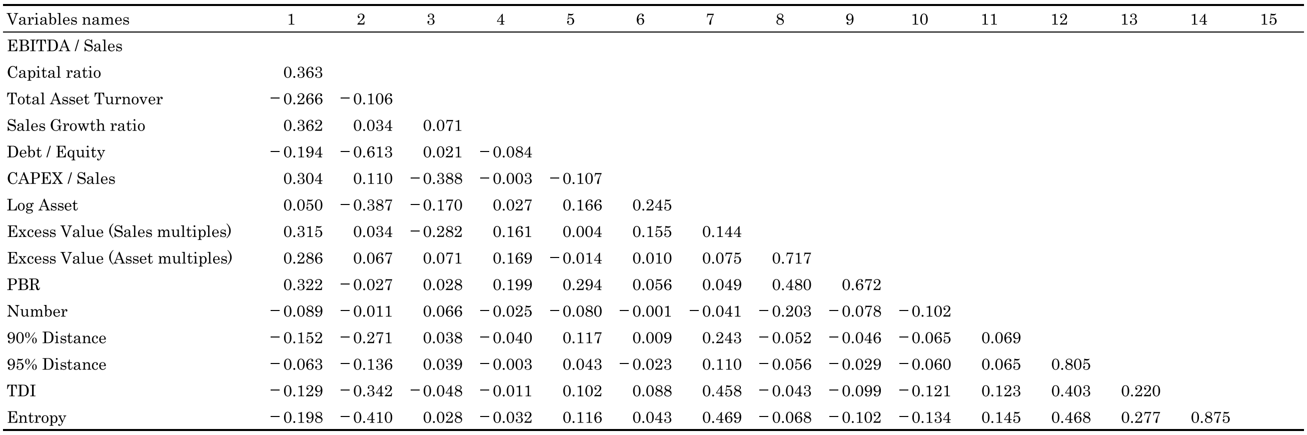
In this chapter, we introduce how to produce the distributed representation of document vectors in Sect. 4.1, calculate two kinds of diversification indexes in Sect. 4.2, compare two kinds of diversification indexes in Sect. 4.3, and calculate the excess value in Sect. 4.4. We will use the distributed representation of document vectors produced in Sect. 4.1 as one of the data to generate a kind of diversification index in Sect. 4.2. The excess firm value, which we will create in Sect. 4.2, is an index of enterprise value. It will be used as one variable in our regression analysis of the relationship between technological diversification and enterprise value (Fig. 2).

[1] The name collation method used in this study does not accurately cover affiliated firms, etc. Analysis of the impact of different identification methods is an issue for the future.



Variable names	Mean	Median	Min	Max	Firm size
EBITDA / Sales	0.069	0.060	−0.645	0.353	2338
Capital ratio	0.492	0.487	0.019	0.939	2338
Total Asset Turnover	0.912	0.888	0.178	2.206	2338
Sales Growth ratio	0.032	0.035	−0.668	1.563	2338
Debt / Equity	1.618	1.052	0.065	51.997	2338
CAPEX / Sales	0.060	0.050	0.000	0.377	2338
Log Asset	13.016	12.877	9.391	17.681	2338
Excess Value (Sales Multiples)	−0.062	−0.042	−2.174	1.974	2338
Excess Value (Asset Multiples)	− 0.005	− 0.016	− 0.936	1.789	2338
PBR	1.463	1.230	0.206	13.991	2338
Number	1.649	2.000	1.000	4.000	2338
90% Distance	0.958	0.960	0.730	1.035	2338
95% Distance	0.996	0.995	0.780	1.085	2338
TDI	19.634	17.108	1.223	60.397	2338
Entropy	5.016	5.088	0.962	6.702	2338

Fig. 1 Descriptive Statistics. This figure shows the descriptive statistics of the variables used in this paper. We use mainly variables15 variables; EBITDA/Sales, Capital ratio, Total Asset Turnover, Sales Growth ratio, Debt/Equity, CAPEX/Sales, Log Asset, Excess value (Based on Sales), Excess value (Based on Asset), PBR, Number, TDI, and Entropy. The first column shows the variable names. After the second column, the mean, median, minimum, maximum, and number of target firms for each variable are described. The variables described in rows 2–5 are used in Sect. 4.4 to calculate the firm's excess value, and the variables described in rows 2 and 6–16 are used in chapter 5 in some regressions



Variables names	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
EBITDA / Sales															
Capital ratio	0.363														
Total Asset Turnover	− 0.266	− 0.106													
Sales Growth ratio	0.362	0.034	0.071												
Debt / Equity	− 0.194	− 0.613	0.021	− 0.084											
CAPEX / Sales	0.304	0.110	− 0.388	− 0.003	− 0.107										
Log Asset	0.050	− 0.387	− 0.170	0.027	0.166	0.245									
Excess Value (Sales multiples)	0.315	0.034	− 0.282	0.161	0.004	0.155	0.144								
Excess Value (Asset multiples)	0.286	0.067	0.071	0.169	− 0.014	0.010	0.075	0.717							
PBR	0.322	− 0.027	0.028	0.199	0.294	0.056	0.049	0.480	0.672						
Number	− 0.089	− 0.011	0.066	− 0.025	− 0.080	− 0.001	− 0.041	− 0.203	− 0.078	− 0.102					
90% Distance	− 0.152	− 0.271	0.038	− 0.040	0.117	0.009	0.243	− 0.052	− 0.046	− 0.065	0.069				
95% Distance	− 0.063	− 0.136	0.039	− 0.003	0.043	− 0.023	0.110	− 0.056	− 0.029	− 0.060	0.065	0.805			
TDI	− 0.129	− 0.342	− 0.048	− 0.011	0.102	0.088	0.458	− 0.043	− 0.099	− 0.121	0.123	0.403	0.220		
Entropy	− 0.198	− 0.410	0.028	− 0.032	0.116	0.043	0.469	− 0.068	− 0.102	− 0.134	0.145	0.468	0.277	0.875	

Fig. 2 Correlation Coefficient Matrix. This figure shows the correlation matrix of the variables used in this paper. We use mainly variables15 variables; EBITDA/Sales, Capital ratio, Total Asset Turnover, Sales Growth ratio, Debt/Equity, CAPEX/Sales, Log Asset, Excess value (Based on Sales), Excess value (Based on Asset), PBR, Number, TDI, and Entropy. The first column shows the variable names. After the second column, the correlation coefficients between each variable are described. The variables described in rows 2–5 are used in Sect. 4.4 to calculate the firm's excess value, and the variables described in rows 2 and 6–16 are used in chapter 5 in some regressions

4.1 Producing the distributed representation of documents

When we use text data, such as patent documents, for analysis, we need to convert the text information into numerical information. In this section, we explain how to convert text data into numeric information, in other words, how to obtain a distributed representation of a document.

Before we obtained a distributed representation of the patent documents, we performed preprocessing, following Gupta et al. [13]. Specifically, we removed common numbers and stop-words from the patent documents. Numbers in the sentences were not considered helpful for topic classification and topic-based adjustment. Therefore, we reduced the vocabulary by removing numbers to preserve the computer's performance. Stop-words are commonly used in the text without contextual meaning, like "the" and "an". Stop-words are often not useful for their high frequency of occurrence. Hence, we removed stop-words to preserve the performance of the calculator. Indeed, there are cases when numbers play an essential role. For example, some patents are related to chemistry or physics. In those fields, numbers are sometimes used to express mathematical expressions, chemical formulas, and speed. Each number may have great significance in indicating the differences. However, in the abstract data used in this study, numbers are often used to indicate section numbers. Few abstract data contain mathematical expressions and chemical formulas. In addition, we performed stemming and converted the letters of the patent documents to lowercase in order to preserve our computer's performance.

We use Sparse Composite Document Vectors (SCDV) to obtain the distributed representation of the document.[2] The SCDV is obtained by adjusting the distributed representation of words based on the topic of each sentence and computing the average of the adjusted distributed representation of words. The data is each patent's abstract data summarized by some DWPI experts in terms of novelty, detailed description, application, and superiority of the invention. To obtain the distributed representation of the patent's document vector through SCDV, we use 688,172 patents, including patents of firms other than 167. The more word or sentence data available, the more elaborate dispersed expression of the word and sentence vectors can be obtained.

First, we obtain a distributed representation of the d dimensional words through the Skip-Gram model for these data. Next, through the GMM, we stochastically classify the whole word vectors into each cluster and weight each cluster with a probability. We transform the $\overrightarrow{wcv}_{ik}$ thus obtained into a distributed representation of the word in the $d \times K$ dimension by combining it K times and then weight this with the inverse document frequency IDF of each word to obtain $\overrightarrow{wtv}_i$. w_i is the ith word, and k is the number of clusters.

$$\overrightarrow{wcv}_{ik} = \vec{wv}_i \times P\left(C_k|w_i\right) \tag{1}$$

$$\overrightarrow{wtv}_i = IDF_t \times \oplus_{(1\sim K)} \overrightarrow{wcv}_{ik} \tag{2}$$

The inverse document frequency in Eq. (2) is measured as shown in Eq. (3). N means the total number of documents, and df_t means the number of documents in which the word t appears.

$$IDF_t = log\frac{N}{df_t} + 1 \tag{3}$$

Based on the distributed representation of word obtained thus, we obtain the distributed representation of document vectors by summing the distributed representation of word $\overrightarrow{wtv}_i$ in document D_n and standardizing it. n denotes the document number. Finally, we obtain the distributed representation of each patent document $SCDV_{D_n}$ by making it sparse at the threshold. We set the number of dimensions by the Skip-Gram model at 200, the number of clusters by the mixture model at 60, and the threshold for sparse at 3%, following Mekala et al. [33]. The SCDV size of each patent is 12,000. Figure 3 shows the visualization of all 6,314 patent vectors which Bridgestone has.

4.2 Calculating two kinds of diversification indexes

We create diversification indexes for each firm using patent data. In this paper, we create two kinds of diversification indexes based on the abstract data of each patent and each IPC index. The descriptive statistics for each diversification index are shown in rows 13–16 of Fig. 1.

First, we explain how to calculate the diversification index using the abstracts of each patent. At first, we collect, for each firm i at time t, the patents from 2002, the first year, to time t. The matrix size is the number of published patents by firm i from 2002 to time t multiplied by 12,000 SCDV size. We then calculate the position vector of the center of gravity g of firm i at time t. The size of g is $1 \times 12,000$. Finally, we calculate the distance, $Distance$, between the position vector of the center of gravity g and the patent a at time t for firm i, as shown in Eq. (4).

[2] Thanks to the rapid progress in natural language processing, many models have been reported to obtain the distributed representation of words or documents in recent years. Analysis using natural language processes, except for SCDV, is a future challenge.

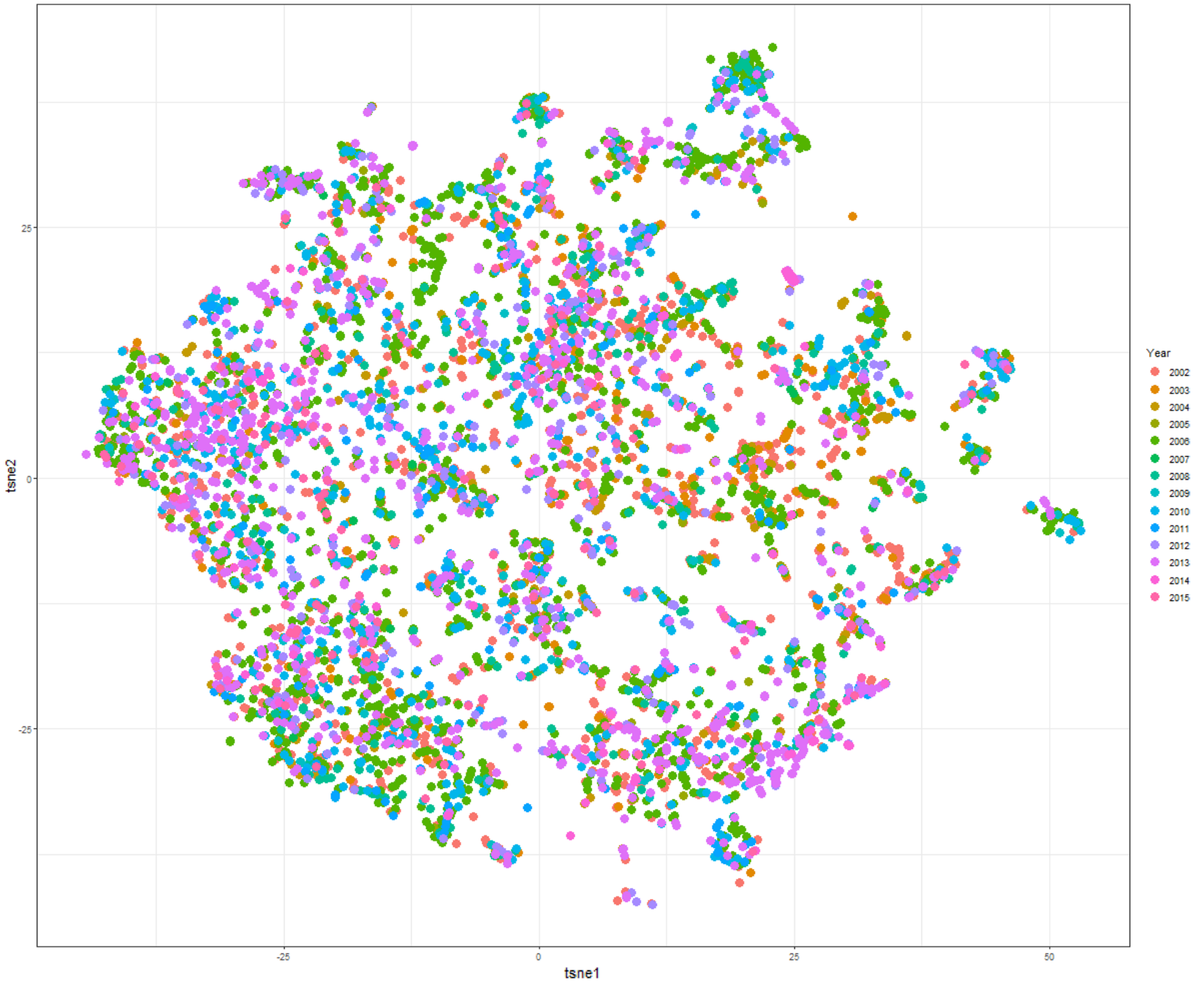


Fig. 3 Visualization of Bridgestone's Patents. This figure shows a two-dimensional visualization of the distributed representation of document vectors obtained from Sect. 4.1 for the patents owned by Bridgestone using t-SNE. Each point represents a patent held by Bridgestone. The legend shows the year of publication of the gazette for each patent held by Bridgestone

$$Distance = \sqrt{\left(patent_{ait} - g_{it}\right)^2} \tag{4}$$

We calculate the position vector of the center of gravity g of firm i at time t by the accumulation of the distributed representation of each patent from the first year, 2002, to time t to express that technological development is generated by the accumulation from the past. The position vector of the center of gravity g is the center of each patent held by firm i from 2002 to time t. It means that the position vector of the center of gravity g can be considered to represent the center of R&D, in other words, the core technology of firm i.

We then sort the distance between the position vector of the center of gravity g of firm i in year t and all patents of firm i from 2002 to year t in ascending order, and outliers are excluded at the 90th and 95th percentile points. We define its 90th and 95th percentile points (90% Distance and 95% Distance) as one kind of diversification index.[3] We can judge the diversification index (90% Distance and 95% Distance) to be a multi-business firm since the longer the distance, the farther away from the center g which indicates the core business and technology of the firm. We can judge the longer the distance, the more multi-business firm is, for the diversification index (90% Distance and 95% Distance).

Next, we explain how to calculate the other kind of diversification indexes from each patent's IPC index data. In this

[3] In this study, we employ 90% and 95% confidence intervals, which are commonly used in statistics field.

paper, we calculate the degree of technological similarity between the technologies of each firm from the IPC index attached to each patent and regard it as a diversification index. In terms of technological similarity, a low degree of similarity means that a firm is a multi-business firm because it expands its technology into multiple fields. A high degree of similarity means that the technology is highly concentrated. The IPC is an index consisting of up to fifth layer structure using numbers and alphabets, and the deeper the layers go from the first layer to the fifth layer, the more detailed the classification is. In this paper, we refer to Iwaki [18] and use "Sub-class" in the third layer. As shown in previous work in Chapter 2, there are two methods of measuring the degree of technological similarity between technologies. Many studies use the HHI or Entropy. Therefore, we attempt to create two kinds of technological similarity of technologies, in other words, diversification indices, (1) TDI, the degree of technological diversification index, which is based on HHI, and (2) Entropy.

Regarding creating the HHI, we follow Leten et al. [28] and Iwaki [18] and calculate the HHI of a firm i at t from the frequency n of Sub-class s of the IPC at firm i at t. Next, the reverse of this, the degree of technological diversification index (TDI), is created as shown in Eqs. (5) and (6). m represents the number of types of Sub-class in the IPC.

$$N_{it} = \sum_{s=1}^{m} n_{ist} \tag{5}$$

$$TDI_{it} = \frac{1}{\sum_{s=1}^{m} \left(\frac{n_{ist}}{N_{it}} \right)^2} \tag{6}$$

The HHI indicates occupancy, meaning that the larger the value, the more concentrated it is in one Sub-class. The degree of the technological diversification index (TDI) is the inverse of the HHI, and the higher the value, the more diversified the firm is.

Regarding the creation of Entropy, we refer to Jacquemin and Berry [20] and Miyazawa [36] and calculate the technical similarity, Entropy, as shown in Eq. (7) below. The meaning of each symbol is the same as the Technological Diversification Index (TDI).

$$Entropy_{it} = -\sum_{s=1}^{m} \frac{n_{ist}}{\sum_{s=1}^{m} n_{ist}} \times \log_2 \left(\frac{n_{ist}}{\sum_{s=1}^{m} n_{ist}} \right) \tag{7}$$

Entropy is a Sub-class's clutter, and the larger the value, the more diversified the firm is.

4.3 Comparing two kinds of diversification indexes

The diversification measure based on patent text, Distance, is considered superior to those on IPC, HHI and Entropy, in that it has the potential to have detailed information. In previous studies, technological diversification indices were created using indices assigned by specific criteria, such as IPC, and financial information, such as R&D expenditures, to measure R&D capabilities. In this study, HHI and Entropy are relevant in Sect. 4.2. The IPC, accounting indicators only reflect the rough state of a firm's technology, as the IPC is a technical overview. It is difficult to say that they accurately and in detail represent the firm’s state of technological diversification and/or its differences from other firms. On the other hand, Distance in Sect. 4.2 is created based on patent documents that show the firm's technologies in detail through natural language processing. Therefore, Distance is superior to HHI and Entropy, in that it contains more detailed information and can accurately and in detail illustrates the firm’s spread of technological capabilities and how it differs from other firms. In addition, as other advantages of our proposal method, we can visualize in a 2- or 3-dimensional space by applying natural language processing technology and grasp the spread of technological capabilities a firm possesses and the trajectory of the movement of its center of gravity.

4.4 Calculating the excess value

We calculate the excess value ($Excessvalue$) of firm i at time t as shown in Eqs. (8) and (9) below by referring to Berger and Ofek [4] in this paper.

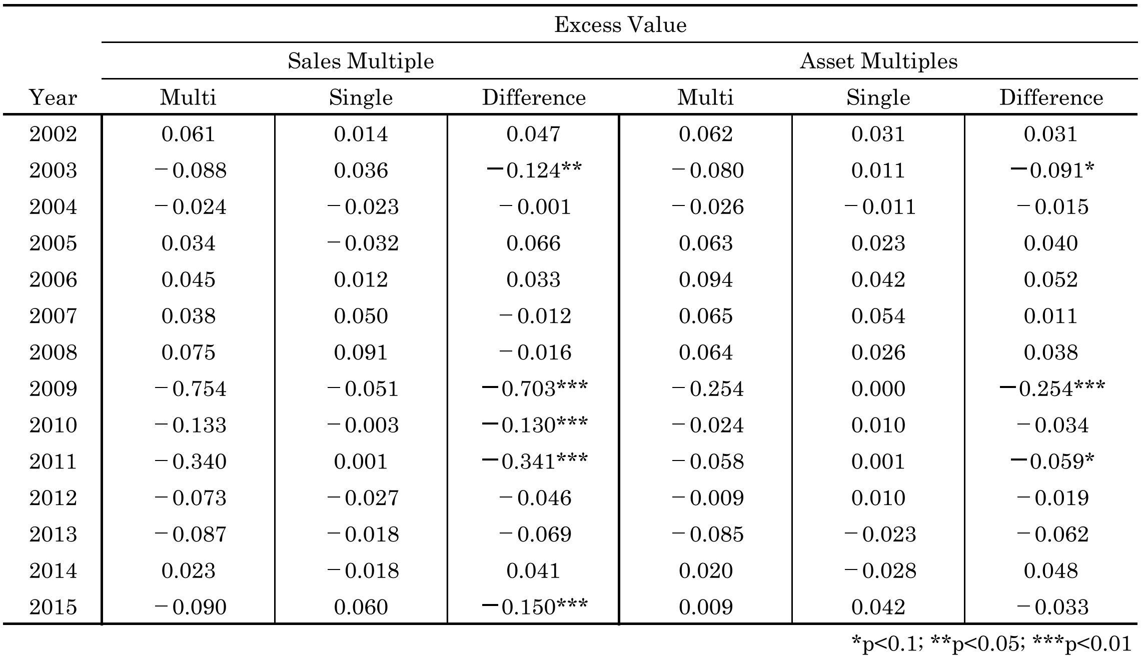
$$Excessvalue_{it} = ln\left(\frac{V_{it}}{IV_{it}} \right) \tag{9}$$

$$IV_{it} = \sum_{k=1}^{n} Index_{ikt} \times median_{kt}\left(\frac{V_t}{Index_t} \right) \tag{9}$$

V represents the enterprise value of firm i at time t, which is the sum of the market value and the book value of debt in Eq. (8). IV represents the firm's imputed value of firm i at time t, calculated by summing the value of segments ($k = 1, 2, \cdots, n$) a firm i has in Eq. (9). Each value of segments is calculated by multiplying the median valuation multiplier calculated by the sales of the single-segment firms which belong to the same industry as segment k of firm i by the $Index$ of segment k at time t of firm i. The segment for single-segment firms means the industry to which single-segment firms belong. Therefore, the excess value is defined as the log of the total enterprise and imputed value ratio using the median segment multiplier based on $theIndex$. This paper uses the firm's sales and total assets as $Index$.[4]

We must figure out in advance whether firm i is a single-segment firm or the multi-business firm and which industry

[4] Detail analysis of the excess value based on other indexes is a future task. For example, Berger and Ofek [4] calculated the excess value based on EBIT other than sales and assets.



	Excess Value					
	Sales Multiple			Asset Multiples		
Year	Multi	Single	Difference	Multi	Single	Difference
2002	0.061	0.014	0.047	0.062	0.031	0.031
2003	−0.088	0.036	−0.124**	−0.080	0.011	−0.091*
2004	−0.024	−0.023	−0.001	−0.026	−0.011	−0.015
2005	0.034	−0.032	0.066	0.063	0.023	0.040
2006	0.045	0.012	0.033	0.094	0.042	0.052
2007	0.038	0.050	−0.012	0.065	0.054	0.011
2008	0.075	0.091	−0.016	0.064	0.026	0.038
2009	−0.754	−0.051	−0.703***	−0.254	0.000	−0.254***
2010	−0.133	−0.003	−0.130***	−0.024	0.010	−0.034
2011	−0.340	0.001	−0.341***	−0.058	0.001	−0.059*
2012	−0.073	−0.027	−0.046	−0.009	0.010	−0.019
2013	−0.087	−0.018	−0.069	−0.085	−0.023	−0.062
2014	0.023	−0.018	0.041	0.020	−0.028	0.048
2015	−0.090	0.060	−0.150***	0.009	0.042	−0.033

*p<0.1; **p<0.05; ***p<0.01

Fig. 4 Time Series of Excess Value created in Sect. 4.3 and Comparing. This figure shows the excess value created in Sect. 4.4 for each multi-business firm and each single-segment firm over time. The name of the variable is written in the first column. The excess value, based on the firm's sales, is shown in the second and third columns, and the excess value, which is based on the firm's total assets, is shown in the fifth and sixth columns. The mean difference between the excess value of multi-business firms and the excess firm value of single-segment and the results after the t-test are shown in the fourth and seventh columns. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

segment k of firm i belongs to if firm i is a multi-business firm. The existing industrial classification systems, such as Japan Standard Industrial Classification (JSIC), Global Industry Classification Standard (GICS), and TOPIX Sector Indices, are often used to recognize the industry to which a firm belongs. However, some problems are pointed out in this method. Kimura [23] and Matsumoto et al. [32] pointed out that the existing industrial classifications may not accurately reflect the firm's current state. They state that even if a firm is multi-business, only one industrial classification code is usually assigned. It is difficult for us to objectively determine whether one firm is a multi-business firm, and the existing industrial classification code is assigned mainly based on a business with the largest share of the firm's sales composition. Therefore, we construct a new industrial classification system in this paper.

We assign industries to each firm by classifying the firms through the Gaussian Mixture Model (GMM). The GMM is one of the clustering methods to group similar data patterns into the same group. One of the characteristics of the GMM is that it assumes that the observed data exist in mixed distributions with different probability distributions and estimates the class label to which each data belongs as a latent variable [5].

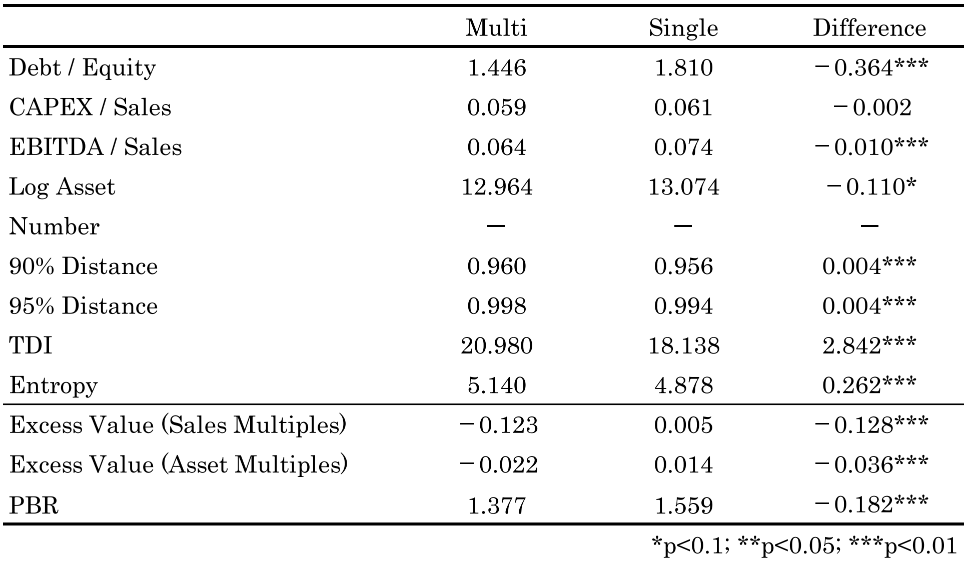
$$p(x) = \sum_{k=1}^{K} \pi_k N\left(x|\mu_k, \sigma_k\right) \tag{10}$$

$$\ln p(X|\pi, \mu, \sigma) = \sum_{n=1}^{N} \ln\left\{ \sum_{k=1}^{K} \pi_k N\left(x_n|\mu_k, \sigma_k\right) \right\} \tag{11}$$

The probability of gaussian mixture distribution is assumed to be the sum of the k th element's normal distribution $N\left(x|\mu_k, \sigma_k\right)$ multiplied by its coefficient π_k, as in Eq. (10). The sum of π_k is 1. Using maximum likelihood estimation, we estimate the parameters π, μ, σ based on the observed data $X = \left\{x_1, \cdots, x_N\right\}$. Specifically, we maximize the likelihood function in Eq. (11) using the EM algorithm [8]. The EM (Expectation Maximization) algorithm is a calculation procedure that approaches a most likely value by giving appropriate initial values to the parameters π, μ, σ and repeating the Expectation and Maximization steps.

We describe the settings in the GMM. We set the initial number of clusters K to 8, the same as our industry classification, the Japanese Standard Industrial Classification.[5]

[5] In this analysis, we set the number of clusters 8 as an initial value since the target firms in this paper belong to 8 industries in the Japan Standard Industrial Classification (JSIC). Detail analysis of the best number of clusters is a future task.



	Multi	Single	Difference
Debt / Equity	1.446	1.810	− 0.364***
CAPEX / Sales	0.059	0.061	− 0.002
EBITDA / Sales	0.064	0.074	− 0.010***
Log Asset	12.964	13.074	− 0.110*
Number	—	—	—
90% Distance	0.960	0.956	0.004***
95% Distance	0.998	0.994	0.004***
TDI	20.980	18.138	2.842***
Entropy	5.140	4.878	0.262***
Excess Value (Sales Multiples)	− 0.123	0.005	− 0.128***
Excess Value (Asset Multiples)	− 0.022	0.014	− 0.036***
PBR	1.377	1.559	− 0.182***

*p<0.1; **p<0.05; ***p<0.01

Fig. 5 Comparing the Indexes used in the Regression. This figure shows the results of the comparison of descriptive statistics for the main measures used in the regression equation. The part above the middle line is used as the independent variable in regression, and the part below the line is used as the dependent variable. The name of the variable is written in the first column. The average values of multi-business firms for each index are shown in the second column, and the average values of single-segment firms for each index are described in the third column. The fourth column shows the difference between the mean values of the multi-business and single-segment firms for each index and the results after the t-test. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

In this paper, we used four types of data x_n; (1) operating margin, which indicates profitability; (2) capital ratio, which indicates safety; (3) total asset turnover, which indicates activity and (4) sales growth, which indicates growth [32, 49, 51].

Each firm has assigned probability for each of the 8 clusters due to the analysis with the GMM. We name firms' first and second industries in order of clusters with the highest attribution probability. This paper considers the term "Cluster" as the industry.

Figure 4 shows the means and results of the t-test of the excess value using sales and asset multipliers over time in this section for multi-business and single-segment firms. The column "Difference" in Figure 4 shows the difference in the mean excess value between the multi-business and single-segment firms. In 2003, 2009, and 2011, the excess value of multi-business firms was statistically significantly lower than the excess firm value of single-segment firms. Especially in 2009, the excess value of multi-business firms was significantly lower at the 1% level than the excess firm value of single-segment firms, both based on sales and total assets. The results in Figure 4 suggest that, in some years, conglomerate discounts may have occurred in which multi-business firms are valued lower than single-segment firms.

Agency problems within firms or between a firm's management and investors are often cited as a theoretical reason why multi-business firms are undervalued compared to single-segment firms [44]. Agency problems are conflict-of-interest problems that arise between principals and agents due to information asymmetries and incomplete contracts. Regarding the agency problem within a firm, the optimal capital allocation to each field is challenging. Diversification of fields such as business, technology, and geography complicates stakeholders' relationships. It can lead to excessive political wrangling and rent-seeking over the allocation of funds between the borrowers of funds (e.g., heads of departments and fields) and the lenders of funds (e.g., management) [42]. Consequently, this leads to inefficiencies in the allocation of funds, where insufficient funds are allocated to sectors with growth potential, and extra funds are allocated to sectors with few investment opportunities. The inefficient allocation of funds is considered to lead to managerial inefficiency and, finally, to a decline in enterprise value [41].

Regarding the agency problem between managers and investors, information asymmetry exists between them when management and ownership are separated. Investors cannot adequately monitor the behavior of managers, which in turn causes managers to act to obtain personal benefits, a moral hazard for the managers [46]. Diversification itself (and the increased complexity caused by it) contributes to information asymmetry between management and investors, leading to the deterioration of the firm's enterprise value.

5 Empirical analysis

This chapter analyzes the relationship between technological diversification and enterprise value and expresses the result.

5.1 Compare the indexes used in the regression

Figure 5 summarizes the indexes used in this paper by multi-business and single-segment firms. The column “Difference" in Figure 5 shows the difference between the average of multi-business firms and the average of single-segment firms for each variable. The difference in the average excess firm value between multi-business and single-segment firms is due to various factors other than diversification. The column “Difference" shows a statistically significant difference between the multi-business and single-segment firms on indexes other than CAPEX/Sales. Besides, as discussed in Sect. 4.4, for excess value based on sales and total assets, the average excess value of multi-business firms is significantly lower than the average excess firm value of single-segment firms at the 1% level.

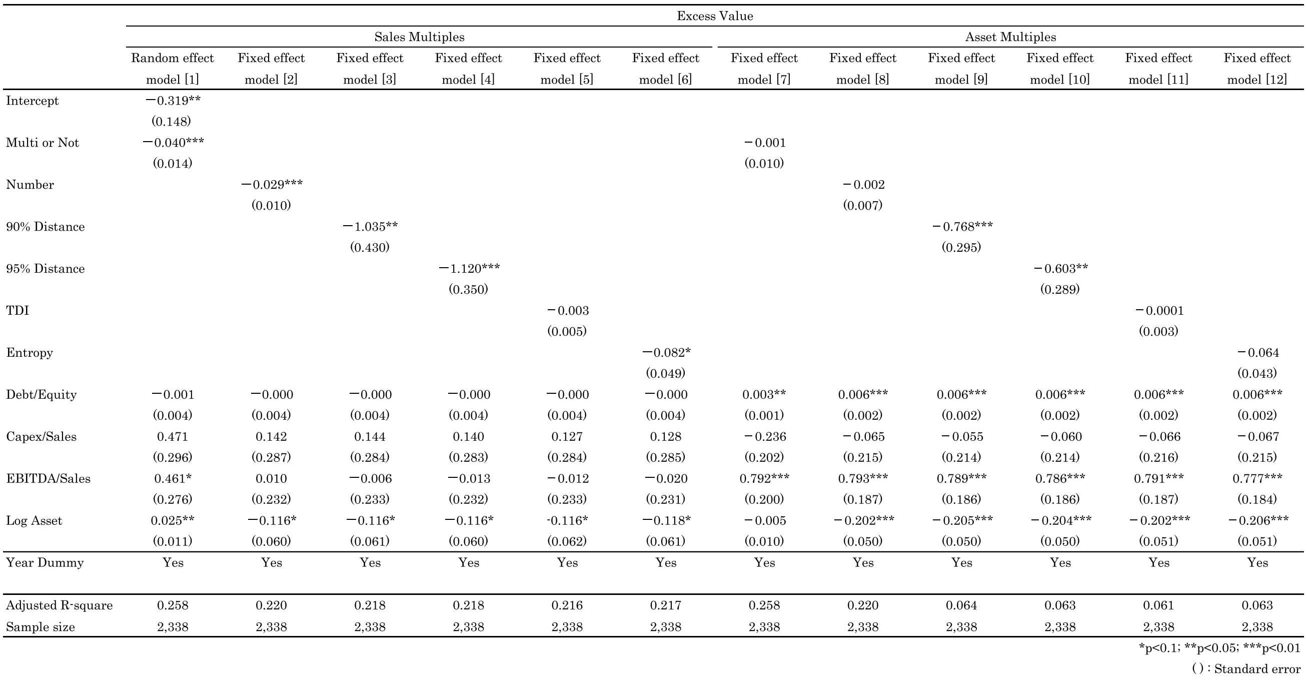
5.2 Analyzing the relationship between enterprise value and technological diversification

Our data is not missing any financial data and some diversification indexes measured by patent information for the 167 firms. In other words, our data is balanced over 14 years for 167 firms. Using panel analysis, we analyze the relationship between a firm's diversification and enterprise value and between technological diversification and enterprise value. We follow Berger and Ofek [4] and Ushijima [47]. The reason for analyzing not only the relationship between a firm's technological diversification and enterprise value, but also between a firm's diversification and enterprise value, is to validate the excess enterprise value generated in Sect. 4.4. For this reason, we consider an analytical framework following previous studies that used excess value. The reason for using a panel analysis is to consider firm-specific factors. A firm's diversification and/or technological diversification results from the firm’s behavior, such as decision-making. The issue of endogeneity is accompanied when we discuss the relationship between technological diversification and enterprise value [7, 47]. The endogeneity problem may stem from factors in other features of the firm rather than the firm's diversification and/or technological diversification. We can consider the firm's fixed factors by using fixed and random effects models in panel analysis, assuming other firm characteristics do not change over time.

The model we estimate is followed by Eq. (12). The dependent variable, *Excessvalue*, is calculated in Sect. 4.4. We use six diversification indexes, including two variables in addition to four variables created in chapter 4, as the independent variable, $Div_{i,t}$. One is a dummy variable that takes the value 1 when firm i is a multi-business firm and 0 otherwise, and the other is the number of industry segments to which firm i belong. $X_{i,t}$ is the control variable. These two variables are indicators of a firm’s diversification. We control for factors that could affect excess value and whose magnitudes are not entirely determined by whether the firm is multi-business or not, based on Berger and Ofek [4] and Ushijima [47]. We use five variables, Capex/Sales, Debt/Equity, EBITDA/Sales, Log Asset, Year Dummy. FE_i denotes the fixed effect for each firm.

$$Excess\ value_{it} = \alpha + \beta_1 \cdot Div_{i,t} + \beta_2 \cdot X_{i,t} + FE_i + \epsilon_{i,t} \tag{12}$$

Figure 6 shows the estimation results of the panel analysis using the excess value based on sales and asset multiples as the dependent variables, respectively. First, when we use the excess value using sales multiples as a dependent variable, the diversification dummy (Multi or Not) in Model [1] is statistically significantly negative at the 1% level, and the number of industries (Number) in Model [2] is also statistically significantly negative at the 1% level. Ushijima [47] states that the significantly negative coefficient on the number of industry segments reflects the difference between multi-business and single-segment firms, and increasing the degree of diversification does not simply lead to a decline in enterprise value. The estimation results in Model [1] and Model [2] suggest that multi-business firms may be undervalued compared to single-segment firms; in other words,



	Excess Value											
	Sales Multiples						Asset Multiples					
	Random effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]	Fixed effect model [7]	Fixed effect model [8]	Fixed effect model [9]	Fixed effect model [10]	Fixed effect model [11]	Fixed effect model [12]
Intercept	−0.319**											
	(0.148)											
Multi or Not	−0.040***						−0.001					
	(0.014)						(0.010)					
Number		−0.029***						−0.002				
		(0.010)						(0.007)				
90% Distance			−1.035**						−0.768***			
			(0.430)						(0.295)			
95% Distance				−1.120***						−0.603**		
				(0.350)						(0.289)		
TDI					−0.003						−0.0001	
					(0.005)						(0.003)	
Entropy						−0.082*						−0.064
						(0.049)						(0.043)
Debt/Equity	−0.001	−0.000	−0.000	−0.000	−0.000	−0.000	0.003**	0.006***	0.006***	0.006***	0.006***	0.006***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Capex/Sales	0.471	0.142	0.144	0.140	0.127	0.128	−0.236	−0.065	−0.055	−0.060	−0.066	−0.067
	(0.296)	(0.287)	(0.284)	(0.283)	(0.284)	(0.285)	(0.202)	(0.215)	(0.214)	(0.214)	(0.216)	(0.215)
EBITDA/Sales	0.461*	0.010	−0.006	−0.013	−0.012	−0.020	0.792***	0.793***	0.789***	0.786***	0.791***	0.777***
	(0.276)	(0.232)	(0.233)	(0.232)	(0.233)	(0.231)	(0.200)	(0.187)	(0.186)	(0.186)	(0.187)	(0.184)
Log Asset	0.025**	−0.116*	−0.116*	−0.116*	-0.116*	−0.118*	−0.005	−0.202***	−0.205***	−0.204***	−0.202***	−0.206***
	(0.011)	(0.060)	(0.061)	(0.060)	(0.062)	(0.061)	(0.010)	(0.050)	(0.050)	(0.050)	(0.051)	(0.051)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.258	0.220	0.218	0.218	0.216	0.217	0.258	0.220	0.064	0.063	0.061	0.063
Sample size	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

Fig. 6 Result: Analyzing the Relationship between Enterprise Value and Technological Diversification. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is excess value, defined as the log of the ratio of enterprise value to imputed value using the median segment multiplier in Model [1] to Model [12]. Column 1 shows the variable names. Multi or Not is a dummy, which takes the value 1 when firm *i* is a multi-business firm and 0 otherwise. The number is the number of industry segments to which firm *i* belong. 90% Distance and 95% Distance are the diversification indexes of the firm *i* created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of a firm *i* created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm *i*, and CAPEX/Sales is the ratio of CAPEX to Sales of firm *i*. EBITDA/Sales is the ratio of EBITDA to Sales of firm *i*. Log Asset is the log of the asset of firm *i*. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

the conglomerate discount may exist. This result is consistent with Berger and Ofek [4] and Ushijima [47]. The fact that the results in Model [1] and Model [2] are similar to those of previous studies suggests that the *Excessvalue* created by the GMM in Sect. 4.4 is practical.

Next, in Model [6], the diversification index's (Entropy) coefficient is significantly negative at the 5% level. The results suggest that the advance of technological diversification may negatively affect enterprise value and indicate the existence of diversification discounts in the fields of technology. Jacquemin and Berry [20] report that Entropy is a better diversification index, comparing Entropy with HHI. The fact that coefficient of diversification index (Entropy) is statistically significant, while the coefficient of diversification index (TDI) means the same suggestion as Jacquemin and Berry [20]. Finally, the coefficients of the diversification index (90% Distance) created from patent documents in Model [3] and Model [9] are significantly negative at the 5% or less level. In addition to this, the coefficients of the diversification index (95% Distance) created from patent documents in Model [4] and Model [10] are also significantly negative at the 5% or less level. The results of Model [3], Model [4], Model [6], Model [9], and Model [10] suggest that technological diversification can damage enterprise value. The fact that the diversification indexes (90% Distance and 95% Distance) created from the patents document show similar results to the diversification index (Entropy) created from the IPC indicates that the index created from unstructured data may be helpful. The analysis in this section shows that the advance of technological diversification potentially damages the enterprise value, and unstructured data such as patent documents may be helpful.

5.3 Analyzing the relationship in case of including additional variables

Campa and Kedia [7] analyze the relationship between firm diversification and enterprise value by adding some control variables to those used in Berger and Ofek's [4] test of the robustness of the conglomerate discount. We refer to Campa and Kedia [7] and add the control variables to examine the robustness. We include lagged Capex/Sales, EBITDA/Sales, log of total assets, and SLTA (Square of Log of Total Asset). The negative coefficient of SLTA means that the effect of firm size on excess value decreases as firm size increases.

Figures 7, 8 show the estimation results when we add control variables to the previous model discussed in Sect. 5.2. Figure 7 shows the results when the excess firm value based on sales is the dependent variable, and Figure 8 shows the results when the excess firm value based on total assets is the dependent variable. Compared to the results of the Model [12] in Figure 6 with the Model [6] in Figure 8, the coefficient of Entropy turns out to be statistically significantly negative at the 10% level due to the addition of the control variable. As a result, the coefficient of Entropy is statistically significantly negative when we measure the excess firm value based on sales or total assets. The addition of control variables brings robustness to our suggestion in Sect. 5.2 that advancing technological diversification reduces enterprise value. The results of all diversification indexes are mostly the same, excluding Model [6] in Figure 8. In Sects. 5.2 and 5.3, higher technological diversification has a possibility to damage the enterprise value, and this result is statistically robust, including other additional variables. We proceed with the analysis, including additional variables, based on Sect. 5.3.

5.4 Analyzing including the diversification index lagged by one period earlier

In 5.2 and 5.3, we indicated that higher technological diversification might damage enterprise value. However, it is also possible that firms with low enterprise value perform innovation activities for various fields. In short, an endogenous problem may occur. Therefore, in order to correct the endogeneity problem, we analyzed the relationship between the technological diversification index one period earlier, $Div_{i,t-1}$, and enterprise value, as shown in Eq. (13).

$$Excess\ value_{it} = \alpha + \beta_1 \cdot Div_{i,t-1} + \beta_2 \cdot X_{i,t} + FE_i + \epsilon_{i,t} \tag{13}$$

Figure 9 shows the results of Eq. (13). The coefficient of the diversification index one period earlier, 95% Distance, in Model [2] is − 0.656, which is significantly negative at the 10% level. The coefficients of the diversification indexes in the other models, except for Model [7], are negative, although not significant. The results of these analyses suggest that technological diversification one period ago may have a negative impact on enterprise value in the following period. Furthermore, it strengthens the implications of Sects. 5.2 and 5.3.

5.5 Comparing two kinds of diversification indexes based on patent data I

The results in Sects. 5.2 and 5.3 show that the diversification indexes (90% Distance and 95% Distance) created from patent documents are statistically significantly negative at a higher level than those created from IPC (TDI and Entropy). This result may lead to a more detailed analysis of the relationship between technological diversification and enterprise value since the indexes using unstructured data, such as patent data, may have more affluent information on technological diversification than those assigned under specific criteria. In this section, we examine the superiority of 90% Distance and 95% Distance compared to TDI and Entropy. Since the correlation coefficients of 90% Distance and TDI/Entropy are 0.403/0.468, and the correlation coefficients of 95% Distance and TDI/Entropy are 0.220/0.277, as shown in the correlation coefficient matrix in Figure 2, the correlation between the diversification indexes created from the patent documents and diversification indexes created from the IPC is not strong. We set the regression equation as shown in Eq. (14), in which we add the diversification index based on the patent document, 90% Distance or 95% Distance, to the independent variable, $Div①$, and the diversification index based on IPC, TDI or Entropy, to the independent variable, $Div②$.

$$Excess\ value_{it} = \alpha + \beta_1 \cdot Div①_{i,t} + \beta_2 \cdot Div②_{i,t} + \beta_3 \cdot X_{i,t} + FE_i + \epsilon_{i,t} \tag{14}$$

Figure 10 shows the estimation results of Eq. (14). While the coefficient of TDI is insignificant in Model [1] and Model [5], the coefficient of 90% Distance is significantly negative at the 5% level. Also, while the coefficients of TDI are not significant in the Model [2] and [6], the 95% Distance coefficient is significantly negative at the 1% and 10% levels, respectively. The estimation results of Model [4] using excess value measured by sales multiple as the dependent

	Excess Value					
	Sales Multiples					
	Random effect model [1]	Random effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]
Intercept	−1.989*	−1.953*				
	(1.078)	(1.071)				
Multi or Not	−0.043***					
	(0.014)					
Number		−0.039***				
		(0.010)				
90% Distance			−0.968**			
			(0.427)			
95% Distance				−1.057***		
				(0.346)		
TDI					−0.004	
					(0.005)	
Entropy						−0.082*
						(0.049)
Debt/Equity	0.0002	0.0001	0.001	0.001	0.001	0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Capex/Sales	0.261	0.264	0.090	0.087	0.085	0.085
	(0.304)	(0.304)	(0.294)	(0.293)	(0.294)	(0.294)
EBITDA/Sales	0.372*	0.369*	−0.035	−0.042	−0.037	−0.045
	(0.219)	(0.218)	(0.239)	(0.239)	(0.239)	(0.239)
Log Asset	0.504***	0.509***	0.164	0.184	0.215	0.223
	(0.184)	(0.183)	(0.466)	(0.469)	(0.469)	(0.460)
Capex/Sales	0.282	0.296	0.170	0.167	0.161	0.160
(1 lag)	(0.245)	(0.241)	(0.248)	(0.248)	(0.249)	(0.249)
EBITDA/Sales	−0.186	−0.188	−0.324	−0.324	−0.342	−0.340
(1 lag)	(0.216)	(0.216)	(0.214)	(0.213)	(0.212)	(0.212)
Log Asset	−0.454***	−0.459***	−0.429***	−0.430***	−0.428***	−0.426***
(1 lag)	(0.109)	(0.110)	(0.114)	(0.113)	(0.114)	(0.114)
Capex/Sales	−0.364	−0.371	−0.390	−0.382	−0.425	−0.419
(2 lag)	(0.265)	(0.266)	(0.266)	(0.266)	(0.264)	(0.263)
EBITDA/Sales	0.682***	0.673***	0.278*	0.276*	0.260	0.256
(2 lag)	(0.169)	(0.167)	(0.161)	(0.161)	(0.160)	(0.160)
Log Asset	0.228**	0.229**	0.102	0.103	0.101	0.098
(2 lag)	(0.097)	(0.097)	(0.099)	(0.099)	(0.101)	(0.101)
SLTA	−0.010	−0.010	0.001	0.001	−0.0004	−0.001
	(0.006)	(0.006)	(0.017)	(0.017)	(0.017)	(0.017)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.270	0.272	0.225	0.226	0.224	0.225
Sample size	2,338	2,338	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

◂**Fig. 7** Result: Analyzing the Relationship in case of including Additional Variables. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is excess value, defined as the log of the ratio of enterprise value to imputed value based on sales in Model [1] to Model [6]. Column 1 shows the variable names. Multi or Not is a dummy, which takes the value 1 when firm i is a multi-business firm and 0 otherwise. The number is the number of industry segments to which firm i belong. 90% Distance and 95% Distance are the diversification indexes of a firm i created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm i created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm i, and CAPEX/Sales is the ratio of CAPEX to Sales of firm i. EBITDA/Sales is the ratio of EBITDA to Sales of firm i. Log Asset is the log of total assets of firm i. SLTA is the square of the log of total assets of firm i. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level respectively

variable shows that the coefficient of 95% Distance is significantly negative at the 10% level, while the coefficient of Entropy is not significant. From these estimation results, it can be found that the diversification index using patent documents tends to show an advantage over the index created from the IPC.[6] Since IPC is an index based on the technology field, diversification indexes based on IPC can be interpreted as an index of diversification in technology. By considering diversification indexes (TDI and Entropy) using IPC code, it can be interpreted that the coefficients of diversification indexes (90% Distance and 95% Distance) based on the patent documents are statistically significantly negative, even when the technical field is considered. As the reason why diversification indices (90% Distance and 95% Distance) show superiority over the diversification indices (TDI and Entropy), we can consider that diversification indices created from the patent documents have information on technological diversification that diversification indices of IPC cannot capture. In other words, the result shows that using not index assigned under specific criteria but unstructured data such as patent documents may lead to a more precise analysis of the relationship between diversification and enterprise value.

5.6 Analyzing the relationship between PBR and technological diversification

As we argued in chapter 2, there is no consensus on the existence of the conglomerate discount and the relationship between technological diversification and enterprise value. In addition, it would be coincident that the coefficients of 90% Distance and 95% Distance showed beneficial output in the analysis in Sects. 5.2 and 5.3. In this section, we change the dependent variable from Excess Value used in Sects. 5.2 and 5.3 to PBR, price to book ratio, and analyze the relationship between the firm's diversification and the enterprise value to show the validity of the diversification index based on the patent document to previous analysis. PBR is a measure of market value divided by net assets. It indicates whether or not the stock market's expectations of a firm are rising. As shown in Figure 5, the difference between the average PBR of multi-business and single-segment firms is significantly negative at the 1% level, indicating that the average PBR of multi-business firms is lower than the average PBR of single-segment firms. It suggests that multi-business firms may be undervalued in the stock market compared to single-segment firms. Therefore, we can expect similar results in Sects. 5.2 and 5.3. For the above reasons, we set Eq. (15) to analyze the relationship between technological diversification and enterprise value. The dependent variable is PBR. Independent variables are the same as in Sect. 5.3.

$$PBR_{it} = \alpha + \beta_1 \cdot Div_{i,t} + \beta_2 \cdot X_{i,t} + FE_i + \epsilon_{i,t} \tag{15}$$

Figure 11 shows the estimation results of Eq. (15). The 90% Distance coefficient in Model [3] is negative and significant at the 10% level. The 95% Distance coefficient in Model [4] is negative and significant at the 10% level. These results suggest that the advance of technological diversification may have a negative impact on enterprise value. This implication in Sect. 5.6 is similar to the suggestions in Sects. 5.2 and 5.3. The analysis result in Sect. 5.6 implies that higher technological diversification may negatively impact enterprise value when the dependent variable changes. It suggests the robustness of the analysis results in Sects. 5.2 and 5.3.

5.7 Comparing two kinds of diversification indexes based on patent data II

As discussed in Sect. 5.5, using unstructured data such as patent documents lead to more precise analysis than index data based on specific criteria. The results in Sect. 5.4 suggest that diversification indexes based on patent documents have more information about technological diversification than indexes using the IPC index. In this section, we examine whether unstructured data has a possibility to be analyzed more precisely than the criteria given under

[6] The result is particularly significant for Model [1], [2] and [4]. Detailed analysis is a future issue.

	Excess Value					
	Asset Multiples					
	Fixed effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]
Intercept						
Multi or Not	−0.003					
	(0.010)					
Number		−0.003				
		(0.007)				
90% Distance			−0.743**			
			(0.297)			
95% Distance				−0.584**		
				(0.294)		
TDI					−0.001	
					(0.004)	
Entropy						−0.073*
						(0.043)
Debt/Equity	0.006***	0.006***	0.007***	0.007***	0.006***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Capex/Sales	−0.131	−0.131	−0.127	−0.130	−0.131	−0.131
	(0.212)	(0.212)	(0.210)	(0.210)	(0.212)	(0.211)
EBITDA/Sales	0.803***	0.803***	0.795***	0.793***	0.799***	0.785***
	(0.207)	(0.207)	(0.206)	(0.206)	(0.207)	(0.205)
Log Asset	0.203	0.204	0.198	0.209	0.214	0.249
	(0.359)	(0.359)	(0.354)	(0.357)	(0.359)	(0.345)
Capex/Sales	0.207	0.209	0.212	0.208	0.205	0.204
(1 lag)	(0.202)	(0.201)	(0.202)	(0.202)	(0.203)	(0.202)
EBITDA/Sales	−0.347**	−0.347**	−0.342**	−0.343**	−0.351**	−0.356**
(1 lag)	(0.165)	(0.165)	(0.164)	(0.163)	(0.166)	(0.166)
Log Asset	−0.128	−0.129	−0.127	−0.127	−0.127	−0.124
(1 lag)	(0.083)	(0.083)	(0.083)	(0.082)	(0.083)	(0.083)
Capex/Sales	−0.267	−0.268	−0.238	−0.241	−0.265	−0.260
(2 lag)	(0.222)	(0.223)	(0.223)	(0.224)	(0.221)	(0.220)
EBITDA/Sales	0.153	0.153	0.155	0.153	0.149	0.135
(2 lag)	(0.125)	(0.125)	(0.125)	(0.125)	(0.126)	(0.124)
Log Asset	−0.048	−0.048	−0.051	−0.050	−0.050	−0.055
(2 lag)	(0.079)	(0.079)	(0.079)	(0.079)	(0.079)	(0.080)
SLTA	−0.009	−0.009	−0.009	−0.010	−0.010	−0.011
	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.013)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.067	0.067	0.069	0.069	0.067	0.070
Sample size	2,338	2,338	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

◂**Fig. 8** Result: Analyzing the Relationship in case of including Additional Variables. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is excess value, defined as the log of the ratio of enterprise value to imputed value based on total assets in Model [1] to Model [6]. Column 1 shows the variable names. Multi or Not is a dummy, which takes the value 1 when firm *i* is a multi-business firm and 0 otherwise. The number is the number of industry segments to which firm *i* belong. 90% Distance and 95% Distance are the diversification indexes of the firm *i* created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm *i* created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm *i*, and CAPEX/Sales is the ratio of CAPEX to Sales of firm *i*. EBITDA/Sales is the ratio of EBITDA to Sales of firm *i*. Log Asset is the log of total assets of firm *i*. SLTA is the square of the log of total assets of firm *i*. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

specific criteria, even when the dependent variable changes from Excess Value to PBR, and the validity of the result of Sect. 5.5. The regression equation is shown in (16). We set 90% Distance or 95% Distance to the independent variable, *Div*①, and TDI or Entropy, to *Div*②.

$$PBR_{it} = \alpha + \beta_1 \cdot Div\textcircled{1}_{i,t} + \beta_2 \cdot Div\textcircled{2}_{i,t} + \beta_3 \cdot X_{i,t} + FE_i + \epsilon_{i,t} \tag{16}$$

Figure 12 shows the estimation results. The coefficient of TDI is insignificant, but 95% Distance coefficient is significantly negative at the 10% level in Model [2]. This result suggests that the 95% Distance may have information on technological diversification, which is not captured by the TDI based on IPC, even when the Excess Value is changed to the PBR.[7] The results in this section are similar to those in Sect. 5.5. They show that the statistical robustness of the possibility that unstructured data can be used more precisely than indicators given under specific criteria is high.

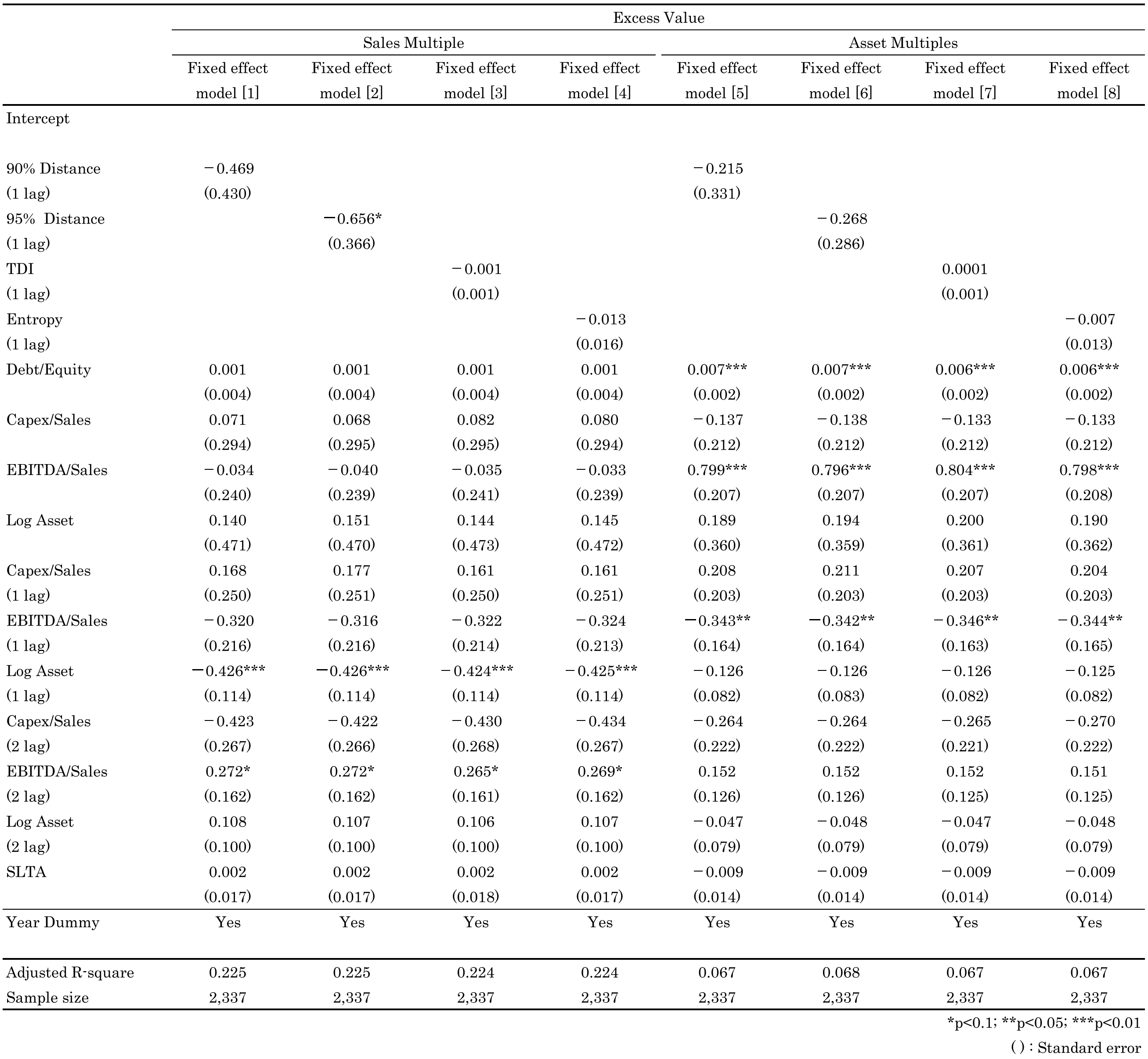
6 Conclusion

Thanks to the development of natural language processing technology, it is now possible to analyze texts efficiently. A notable field in which natural language processing technology is being leveraged is finance. In finance, the evaluation of non-financial information, such as human resources and technology possessed by each firm, is becoming active, as in the case of ESG investment. Patents, which are often used to evaluate a firm's technology, contain not only structured data such as IPC but also text and figures. Until now, value evaluation analysis has mainly been conducted using structured data such as IPC. However, through the development of natural language processing, efficient analysis using patent text has become possible, and more detailed value evaluation analysis than ever before is expected.

Diversification through technological expansion is an important decision for a firm. Now that consumer preferences, technological innovation, and economic change are changing at a rapid pace, firms need to constantly consider the configuration of their businesses and technologies and respond to these changes. Suppose the firm decides to diversify to respond to these topics. In that case, the discussion on technological diversification is vital because the portfolio's composition may positively or negatively impact the enterprise value.

This study analyzed the relationship between technological diversification and enterprise value. This topic has attracted attention in recent years and is vital in finance. In addition, we examined the usefulness of natural language processing technology in the analysis. Specifically, we obtained distributed representations of patent documents and generated technological diversification indexes. We analyzed the relationship between technological diversification and enterprise value. Our analysis indicated that an increase in the degree of technological diversification may damage enterprise value. The result strengthens the previous opinions that technological diversification has a negative

[7] We conduct the same analysis when we change the dependent variable to Tobin Q by referring to Lang and Stulz [25]. As a result, we get the same result that the diversification would damage the firm's enterprise value, and the distance from patent documents would have more information than the distance from the IPC of the patent.



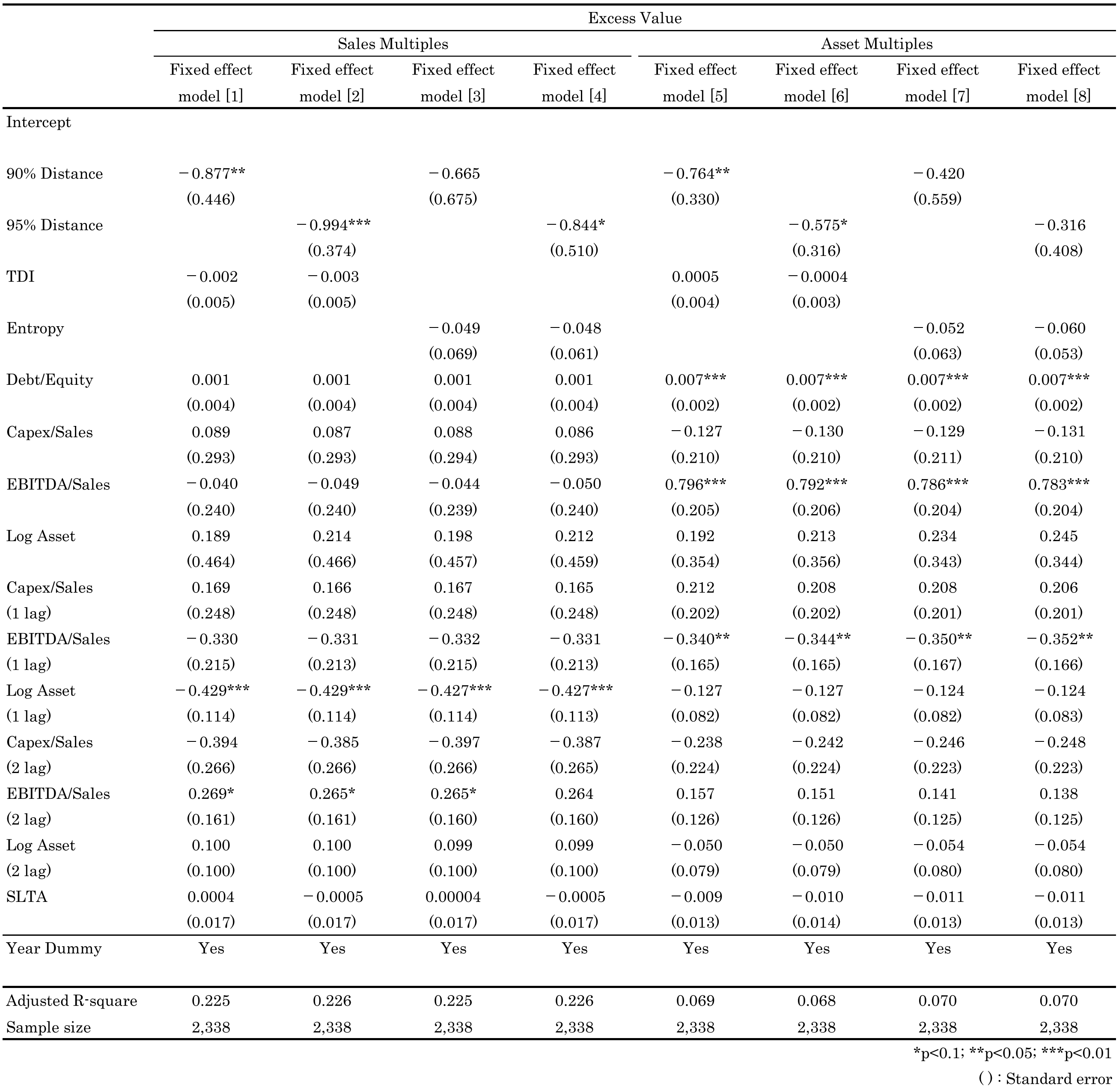
	Excess Value							
	Sales Multiple				Asset Multiples			
	Fixed effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]	Fixed effect model [7]	Fixed effect model [8]
Intercept								
90% Distance	−0.469				−0.215			
(1 lag)	(0.430)				(0.331)			
95% Distance		−0.656*				−0.268		
(1 lag)		(0.366)				(0.286)		
TDI			−0.001				0.0001	
(1 lag)			(0.001)				(0.001)	
Entropy				−0.013				−0.007
(1 lag)				(0.016)				(0.013)
Debt/Equity	0.001	0.001	0.001	0.001	0.007***	0.007***	0.006***	0.006***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
Capex/Sales	0.071	0.068	0.082	0.080	−0.137	−0.138	−0.133	−0.133
	(0.294)	(0.295)	(0.295)	(0.294)	(0.212)	(0.212)	(0.212)	(0.212)
EBITDA/Sales	−0.034	−0.040	−0.035	−0.033	0.799***	0.796***	0.804***	0.798***
	(0.240)	(0.239)	(0.241)	(0.239)	(0.207)	(0.207)	(0.207)	(0.208)
Log Asset	0.140	0.151	0.144	0.145	0.189	0.194	0.200	0.190
	(0.471)	(0.470)	(0.473)	(0.472)	(0.360)	(0.359)	(0.361)	(0.362)
Capex/Sales	0.168	0.177	0.161	0.161	0.208	0.211	0.207	0.204
(1 lag)	(0.250)	(0.251)	(0.250)	(0.251)	(0.203)	(0.203)	(0.203)	(0.203)
EBITDA/Sales	−0.320	−0.316	−0.322	−0.324	−0.343**	−0.342**	−0.346**	−0.344**
(1 lag)	(0.216)	(0.216)	(0.214)	(0.213)	(0.164)	(0.164)	(0.163)	(0.165)
Log Asset	−0.426***	−0.426***	−0.424***	−0.425***	−0.126	−0.126	−0.126	−0.125
(1 lag)	(0.114)	(0.114)	(0.114)	(0.114)	(0.082)	(0.083)	(0.082)	(0.082)
Capex/Sales	−0.423	−0.422	−0.430	−0.434	−0.264	−0.264	−0.265	−0.270
(2 lag)	(0.267)	(0.266)	(0.268)	(0.267)	(0.222)	(0.222)	(0.221)	(0.222)
EBITDA/Sales	0.272*	0.272*	0.265*	0.269*	0.152	0.152	0.152	0.151
(2 lag)	(0.162)	(0.162)	(0.161)	(0.162)	(0.126)	(0.126)	(0.125)	(0.125)
Log Asset	0.108	0.107	0.106	0.107	−0.047	−0.048	−0.047	−0.048
(2 lag)	(0.100)	(0.100)	(0.100)	(0.100)	(0.079)	(0.079)	(0.079)	(0.079)
SLTA	0.002	0.002	0.002	0.002	−0.009	−0.009	−0.009	−0.009
	(0.017)	(0.017)	(0.018)	(0.017)	(0.014)	(0.014)	(0.014)	(0.014)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.225	0.225	0.224	0.224	0.067	0.068	0.067	0.067
Sample size	2,337	2,337	2,337	2,337	2,337	2,337	2,337	2,337

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

Fig. 9 Result: Analyzing including the diversification index lagged by one period earlier. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is excess value, defined as the log of the ratio of enterprise value to imputed value using the median segment multiplier in Model [1] to Model [8]. Column 1 shows the variable names. 90% Distance and 95% Distance are the diversification indexes of the firm *i* created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm *i* created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm *i*, and CAPEX/Sales is the ratio of CAPEX to Sales of firm *i*. EBITDA/Sales is the ratio of EBITDA to Sales of firm *i*. Log Asset is the log of total assets of firm *i*. SLTA is the square of the log of total assets of firm *i*. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

impact on enterprise value. Then, in analyzing the relationship between technological diversification and enterprise value, we verified the usefulness of the diversification index generated from the text by comparing it with diversification indexes created from existing indexes that have been used in many studies, such as IPC. We showed that the use of diversification indexes created from text data has the potential to provide a more detailed analysis than the use of diversification indexes created from indicators assigned to specific



	Excess Value							
	Sales Multiples				Asset Multiples			
	Fixed effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]	Fixed effect model [7]	Fixed effect model [8]
Intercept								
90% Distance	−0.877**		−0.665		−0.764**		−0.420	
	(0.446)		(0.675)		(0.330)		(0.559)	
95% Distance		−0.994***		−0.844*		−0.575*		−0.316
		(0.374)		(0.510)		(0.316)		(0.408)
TDI	−0.002	−0.003			0.0005	−0.0004		
	(0.005)	(0.005)			(0.004)	(0.003)		
Entropy			−0.049	−0.048			−0.052	−0.060
			(0.069)	(0.061)			(0.063)	(0.053)
Debt/Equity	0.001	0.001	0.001	0.001	0.007***	0.007***	0.007***	0.007***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
Capex/Sales	0.089	0.087	0.088	0.086	−0.127	−0.130	−0.129	−0.131
	(0.293)	(0.293)	(0.294)	(0.293)	(0.210)	(0.210)	(0.211)	(0.210)
EBITDA/Sales	−0.040	−0.049	−0.044	−0.050	0.796***	0.792***	0.786***	0.783***
	(0.240)	(0.240)	(0.239)	(0.240)	(0.205)	(0.206)	(0.204)	(0.204)
Log Asset	0.189	0.214	0.198	0.212	0.192	0.213	0.234	0.245
	(0.464)	(0.466)	(0.457)	(0.459)	(0.354)	(0.356)	(0.343)	(0.344)
Capex/Sales	0.169	0.166	0.167	0.165	0.212	0.208	0.208	0.206
(1 lag)	(0.248)	(0.248)	(0.248)	(0.248)	(0.202)	(0.202)	(0.201)	(0.201)
EBITDA/Sales	−0.330	−0.331	−0.332	−0.331	−0.340**	−0.344**	−0.350**	−0.352**
(1 lag)	(0.215)	(0.213)	(0.215)	(0.213)	(0.165)	(0.165)	(0.167)	(0.166)
Log Asset	−0.429***	−0.429***	−0.427***	−0.427***	−0.127	−0.127	−0.124	−0.124
(1 lag)	(0.114)	(0.114)	(0.114)	(0.113)	(0.082)	(0.082)	(0.082)	(0.083)
Capex/Sales	−0.394	−0.385	−0.397	−0.387	−0.238	−0.242	−0.246	−0.248
(2 lag)	(0.266)	(0.266)	(0.266)	(0.265)	(0.224)	(0.224)	(0.223)	(0.223)
EBITDA/Sales	0.269*	0.265*	0.265*	0.264	0.157	0.151	0.141	0.138
(2 lag)	(0.161)	(0.161)	(0.160)	(0.160)	(0.126)	(0.126)	(0.125)	(0.125)
Log Asset	0.100	0.100	0.099	0.099	−0.050	−0.050	−0.054	−0.054
(2 lag)	(0.100)	(0.100)	(0.100)	(0.100)	(0.079)	(0.079)	(0.080)	(0.080)
SLTA	0.0004	−0.0005	0.00004	−0.0005	−0.009	−0.010	−0.011	−0.011
	(0.017)	(0.017)	(0.017)	(0.017)	(0.013)	(0.014)	(0.013)	(0.013)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.225	0.226	0.225	0.226	0.069	0.068	0.070	0.070
Sample size	2,338	2,338	2,338	2,338	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

Fig. 10 Result: Comparing two kinds of Diversification Indexes based on Patent Data I. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable in Model [1] to Model [4] is the excess value based on sales, and the dependent variable in Model [5] to Model [8] is the excess value based on total assets. Column 1 shows the variable names. 90% Distance and 95% Distance are the diversification indexes of the firm *i* created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm *i* created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm *i*, and CAPEX/Sales is the ratio of CAPEX to Sales of firm *i*. EBITDA/Sales is the ratio of EBITDA to Sales of firm *i*. Log Asset is the log of total assets of firm *i*. SLTA is the square of the log of total assets of firm *i*. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

	PBR (Market Capitalization / Net Asset)					
	Fixed effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]	Fixed effect model [5]	Fixed effect model [6]
Intercept						
Multi or Not	−0.010					
	(0.027)					
Number		−0.006				
		(0.021)				
90% Distance			−1.431*			
			(0.833)			
95% Distance				−1.509*		
				(0.773)		
TDI					−0.005	
					(0.010)	
Entropy						−0.111
						(0.122)
Debt/Equity	0.180***	0.180***	0.180***	0.180***	0.180***	0.180***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)
Capex/Sales	0.025	0.025	0.031	0.027	0.025	0.024
	(0.554)	(0.554)	(0.552)	(0.551)	(0.554)	(0.552)
EBITDA/Sales	3.844***	3.844***	3.828***	3.818***	3.829***	3.816***
	(1.243)	(1.244)	(1.240)	(1.240)	(1.242)	(1.238)
Log Asset	0.640	0.637	0.625	0.625	0.686	0.704
	(0.926)	(0.927)	(0.917)	(0.920)	(0.922)	(0.912)
Capex/Sales	0.948*	0.947*	0.952*	0.947*	0.938*	0.938*
(1 lag)	(0.527)	(0.527)	(0.529)	(0.528)	(0.527)	(0.526)
EBITDA/Sales	−0.190	−0.191	−0.181	−0.181	−0.205	−0.204
(1 lag)	(0.735)	(0.735)	(0.733)	(0.733)	(0.739)	(0.737)
Log Asset	−0.570**	−0.569**	−0.566**	−0.567**	−0.565**	−0.561**
(1 lag)	(0.247)	(0.247)	(0.246)	(0.246)	(0.246)	(0.245)
Capex/Sales	−1.230**	−1.229**	−1.173*	−1.163*	−1.224**	−1.216**
(2 lag)	(0.618)	(0.619)	(0.625)	(0.628)	(0.616)	(0.616)
EBITDA/Sales	0.863**	0.863**	0.867**	0.863**	0.844**	0.836**
(2 lag)	(0.381)	(0.381)	(0.380)	(0.380)	(0.390)	(0.387)
Log Asset	−0.296	−0.296	−0.302	−0.301	−0.303	−0.307
(2 lag)	(0.192)	(0.191)	(0.192)	(0.191)	(0.193)	(0.194)
SLTA	−0.030	−0.029	−0.029	−0.030	−0.031	−0.032
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.033)
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-square	0.427	0.427	0.428	0.428	0.427	0.428
Sample size	2,338	2,338	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

◂**Fig. 11** Result: Analyzing the Relationship between PBR and Technological Diversification. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is PBR in Model [1] to Model [6]. Column 1 shows the variable names. Multi or Not is a dummy, which takes the value 1 when firm i is a multi-business firm and 0 otherwise. The number is the number of industry segments to which firm i belong. 90% Distance and 95% Distance are the diversification indexes of the firm i created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm i created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm i, and CAPEX/Sales is the ratio of CAPEX to Sales of firm i. EBITDA/Sales is the ratio of EBITDA to Sales of firm i. Log Asset is the log of total assets of firm i. SLTA is the square of the log of total assets of firm i. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

criteria. It suggests that patent documents' textual information may contain more than indexes assigned to specific criteria.

The contribution of this paper is 3 points. The first is that our results support previous research on the relationship between technological diversification and enterprise value. We analyzed their relationship in Sects. 5.2, 5.3 and 5.6. The results suggest that increasing the degree of the firm's technological diversification may damage the enterprise value. It supports previous studies that technological diversification negatively impacts enterprise value and helps guide a consistent view of their relationship. Second, our study demonstrates the effectiveness of text data in finance. We compared the index created from patent texts with those created from IPC in Sects. 5.5 and 5.7. As a result, we showed that the use of text information, such as patent documents, has the potential to provide a more detailed analysis than previous studies that leverage indicators assigned based on specific criteria. It means that using unstructured data can be helpful in corporate finance. Third, by obtaining distributed representations of texts and their visualization, we have shown that new features, trends, and insights may be obtained. We obtained the distributed representation of the patent's abstract text data. By obtaining the distributed representation, we could understand the diversification of firms' R&D activities and trends, primarily through the visualization shown in Figure 3. The fact that we could gain deeper insights into what we had previously grasped through the acquisition of distributed representations or visualization would lead to the future development of research in finance.

There is still room in this analysis for creating variables and comparing between variables. In the present analysis, we created a diversification indicator, distance from the center of gravity, based on patent abstracts, which we used to analyze its relationship with enterprise value and to compare it with the diversification indicator created by the IPC. In addition to the IPC indicators and abstracts used in this analysis, patents include other information such as cited and uncited information, patent families, and application and publication dates. From such information we can also extract or create information about a firm's technological strategy, such as innovation indicators. Diversification indicators can also be created. Of course, we can also create diversification indicators. Creation and comparison of diversification indicators using this information, and analysis using these indicators as control variables in regression equations, is a subject for future work.

Another future issue is to confirm the robustness of the results of this analysis through a more detailed analysis. For example, the robustness of this analysis should be tested by increasing the number of firms and/or extending the period. In addition, we attempted to address the endogeneity issue in Sect. 5.4 by using the diversification index lagged by one period earlier. However, this treatment may not adequately address the endogeneity issue. Therefore, we need to check the robustness of the claims of this analysis more using generalized method of moments (GMM), the instrumental variables method [14], and propensity score matching.

	PBR (Market Capitalization / Net Asset)			
	Fixed effect model [1]	Fixed effect model [2]	Fixed effect model [3]	Fixed effect model [4]
Intercept				
90% Distance	−1.350		−1.079	
	(0.970)		(1.673)	
95% Distance		−1.441*		−1.240
		(0.841)		(1.106)
TDI	−0.002	−0.003		
	(0.010)	(0.010)		
Entropy			−0.057	−0.060
			(0.187)	(0.147)
Debt/Equity	0.180***	0.180***	0.180***	0.180***
	(0.034)	(0.034)	(0.034)	(0.034)
Capex/Sales	0.031	0.027	0.030	0.026
	(0.551)	(0.550)	(0.551)	(0.550)
EBITDA/Sales	3.823***	3.811***	3.818***	3.808***
	(1.238)	(1.237)	(1.237)	(1.236)
Log Asset	0.647	0.685	0.664	0.689
	(0.916)	(0.917)	(0.908)	(0.907)
Capex/Sales	0.951*	0.946*	0.948*	0.945*
(1 lag)	(0.529)	(0.528)	(0.530)	(0.528)
EBITDA/Sales	−0.187	−0.189	−0.190	−0.190
(1 lag)	(0.738)	(0.738)	(0.737)	(0.736)
Log Asset	−0.565**	−0.566**	−0.563**	−0.564**
(1 lag)	(0.246)	(0.246)	(0.244)	(0.245)
Capex/Sales	−1.176*	−1.166*	−1.181*	−1.169*
(2 lag)	(0.626)	(0.628)	(0.628)	(0.629)
EBITDA/Sales	0.859**	0.851**	0.852**	0.848**
(2 lag)	(0.391)	(0.391)	(0.391)	(0.389)
Log Asset	−0.304	−0.304	−0.306	−0.306
(2 lag)	(0.193)	(0.193)	(0.194)	(0.194)
SLTA	−0.030	−0.031	−0.031	−0.032
	(0.034)	(0.034)	(0.033)	(0.033)
Year Dummy	Yes	Yes	Yes	Yes
Adjusted R-square	0.428	0.428	0.428	0.428
Sample size	2,338	2,338	2,338	2,338

*p<0.1; **p<0.05; ***p<0.01

( ) : Standard error

◂**Fig. 12** Result: Comparing two kinds of Diversification Indexes based on Patent Data II. This figure shows the result supported by the Hausman test after panel analysis. The term runs from 2002 to 2015. The dependent variable is PBR in Model [1] to Model [4]. Column 1 shows the variable names. 90% Distance and 95% Distance are the diversification indexes of the firm i created by the text information of the patent in Sects. 4.1 and 4.2. TDI and Entropy are the diversification indexes of the firm i created by each patent's International Patent Classification (IPC) code in Sect. 4.2. Debt/Equity is the ratio of Debt to Equity of firm i, and CAPEX/Sales is the ratio of CAPEX to Sales of firm i. EBITDA/Sales is the ratio of EBITDA to Sales of firm i. Log Asset is the log of total assets of firm i. SLTA is the square of the log of total assets of firm i. The figure in parentheses means the standard deviation adjusted by Cluster Robust Standard Error. We show Adjusted R-square and sample size in each regression at the bottom. ***, ** and * denote significance at the 1%, 5% and 10% level, respectively

Acknowledgements This research was supported by Telecommunications Advancement Foundation, JSPS KAKENHI Grant No. JP20K01751and Keio University Academic Development Funds.

Declarations

Competing interests The authors have declared that they do not have any conflict of interest.

References

1. Barney JB (2001) Resource-based theories of competitive advantage: a ten-year retrospective on the resource-based view. J Manag 27(6):643–650
2. Bellstam G, Bhagat S, Cookson JA (2021) A text-based analysis of corporate innovation. Manage Sci 67(7):4004–4031
3. Bergek A, Berggren C, Tell F (2009) Do technology strategies matter? A comparison of two electrical engineering corporations, 1988–1998. Technol Anal Strateg Manag 21(4):445–470
4. Berger PG, Ofek E (1995) Diversification's effect on firm value. J Financ Econ 37(1):39–65
5. Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning, vol 4. Springer, New York, p 738
6. Brealey RA, Myers SC, Allen F, Mohanty P (2012) Principles of corporate finance. Tata McGraw-Hill Education, New York
7. Campa JM, Kedia S (2002) Explaining the diversification discount. J Financ 57(4):1731–1762
8. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B (Methodol) 39(1):1–22
9. Fai F (2003) Corporate technological competence and the evolution of technological diversification. In Corporate Technological Competence and the Evolution of Technological Diversification. Edward Elgar Publishing, Cheltenham
10. Fan F, Li B, Yang Y (2017) Study of the impact of TMT characteristics on the technology diversification and performance relationship of high-tech enterprise. In: The 2017 international conference on grey systems and intelligent services (GSIS). IEEE, pp 393–399
11. Fukui Y, Ushijima T (2007) Corporate diversification, performance, and restructuring in the largest Japanese manufacturers. J Jpn Int Econ 21(3):303–323
12. Granstrand O, Oskarsson C (1994) Technology diversification in" MUL-TECH" corporations. IEEE Trans Eng Manage 41(4):355–364
13. Gupta V, Saw A, Nokhiz P, Gupta H, Talukdar P (2019) Improving document classification with multi-sense embeddings. arXiv preprint arXiv:1911.07918
14. Hayashi F (2011) Econometrics. Princeton University Press, Princeton
15. Hoberg G, Phillips G (2010) Product market synergies and competition in mergers and acquisitions: a text-based analysis. The Review of Financial Studies 23(10):3773–3811
16. Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. J Polit Econ 124(5):1423–1465
17. Hoechle D, Schmid M, Walter I, Yermack D (2012) How much of the diversification discount can be explained by poor corporate governance? J Financ Econ 103(1):41–60
18. Iwaki Y (2017) Patent portfolio and firm value: an empirical study on the technological diversification and firm performance. Bus Account Rev 19:61–76
19. Iwaki Y, Okada K (2018) Experiment analysis between the technological breadth firm overs and firm's performance. Japan Finance Association
20. Jacquemin AP, Berry CH (1979) Entropy measure of diversification and corporate growth. J Ind Econ 27(4):359–369
21. Jaffe AB (1986) Technological opportunity and spillovers of R&D: evidence from firms' patents, profits, and market value
22. Kim J, Lee CY, Cho Y (2016) Technological diversification, core-technology competence, and firm growth. Res Policy 45(1):113–124
23. Kimura F (2009) A comparison of reliability of industrial classifications in Japan. Contemp Discl Res 9:33–42
24. Laeven L, Levine R (2007) Is there a diversification discount in financial conglomerates? J Financ Econ 85(2):331–367
25. Lang LH, Stulz RM (1994) Tobin's q, corporate diversification, and firm performance. J Polit Econ 102(6):1248–1280
26. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning. PMLR, pp 1188–1196
27. Lerner J (1994) The importance of patent scope: an empirical analysis. RAND J Econ 25(2):319–333
28. Leten B, Belderbos R, Van Looy B (2007) Technological diversification, coherence, and performance of firms. J Prod Innov Manag 24(6):567–579
29. Li K, Mai F, Shen R, Yan X (2021) Measuring corporate culture using machine learning. Rev Financ Stud 34(7):3265–3315
30. Lin C, Chang CC (2015) The effect of technological diversification on organizational performance: an empirical study of S&P 500 manufacturing firms. Technol Forecast Soc Chang 90:575–586
31. Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Financ 66(1):35–65
32. Matsumoto Y, Suge A, Takahashi H (2018) Capturing corporate attributes in a new perspective through fuzzy clustering. In: JSAI international symposium on artificial intelligence. Springer, Cham, pp 19–33
33. Mekala D, Gupta V, Paranjape B, Karnick H (2016) SCDV: sparse composite document vectors using soft clustering over distributional representations. arXiv preprint arXiv:1612.06778
34. Miller DJ (2004) Firms' technological resources and the performance effects of diversification: a longitudinal study. Strateg Manag J 25(11):1097–1119
35. Miller DJ (2006) Technological diversity, related diversification, and firm performance. Strateg Manag J 27(7):601–619

36. Miyazawa T (2017) Diversification and technological proximity of R&D investment. Tokyo Seitoku Univers J Bus Adm 6:1–23
37. Nishi Y, Suge A, Takahashi H (2021) Construction of a news article evaluation model utilizing high-frequency data and a large-scale language generation model. SN Bus Econ 1:104
38. Oikawa K, Takahashi H (2019) Innovation and technological locations of firms: an agent-based approach. In: Eastern economic association 45th annual conference
39. Palepu K (1985) Diversification strategy, profit performance, and the entropy measure. Strateg Manag J 6(3):239–255
40. Pugliese E, Napolitano L, Zaccaria A, Pietronero L (2019) Coherent diversification in corporate technological portfolios. PLoS ONE 14(10):e0223403
41. Rajan R, Servaes H, Zingales L (2000) The cost of diversity: The diversification discount and inefficient investment. J Financ 55(1):35–80
42. Scharfstein DS, Stein JC (2000) The dark side of internal capital markets: divisional rent-seeking and inefficient investment. J Financ 55(6):2537–2564
43. Silverman BS (1999) Technological resources and the direction of corporate diversification: toward an integration of the resource-based view and transaction cost economics. Manage Sci 45(8):1109–1124
44. Stein JC (2003) Agency, information and corporate investment. Handb Econ Finance 1:111–165
45. Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: quantifying language to measure firms' fundamentals. J Financ 63(3):1437–1467
46. Tirole J (2010) The theory of corporate finance. Princeton University Press, Princeton
47. Ushijima T (2015) Diversification discount and corporate governance. Policy Research Institute, Ministry of Finance, Japan, "Financial Review" (121), pp 69–90
48. Watanabe C, Matsumoto K, Hur JY (2004) Technological diversification and assimilation of spillover technology: Canon's scenario for sustainable growth. Technol Forecast Soc Chang 71(9):941–959
49. Weiner C (2005) The impact of industry classification schemes on financial research
50. Yamaguchi T (2009) Diversification of R & D investment and profitability. J Sci Policy Res Manag 24(1):89–100
51. Yong LS, Ingham H (2013) A latent class cluster analysis study of financial ratios and industry characteristics. Aust J Basic Appl Sci 7(11):46–53
52. Zabala-Iturriagagoitia JM, Gómez IP, Larracoechea UA (2020) Technological diversification: a matter of related or unrelated varieties? Technol Forecast Soc Chang 155:119997

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.