

→ useful visualizations & methods for TDA  
→ can draw conclusion re word count & sentiment analysis

## RESEARCH

## Open Access



# About relationship between business text patterns and financial performance in corporate data

BangRae Lee<sup>1</sup>, Jun-Hwan Park<sup>1</sup>, Leenam Kwon<sup>1</sup>, Young-Ho Moon<sup>1</sup>, YoungHo Shin<sup>1</sup>, GyuSeok Kim<sup>1</sup>  
and Han-joon Kim<sup>2\*</sup>

\* Correspondence: khj@uos.ac.kr

<sup>2</sup>The Graduate School of the University of Seoul, Seoul, Republic of Korea

Full list of author information is available at the end of the article

## Abstract

This study uses text and data mining to investigate the relationship between the text patterns of annual reports published by US listed companies and sales performance. Taking previous research a step further, although annual reports show only past and present financial information, analyzing text content can identify sentences or patterns that indicate the future business performance of a company. First, we examine the relation pattern between business risk factors and current business performance. For this purpose, we select companies belonging to two categories of US SIC (Standard Industry Classification) in the IT sector, 7370 and 7373, which include Twitter, Facebook, Google, Yahoo, etc. We manually collect sales and business risk information for a total of 54 companies that submitted an annual report (Form 10-K) for the last three years in these two categories. To establish a correlation between patterns of text and sales performance, four hypotheses were set and tested. To verify the hypotheses, statistical analysis of sales, statistical analysis of text sentences, sentiment analysis of sentences, clustering, dendrogram visualization, keyword extraction, and word-cloud visualization techniques are used. The results show that text length has some correlation with sales performance, and that patterns of frequently appearing words are correlated with the sales performance. However, a sentiment analysis indicates that the positive or negative tone of a report is not related to sales performance.

**Keywords:** Corporate annual report, 10-k, Text mining, Business keyword, Financial performance, Keyword trends, Word cloud, Sentiment analysis, Correlation coefficient, Hierarchical clustering, Dendrogram

## Introduction

Korea's small and medium-sized enterprises' (SMEs) global competitiveness is relatively weak due to a national economic structure favoring larger corporations. To strengthen the competitiveness of SMEs, policy discourages large companies from entering industries that are appropriate for SMEs. In most industries, large companies and SMEs collaborate and compete at the same time; however there may be industries to which SMEs are more suited. Park et al. studied the selection of industries suitable for SMEs in Korea (Park et al., 2016). However, for SMEs to be able to compete internationally, they need to be competitive with global companies. To strengthen their competitiveness, SMEs that are not rich in human resources and financial status must be

innovative. Open innovation can be a way of achieving this. Open innovation benefits from both internal and external knowledge. Research on innovation, or open innovation, has been actively pursued within academia. Witt has studied innovation that is necessary for our future (Witt, 2016). Yun et al. have studied dynamics that lead from open innovation to evolutionary change (Yun et al., 2016). Svirina, et al. conducted a study applying the concept of open innovation to social businesses (Svirina et al., 2016). By utilizing both internal knowledge and external knowledge, open innovation can strengthen SME competitiveness. This study analyzes the annual reports of US listed companies, with the aim of helping Korean SMEs absorb foreign knowledge and open innovation.

Bibliometric analysis studies using patent or paper data in the field of technology intelligence research are being studied for the purpose of discovering new opportunities from the technical planning point of view. In bibliometric research, the areas where many papers and patents are produced or cited are defined as promising technologies or areas. According to Jeong et al., patent analysis can identify new free technology and utilize it as a seed technology for SMEs and as a means of growth engine (Jeong et al., 2014). In another study, Jeong et al. extracted future promising research areas by using bibliographic coupling and clustering technique for the top 1% of cited papers in SCOPUS data (Jeong et al., 2008). These techniques can be used as a way to shorten the effort and time of R & D and technology development in countries with limited resources.

On the other hand, the analysis of the correlation between the business text and the financial analysis, which is the subject of this study, is interested in exploring the characteristics of the business documents of the companies with good financial performance and exploring the companies and the areas with the good financial performance. As a result, the bibliometric technique for papers and patents focuses on the area of people's interest. On the other hand, this study is interested in finding companies or business area that is expected to perform well in the future in terms of financial or operating performance.

The purpose of this study is to find out whether there is a correlation between business sales patterns and business text patterns. This is based on the research that other researchers in the past have found that estimating future financial performance with a time-series pattern of financial performance is limited and business texts better represent future prospects (Shirata et al. 2011). In this paper, we use text mining techniques to create business text patterns and explore how they relate to corporate financial performance. In other words, if the financial performance of a company is good, we want to find out what patterns appear in the business text.

The study is a field of business intelligence that predicts the future financial performance of a company through business text analysis. In this field, robots that predict future stock prices through various data analysis have recently appeared, but there is not much research in academic aspect. This may be because new areas such as Big Data, as mentioned by Celia, SatikoIshikiriyama et al. (2015), have attracted much more research interest among researchers. In addition, traditional financial information researchers seem to be lacking knowledge about methodology that utilizes text mining techniques to forecast company financial performance.

This field is a fusion field where text mining knowledge and financial knowledge are combined. However, as the interest of academia is concentrated elsewhere, it is considered that there is not much research in this field and it is sporadic. Although this study does not comprehensively deal with text mining and sales performance, it is expected to be a good alternative to predict future business performance by analyzing the correlation between business text pattern and business performance.

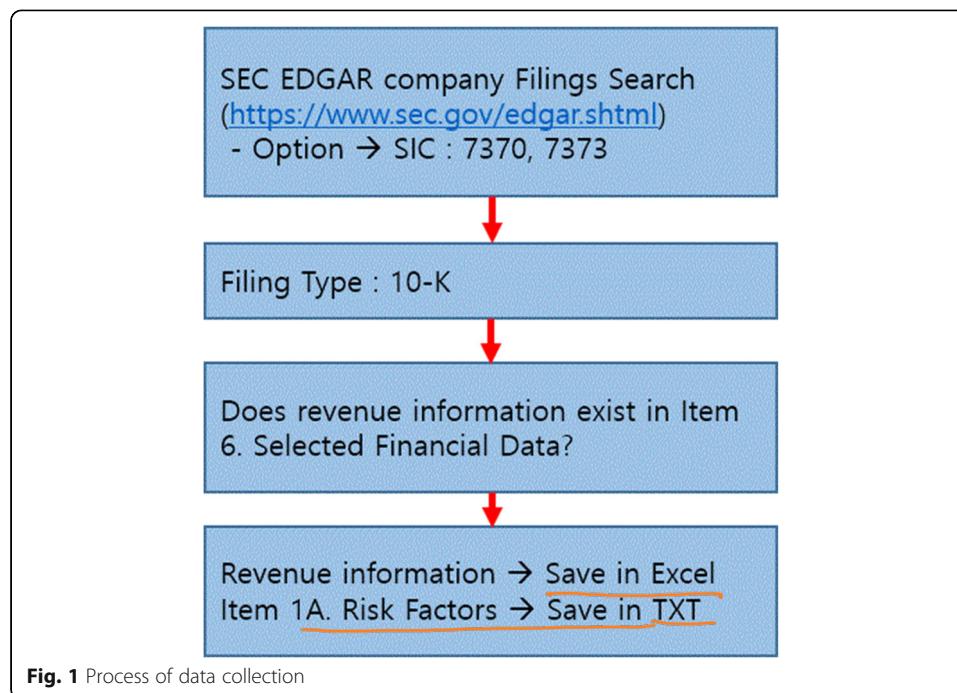
The annual reports of US listed companies is publicly available in 10-K format. These reports can contain information that illustrate the companies' global economic conditions. Furthermore, these reports contain not only the financial information of a company but also the main business content, competitiveness and risk factors within each industry. This basic information is useful for grasping global industry trends. Therefore, in addition to traditional financial analysis, text analysis on annual reports has been attempted using text-mining techniques. Lee et al. attempted to analyze a company's business model by applying text mining to an annual report (Lee et al., 2014). The study analyzed the report's business descriptions and future strategies. Specifically, business keywords were extracted, and the average frequency and growth rate of each keyword were calculated. The position of the business keywords in the "business model evolution map" based on these two values was shown. Additionally, it is thought that annual reports have richer information in the text than in the numerical values. Consequently, there have been attempts to apply text-mining techniques to annual reports (Kloptchenko et al., 2002). Kloptchenko et al. study the implications for future financial performance that can be drawn from the textual portion of quarterly reports using self-organizing maps and text clustering techniques. Self-organizing maps and text clustering were used to analyze financial quantitative information and qualitative data analysis respectively. In the text-clustering process, word histograms were created from the text documents and made into distribution functions. Subsequently, Euclidean distances between histograms were obtained to measure the similarity between reports. A comparison of the quantitative result of financial information and the qualitative results of text showed inconsistencies. Thus, they argued that the style of writing changes before a dramatic change occurs in financial performance. For example, if the firm's position is expected to deteriorate in the next quarter, the current quarter's report will be more pessimistic, even though the financial performance is, as yet, unchanged. Another study aimed to predict the possibility of a firm defaulting, by applying text-mining techniques to annual reports (Shirata et al. 2011). Shirata et al. used textual information from Japanese companies' annual reports to establish key phrases that could be used to predict corporate default. Prediction of default is one of the most important research topics among accounting researchers, with most of them using financial ratios. However, a change in the business performance of a company is actually articulated earlier in the nonfinancial information. Their study argued that "dividend" or "retained earnings" appearing in the same paragraph is effective for distinguishing between a defaulting company and a non-defaulting company. Lee et al. conducted a text- mining technique on 10-K reports to extract and visualize industrial service portfolios (Lee et al., 2016). In this study,

they used a self-organizing map (SOM) technique to visualize the service portfolio and check word- usage patterns. Lee. et al.'s study predicted stock price by analyzing financial events of 8-K documents (Lee et al., 2014). In this study, the accuracy of the stock price prediction for the following day is improved by 10% when the text is considered. Of course, they emphasized that the analysis should not be considered a comprehensive trading strategy, but stressed that text analysis has predictive power in terms of stock price volatility. SatokoIshikiriyama et al., 2015 attempted to ascertain topics of interest in business intelligence using sample analysis on the top 35 read papers in the field. Text-mining techniques determining the frequency of words, or word clouds, were created to identify the main topic of discussion in the paper. The research showed a decreasing interest in business intelligence in academia, suggesting that new areas, such as Big Data, may have received more attention. Pulliza's study used sentiment analysis in the modeling of speculation in 10-K documents (Pulliza, 2015). This model was applied to the MPQA corpus, to extract features with high correlation with speculative sentences in the 10-K document set. As part of the results, they argued, that "regulation, fund, and supplier were ranked higher than in the documents with the highest amount of speculative sentences".

The studies reviewed indicate that applying text-mining techniques to annual reports is more effective for predicting the future value of business or business trends than analyzing financial information alone. The methods are diverse: 'Term frequency-inverse document frequency' (TF-IDF) is calculated to extract important words from text, followed by a clustering method; a SOM or word cloud can be produced as a visualization technique; and techniques such as sentiment analysis were also used. Sentiment analysis is often referred to as opinion mining, which is generally defined as being "aimed at determining the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event." (Wikipedia 2018). In this study, we apply various text or data-mining techniques to text referring to risk factors and sales information in the annual reports of US listed companies. Sentence and word counts, sentiment analysis, keyword extraction and clustering are applied. In addition, we examine the correlation between these text patterns and sales.

## Data

This study uses the annual reports of US listed companies (10-K) retrieved from the EDGAR (2018) operated by the SEC (U.S. Securities and Exchange Commission). In the EDGAR's company filings search, we use the Standard Industry Classification to search the list of companies particular to that industry [SEC EDGAR website] (Fig. 1). This study analyzes the annual reports of the companies belonging to two categories of US SIC (Standard Industry Classification) in the IT sector: 7370 and 7373 (Table 1). This provides a list of all companies that have published annual reports, in the past and present, in these categories. Specifically, it provides company names, state/country identification codes, and company identification numbers and central index keys (CIK), which are linked to a company's filing report. However, the list also includes those companies that have



published reports but no longer exist, or companies that have diminished in size and are consequently no longer obliged to submit reports. In this study, among those companies that submitted a 10-K report after 2015, only those companies that provide information on sales for the last three consecutive years were analyzed. While many companies that submitted 10-K reports provide sales information for the last five years, some only provide information for the last three. Additionally, some smaller companies withhold sales information, specifically sales in “Item 6. Selected Financial Data”, on the grounds of their size. Most of the sales information is shown as “revenue”, but is sometimes marked as “sales”. We do not distinguish between the two notations. Companies that did not provide sales information because of their small size used the following expressions: “As a smaller reporting company, we are not required to provide the information in response to this Item.” This study excludes these firms from the analysis. As a result, we analyze 54 companies, accounting for 7.5% of 717 companies within categories 7370 and 7373, broken down further in Table 2. We use the entire text of “Item 1A. Risk Factors” and the revenue information of “Item 6 Selected Financial Data” (both highlighted in green) for the analysis.

**Table 1** Description of SIC codes used in our analysis

SIC code	Description	Representative companies	Number of Companies disclosed	Number of companies analyzed
7370	Services-Computer Programming, Data Processing, etc.	Facebook, Twitter, Google(Alphabet), LinkedIn	267	26
7373	Services-Computer Integrated Systems Design	Yahoo	450	28

**Table 2** Table of contents of annual report

Section	Description
<b>PART I</b>	
Item 1.	Business
Item 1A.	Risk Factors
Item 1B.	Unresolved Staff Comments
Item 2.	Properties
Item 3.	Legal Proceedings
Item 4.	Mine Safety Disclosures
<b>PART II</b>	
Item 5.	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
Item 6.	Selected Financial Data
Item 7.	Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 7A.	Quantitative and Qualitative Disclosures About Market Risk
Item 8.	Financial Statements and Supplementary Data
Item 9.	Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A.	Controls and Procedures
Item 9B.	Other Information
<b>PART III.</b>	
Item 10.	Directors, Executive Officers and Corporate Governance
Item 11.	Executive Compensation
Item 12.	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
Item 13.	Certain Relationships and Related Transactions, and Director Independence
Item 14.	Principal Accountant Fees and Services
<b>PART IV.</b>	
Item 15.	Exhibits and Financial Statement Schedules

Source: SEC EDGAR website: <https://www.sec.gov/edgar/searchedgar/companysearch.html>

## Methodology

This study uses text and data-mining on items describing risk factors in annual reports to investigate the relationship between the item's wording and the Compound Annual Growth Rate (CAGR) of revenue in the last three to five years. This study establishes several hypotheses and examines the validity of the hypothesis through data analysis to examine the aforementioned relationships. The hypotheses used in this experiment are shown in Table 3.

Hypothesis 1 can be verified by simply comparing two CAGR values of sales (Fig. 2).

Hypothesis 2 is to verify whether there is a correlation between sales performance and text length of 'Risk Factors' of 10-K report. The text length was defined as the number of sentences, the number of words, and the number of words per sentence. This is a direct way to see how each of the three variables representing text length correlates with sales performance. In addition, we examined whether the clustered firms are correlated with the sales performance by clustering firms with three variables such as the number of sentences, the number of words, and

**Table 3** Hypotheses of our study

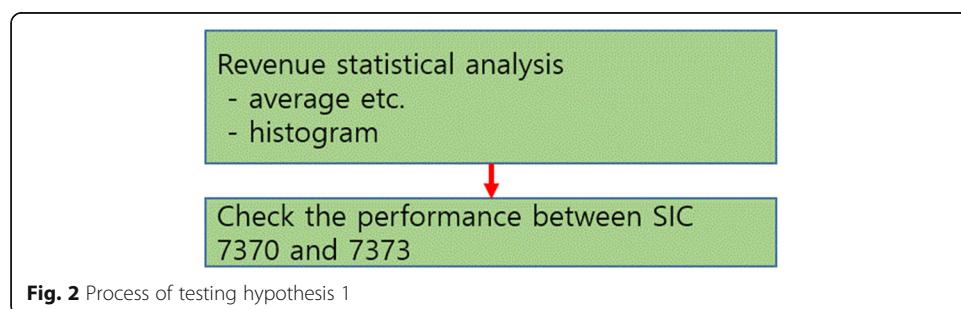
No.	Hypotheses
1	Category 7370 containing Google, Facebook, Twitter, etc. is better than category 7373 containing Yahoo in revenue performance.
2	(2-1) Companies with low revenue performance tend to write a shorter "Risk Factors" item section, to skip over risks. (2-2) Companies with high revenue performance tend to write a shorter "Risk Factors" item section because they think there are fewer risks in their business.
3	(3-1) In item "Risk Factors", a positive/negative tone correlates with sales performance (3-2) Using the positive/negative text analysis results in the "Risk Factors" item section, we can group companies by sales performance.
4	The occurrence patterns of words in "Risk Factors" are correlated with sales performance.

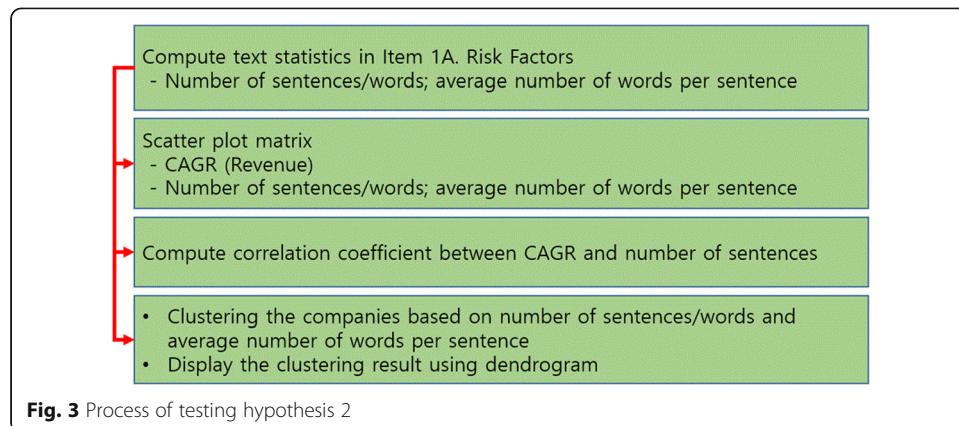
! of interest →

the number of words per sentence. That is, it is a method to check whether the clustering result considering the three variables representing the text length is correlated with the sales performance.

To test hypothesis 2, we calculate the number of sentences, words and words per sentence in text regarding "Risk Factors" and compare it with the CAGR values. After obtaining the four values, plotting a 'scatter plot' can be used to ascertain the likelihood of a correlation. Subsequently, in the indication of a correlation, the correlation coefficient can be calculated. In addition, using the statistical values from the texts, we use clustering to check whether clusters are formed among companies with similar sales results (Fig. 3).

Hypothesis 3 is a method of examining whether the positive or negative tone in the business text correlates with sales performance. The technique of analyzing the tone of affirmation or negation of text is called 'sentiment analysis'. According to Taboada et al., this is defined as follows: "Sentiment analysis refers to the general method to extract subjectivity and polarity from text (potentially also speech)" (Taboada et al., 2011). In sentiment analysis, for example, if the sentence contains a positive expression such as 'good', the tone of the sentence is evaluated as a positive sentence, and a sentence containing an expression such as 'bad' is determined as a negative sentence. In this study, we examine whether the number of positive statements or negative statements correlates with the sales performance of individual companies. For example, if companies with a large number of affirmative statements have good sales performance, companies with a lot of positive sentences will be expected to have good sales performance in the future.

**Fig. 2** Process of testing hypothesis 1

**Fig. 3** Process of testing hypothesis 2

To test hypothesis 3, we conduct a 'sentiment analysis' on each sentence in the "Risk Factors" item. Sentiment analysis, or opinion mining, is a method for determining whether a sentence's tone is positive or negative. We use the 'RSentiment' package in R [Package 'RSentiment manual']. Using the 'calculate\_total\_presence\_-sentiment' function in this package, the entire sentence is classified as positive, negative, very positive, very negative, neutral, or sarcasm. The number of sentences in each of these six categories is calculated from each company's "Risk Factors" item. A scatter plot is used to determine the existence of a correlation between the CAGR values and the sentiment analysis. In addition, we apply a clustering technique to the sentiment analysis result to determine whether the companies are grouped by sales performance.

Hypothesis 4 assumes that the words in the business text will appear differently depending on sales performance. In other words, it is assumed that words that have a positive meaning in a group with good sales performance are shown, whereas in the opposite case, a lot of words with a negative meaning are expected to appear. However, this method differs from the sentiment analysis used in the hypothesis 3. The sentiment analysis analyzes the tone of the sentence, while the analysis focuses on what words are used directly. For example, a company with a good sales performance is expected to talk more about future prospects and plans, while a company with a poor sales performance is expected to talk more about immediate sales and risk reduction. Based on these assumptions, we check whether word patterns are correlated with sales performance. In the meantime, we extracted words using TF-IDF, which is widely used in the field of information retrieval, in order to remove unnecessary words and select good key-words that reflect the core of contents.

To test hypothesis 4, texts from the highest and lowest three ranking companies, based on sales' annual growth rate, are grouped into two files, respectively, and the two text files subsequently compared. Prior to comparison, textual refinement is performed using natural language processing techniques. Refining removes whitespace, punctuation, numbers, and stop-words, and changes uppercase letters to lower case. After that, term frequency-inverse document frequency (TF-IDF) is calculated to select meaningful words, which are then used to create a term-document matrix (TDM). Note that in this process, the two groups of

text files are placed in corpus, and the unit of ‘document’ in TF-IDF calculation is a paragraph. In other words, since, from the viewpoint of CAGR of sales, the upper group is represented as one document and the lower group is also represented as one document, in this study, the unit of ‘document’ was set to the paragraph in the text instead of the text per company for TF-IDF calculation. R language was used for data analysis. Specifically, after loading text with ‘readLines’ function of ‘tm’ package, we made it into a corpus by using ‘VectorSource’ function. After that, the TF-IDF is calculated and the TF-IDF values are added for each term and finally the top 200 words with high total values were selected. 200 words are selected for each of the upper and lower groups, so if the word sets in each group are completely different, 400 are selected. We checked the characteristics of the selected words that do not overlap between the two matrices. In addition, a word cloud is created using the TF-IDFs of the words in the two text files (Fig. 4).

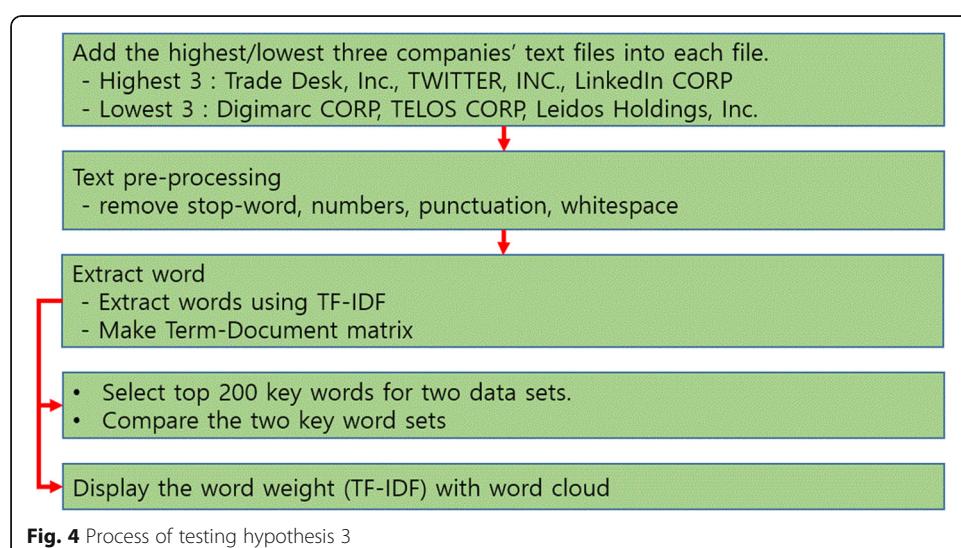
## Results

### Comparison of CAGR values

This analysis verifies hypothesis 1. According to hypothesis 1, the CAGR value of category 7370, which includes many companies dealing with SNS (social network service) data, should be higher in the IT sector. The CAGRs are calculated using the last 3, 4 or 5 years of sales, and the statistical values are calculated for each industrial classification. The results, shown in Table 4, confirm that the hypothesis is appropriate. Figure 5 shows the histogram of the distribution of firms in two groups.

### Relation between text statistics and CAGR

The text in “Item 1A. The Risk Factors” is a part of a company’s disclosure of current and future risk factors. Publishing such risk factors in an annual report can be



**Table 4** CAGR statistics by SIC code

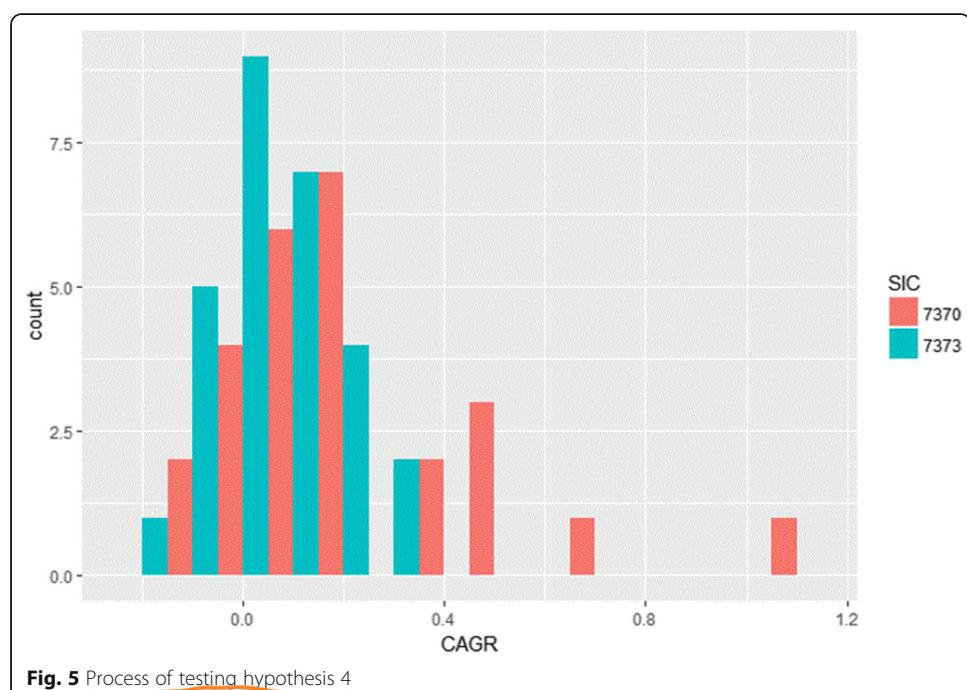
SIC code	average	max	min
7370	22.5%	113.4%	-9.2%
7373	5.0%	32.7%	-16.3%

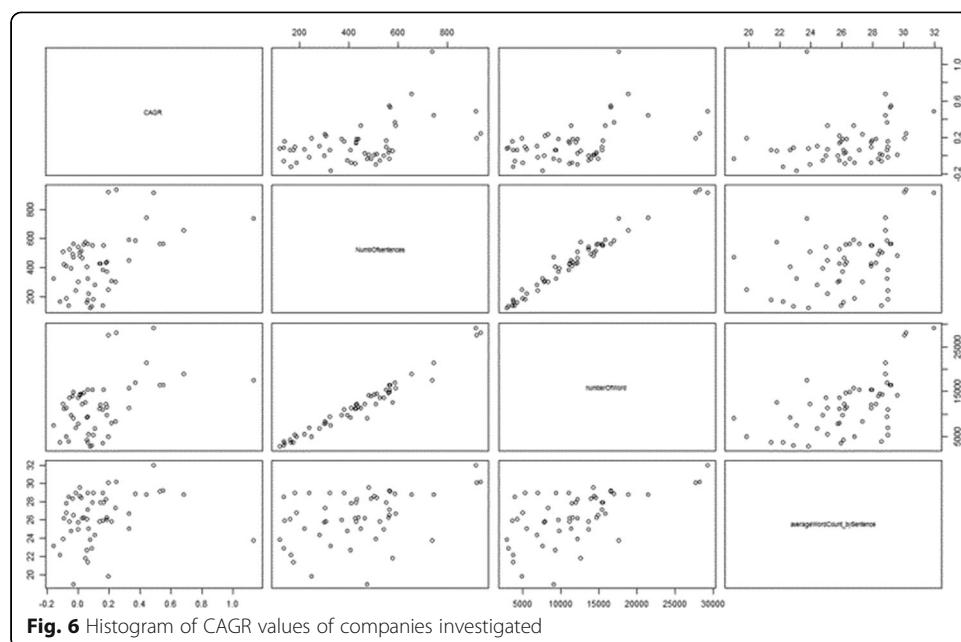
embarrassing. Therefore, we assume companies with low sale performance tend to publish shorter Risk Factor items (hypothesis 2–1). However, if the company's sales performance is good, it can be assumed this will also result in shorter Risk Factor items (hypothesis 2–2).

To verify the two hypotheses, the number of sentences, words and words per sentence in the "Risk Factors" text from each company are compared with their yearly CAGR values. First, we obtain a scatter plot matrix of the four variables (number of sentences, words and words per sentence and CAGR values), and gauge which variables are correlated (Fig. 6).

We find a weak linear relationship between CAGR values and number of sentences (Fig. 7). The correlation coefficient between the number of sentences and CAGR values is 0.474. However, if the data are analyzed separately by SIC classification 7370 and 7373, the correlation coefficients are 0.575 and 0.358 respectively (data not shown).

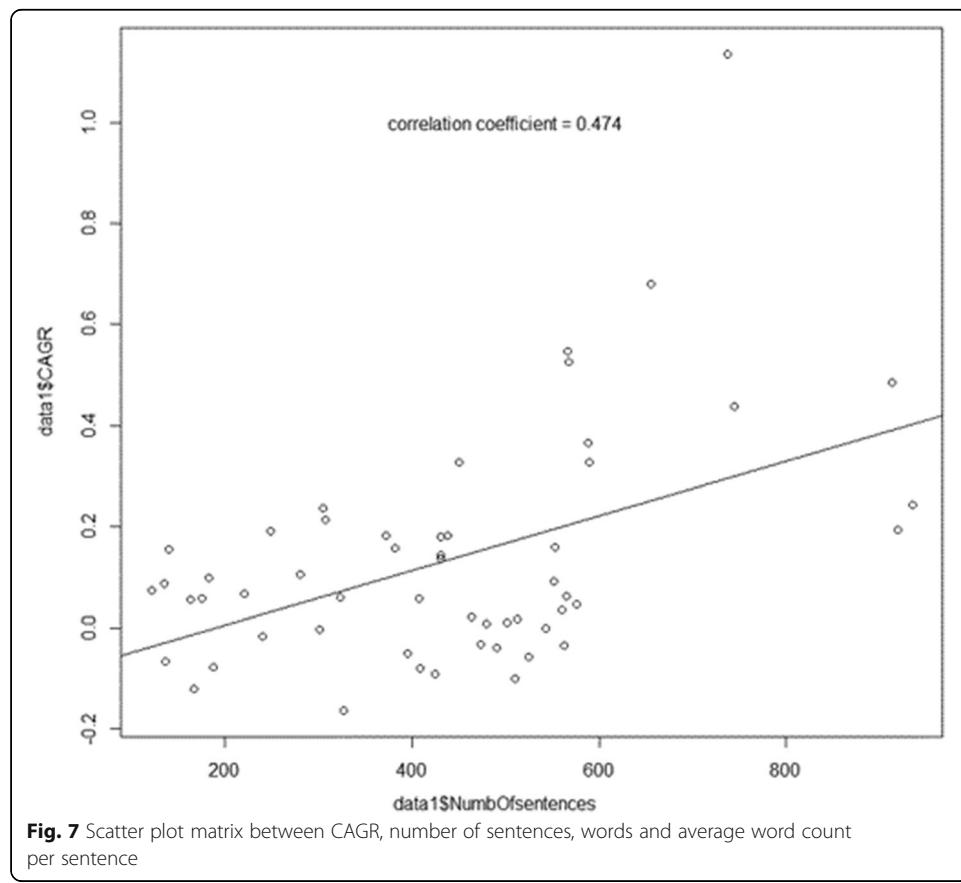
While none of the variables are strongly correlated, clustering the companies using three statistics of text may reveal a pattern. In other words, of the clusters derived by clustering analysis, certain clusters may be similar. In this study, we apply the WARD method as a hierarchical clustering technique and the results are visualized as dendograms (Fig. 8). The CAGR value of each company is shown in the node (label) of the data. CAGR values are not very similar among groups, but the group encircled in red has the highest value of CAGR values (Fig. 8). This cluster contains six companies. Among them, the lowest value of CAGR is 0.365, which is a company ranked seventh

**Fig. 5** Process of testing hypothesis 4

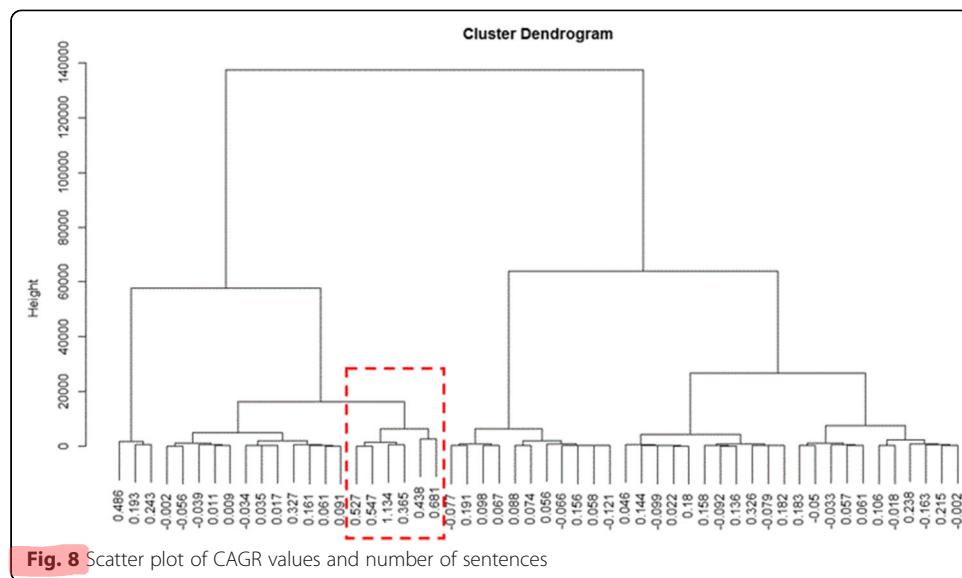


**Fig. 6** Histogram of CAGR values of companies investigated

among the total of 54 companies. Only one firm with a CAGR value higher than 0.365 is excluded from this cluster. This implies that, based on the statistical properties of a text, such as quantifying the number of sentences, clustering techniques can be used to cluster the companies according to the size of the CAGR of sales. As a result, in this



**Fig. 7** Scatter plot matrix between CAGR, number of sentences, words and average word count per sentence

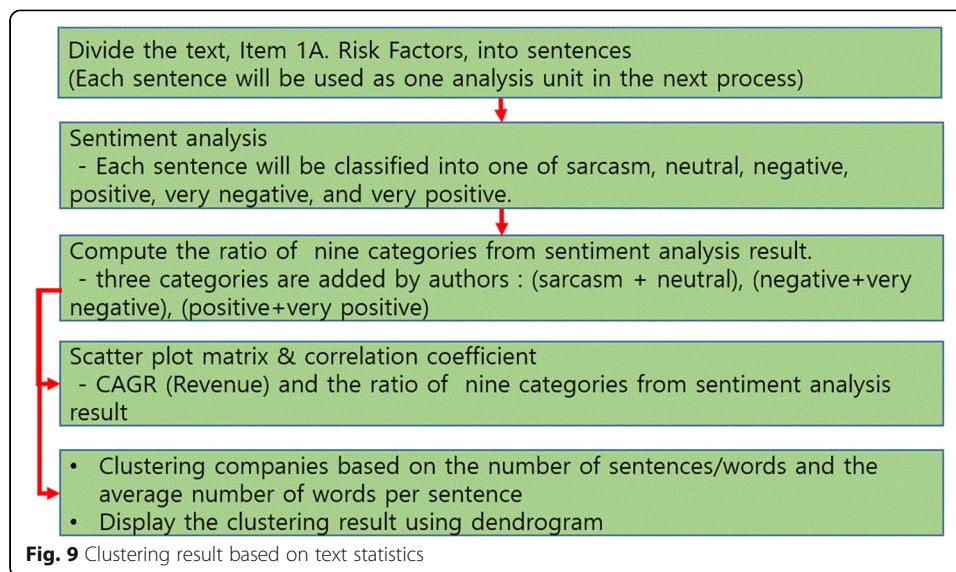


initial study, we do not find any clear evidence to support the hypotheses (2–1) and (2–2). However, the feasibility of correlating sales information and statistics of text should be further investigated in subsequent studies.

#### Relation between sentiment analysis result and CAGR

Sentiment analysis is a technique for determining whether a sentence's tone is positive, negative or neutral. In our study, sentiment analysis was introduced with the assumption that the tone of the text written in "Risk Factors" will be different depending on sales performance (hypotheses (3–1) and (3–2)). First, the full text is a company's Risk Factors part in the 10-K annual report. The full text is decomposed into sentences and sentiment analysis was performed for each sentence. The 'calculate\_total\_presence\_sentiment' function in the 'RSentiment' package categorizes the input sentence as positive, negative, very positive, very negative, neutral or sarcasm. This study adds three categories to the six existing categories as follows: sarcasm + neutral (S + N); positive + very positive (P + VP); negative + very negative (N + VN). In addition, the results of the sentiment analysis contain the number of sentences belonging to each category. The number of sentences in each category is normalized by the total number of sentences in the corpus. By plotting the ratio of each category against CAGR values, we determine which variables are correlated (Fig. 9). The correlation between the CAGR values and the other nine categories, shown in the first column, are not particularly linear (Fig. 10). Thus, none of the nine categories of sentiment analysis appear to be correlated with CAGR. The correlation coefficients are equally poor (Table 5). As a result, hypothesis (3–1) was rejected.

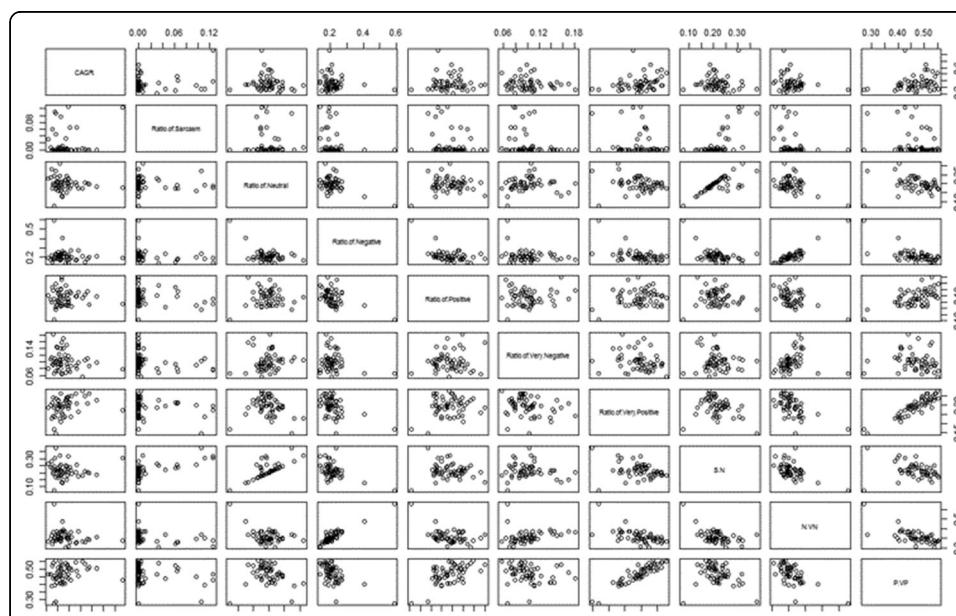
Although the nine categories of sentiment analysis do not correlate with CAGR values, we are able to group companies by sales performance by applying clustering. We apply hierarchical clustering using nine categories of data and visualize the results as a dendrogram. Figure 11 shows that the CAGR values of the nodes (labels) are not grouped into similar values but randomly clustered. Therefore, the result of clustering



of sentiment analysis results is not related to the growth rate of business performance.  
In conclusion, hypothesis (3–2) is rejected. !

#### Relation between keywords and CAGR value

The texts of the three highest and lowest ranking companies are grouped into one text file respectively, and analyzed. First, the pre-processing technique that removes stop-words, numbers, punctuations, and whitespaces are performed on the two texts. To select important words, a TDM is created applying TF-IDF weighting. After creating the TDM from each of the two documents, the top 200 terms, from each matrix were



**Fig. 10** Scatter plot matrix between CAGR values and nine categories of sentiment analysis

**Table 5** Correlation coefficients between CAGR values and nine categories of sentiment analysis

	Sarcasm	Neutral	Negative	Positive	Very Negative	Very Positive	S + N	N + VN	P + VP
CAGR	0.170	-0.125	-0.135	-0.046	-0.191	0.289	0.023	-0.213	0.221

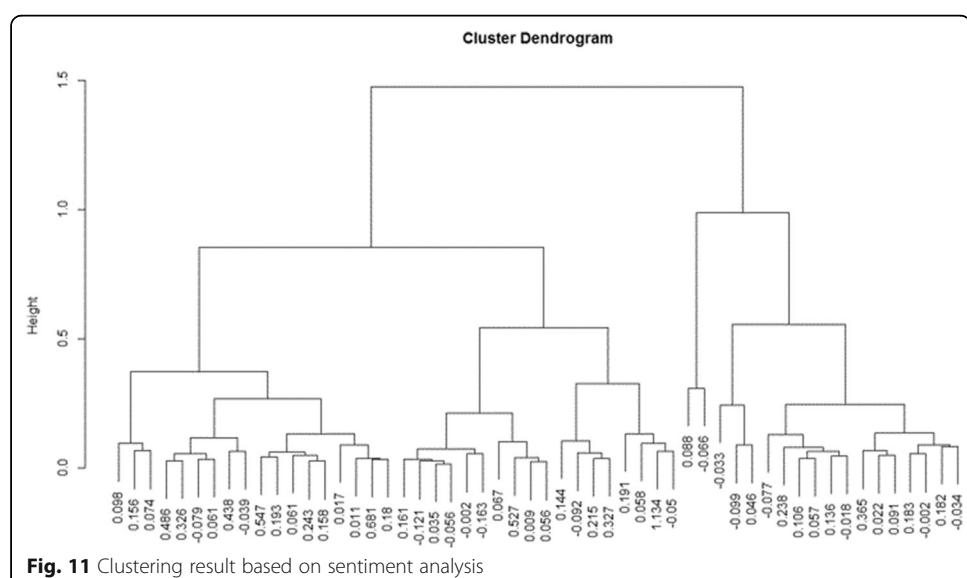
extracted. This analysis looks at the characteristics of words that do not overlap the two text files. Table 6 shows the top 50 terms of the top 200 terms for the highest and lowest ranking companies, and those that overlapped. In the table, in the list of the upper 200 words extracted from each, 119 are overlapping and 81 are unique words.

Non-overlapping keywords from the bottom three companies include many words related to finance, contract/law, and risk. Specifically, words related to finance were as follows: *revenues, assets, fiscal, cost, capital, budget, spending, and cash*; for law: *contracts, government, contract, claims, legal*; and risk: *failure, loss, risk*. By comparison, the words derived from the top three companies in sales are very noticeable in terms of clients and advertisement. Words related to client were as follows: *users, use, user, client, members*; and for advertising: *advertising, advertisers*. As such, we find that companies with poor operating results tend to use terms relating to finance, contract and law and risk words more frequently, owing to discussions of financial pressure and risk of default. On the other hand, companies with good sales performance are thought to focus more on users and advertising. It becomes clear that, after analyzing the reports of these six companies, there is some correlation between the use of language and sales growth rate. Therefore, in this context, hypothesis 4 can be adopted.

Word clouds are created from the corpus of the two highest and lowest three ranking companies in terms of sales growth rate (Fig. 12). Figure 13 shows word clouds of both overlapping and non-overlapping keywords.

## Conclusions

In this study, we apply text mining to the annual reports of US companies. The aim was to investigate whether word patterns found in selected texts were related to the



**Table 6** Non-overlapped and overlapped top 50 keywords from the two cases

No.	H-word	TF-IDF*	Inter-word	TF-IDF*	L-word	TF-IDF*
1	users	1052	may	3510	contracts	684
2	advertising	975	business	2184	government	619
3	platform	880	services	1702	leidos	354
4	data	854	products	1569	contract	346
5	user	679	results	1351	revenues	317
6	content	640	including	1228	assets	242
7	advertisers	601	future	1221	profitability	232
8	members	599	operating	1221	customer	230
9	notes	568	will	1201	technologies	229
10	access	558	result	1201	process	228
11	class	523	new	1199	programs	222
12	clients	522	stock	1169	fiscal	205
13	internet	505	ability	1167	part	204
14	mobile	496	financial	1148	threats	201
15	inventory	479	revenue	1028	cost	197
16	internal	421	significant	991	competitive	197
17	parties	416	adversely	961	budget	186
18	united	406	information	914	annual	177
19	privacy	397	common	897	obtain	176
20	engagement	397	laws	886	lockheed	176
21	states	386	use	884	spending	176
22	international	384	growth	878	requirements	172
23	twitter	372	operations	852	digimarc	166
24	change	368	changes	834	liability	166
25	software	362	customers	819	include	161
26	expect	357	technology	810	years	159
27	stockholders	351	market	809	agreement	158
28	reporting	348	also	807	delays	158
29	features	347	companies	804	patents	155
30	protection	338	costs	794	year	155
31	effective	337	property	790	prospects	148
32	harmed	326	addition	776	generally	146
33	brand	324	subject	774	funding	146
34	countries	319	affect	767	patent	144
35	practices	317	time	753	levels	144
36	devices	317	employees	751	debt	142
37	rate	312	intellectual	747	contractual	142
38	service	303	rights	715	estimates	142
39	shares	303	continue	698	report	142
40	foreign	302	tax	697	Martin	141
41	negatively	301	able	692	financing	141
42	base	299	certain	686	impairment	141
43	online	297	solutions	676	losses	141
44	source	297	risks	674	current	138

unique words in high-performance group  
(i.e. non-overlapping w/ low-group)

common words

✓ interesting method  
might be useful  
for exploration  
data analysis  
EDA?

**Table 6** Non-overlapped and overlapped top 50 keywords from the two cases (Continued)

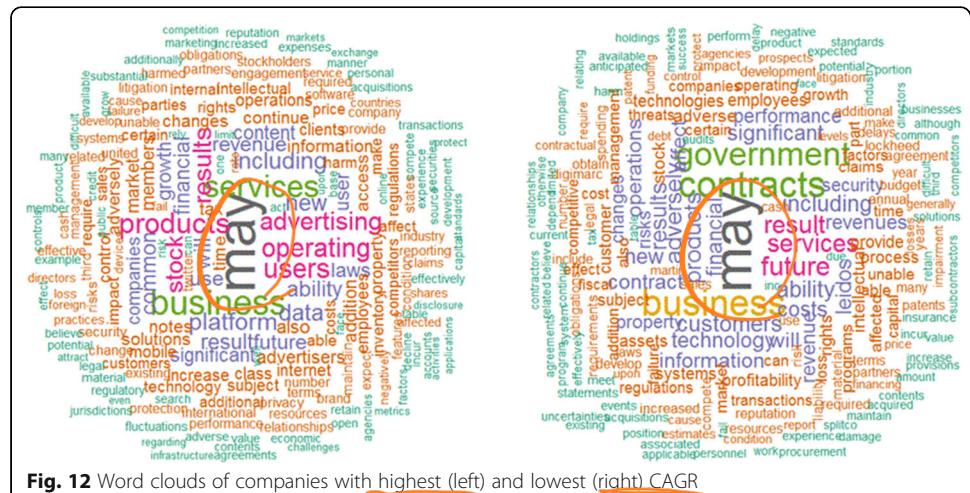
No.	H-word	TF-IDF	Inter-word	TF-IDF	L-word	TF-IDF
45	decline	295	security	656	insurance	136
46	credit	291	impact	656	depend	136
47	securities	286	control	650	meet	136
48	example	286	increase	648	perform	135
49	regulatory	285	performance	632	inc	135
50	fluctuations	284	regulations	631	expected	134

'H-word' keywords occurring only in the top three companies of CAGR of revenue, 'Inter-word' keywords occurring in the corpus of both, 'L-word' keywords occurring only in the three companies with the lowest CAGR values

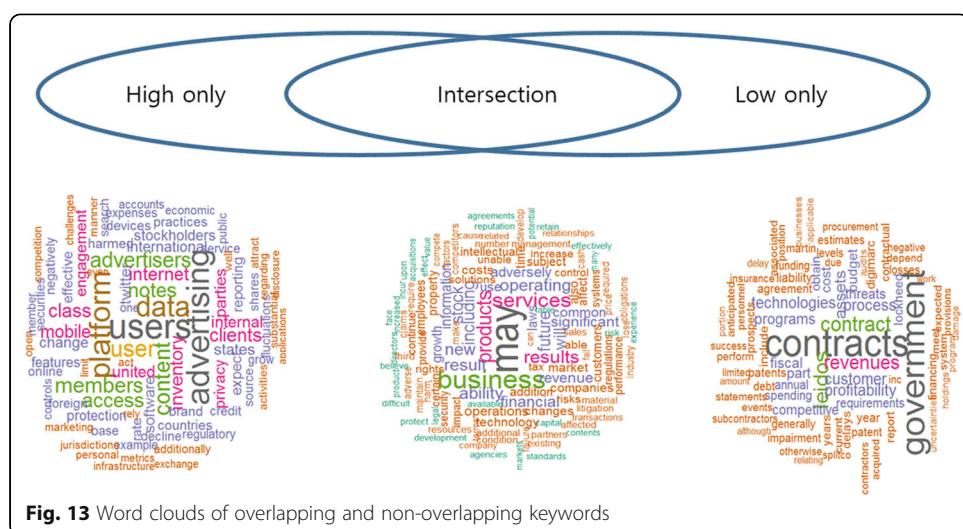
\*The TF-IDF value generally includes sub-decimal values, but only the integer value is represented in this table.

business performance of the company. We test four hypotheses: hypothesis 1 postulates that category 7370 companies, which include a large number of companies engaged in SNS activities, will have a better business performance than category 7373 companies, such as Yahoo. Hypothesis 1 is verified. Hypotheses 2 through 4 are analyzed by applying text and data-mining techniques to the risk factors of annual reports. Hypothesis 2 postulates that sales performance affects text statistics such as number of sentences. There is some evidence of correlation between sales performance and text statistics, however, further research is required. Hypothesis 3 postulates that the tone of the text correlates with sales performance. Applying sentiment analysis, no correlation was found and, thus, hypothesis 3 is rejected. Hypothesis 4 postulates that word usage in the text is correlated with the sales performance, and the hypothesis is temporarily adopted.

In summary, we identify a number of correlations between sales performance and the text pattern of company reports by applying text-mining technology. We expect to have more themes to be studied in the future. For hypothesis 2, better results can be expected by changing the data-processing method. Hypothesis 3 is rejected in this instance, but it is expected that better results can be obtained if the



This is useful visualization



classification method of sentiment analysis is optimized for the text of annual reports. The analysis framework for hypothesis 4 needs to be designed to cover all the data, not only the highest and lowest three ranking companies in terms of sales performance. Also, if key phrases are extracted rather than the number of words, results may be more meaningful.

This study provides the following conclusions. Companies with good financial performance and bad companies often use different words. Therefore, it is very meaningful to analyze the words that appear predominantly in the business text in predicting a company's future sales performance. However, the positive or negative tone of the business text is not relevant to forecasting the future financial performance of the company, as it appears to be lacking in correlation with sales performance. As a result, we hope that this study will serve as a stepping stone to develop the research contents by predicting future financial performance of companies and finding promising business areas.

## Additional file

**Additional file 1: Table S1.** Revenue of the US listed companies classified to SIC 7370 and 7373. (DOCX 22 kb)

### Acknowledgements

This paper was received 'best paper award' at SOlmc 2017 conference.

The main ideas in this paper were previously reported in the conference (Lee et al., 2017).

### Funding

This research was supported by Korea Institute of Science and Technology Information(KISTI) and Basic Science Research Program through the National Research Foundation of Korea (NRF-2015R1D1A1A09061299) funded by the Ministry of Education.

### Availability of data and materials

All data can be obtained by manually querying the SEC EDGAR system (URL: <https://www.sec.gov/edgar/searchedgar/companysearch.html>). However, financial information for each company is attached as an Additional file 1 separately at the end of the manuscript.

### Authors' contributions

BRL, JHP, LNK, and YHM carried out a systematic literature review. BRL, YHS, and GSK collected data and programmed algorithms to analyse the data. BRL and HJK wrote and revised the final manuscript. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Publisher's Note**

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

**Author details**

<sup>1</sup>Korea Institute of Science and Technology Information, Seoul, Republic of Korea. <sup>2</sup>The Graduate School of the University of Seoul, Seoul, Republic of Korea.

Received: 15 October 2017 Accepted: 4 January 2018

Published online: 02 February 2018

**References**

- EDGAR. 2018 <https://www.sec.gov/edgar/searchedgar/companysearch.html>. Accessed 01 April 2017
- Jeong, E. S., Kim, Y. G., Lee, S. C., Kim, Y. T., & Chang, Y. B. (2014). Identifying Emerging free technologies by PCT patent analysis. *JKIFCS*, 9(1), 111–122.
- Jeong, E. S., Yeo, W. D., Cho, D. Y., & Suh, I. W. (2008). Emerging research field selection of Construction & Transportation Sectors using Scientometrics. *The Journal of the Korea Contents Association*, 8(2), 231–238.
- Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., & Visa, A. (2002). Combining data and text mining techniques for analyzing financial reports, *Eighth Americas Conference on Information Systems*.
- Lee, B., Park, J.H., Kwon, L., Moon, Y.H., & Kim, H.J. (2017). An analysis on the relation between business text pattern and financial performance in US listed companies, SOltnC 2017, Latvia.
- Lee, H., Surdeanu, M., MacCartney, B., & Jarafsky, D. (2014). On the importance of text analysis for stock price prediction. Proceedings of the 9th edition of the language resources and evaluation conference (LREC). 1170-1175.
- Lee, J., & Hong, Y. S. (2014). Business model mining: Analyzing a Firm's business model with text Mining of Annual Report. *Industrial Engineering & Management Systems*, 13(4), 432–441.
- Lee, J., & Hong, Y. S. (2016). Extraction and visualization of industrial service portfolios by text mining of 10-K annual reports. *Flexible Services and Manufacturing Journal*, 28, 551–574.
- Park, J. H., Lee, B., Moon, Y. H., & Kwon, L. N. (2016). Study for selection of industrial areas suitable to small and medium-sized enterprises (SMEs) in Korea. *Journal of Open Innovation: Technology, Market, and Complexity*, 2016, 2(19).
- Pulizza, J. L. (2015). An analysis of speculative language in SEC 10-K filings. Chapel Hill, North Carolina: A master's paper for the M.S.
- Satikolshikiriyama, C., Miro, D., & SimõesGomes, C. F. (2015). Text mining business intelligence: A small sample of what words can say. *Procedia Computer Science*, 55, 261–267.
- Shirata, C. Y., Kakeuchi, H., & Watanabe, H. (2011). Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of emerging technologies in accounting*, 8.
- Svirina, A., Azbarova, A., & Oganisjana, K. (2016). Implementing open innovation concept in social business. *Journal of Open Innovation: Technology, Market, and Complexity*, 2016, 2(20).
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307.
- Wikipedia, Sentiment analysis. 2018. [https://en.wikipedia.org/wiki/Sentiment\\_analysis](https://en.wikipedia.org/wiki/Sentiment_analysis) Accessed 01 April 2017.
- Witt, U. (2016). What kind of innovations do we need to secure our future. *Journal of Open Innovation: Technology, Market, and Complexity*, 2016, 2(17).
- Yun, J. J., Won, D., & Park, K. (2016). Dynamics from open innovation to evolutionary change. *Journal of Open Innovation: Technology, Market, and Complexity*, 2016, 2(7).

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)