経営論集　第 32 巻第 1 号　2022 年　83–99 頁

Fraud Detection Method by Textual Analysis of CEO Letters in the Perspective of Obfuscation Hypothesis: Evidence from Japanese Firms Listed on the U.S. Stock Exchange

Masumi Nakashima

〈**Abstract**〉

This study uses a sample of Japanese firms listed in the U.S. and conduct textual analysis of Chief Executive Officer's (CEO) letters to test the obfuscation hypothesis by examining whether there are differences in text features focusing the on Jaccard coefficients, the frequency of specific words, and distance from the origin between fraudulent and non-fraudulent firms, as well as during and after the detection of fraud periods. First, the finding is that the Jaccard coefficient during fraud execution period is higher than that after fraud is detected, indicating that fraudulent firm uses more distinctive words during fraud execution period. Second, the results of the correspondence analysis show that the words are far from zero, indicating unrelated words are placed during fraud-execution period, and the words are closer to zero and the words are placed close to each other after fraudulent detection period, and the words of non-fraudulent firms are plotted close to the origin. In other words, the Jaccard coefficients of fraudulent firm are higher during the fraudulent period than after fraud is detected period, and the words are plotted farther from the origin. This suggests that, managers may have misled investors by using impression management to make the text more difficult to read during fraud-execution period. This study can be presented as a possibility for fraud detection when it is observed that CEO letters have the characteristic words from those of non-fraudulent firms through textual analysis.

〈**Keywords**〉

text-mining; correspondence analysis; CEO letters, impression management, Jaccard coefficient.

I am grateful to moderator Keisuke Takeguchi and the participants at the 38th Annual Meeting of Accounting and Economics Association of Japan (AEAJ), moderator and discussant Paul Scarbrough and the participants at 2021 (16th) Annual Conference of Asia-Pacific Management Accounting Association, and Noriyuki Tsunogaya and his students at Hitotsubashi University, School of Business Administration for their precious comments. This study is supported by grant-in-aid as a scientific study of the fiscal year 2022-2025 by the Ministry of Education, Culture, Sports, Science and Technology (Representative: Masumi Nakashima 22K01813: Theoretical and Empirical Studies of Earnings Management for Misrepresentation and Tone Management).

2022年10月24日受理

I. INTRODUCTION

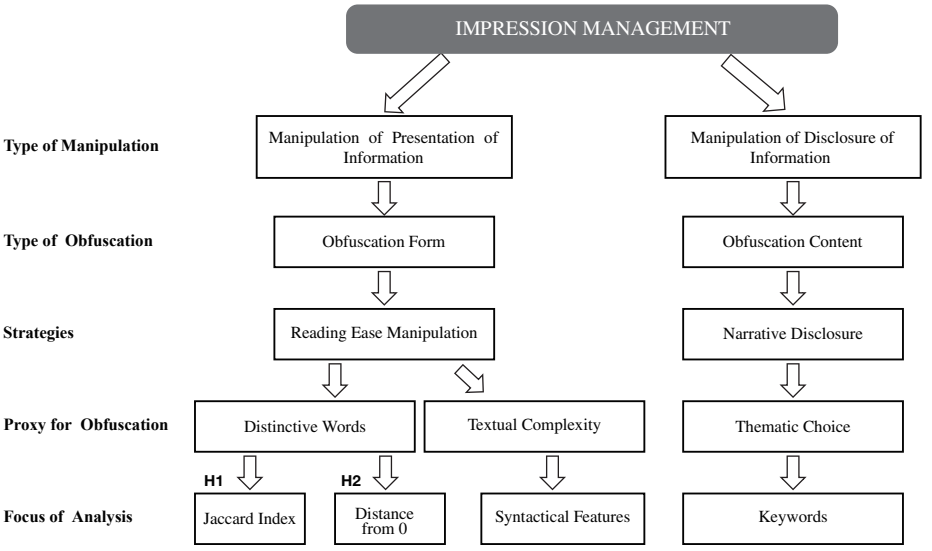
In recent years, there has been an increase in research focusing on narrative information. Especially, fraud detection studies using qualitative information (Churyk et al, 2009; Chen et al 2014; Nakashima et al. 2022) have been conducted, but they have not yet accumulated much. Nakashima et al. (2019) find mixed results in the readability of fraudulent firms, which may be one of the reasons why fraud detection method research has not increased. Although it is not fraud research, the obfuscation hypothesis has been tested in many studies that have examined the relationship between readability and corporate performance by using readability, which is considered a concept of textual complexity, of narrative information as a measurement proxy (Ajina et al. 2018; Clatworthy and Jones 2003 and 2006; Courtis 1998; Davis and Tama-Sweet 2012; Hassan and Habib 2020; Hirose et al. 2017; Leary and Kowalski 1990; Li 2008; Linsley and Lawrence 2005; Lo et al. 2017; Patelli and Pedrini 2014; Souza et al. 2017).

Impression management that is defined as the process by which people attempt to control the impressions others form of them (Leary and Kowalski 1990, 34) is discussed by testing of the obfuscation hypothesis that asserts how managers attempt to confuse investors for bad news of the firm (Markle-Davis 2007, 8). In this study, I conduct textual analysis of Chief Executive Officer's (CEO) letters of Toshiba to test the obfuscation hypothesis that means whether Toshiba managers attempt to confuse investors. In other words, I use a sample of Japanese firms listed in the U.S. and conduct textual analysis of their CEO letters to test the obfuscation hypothesis by examining whether there are differences in text features focusing on the Jaccard coefficient and distance from the origin as readability measurements between Toshiba and non-fraudulent firms, as well as during and after the detection of fraud periods.

Fraud studies that have used the KH coder to conduct textual analysis include Ohno (2020) and Nakashima et al. (2021). Ohno (2020) finds that the words are located far from the origin based on the correspondence analysis of the fraudulent firms. Nakashima et al. (2021) find that the co-occurrence network results show that the words of the fraudulent firms are not connected to each other during fraud-execution period.

The diagram of Figure 1 illustrates a theoretical framework of this study. There are two manipulations that managers engage in impression management. One is about the format of information presentation, and the other is about the content of disclosure. There are strategies for each of these. (1) obfuscation is the complexity of the text, and (2) the choice of themes for narrative information (Mercle-Davis 2007, Figure 2-3). In this study, I test whether Toshiba managers manage impressions in CEO letters by testing the Jaccard coefficient in H1 and the distance from the origin in H2 for the readability manipulation to see if they manage impressions in narrative information.

FIGURE 1
Theoretical Framework



Note: This figure is prepared by author following Merkle-Davis (2013, Figure 2.3)

Fraud research has focused on Management's Discussion and Analysis (MD&A) disclosure, but this study focuses on CEO letters. Why do I focus on CEO letters? Because CEO letters play an important role for communicating the attitudes, values, and behaviors of those in the roles of senior management (Amernic et al. 2019, v). CEO letters, like MD&A disclosures, are not subject to audit; however, in contrast to the MD&A disclosure, CEO letters are not subject to special disclosure rules by the U.S. Securities and Exchange Commission (SEC) or Financial Service Agency (FSA).[1] Since CEO letters possess a higher degree of discretion than MD&A disclosure, I consider that fraud may be detected through CEO letters by examining whether there is a difference in vocabulary such as specific words or spatial relationship between words in CEO letters.

This study contributes to the literature in several ways. Previous fraud detection studies have focused on MD&A disclosure. First, this study presents fraud detection method to conduct textual

[1] The FSA presents the "Principles on Disclosure of Descriptive Information" in 2018. https://www.fsa.go.jp/en/news/2019/20190606-3/01.pdf According to the "Principles on Disclosure of Descriptive Information" (2018, 12), as a matter required by law, "In disclosing management's analysis of the state of financial position, results of operations, and cash flows (results of operations, etc.), the content of the analysis and discussion of the state of operations, etc. from management's perspective should be described in a specific and understandable manner. In disclosing the analysis of the status of financial position, operating results, and cash flows (operating results, etc.) by the management, the analysis and discussion of the status of operating results, etc. from the management's perspective must be stated in a concrete and easy-to-understand manner. In doing so, for each category described in the overall business and segment information, the content of recognition and analysis and discussion from management's perspective (e.g., analysis of factors that have a significant impact on operating results) should be described in relation to the content of other items described in the securities report, in addition to the content of management policies and management strategies, etc." (FSA 2018, 12).

analysis of CEO letters, which are highly discretionary documents by managers and are more likely to include their values and beliefs, through revealing the differences between fraudulent and non-fraudulent firms and the differences in the pre-and post- fraud detection.

Nakashima (2022) is textual analysis study that examines the relationship between financial statement figures and narrative information in terms of management's consistency with financial statement figures and compliance with the Sincerity Principle.[2] Second, this study examines whether Toshiba manager uses impression management (Leary and Kowalski 1990, 34) based on the obfuscation hypothesis, focusing on textual characteristics of CEO letters of fraudulent firm, that is the Jaccard coefficients and the distance from zero as readability. Textual analysis of narrative information may help regulators and auditors in their investigations and audits by suggesting that fraud can be detected by the Jaccard coefficients and the distance from zero. Although more studies focus on total number of words and lexical diversity or grade in Japanese language as readability, there have been a mixed result. Third, this study presents the Jaccard coefficient as new measure of text complexity.

This study is organized as follows. Section 2 presents the theoretical framework, and Section 3 reviews the previous studies. Section 4 develops the hypotheses, Section 5 presents the research design, and Section 6 presents the results of the analysis. Finally, conclusions and future research are presented.

II. LITERATURE LEVIEW

Based on agency theory (Jensen and Meckling 1976), information asymmetry occurs between Principle and Agent, and managers have incentives to use information asymmetry in ways that are beneficial to managers, not necessarily to investors.[3] Bloomfield (2002, 238) shows that many decisions are motivated by management's desire to make it difficult for investors to hide information that management itself does not want to influence the company's stock price.

Studies that have examined readability in MD&A information include Li (2008), Bushee et al. (2008), Lo et al.(2017), and Demaline (2020). Li (2008), Hirose et al. (2017), and Nakashima et al. (2022a) find that annual reports of low-profit firms are more difficult to read, while annual reports of

---

[2] The Financial Accounting Standards Board (FASB 2010), in its Statement of Financial Accounting Concepts No. 8, Conceptual Framework for Financial Reporting. Chapter 3, Qualitative Characteristics of Useful Financial Information, states that "Financial statements present economic phenomena in words and numbers. In order to be useful, financial statements must not only present phenomena that are relevant to the purpose, but they must also be faithfully presented. To be fully faithful, they must combine three characteristics: completeness, neutrality, and absence of error" (FASB 2010, QC12).

[3] On the other hand, it is said that managers disclose information for the purpose of communicating inside information in order to reduce information asymmetry; Li et al. (2013) find that when the prediction of the next year's return on net operating assets is conditional on the level of competition, the resulting significant and economically meaningful difference is between high and low levels of competition. The result is that significant and economically meaningful differences arise between firms with high and low levels of competition. The results suggest that disclosure strategies can be useful for financial statement analysis.

high-profit firms are more readable and more persistent. Nakashima et al. (2022) indicates that the MD&A disclosure of fraudulent firms differs in their Japanese language readability from the MD&A disclosure of nonfraudulent firms. Bushee et al. (2008) decompose managerial linguistic complexity into information and the obfuscation components and find that the information component of linguistic complexity is associated with low information asymmetry and the obfuscation component is associated with high information asymmetry.

Lo et al. (2017) find that the readability of MD&A information examines whether it varies with earnings management. Firms with earnings management for the previous year's earnings had more complex MD&A information. Li (2008) indicates that good news is easier to communicate, and that the obfuscation complicates disclosure. Demaline (2020) examines the readability between MD&A information of SEC investigated firms and MD&A information of average trading firms and finds that MD&A information of SEC investigated firms are more readable and the result supports the obfuscation hypothesis of impression management.

Studies that have examined the readability of the CEO letters include Courtis (1998) and Clatworthy and Jones (2003; 2006). Courtis (1998) investigates the variation in readability of CEO letters in a sample of 120 listed firms in Hong Kong and finds that readability scores are variable between high profitable and low profitable firms. Clatworthy and Jones (2003) present that in both the top 50 and bottom 50 CEO letters of the UK firms, managers emphasizes the positive aspects of performance, taking credit for good news and blaming the environment for bad news. Ajine et al. (2016) examine the relationship between earnings management and annual report readability for listed firms in France and find that the annual reports of firms with earnings management increase in complexity.

Clatworthy and Jones (2006) examine the narrative information practices of 50 high-profit and 50 low-profit firms and find that textual feature of CEO letters differs by profitability. Low-profit firms are characterized by fewer important financial indicators, few quantitative results, more passive text, and more statements about the future. This indicates that the narrative information of low-profit firms is not completely consistent with their financial performance and does not present a balanced and objective view. From the above, Clatworthy and Jones (2006) indicate that narrative information can be used for impression management.

Hassan et al. (2019) show that listed firms in Qatar use the complexity of narrative information as a disclosure strategy to strengthen impressions and maintain social legitimacy because firms with high annual report readability have higher profitability and lower agency costs. Hassan and Habib (2020) examine the relationship between readability of narrative information in 10K reports and firms' liquidity and payout production and find that firms with lower readability have more cash and little cash dividends. The above shows that the results converge with the research findings that managers mislead investors by decreasing readability of narrative information and using complexity to reinforce their impressions.

III. HYPOTHESIS DEVELOPMENT

Churyk et al. (2009) find qualitative differences in the language characteristics of the MD&A section of the annual reports of firms for which amended financial statements are requested by the SEC. They also show that positive sentiment is lower in the reports of firms for which amended financial statements are requested and is associated with a higher total number of words in the MD&A. Lee et al. (2013) develop the approach of Churyk et al. (2009) and conduct stepwise regressions to show positive sentiment terms, fewer verbs in the present tense, the presence or absence of an increase in word count, and fewer colons. However, these previous studies are conducted to analyze narrative information in English.

Nakashima et al. (2022a) is the fraudulent study that analyzes MD&A information in Japanese. Nakashima et al. (2022a) examines whether there are commonalities and peculiarities in Japanese narrative information by comparing the MD&A disclosures of a paired sample of fraudulent and non-fraudulent firms. They find that there is a significant difference in the readability of the Japanese language and the number of predicates per sentence. This result indicates that fraudulent firm managers may obfuscate MD&A disclosure to confuse stakeholders and hide fraud. In light of the results of these previous studies, this study can expect that there is a difference between the distinctive words of fraudulent firm in the pre-and post-fraud detection periods and those of non-fraudulent firms. Therefore, this study formulates the following hypothesis:

H1a: There is a difference in the word use of fraudulent firm in the pre-and post-the fraud detection periods.

H1b: There is a difference in the word use between fraudulent firm and non-fraudulent firms.

Ohno (2020) conducts textual analysis of MD&A disclosure of firms in the chemical industry in Japan using KH Coder, and the results of correspondence analysis of MD&A disclosure of fraudulent firms show that the words are plotted far from the origin. Nakashima et al. (2022b) implement text-mining analysis of MD&A disclosure of construction and retail industry. The results of correspondence analysis indicates that the words are plotted far away from the central point during fraud execution period. This suggests that many of them are distinctive with weak relevance between words.

Based on this result, I predict that fraudulent firm managers use distinctive words, unlike non-fraudulent firm managers, to mislead investors and other financial statement users. Therefore, the following hypothesis is developed:

H2a: There is a difference in the distance from zero in the pre-and post-fraud detection periods for fraudulent firm.

H2b: There is a difference in the distance from zero between fraudulent firm and non-fraudulent firms.

IV. METHODOLOGY

4.1 Sample Selection

This study examines CEO letters in annual reports or shareholder newsletters of the Japanese firms listed in the U.S. Of the firms, Toshiba is fraudulent firm and other nine firms are non-fraudulent firms. This study investigates CEO letters of Toshiba during the fraud execution period and after fraud-detection period. CEO letters from 2009 to 2014 are analyzed during the fraud-execution period, CEO letters from 2015 to 2019 are analyzed after the fraud is detected, and CEO letters from 2009 to 2014 are analyzed for non-fraudulent firms. Table 1 shows the sample selection procedures. As of March 2021, there are fourteen Japanese firms listed in the U.S. Of these, nine firms are tested as the final sample, excluding two delisted firms, two financial institutions, and one firm for which data is not available.

TABLE 1
Sample Selection

Criteria	Number
SEC Standard Japanese Firms	14
△Delisted firms	(2)
	12
△Financial Institution	(2)
△Data Not Available	(1)
Total	9

The sample firms are following: WACOL HOLDINGS CORP, FUJIFILM Holdings Corporation, KOMTATSU LTD, TOSHIBA CORPORATION, OMRON Corporation, SONY CORPORATION, TDK COPORATION, SANYO Electric Co., Murata Manufacturing Co., Ltd, CANON INC.

4.2 Analysis Methodology

This study conducts overview analysis, distinctive word analysis, and correspondence analysis using KH Coder.[4] Before each analysis, the target parts of speech are selected as preparatory work.[5] For example, nouns (profit), sub-variant nouns (strengthen), adjectives (good), tags (compliance), verbs (tackle), and adjectives (severe) are selected. In addition, the words such as “representative” “director,” “president,” “everyone,” and “our company.” are removed. The rationale behind the selection of unused words is that universal words that appear in large numbers have co-occurrence relations with various

[4] KH Coder is a free software for statistical analysis of text information invented by Koichi Higuchi. For details, please refer to the following Web site. https://khcoder.net/

[5] Data preparation, cleansing, and preprocessing are carried out as preparatory work prior to the analysis, especially preprocessing using MeCab, a morphological analysis engine, since it supports many business terms. Next, I create a dictionary. First, for words whose meanings change significantly depending on the division, I change cash/flow → cash flow stake/holder → stakeholder. In addition, I add: unknown word compliance that may be necessary for the research project and interpretation of the results.

words, and the community of the co-occurrence network becomes large, making it unsuitable for topic extraction. Therefore, these words are excluded from the present analysis.

4.3 Measure of Readability

In this analysis, readability is defined as the Jaccard coefficient in the characteristic term analysis and the distance from the origin in the correspondence analysis. The Jaccard coefficient is calculated by dividing the number of co-occurrences by the number of occurrences in either direction, and indicates the strength of the co-occurrence. Terms with high Jaccard coefficients are regarded as distinctive terms that are not common words. To test hypotheses 1 and 2, this study uses characteristic terms analysis and correspondence analysis, respectively.

V. OVERVIEW ANALYSIS

In order to capture the characteristic terms in CEO letters of the Japanese firms listed in the U.S. Stock Exchange, this study conducts an overview analysis of Toshiba's fraud during the fraud-execution, after the fraud was detected, and non-fraudulent firms, respectively. Table 2 shows the descriptive statistics of Toshiba's fraud-execution period, post-fraud detection period, and non-fraudulent firms. The target sentences of Toshiba's fraud-execution period, after fraud-detection period, and non-fraudulent firms are 146, 370, and 4456, respectively. The total extracted words for Toshiba's fraudulent, post-fraudulent, and non-fraudulent firms are 1021, 2791, and 27407, respectively. The total number of extracted words is the total number of all words. The number of different words in Toshiba's fraudulent, post-fraudulent periods, and non-fraudulent firms is 383, 844, and 2863, respectively.

TABLE 2
Descriptive Statistics

	Fraudulent Firm		
	Fraud-Execution Period	Fraud-Detection Period	Non-Fraudulent Firms
Number of Sentences	146	370	4456
Total Number of Words Extracted	1021	2791	27407
Number of Different Words	383	844	2836

Table 2 shows the results of the number of target sentences, the total number of extracted words, and the number of different words for fraud-execution and fraud-detection periods of Toshiba, and non-fraudulent firms.
Total number of words extracted = Total number of all words
Number of different words = Number of different words = number of words without duplication

The number of different words is the number of words after eliminating duplicates. After the fraud is revealed, the total number of extracted words increase in Toshiba. This may be due to the use of words that are not used in normal times, such as apology for accounting fraud problems and handling

rules in the stock market.

Table 3 presents the top 50 extracted words from fraudulent firm, during and after the detection of fraud from non-fraudulent firms. In the next section, I capture the distinctive words from the extracted words.

TABLE 3　Top 50 Words Extracted from During and After Fraud Detection Periods of Fraudulent Firm and from Non-fraudulent Firms

During Commtting Fraud Period

Extracted Words	Number	Extracted Words	Number	Extracted Words	Number
Management	24	Contribution	5	Great	3
Executive	19	Financial	5	Effort	3
Company	16	Sales	5	Investment	3
Sector	15	Profit increase	5	Transparent	3
Profit and Loss	14	Dividend	5	Annual	3
Group	13	Device	4	Field	3
Growth	13	Program	4	Change	3
Body	13	LCD	4	Position	3
Global	12	Fulfillment	4	Area	3
Value	12	Company	4	System	2
Improvement	12	Recovery	4	Total	2
Prior period	12	Thick	4	Negative	2
Consolidated	12	Do	4	Manufacturers	2
Improvement	11	Market	4	Level	2
Structure	11	Realization	4	Deterioration	2
Support	11	More than	4	Projects	2
Profit	11	Situation	4	Intention	2
Acceleration	10	New	4	Transition	2
Director	10	Progress	4	Overseas	2
Commissioner	9	Proceed	4	Hold	2
Reform	9	All-out	4	Distinctive	2
Environment	9	Revenue Increase	4	Activity	2
Development	9	Substantial	4	Based	2
Towards	8	Electronic	4	Base	2
Black	8	Current term	4	Scale	2
Society	8	Accompanying	4	Discussion	2
Center	8	Towards	4	Competition	2
Policy	8	Thank you	3	Emerging countries	2
Performance	7	Governance	3	Equilibrium	2
Business	7	Corporate	3	Air Conditioning	2
Decision	7	TV	3	Pertaining	2
External	7	Balance	3	Tilt	2
Promotion	7	Memory	3	Economic	2
Sales	7	Add	3	Severe	2
Profit	7	Home	3	Decline	2
Innovation	6	Challenges	3	Decline	2
Infrastructure	6	Expansion	3	Further	2
Basic	6	Secure	3	Think	2
Function	6	Expectation	3	Increase	2
Strengthen	6	Plan	3	High level	2
Concurrent	6	Expectation	3	Domestic	2
Strong	6	Effect	3	Nomination	2
Efforts	6	Efficiency	3	Age	2
World	6	Maximum	3	Implementation	2
Achieve	6	Sustain	3	Implement	2
Transformation	6	Important	3	Achievements	2
Top	5	Evolution	3	Take	2
Sales	5	Installation	3	Demand	2
Impact	5	Selection	3	Appointment	2
Continuation	5	System	3	Group	2

After Fraud Detection Period

Extracted Words	Number	Extracted Words	Number	Extracted Words	Number
System	40	Necessary	11	Limited	6
Management	37	Report	11	High	6
Reform	32	Consolidated	11	Debt	6
Company	32	Accounting	10	Maximum	6
Growth	28	Confirmation	10	Support	6
Designation	26	Structure	10	Blockage	6
Improvement	25	Contribution	10	Acquisition	6
Management	25	Funds	10	Listing	6
Group	24	Execution	10	Deep	6
Strengthening	24	Initiatives	10	Progress	6
Market	24	Trust	10	All-out	6
Technology	23	Target	10	Profit and Loss	6
Issues	23	Profit	10	Large	6
Stock	22	Energy	9	Focus	6
Society	22	Cyber	9	Excess	6
Proceed	22	Physical	9	Acknowledge	6
Investment	22	Unity	9	Dividend	6
Towards	21	Company	9	Sale	6
Measures	21	Plan	9	Announcement	6
Plan	20	Finance	9	Guarantee	6
Securities	20	Examination	9	Infrastructure	5
Value	19	Companywide	9	Technology	5
Internal	19	Judgment	9	Process	5
Attention	18	Area	9	Negative	5
Special	18	Company	8	Model	5
Continuation	17	Corporate	8	Nurturing	5
Profit	17	Sales	8	Overseas	5
Transactions	16	Add	8	Certainty	5
World	15	Capital	8	Audit	5
Possible	14	Execution	8	Related	5
Foundation	14	Procurement	8	Crisis	5
Transfer	14	Thorough	8	Foundation	5
Submission	14	Valuable	8	Performance	5
Risk	13	System	7	Results	5
Revival	13	Stakeholders	7	Construction	5
Memory	12	Stability	7	Consideration	5
Development	12	Opinion	7	Prospect	5
Do	12	Recover	7	Build	5
Implementation	12	Completion	7	Think	5
Efforts	12	Scale	7	Assets	5
Change	12	Improvement	7	Sustained	5
Challenge	11	New	7	Self	5
Release	11	Third Party	7	Retrieve	5
Nuclear	11	Climate	7	Procedure	5
Think	11	Report	7	Worry	5
Realization	11	Apology	6	New	5
Focus	11	Governance	6	Sincerity	5
Situation	11	Management	6	Parent Company	5
Appropriate	11	Based	6	Explanation	5
Same firm	11	Accounting	6	Strategy	5

Non-Fraudulent Firms

Extracted Words	Number	Extracted Words	Number	Extracted Words	Number
Group	378	Electronics	85	Achievement	58
Growth	358	Expectation	84	Medium Term	58
Market	307	High	83	Domestic	57
Profit	306	Recovery	82	Need	57
Product	297	Automobile	82	Overseas	56
Management	229	Progress	82	Important	56
Sales	221	Center	81	Communication	56
Expansion	215	Economy	79	Cost	55
Environment	207	New	79	Foundation	54
World	200	Promotion	79	Film	53
Technology	197	Add	78	Return	53
Companies	195	Provision	77	Business	52
Development	188	Challenges	74	Customers	51
Performance	177	System	74	Focus	51
Production	177	Expansion	74	Bases	50
Strengthen	171	Target	74	Strength	50
Current period	171	Efficiency	73	Economy	50
Parts	167	Efforts	73	Materials	50
Dividend	157	Reduction	72	New	50
Structure	154	Substantial	72	Energy	49
Sales	150	Forecast	71	Exchange	49
Society	146	Implementation	71	Industry	49
Field	146	Continued	70	Effect	48
Consolidated	145	Support	70	Manufacturing	48
Reform	140	Network	68	Input	48
Products	139	Research	68	Various	47
Sales	126	Severe	67	Stable	47
Plan	125	Policy	67	Area	47
Previous period	119	Functions	66	Accounting	46
Revenue	114	Region	66	Trust	46
Improvement	112	Large	65	Terminal	46
Figure	111	Change	65	Compare	46
Towards	108	Doing	64	Forecast	46
Service	107	Brand	63	Head	45
Value	107	Price	63	Sustain	44
Impact	105	Related	63	Focus	44
Realization	104	Situation	63	Contents	43
Invest	103	TV	62	Industry	43
Improvement	101	Company	62	Equipment	43
Demand	98	HDD	61	Quality	43
Response	98	Global	61	Game	42
Activity	95	Everybody	61	Spirit	42
Equipment	91	Measures	61	Constitution	42
Strategy	91	Have	61	All over the world	41
Aim	91	Competition	60	Loss	41
Thinking	88	Proactive	60	Hitting	41
Increase	88	Fiscal Year	60	Utilization	40
Tackle	87	Possible	59	Decrease	40
Electronic	87	Women	58	Information	40
Contribution	86	Transition	58	Steady	40

The table shows the top 50 words extracted from during committing fraud and after fraud detection of the fraudulent firm and non-fraudulent firms, respectively.

VI. HYPOTHESIS TESTING RESULTS

Test of Hypothesis 1

Table 4, Panel A presents the characteristic words in the pre-and post-fraud detection periods. It can be seen that many positive words such as "value improvement" and "profit structure improvement" are used during fraud. After fraud is detected, words related to fraud prevention such as internal management system and reform are found. Table 4, Panel B shows the characteristic words of Toshiba and non-fraudulent firms. In the case of Toshiba, more positive words such as "execution", "acceleration", and "improvement of earnings structure" are often used during fraud-execution period, while in the case of non-fraudulent firms, words for shareholders such as "product" and "market" are often used.

TABLE 4
Characteristic Words Analysis

Panel A:

Fraud-Execution Period of Toshiba		After Fraud Detection Period of Toshiba	
Terminology	Jaccard Index	Terminology	Jaccard Index
Management	0.1080	System	0.0912
Enterprise	0.0862	Reform	0.0714
Sector	0.0811	Promote	0.0588
Value	0.0745	management	0.0568
Global	0.0743	strengthen	0.0559
Quality	0.0728	Aim at	0.0556
Progress	0.0724	Stock Market	0.0541
Support	0.0724	Market	0.0535
Group	0.0710	Plan	0.0514
Improvement	0.0710	Measures	0.0514

Panel B:

Fraud Execution Period of Toshiba		Non-Fraudulent Firms	
Terminology	Jaccard Index	Terminology	Jaccard Index
Division	0.0755	Growth	0.0735
Execution	0.0588	Market	0.0598
Quality	0.0585	Products	0.0559
Profit and Loss	0.0563	Profit	0.0493
Supervision	0.0548	Expansion	0.0457
Management	0.0541	World	0.0432
Global	0.0539	Development	0.0378
Support	0.0519	Technology	0.0353
Acceleration	0.0486	Parts	0.0314
Improvement	0.0486	Production	0.0312

The Jaccard coefficient is calculated by dividing the number of co-occurrences by the number of occurrences in either direction, and indicates the strength of the co-occurrence. Words with high Jaccard coefficients are regarded as characteristic terms. It shows the characteristic words during and after detection of fraud periods. During fraud execution period, there are many positive words such as "value improvement" and "profit structure improvement. After fraud is detected, words related to fraud prevention such as internal management system and reform are seen.

Table 5 shows the time-series transition of the characteristic words during and after detection of fraud. During the period of fraud execution, many words related to supervision, governance, and profit structure are found as distinctive words. On the other hand, after fraud-detection, words related to fraud prevention such as internal management system and reform are found.

TABLE 5

Time-Series of Feature Words

Fraud Execution Period of Fraudulent Firm

2009		2010		2011		2012		2013		2014	
the term	Jaccard index	the term	Jaccard index	the term	Jaccard index	the term	Jaccard index	the term	Jaccard index	the term	Jaccard index
program	0.1304	Profit and loss	0.1429	Department	0.1389	Department	0.1765	directed by	0.1429	Outside the company	0.2069
center	0.1034	Exceed	0.1429	Surplus	0.1250	acceleration	0.1250	Fulfill	0.1154	Decision	0.1613
recovery	0.1000	Constitution	0.1250	Achievement	0.1176	Deployment	0.1212	Execution	0.1143	Execution	0.1351
performance	0.0968	Promotion	0.1200	Prophase	0.0938	global	0.1143	function	0.1111	directed by	0.1290
group	0.0962	Start	0.1000	acceleration	0.0882	Increased profit	0.1071	business	0.0968	Basic	0.1250
Execution	0.0909	Finance	0.0968	Improvement	0.0877	Strong	0.1034	management	0.0896	policy	0.1212
Early	0.0870	Decrease in sale	0.0952	Profit and loss	0.0833	infrastructure	0.0882	Group	0.0800	Committee member	0.1000
Constitution	0.0857	Construction	0.0909	global	0.0811	environment	0.0833	position	0.0800	important	0.1000
Strict	0.0833	Strive	0.0909	Kokoku	0.0741	Constitution	0.0769	rich	0.0741	management	0.0857
Age	0.0833	Improvement	0.0882	Electric applianc	0.0741	demand	0.0769	Governance	0.0645	the company	0.0833

After Fraud-Detection Period of Fraudulent Firms

2015		2016		2017		2018		2019	
the term	Jaccard index	the term	Jaccard index	the term	Jaccard index	the term	Jaccard index	the term	Jaccard index
management	0.1096	Advance	0.1500	System	0.0963	internal	0.1111	plan	0.1353
Trust	0.1053	market	0.0984	management	0.0820	The company	0.1017	investment	0.0922
Take	0.0938	Foundation	0.0962	Funding	0.0796	group	0.1013	Technology	0.0876
deep	0.0811	management	0.0952	stock	0.0738	Improvement	0.1013	Enterprise	0.0736
Accounting	0.0769	Operation	0.0870	caution	0.0667	System	0.0988	reform	0.0705
performance	0.0732	Thorough	0.0851	Special	0.0667	stock	0.0909	growth	0.0637
Contribution	0.0682	region	0.0816	Brand	0.0667	management	0.0896	Measures	0.0634
Work on	0.0638	Aim	0.0746	Designation	0.0661	Securities	0.0820	Revenue	0.0608
System	0.0615	Do	0.0741	Improvement	0.0657	Measures	0.0758	cyber-	0.0606
Reflection	0.0606	caution	0.0714	market	0.0630	Continue	0.0746	physical	0.0606

This table shows the time-series transition of characteristic words during and after fraudulent execution periods. The Jaccard index is calculated by dividing the number of co-occurrences by the number of occurrences of either one, and indicates the strength of co-occurrence. Words with a high Jaccard coefficient are regarded as characteristic words. Oversight, governance, and profitability can be seen during fraudulent execution period. After detection of fraud period, there are words related to fraud prevention such as internal control system and reform.

Table 6 shows the t-test results for the Jaccard coefficients during and after fraud detection and for Toshiba and non-fraudulent firms. From Panel A of Table 6, I observe a significant difference in the Jaccard coefficients between during fraud execution period and after detection of fraud period of Toshiba, and from Panel B of Table 6, I observe a significant difference in the Jaccard coefficients between Toshiba and non-fraudulent firms. The total number of sentences in during fraud execution periods is 146, and the total number of sentences of occurrences of "management" both during fraud execution period and after fraud detection period is 49. Focusing on the number of sentences of occurrences of "management" during fraud execution period is 19, and the co-occurrence of "management" during fraud execution periods is 19. In this case, the Jaccard coefficient is computed as follows:

$$\text{Jaccard Coefficient} = 19/49 + 146 - 19 = 0.108.$$

TABLE 6

The Results of t-test

Panel A: In the Pre-and Post-Fraud Detection Periods of Toshiba

	Fraud Execution Period		After Fraud Detection Period				
	MEAN	S.D.	MEAN	S.D.	t-value	significance	
Jaccard Index	0.105	0.027	0.083	0.019	5.083	0.000	***

Panel B: Fraudulent Firms and Non-Fraudulent Firms

	Fraudulent Firm		Non-Fraudulent Firms				
	MEAN	S.D.	MEAN	S.D.	t-value	significance	
Jaccard Index	0.056	0.008	0.046	0.014	1.990	0.066	*

*, **, *** indicate significance at the $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively.

Table 7 presents the descriptive statistics of the Jaccard coefficients in the pre-and post-fraud detection periods and non-fraudulent firms. The Jaccard coefficients during fraud execution period are highest among them. Taken together, I find that the Jaccard coefficients of fraud execution period are higher than those after fraud detection period, and more distinctive words are used during fraud execution period. Therefore, H1a "There is a difference in the word use of Fraudulent firm in the pre-and post fraud detection periods" and H1b "there is a difference in the word use between Fraudulent firm and non-fraudulent firms." are supported.

TABLE 7
Descriptive Statistics of The Jaccard Index for Each Period

	Mean	Median	S.D.	Min.	Max	Lower Quartile	Upper Quartile
Post-Fraud Detection Period of Toshiba	0.0532	0.0510	0.0086	0.0480	0.0770	0.0488	0.0530
Fraud Execution Period of Toshiba	**0.0549**	**0.0540**	**0.0078**	**0.0480**	**0.0750**	**0.0495**	**0.0570**
Non-Fraudulent Firms.	0.0526	0.0475	0.0125	0.0390	0.0750	0.0430	0.0623

The Jaccard index is calculated by dividing the number of co-occurrences by the number of occurrences in either direction, and indicates the strength of the co-occurrence. Words with high Jaccard coefficients are regarded as characteristic words.

Test of Hypothesis 2a and Hypothesis 2b

In order to test H2, correspondence analysis is conducted. Figure 1 shows the results of the correspondence analysis for Toshiba's fraud-execution period, post-fraud detection period, and non-fraudulent firms. The results of the correspondence analysis show that during the fraud-execution period of Toshiba, the words are far away from the origin and the words are also far away from each other, indicating unrelated words, while after the fraud is detected, the words are close to the origin and the words are placed close to each other.

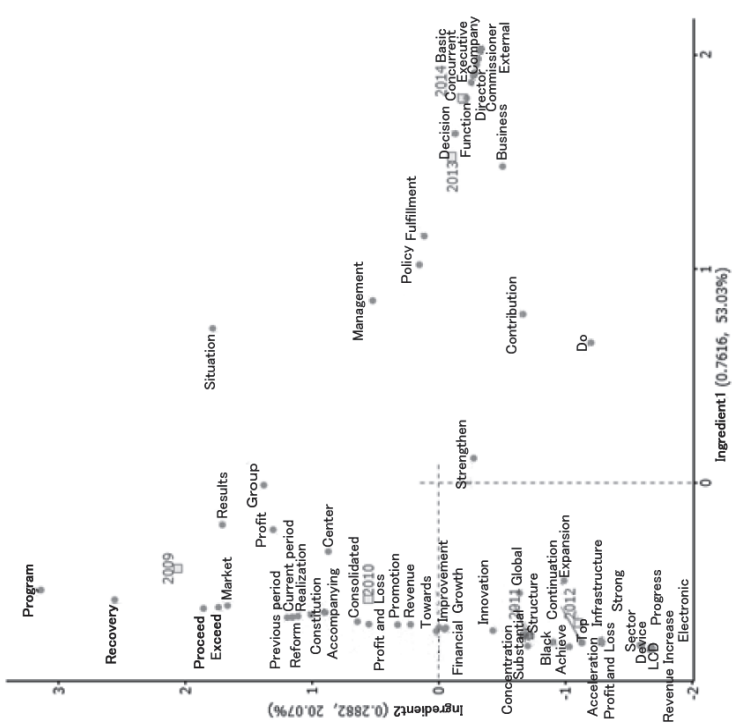
Figure 2 shows the results of the correspondence analysis of Toshiba and non-fraudulent firms by year.[6] It can be seen that the non-fraudulent firms use general words, close to the origin from 2009 to 2014. From Figure 3, it can be observed that during fraud execution, the words are far from the origin and distinctive words are used instead of general words. After fraud-detection period, the plots of the words are closer to the origin. Through the results of the correspondence analysis described above, it is observed that while more distinctive words are used during fraud execution period, more common words are used after fraud-detection period. In addition, the words used by Toshiba are more distinctive than those used by non-fraudulent firms. Taken together, there are differences in the positions of the plots of the words before and after fraud detection is detected periods and between Toshiba and non-fraudulent firms. Thus, H2a and 2b are supported. In other words, it can be said that management may have misled investors by using the impression management during fraud-execution period.

[6] The results of correspondence analysis by firm also plot closer to the origin, as do the results of correspondence analysis by year. The reason why Canon is far from the origin is that CEO's message is included in the Sustainability Report, and therefore, the focus of textual analysis is slightly different from that of other firms' Annual Reports.

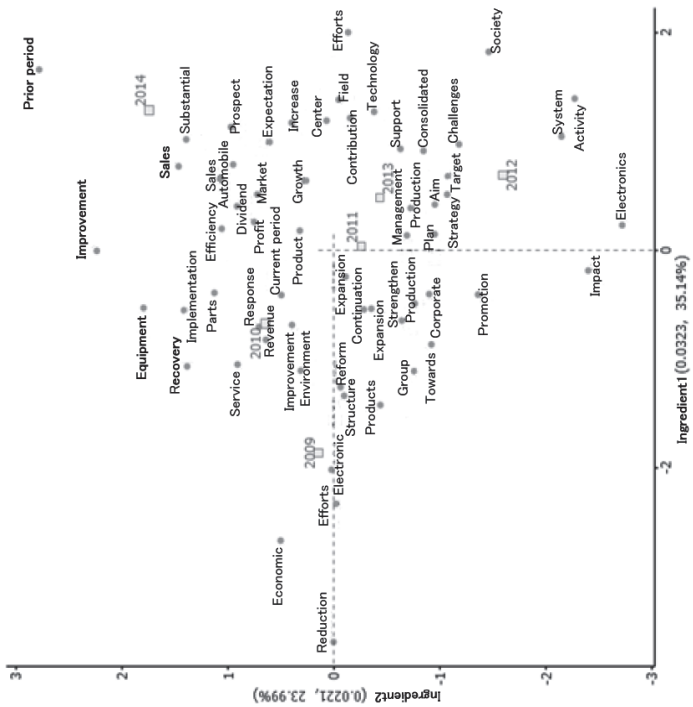
FIGURE 2

The Results of Correspondence Analysis

Fraud Execution Period (Per Year)



Non-Fraudulent Firms (Per Year)

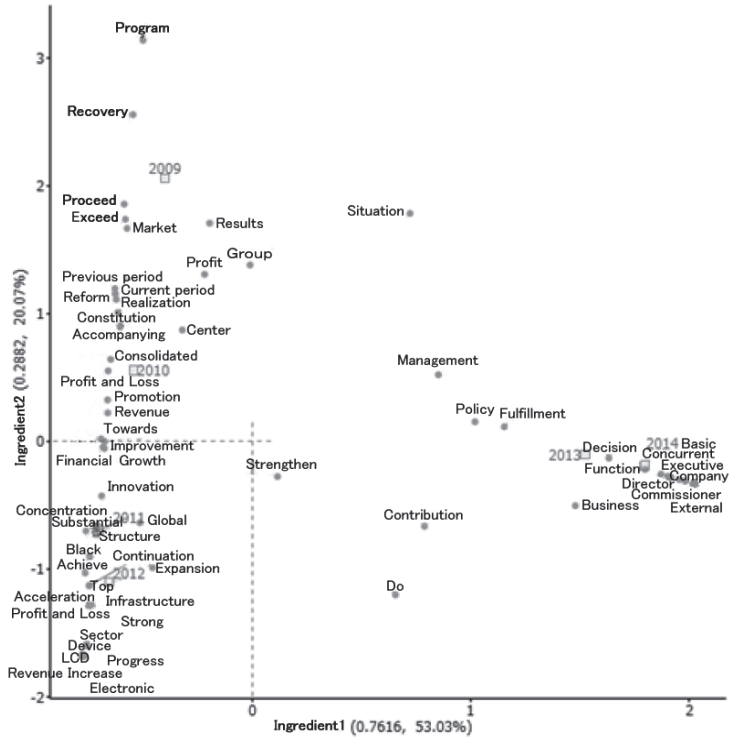


This figure shows a comparison of the correspondence analysis results of Toshiba's fraudulent execution and non-fraudulent firm. From the results during fraud execution periods, it is observed that the words of each year are located far from the origin, and that the words of non-fraudulent firms are located close to the origin.

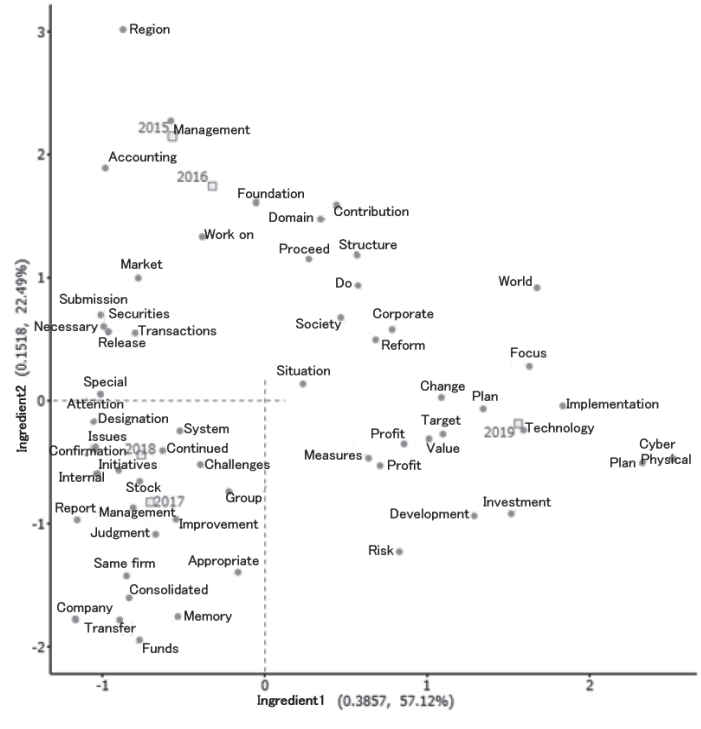
FIGURE 3

The Results of Correspondence Analysis in the Pre-and Post-Fraud Detection Period

Fraud Execution Period



After Fraud Detection Period



This figure presents a comparison of correspondence analysis results before and after fraud detection periods. From the results during fraud execution, it is observed that the words of each year are located far from the origin, and that they are located close to the origin before and after fraud is detected.

VII. CONCLUSION AND FUTURE RESEARCH

This study determines whether managers manage text by examining whether the textual characteristics of CEO letters differs in the pre-and post-fraud detection periods of Toshiba, and between Toshiba and non-fraudulent firms. First, I find that the Jaccard coefficient during fraud execution period of Toshiba is higher than that after fraud-detection period, indicating that Toshiba uses more distinctive words during fraud execution period. Second, the results of the correspondence analysis show that the words are far from the origin, indicating unrelated words are placed during fraud-execution period of Toshiba, and the words are closer to the origin and the words are placed close to each other after fraudulent detection period of Toshiba, and the words of non-fraudulent firms are plotted close to the origin. In other words, the Jaccard coefficients of fraudulent firm are higher during the fraudulent period than after fraud is detected, and the words are plotted farther from the origin.

This suggests that during fraud-execution period, managers may have misled investors by using impression management theory to make the text more difficult to read. This study can be presented as a possibility for fraud detection when it is observed that CEO letters have distinctive words from those of non-fraudulent firms through text analysis. Nakashima (2022) presents a method for detecting fraud by implementing a combination of both textual analysis and financial statement figures, but this study is able to present that fraud can be detected by textual analysis alone. There are several limitations to this study. Due to the small sample size, this study is only a presentation of a pilot study. In the future, it is necessary to test whether managers implement impression management by measuring tone from narrative disclosures.

REFERENCES

Ajina,A. M. Laouiti, and B. Msolli. 2016. Guiding through the Fog: Does annual report readability reveal earnings management? *Research in International Business and Finance* 38:509-516.

Amernic, J., Craig, R., Tourish, D. 2010. Measuring and Assessing Tone at the Top Using Annual Report CEO Letters. Edinburgh: Institute of Chartered Accountants of Scotland.

Bloomfield, R. J. 2002. The "incomplete revelation hypotheses" and financial reporting. *Accounting Horizons*, 16:233-243.

Bushee, B., I. D. Gow., and D. J. Taylor. 2017. Linguistic complexity in firm disclosure: obfuscation or information? *Journal of Accounting Research* 56 (1): 1-49.

Carlsson, S. and R.Lamti. 2015. Tone management and earnings management: A UK evidence of abnormal tone in CEO letters and abnormal accruals. Master Thesis, University of Gothenburg, School of Business Economics and Law.

Clatworthy, M.and M.J.Jones.2003.Financial reporting of good news and bad news: evidence from accounting narratives, *Accounting and Business Research* 33 (3):171-185.

Clatworthy, M.and M.J.Jones.2006. Differential patterns of textual characteristics and company performance in the chairman's statement, *Accounting & Accountability Journal* 19 (4):493-511.

Clausen, S.E.1998. *Applied Correspondence Analysis: An Introduction*, Sage Publication, inc.

Committee of Sponsoring Organizations of the Treadway Commission (COSO). 1994. *Internal Controls Integrated Framework, Executive Summary Framework Reporting to External Parties Addendum to "Reporting to External Parties*." Jersey City, NJ: American Institute of Certified Public Accountants.

Courtis, J.K.1998. Annual report readability variability: tests of the obfuscation hypothesis, *Accounting, Auditing & Accountability Journal* 11(4): 459-471.

Cunningham, C 2005. Section 404 compliance and 'tone at the top'. *Financial Executive* 21(5): 6.

Demaline, C. 2020. Disclosure readability of firms investigated for books-and-records infractions An impression management perspective, *Journal of Financial Reporting and Accounting* 18 (1): 131-145.

Financial Accounting Standard Board (FASB). 2010. *Statement of Financial Accounting Concepts No. 8 September 2010 Conceptual Framework for Financial Reporting. Chapter 3, Qualitative Characteristics of Useful Financial Information*.

Habermas, J. 1984. *The Theory of Communicative Action*, Volume One, Reason and The Rationalization of Society translated by Thomas McCarthy, Beacon Press Boston.

Hasana, M.M., B.A.Abbas and S.N Garas..2018. Readability, governance and performance: a test of the obfuscation hypothesis in Qatari listed firms, *Corporate Governance* 19 (2):270-298.

Hasana, M.M. A.Habib.2020. Readability of narrative disclosures, and corporate liquidity and payout policies, *International Review of Financial Analysis* 68: 1-15.

Henry, E. 2008. Are investors influenced by how earnings press releases are written? *Journal of Business Communication* 45 (4):363-407.

Higuchi, K. 2020. *Quantitative Text Analysis for Social Research*, Nakanishiya Publisher. (in Japanese)

Hirose, Y., H. Hirai.and K.Arai 2017. Effects of readability of MD&A information on future performance: The analysis of text mining. *The Journal of Business Analysis* 33:87-101. (in Japanese)

Huang, X., S. H. Teoh., Y. Zhang. 2014. Tone management, *The Accounting Review* 89 (3):1083-1113.

Jensen, M. C. and W.H. Mecking1976. The theory of the firm: managerial behavior, agency costs, and ownership structure, *Journal of Financial Economics* 3 (4):305-360.

Leary, M.R. and R. M. Kowalski.1990. Impression management: A literature review and two component model, *Psychological Bulletin* 107 (1): 34-47.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of accounting and economics* 45 (2):221-247.

Linsley, P.M. and Lawrence, M.J.2007. Risk reporting by the largest UK companies: readability and lack of obfuscation, *Accounting, Auding & Accountability Journal* 20 (4): 620-627.

Lo, K., F. Ramos, and R.Rogo. 2017. Earnings management and annual report readability, *Journal of Accounting and Economics* 63: 1-25.

Merkle-Davies, D. 2007. The obfuscation hypothesis re-examined: Analyzing impression management in corporate narrative report documents, *Bangor University (United Kingdom). ProQuest Dissertations Publishing*, 2007.

Merkle-Davies, D. and Brennan, N. 2007. Discretionary disclosure strategies in corporate narrative incremental information or impression management? *Journal of Accounting Literature* 26: 116-122.

Nakashima, M. 2020. Relationship between perceived managerial ethical value and earnings quality: Evidence from Japan, *Proceeding of 2020 Taiwan Accounting Association Annual Conference*.

Nakashima, M. 2022. Fraud dection from tone and financial fundamentals from CEO letters: Evidence from the Japanese firms listed on the U. S. Stock Exchanges, *Accounting and Audit Journal* 808 (19):104-114 (in Japanese).

Nakashima, M., H. Hirose, and H. Hirai. 2022a. Fraud detection by focusing on readability of MD&A disclosur: Evidence rom Japan, *Journal of Forensic and Investigative Accounting* 14 (2): 276-298.

Nakashima, M., Y. Nose, H. Ujiie, and S. Yoshida. 2022b. Accounting fraud detection through a text-mining analysis: Evidence from construction and retail industries in Japan, *Business Review, Faculty of Business Administration, Bunkyo Gakuin University* 31 (1): 117-136.

Ohno S. 2020. Are you really telling the truth?" Reading between the lines of fraud MD&A information. *Proceedings of The 2020 Forensic Accounting Section Conference, American Accounting Association*.

Patelli, L.and M.Pedrini. 2014. Is the optimism in CEO's letters to shareholders sincere? Impression management versus communicative action during the economic crisis, *Journal of Business Ethics* 124: 19–34.

de Souza, JAS. And J. C. Rissatti, S. Rover, and J.A. Borba. The linguistic complexities of narrative accounting disclosure on financial statements: An analysis based on readability characteristics, *Research in International Business and Finance* 48: 59-74.

Takamura, H., T.Inui, M. Okumura. 2005. Extracting semantic orientations of words using spin model, In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL2005) :133-140.

Thomas, J. 1997. Discourse in the Marketplace: The making of meaning in annual reports, *Journal of Business Communication* 34 (1): 47-66.

Yuthas, K., R. Rogers, and J. F. Dillard. 2002. Communicative action and corporate annual reports *Journal of Business Ethics* 41: 141–157.