DEEP LEARNING APPLICATIONS IN AUDIT DECISION MAKING

By TING SUN

A dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

in partial fulfillment of requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Professor Miklos A. Vasarhelyi

and approved by

_____

Dr. Miklos A. Vasarhelyi

_____

Dr. Alexander Kogan

_____

Dr. Helen Brown-Liburd

_____

Dr. Rajendra P. Srivastava

Newark, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Deep Learning Applications to Audit Decision Making

By Ting Sun

Dissertation Director:

Professor Miklos A. Vasarhelyi

The objective of this dissertation is to investigate whether the sentiment features of business communication documents or social media information extracted by deep learning techniques deliver relevant and reliable information to auditors.

The first essay investigates the incremental informativeness of sentiment features of earnings conference calls for the prediction of internal control material weaknesses (ICMW). With the help of a deep learning textual analyzer provided by IBM Watson, Alchemy Language API, this essay obtains the overall sentiment score of the text and the confidence score of the emotion "joy." These sentiment features are then used as additional predictors along with other determinants of ICMW suggested by prior literature (i.e., Doyle, Ge, and McVay, 2007a; Ashbaugh-Skaife, Collins, and Kinney, 2007). The results indicate that these sentiment features, especially the score of joy, improve the explanatory ability and the prediction accuracy of the model.

The second essay compares deep learning to the "bag of words" approach and demonstrates the effectiveness and efficiency of deep learning-based sentiment analysis for MD&A sections of 10-K filings in the context of financial misstatement prediction. The findings include (1) sentiment features provide insights for financial misstatement prediction, primarily for fraud detection; (2) the model using deep learning-based

sentiment features generally performs more effectively than the model using sentiment features extracted by the "bag of words" approach.

The third essay examines how the information of tweeting activities about the client company is associated with the audit fee. It examines the relationship between the audit fee of U.S. public firms in 2015 and the properties of tweets about the client firm: the sentiment of tweets, the volume of tweets, and the popularity of tweets. All tweet information is obtained using IBM Twitter Insights, a Twitter data analysis tool that provides sentiment and other enrichments relying on deep learning algorithms. It finds that for companies without going-concern audit opinions and companies with a median level of restatement risk, the audit fee is positively associated with the frequency of negative tweets, and this association is strengthened for companies receiving more retweets than those receiving less retweets.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of my committee members, my colleagues, and my family.

I would like to express my deepest gratitude to my advisor, Dr. Miklos A. Vasarhelyi for his guidance, support, and encouragement over the years. He led me to the world of Accounting Information Systems, opened my eyes to new stages of opportunity and strength, and guided me towards the right path. It would be impossible to count all the ways that he has helped me during my Ph.D. studies. I wish to express my sincere thanks to Dr. Alexander Kogan for all the time and effort he gave to me to help me solving the technical problems of my research. I deeply thank Dr. Helen Brown-Liburd for her constant help in developing research ideas and her invaluable suggestions for this dissertation. I would like to express my heartfelt thanks to Dr. Raj Srivastava for providing such an informative database, SeekiNF, for my research and serving on the committee with insightful comments. I would like to thank Dr. Michael Alles, Dr. Kevin C. Moffitt and Dr. Soo Hyun Cho for their important remarks. Especially, I would like to express my gratitude to Dr. Dan Palmon for his kindest support as department chair and Barbara Jensen for her help and kindness.

Thanks also go to my friends and colleagues at Rutgers: Ahmed Al-Qassar, Deniz Applebaum, Tiffany Chiu, Mauricio Codesso, Jun Dai, Jamie Freiman, Feiqi Huang, Hussein Issa, Qiao Li, Yue Liu, Andrea Rozario, Yunsen Wang, Zhaokai Yan, Cheng Yin, et al. It was a great pleasure to work with all of them.

To my husband, thank you for your unconditional love and support. Without you, I would not be where I am today. To my parents, thank you for believing in me and for all

of your help with my daughter when I needed it the most. To my daughter, Skylar, you are my inspiration to achieve greatness. No words can describe my great gratitude to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1 Introduction

This dissertation consists of three applications of deep learning, an innovative Artificial Intelligence technique, to auditors' decision making. The first chapter provides a brief introduction to deep learning, analyzes the need for deep learning for audit decision making, and discusses the motivation as well as the main research questions of this thesis. Chapters 2 through 4 examine whether and how deep learning assists auditors in assessing the risk of internal control material weakness and financial misstatement, and to determine the audit fee. The last chapter concludes the thesis by summarizing the main findings, discussing limitations, and providing directions for future research.

1.1. An Overview of Deep Learning

Deep learning, also called deep neural network (DNN), develops hierarchical artificial neural networks consisting of layers of neurons. Many tasks, such as image recognition and natural language processing(NLP), that are easy for human beings were extremely hard for a computer (Goodfellow, Bengio, and Courvile, 2016). Recently, due to the accelerated improvement in data storage and the computational capability of a modern computer (e.g., cloud computing), a DNN trained with a large volume of data can represent more and more complex functions. Compared to a traditional neural network, a DNN has more consecutive hidden layers and more neurons within each layer. This structure allows the neural network to identify high-level and abstract data features from the raw data. Specifically, the more complex data features identified by a successive layer are built upon the other, simpler data features extracted by the predecessor layer. Such a

data transition and transformation process through multiple layers of neurons makes a DNN a "thinking" machine. A simplified example of DNN is presented in Figure 1.1. It consists of one input, three hidden, and one output layers. Each layer applies a nonlinear transformation to its input layer and provides a representation. In other words, the output representation of each input layer is provided as input to its next layer. As the input data goes deeper, the nonlinear transformation constructed becomes more complex, and the representation becomes more abstract. The output of the last layer is the final representation of the raw input data, which is the high-level features extracted from the data. The extracted features are useful for further classification, association, and other future tasks (Najafabadi, et al., 2015).

Besides applications of text understanding, image identification, and speech recognition, deep learning technology has led to more complex breakthroughs. For example, AlphaGo, a deep learning system developed by Google, defeated professional champions from Europe, South Korea, and China at the Game of Go by learning from thousands of human amateur and professional games[1]. Most recently, a new version of AlphaGo, AlphaGo Zero learned how to play the game simply by playing games against itself, starting from completely random play rather than learning from past examples.

For sentiment analysis of text data, an example of data processing is as follows. Firstly, the text is transferred through a "shallow" neural network called *Word2vec* to vector sets that numerically represent the content of each word to make the text machine-readable. During this process, the vectors are classified into clusters based on the

[1] For more information about AlphaGo, refer to https://deepmind.com/research/alphago/

mathematical similarities, which facilitates the follow-up sentiment analysis by a DNN. The output of the analysis of *Word2vec* is a vocabulary in which each word of the text is attached with a vector (also called neural word embedding) (deeplearning4j, 2017). Then the vectors are fed into a DNN (e.g., a temporal convolutional network) that further extracts the features of the input data layer by layer and finally classifies the sentiment (e.g., positive, negative, and neural) within a text document (Zhang, Zhao, and LeCun, 2015).

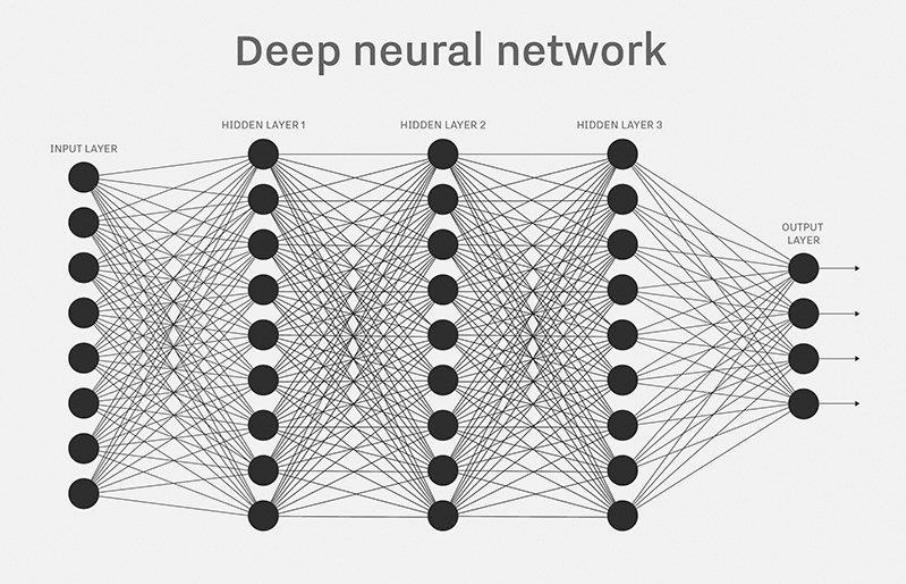

Figure 1.1 A Simplified Deep Neural Network

(Adopted from Nielsen, 2015)

1.2. The Need of Deep Learning for Audit Decision Making

The increasingly developed world necessitates openness to adopt modern data-intensive technologies. Business applications (such as ERP systems), RFID readers, sensors, cloud storage, social media, remote communication tools (such as Skype, live

streams, podcasts), and other front-line technologies have been integrated into the business daily life. It contributes heavily to the production and maintenance of massive amounts of unstructured or semi-structured big data (National Research Council, 2013). As a supplement to traditional structured financial data, unstructured data contains information from various perspectives and sources, facilitating business to explore the status of their products, services, and operations. This is because big data provides more reliable evidence and makes auditors less client data-dependent (Yoon, Hoogduin, and Zhang, 2015). Consequently, examining and extracting meaningful patterns from big data offer insights for audit decision making (Sun and Vasarhelyi, 2017). However, big data analytics is not easy due to the following reasons: (1) the vast majority of big data is semi-structured or unstructured, requiring human experts' efforts of labelling and classification; (2) the volume of the data is too large to be processed manually; (3) big data is usually generated on a real-time basis, which requires timely responses; and (4) big data is complex as it has a variety of data types and comes from different sources. Thus, to understand the data, one must have a related background of knowledge or skills. A survey conducted by AICPA (2014) shows that big data analysis is regarded as one of the top challenges in the future by a quarter of 180 CPA participants. Therefore, to make big data useful and usable for decision making auditors, who usually lack professional data mining and information system knowledge and skills, need an efficient and effective approach to automate audit procedures (Sun and Vasarhelyi, 2017). With deep learning, they could simply use a DNN pre-trained and tested by deep learning specialist along with their own professional accounting judgment and enjoy its benefits.

Some audit tasks are tedious and complex. The automation of such tasks will significantly enhance effectiveness and efficiency of audit work (Raphael, 2015). Trained with sufficiently large samples about how auditors make decisions under different circumstances (which can be realized by providing different values of data attributes), a DNN enables auditors to automate many structured or semi-structured tasks that have been conducted manually for decades, like checking inventories, processing paperwork, reviewing contracts, and drafting audit reports. Even for certain risk assessment activities requiring professional judgment (also called unstructured audit tasks), deep learning provides a new way to support audit decisions. For instance, items in a financial statement or other financial records can be scanned and automatically linked to related evidence, such as images of inventory captured by the webcam, shipping documents, sales invoices, bank confirmations, auditor working papers, and other supporting documents that have been identified and classified by deep learning systems. Furthermore, a list of risky items or even recommended responses can be offered.

1.3. Motivations, Research Questions, and Methods

Conventional data mining techniques are often found inadequate to analyze and extract insightful features from big data, due to its massive size and high dimensionality (Jin, Wah, Cheng, and Wang, 2015). In the audit profession, the Big Four accounting firms have invested hundreds of millions of dollars in deep learning and other AI techniques. KPMG formed an alliance with IBM Watson to develop AI tools for bank loan evaluation. Other auditing firms also have their own high-tech tools, such as Argus and Optix for Deloitte (Rapoport, 2016). In the academic area, however, limited research

examines the issue of applying deep learning to auditing although the last several years have seen many successful applications of deep learning in the area of big data analytics (Google speech team, 2015; Silver et al., 2016; Hinton et al., 2011, Zhang, Zhao, and Lechun, 2015). This thesis aims to extend the application of deep learning to the audit domain and bridge the research gap by exploring the potential of this technique to ascertain valued insights for enhanced decision making of auditing. Specifically, it examines the role that deep learning plays in audit decision making by investigating how sentiment features of audit-related textual data extracted by deep learning algorithms help identify internal control material weakness (ICMW), financial misstatement, and audit fees. Figure 1.2 presents the research design of this dissertation.

The first essay (chapter 2) applies a DNN provided by IBM Watson and trained with more than 200 billion words within a broad domain coverage (Turian, 2015) to extract the overall sentiment and the strength of emotion joy in earnings conference calls. This paper aims to investigate whether the sentiment features provide incremental information for the prediction of ICMW. It examines the explanatory ability of the prediction model and provides empirical evidence that the sentiment features significantly increases the model fitness. Next, using Logistic Regression, Random Forest, and traditional Artificial Neural Network, this research builds classification models and reports that the prediction accuracy, as measured by AUC, false positive rate, false negative rate, and the overall accuracy, improves after using the sentiment features. Besides the existence of ICMW, this essay also explores whether the sentiment features of conference calls are related to the number of ICMW, and whether companies that persistently report ICMW have lower overall sentiment score and joy score. It develops multinomial logistic regression to test

the association between the sentiment features and the existence of single ICMW vs. multiple ICMWs. It also establishes logistic regressions with alternative dependent variables (representing first year ICMW and persistent ICMW) to analyze the effect of sentiment on the persistency of ICMW. The results of both tests support the conjecture that the sentiment features function more effectively in identifying companies with more than one ICMW than companies with only one ICMW. In addition, the sentiment features are more likely to be associated with companies that persistently have ICMW.

Using the same deep learning-based sentiment analysis tool, essay 2 (chapter 3) focuses on the MD&A of 10-K filings and mainly answers three research questions: (1) do the sentiment features of MD&As add information for financial misstatement prediction? (2) if the answer is yes, are they effective for fraud prediction only or both fraud and error? (3) how effective is the model, with sentiment features obtained with deep learning technique, compared to the model using sentiment feature calculated with "bag of words" approach?

Utilizing five machine learning algorithms, essay 2 develops 45 classification models under three types of model structures to conduct three predictions tasks (including predict frauds, errors, and misstatements). Other than the sentiment attributes, 82 misstatement predictors suggested by prior literature (Perols, Bowen, Zimmermann, and Samba, 2017; Dechow et al., 2011; Perols, 2011; Cecchini et al, 2010; Beneish, 1999; Huang, Rose-Green, and Lee, 2012; Churyk, Lee, and Clinton, 2009) are added into those models. The prediction results show that the sentiment features of MD&As enhance the predictive performance of the classification models. Furthermore, for the task of

predicting frauds, the classification model with deep learning-based sentiment features, especially the emotion, outperforms the one using bag-of-words. The results indicate that deep learning is an effective sentiment analysis technique for financial fraud detection. However, for the task of error prediction and hence the misstatement identification, it does not perform as effectively as it does for fraud detection.

While the first two essays analyze finance-specific text, the last essay (chapter 4) sheds lights on a major social media platform, Twitter, which is a publicly available information source. Due to the ease of use, high speed, and wide reach, social media plays an increasingly important role in information sharing and social networking (Asur and Huberman, 2010). It is gradually changing the nature of communication among users (Cong and Du, 2007; Kaplan and Haenlein, 2010; Du and Jiang, 2015). In the business area, social media platforms, such as Facebook, Twitter, Pinterest, LinkedIn, Tumblr, Google+, as well as their competitors, allow stakeholders to create, bookmark, share, and comments on content, which creates enormous, various, and valuable data. Certain information on Twitter reveals the company's potential litigation, deteriorating reputation, increased business risks, internal control deficiencies, unethical behaviors of the chief executives, inappropriate business strategies, and etc., For instance, customers' complaints on a product's quality or poor customer services can predict a downward sales revenue or profitability, which creates incentives for the company to commit financial fraud (Kreutzfeldt and Wallace, 1986). Therefore, social media provides a wealth of useful information for the auditor to establish the "frame of reference." The third essay aims to explore the value of information delivered or suggested by tweets for the identification of companies' risk related to the audit engagement. In particular, it analyzes

the sentiment feature and other properties of tweets and the association between the properties of tweets and the audit fee. This research hypothesizes that the more negative tweets which are posted discussing the client company, the higher the audit fee will be. Furthermore, as the number of retweets measures the popularity of certain topics about the company on Twitter. The second hypothesis is that the association between negative tweets and audit fees is stronger for companies with more retweets. This study uses a tool powered by deep learning, Twitter Insights, to test these two hypotheses. The empirical results support the hypotheses for companies without going-concern opinion and companies with a median level of financial misstatement risk. In other words, tweets are less likely to accurately reflect the audit risk of a company when it is considered to be extremely risky or when its going-concern status is threatened. The last chapter draws conclusions, summarizes the limitation of this dissertation, and provides directions for future research.

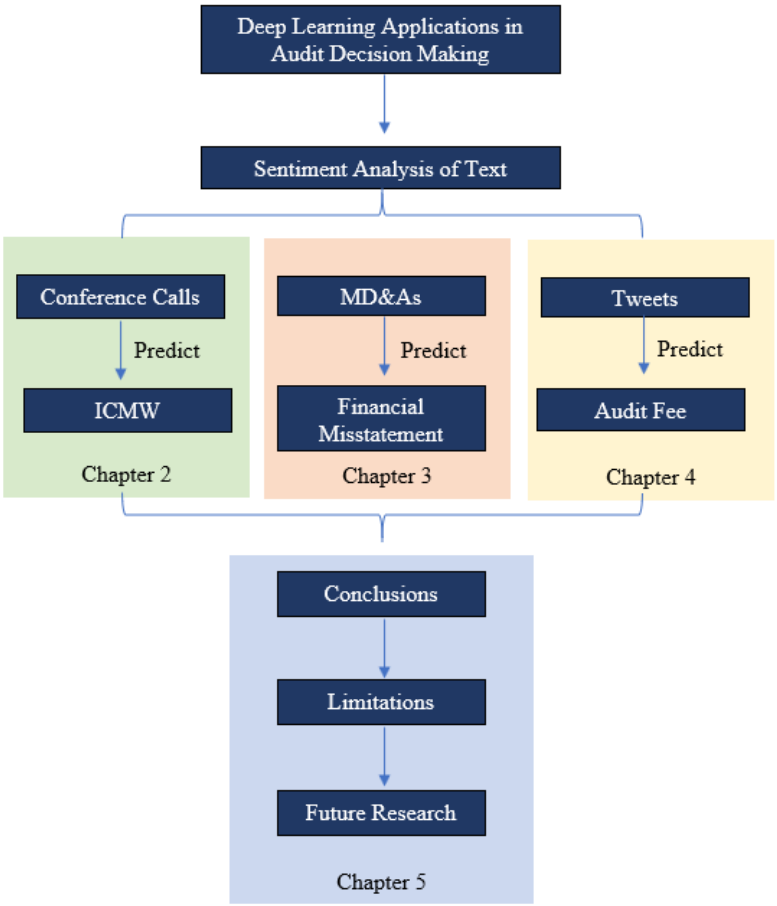

Figure 1.2 The Research Design of this Dissertation

Chapter 2 The Incremental Informativeness of Management Sentiment in Conference Calls for Internal Control Material Weaknesses

2.1. Introduction

Effective internal control provides reasonable assurance for the reliability of financial reporting (PCAOB, 2007). A material weakness in a company's internal control over financial reporting (ICFR) indicates that there are one or more control deficiencies in the design or operation of internal control that create a reasonable possibility of a material misstatement (PCAOB, 2004). Section 404 of Sarbanes–Oxley Act of 2002 requires the disclosure of any material weaknesses identified by auditors in the annual report and, more specifically, the report of management on the company's ICFR should be concluded ineffective by the auditor. In order to form a basis for the opinion on the effectiveness of the ICFR, auditors plan and perform procedures to obtain appropriate evidence regarding the existence of material weakness in internal control (PCAOB, 2007). On the other hand, since the audit of internal control should be integrated with the financial statement audit, deficiencies in the former will affect the effectiveness of the latter as auditors often rely on controls to reduce the substantive testing of financial statement accounts and disclosures. Therefore, the design of an audit plan regarding internal controls, especially determining what evidence auditors will collect and how to use them, plays a key role in helping auditors form a fair opinion.

Despite the fact that Auditing Standard No. 5 provides guidance in conducting ICFR audits, the quality of the ICFR audit is unsatisfactory due to information asymmetry.

Observations from SEC's annual reviews of registrants' disclosures of material weaknesses show that while the percentage of companies with ineffective ICFR and a material weakness has been increasing, the percentage of companies reporting ineffective ICFR has been decreasing (Besch, 2009). The PCAOB points out that, the ICFR audit deficiencies are the most frequent findings in their inspections of audit work over the last few years. Even for the audits by the Big Four firms, they have observed an upward trend in the percentage of ICFR audit with deficiencies from 2010 to 2013. For instance, the PCAOB discovered 36% of integrated audits have ICFR audit deficiencies in 2013 (Franzel, 2015). To improve the effectiveness of ICFR, it is necessary to reduce information asymmetry by exploring the value of the massive volume of data that resides inside and outside of corporate boundaries. Thus, big data, once it has been processed and analyzed, can be considered as supplementary audit evidence (Yoon, Hoogduin, and Zhang, 2015).

Earnings conference calls are considered as the main communication venue between companies and all interested parties (i.e., Frankel, Johnson, and Skinner, 1999; Skinner, 2003; Corbin Perception, 2015; Bushee, Matsumoto, and Miller, 2003; Chen, Demers, and Lev, 2016). In an earnings conference call, senior executives (e.g., CEO and CFO) brief the participants on earnings and information that is relevant to the industry. During a conference call, the management provides an overview of all major issues that affect the company's performance, highlights the business successes, and finally answers informed questions from analysts and investors. In particular, management's presentation conveys and reinforces their opinion on the current business situation and its implications for future performance (Allee and DeAngelis, 2015; Sedor, 2002). Existing research has

identified the importance of management's speeches. An article in *The Atlantic* discusses the hidden messages in earnings conference calls and states that "earnings-report presentations supposedly present hard numbers but listening for the right words can be much more revealing" (Lam, 2015).

Previous studies show that conference calls contain incremental information beyond mandated disclosures for the current and future situation of the company. For example, Druz, Wagner, and Zeckhauser (2015) find that the "tone surprise" of the conference call, which is the residual when negativity in managerial tone is regressed on the firm's recent economic performance and CEO fixed effects, can predict a company's future earnings and analyst uncertainty. The information in conference calls is also used to help auditors predict financial misstatements. Hobson, Mayew, and Venkatachalam (2012) document that cognitive dissonance in the speech of CEOs on conference calls is diagnostic of adverse misreporting. Despite a large number of studies exploring the use of conference calls for the prediction of future performance and financial reporting quality, prior literature has not examined whether the information in conference calls implies the material weakness of internal control. This may be due in part to the fact that internal control issues are rarely mentioned in an earnings conference call[2].

Although internal control is rarely mentioned directly in a conference call, the effectiveness of ICFR concerns both investors and managers, because the existence of internal control material weakness (ICMW) implies the financial reporting of the company is problematic. Researchers assert that "an adverse ICFR opinion signals that a

[2] For example, among the 1651 observations in the final sample in this study, only 28 observations mentioned internal control or internal control related words (or phrases).

misstatement may exist in the financial statements investor reply on to make investor decisions" (Jennings et al. 2008; Ashbaugh-Skaife et al. 2009; Wu and Tuttle 2014; Barr-Pulliam, Brown-Liburd, and Sanderson, 2017). For those with ICMW, investors perceive higher information asymmetry, lower financial statement transparency, higher risk premium, lower sustainability of earnings, and lower earnings predictability (Lopez, Vandervelde and Wu 2009). Consequently, the market negatively reacts to the disclosure of internal control weakness, in terms of reduced share prices (Hammersley, Myers, and Shakespeare 2008) and higher cost of capital (Ashbaugh-Skaife et al. 2009). As for the management, since it is primarily the management's responsibility to design and maintain the internal control system (PCAOB, 2007), the presence of ICMW suggests that the management failed to fulfill their responsibilities, and the senior executives (e.g., CEOs and CFOs) are held accountable for their actions. Research shows that an adverse ICFR opinion leads to increased management turnover (Johnstone, Li, and Rupley 2011). Also, Hoitash, Hoitash, and Johnstone (2012) provide evidence that ICMW disclosures are negatively related to the change in CFO total compensation, bonus compensation, and equity compensation, especially for firms with stronger governance oversight.

Since the effectiveness of ICFR is important to both the investor and the management, the knowledge of the existence of ICMW in the company may affect the way the management speaks. Thus, capturing linguistic clues underlying the conference calls is important for ICMW prediction. Prior research in social psychology supports this. The leakage hypothesis (Ekman and Friesen, 1969) states that the act of deception makes a person feel guilty, stressed, and fearful of detection. DePaulo, Rosenthal, Rosenkrantz, and Green (1982) and Kraut and Poe (1980) assert that a person may experience

relatively heightened cognitive processing when telling a lie than telling the truth. This heightened cognitive processing can be revealed by some linguistic characteristics of the speaker (Burgoon et al., 2016). Since conference calls involve many analysts and institutional investors and may even be open to anyone who is interested in participating (Galant,1994; Feldman, 1996; Waroff, 1994), the management team of a company is less likely to prepare the responses to questions asked by the participants, and consequently it is easier to find linguistic clues for the heightened cognitive processing.

Due to the massive size and the high dimensionality of big data, conventional data mining techniques are often found inadequate to analyze and extract useful features effectively and efficiently (Jin, Wah, Cheng, and Wang, 2015). This essay applies an emerging AI technology, deep learning (also called deep neural network, DNN), to analyze the transcripts of conference calls and extract sentiment features from them. Deep learning algorithms enable automated extraction of complex data features at high levels of abstraction (Najafabadi et al., 2015). With its hierarchical architecture of artificial neural network consisting of multiple layers and nodes, a deep neural network automatically extracts features from the input data. In this process, the output features of a preceding layer (which is less abstract) are immediately fed into the successive layer as input data and more abstract features are defined based on a complex non-linear computation in the node. Due to the deep hierarchical architecture and the complex non-linear computation, deep learning algorithms are beneficial when analyzing big data, such as text, videos, and audios (Sun and Vasarhelyi, 2017).

The objective of this study is to (1) examine the relationship between sentiment features of management from conference calls and the likelihood of ICMW; (2)

demonstrate that the sentiment features contain incremental information for the prediction of ICMW by providing empirical evidence for the significant improvement of the explanatory as well as predictive power of the models with sentiment predictors as compared to the models that merely use financial fundamentals.

The transcripts of conference calls are obtained from SeekiNF[3]. The size of the final transcript sample is 1651 corresponding to fiscal years from 2004 to 2014, among which, 201 firm-years are related to ICMW under SOX 404. This research employs Alchemy language API[4], a deep learning based textual analysis tool provided by IBM Watson, to extract the sentiment features (including the overall sentiment score and the joy score) within the document. Four logistic models are developed and grouped into two classes: group A and group B. Each group includes one baseline model and one sentiment model. The baseline model is built as the starting point and uses a list of ICMW determinants as suggested by previous literature (i.e., Doyle, Ge, and McVay, 2007a; Ashbaugh-Skaife, Collins, and Kinney, 2007), while the sentiment model integrates two sentiment features into the analysis. Models in group B involve all variables used in models of group A as well as a new variable called *Growth*, which is the average of sales growth over 3 years. As this variable involves data from three years, the sample size in models of group B is reduced. The empirical analysis result supports the hypothesis that the joy score is negatively and significantly associated with the existence of ICMW and that after introducing both sentiment features into the baseline model, the explanatory and predictive performance of the model is significantly improved. The findings of the

[3] https://www.seekedgar.com:8443/seekinf.html
[4] https://www.ibm.com/watson/developercloud/alchemy-language.html

additional analysis reinforce the hypotheses and extend the effectiveness of sentiment features to the prediction of the number and the persistency of material weakness.

The remainder of this essay is organized as follows. The next section reviews the related literature and develops hypotheses. Section 3 describes the sentiment features of conference calls. Section 4 describes the research design, details the sample selection, and provides descriptive statistics of independent variables. Section 5 reports the main results. Additional analysis is conducted in Section 6. The last section concludes and presents the limitations.

2.2. Prior Research and Hypotheses Development

2.2.1. Internal Control over Financial Reporting

The presence of ICMW signals that the financial reports of a company may contain material misstatements as the ineffective internal controls allow or introduce errors and frauds into the financial reporting process (Barr-Pulliam, Brown-Liburd, and Sanderson, 2017; Ashbaugh-Skaife, Collins, Kinney, and LaFond, 2009; Jennings, Pany, and Reckers, 2008; Asare et al., 2013).

Consistent with this conjecture, researchers find that companies with internal control weaknesses exhibit lower quality of accrual (Ashbaugh-Skaife et al., 2008; Doyle, Ge, and McVay, 2007b). Hammersley, Myers, and Shakespeare (2008) examine the market reaction to management's disclosure of internal control weaknesses under section 302 of the Sarbanes Oxley Act. They find that the stock price decreases following the disclosure of ICMW and posit that the disclosure of the existence of ICMW causes investors to reevaluate their perceptions of the quality of the accounting information system (Francis

and Ke, 2006), which leads to a negative market reaction. Specifically, their subsample analysis for the companies reporting no other news shows that the size-adjusted returns decrease 0.95% when ICMW is disclosed. Ashbaugh-skaife et al (2009) provide evidence that firms with internal control deficiency exhibit significantly higher idiosyncratic risk and non-diversifiable market risk that affects the market's assessment of firms' cost of equity. Their results show that firms reporting internal control deficiency experience a significant increase in market-adjusted cost of equity, averaging about 93 basis points. Unlike prior literature which has primarily focused on the market reaction to the disclosure of ICW, Lopez, Vandervelde and Wu (2009) conduct a behavior study with 81 MBA students. They find that an adverse opinion on the internal controls over financial reporting leads investors perceive higher risk of material misstatement, higher risk of future financial statement restatement, higher risk premium, increased cost of capital, lower sustainability of earnings, lower earnings predictability, greater information asymmetry, or lower financial statement transparency.

The existence of ICMW also has great impact on top management. Since the issuance of SOX, top management has been held more accountable for the quality of financial reporting (Hoitash, Hoitash, and Johnstone, 2012; Collins, Masli, Reitenga, and Sanchez, 2009). For example, SOX 304 requires that if a listed company restates its financial statements due to material noncompliance as a result of misconduct, the CEO and CFO must reimburse bonuses or other related compensation received during the 12-month period following the filing of the noncompliance financial statement and any profit realized from the sales of securities of the issuer during that period. Moreover, SOX 906 addresses criminal penalties for CEOs and CFOs for certifying a misleading or fraudulent

financial report. The penalties can be upwards of $5 million in fines and 20 years in prison.

According to Auditing Standard No.5, internal control over financial reporting (ICFR) is "*designed by, or under the supervision of, the company's principal executive and principal financial officers, or persons performing similar functions, and effected by the company's board of directors, management, and other personnel*" (PCAOB, 2007). Top management (e.g., CEO and CFO) plays a leading role in the oversight of the effectiveness of internal control systems (McConnell and Banks, 2003; COSO, 2004; Sinnett, 2007; Hoitash, Hoitash, and Johnstone, 2012). The presence of ICMW indicates the incompetence of management regarding designing and maintaining an effective ICFR to provide reasonable assurance regarding the reliability of financial reporting, which serves as an impetus to change governance mechanisms (Larcker, Richardson, and Tuna, 20075). Johnstone, Li, and Rupley (2011) documents the existence of ICMW is associated with increased top management (including CEOs and CFOs) turnover. They argue that the change of CEOs and CFOs helps improve the top management composition and oversight, and this is associated with the remediation of ICMW. Hoitash, Hoitash, and Johnstone (2012) provide evidence that ICMW disclosures are negatively related to the change in CFO total compensation, bonus compensation, and equity compensation, especially for firms with stronger governance oversight.

2.2.2. The Determinants of ICMW

[5] Larcker, D., S. Richardson, and I. Tuna. 2007. Corporate governance, accounting outcomes, and organizational performance. The Accounting Review 82 (4): 963–1008.

On the purpose of detecting ICMW, extant literature examines a series of determinants. Doyle, Ge, and McVay (2007a) find smaller, younger, financially weaker, more complex, growing rapidly, or experiencing restructuring companies are more likely to have material internal control weaknesses. In the same year, they investigate the impact of accruals quality on internal control quality and conclude that internal control weakness is related to lower accruals. While Doyle, Ge, and McVay (2007 b) focus on material weaknesses of internal control, Ashbaugh-Skaife, Collins, and Kinney (2007) consider all significant deficiencies prior to mandated internal control audits. Auditing Standard No.5 defines the deficiency in ICFR as the problem existing in the design or operation of a control that hinders management or employees from preventing or detecting misstatements on a timely basis. While a significant deficiency is not as severe as a material weakness, it is important enough to merit attention by the stakeholder. Ashbaugh-Skaife, Collins, and Kinney's finding is consistent with that of Doyle, Ge, and McVay (2007a): internal control is weaker in companies with complex operations, higher growing speed, greater financial distress, and so forth. In addition, they assert that there is high incidence of auditor resignations prior to internal control deficiencies disclosures. Zhang, Zhou, and Zhou (2007) provide further evidence in the aspect of auditing. They indicate that audit committee quality is positively related to internal control weaknesses (ICW) and that auditor independence is negatively associated to ICW. Similarly, Krishnan (2005), by examining the disclosure provided by companies changing auditors, documents the relationship between audit committee quality and the internal control effectiveness. Recent studies examine other determinants including auditor tenure, auditor-client geographic distance (Chen, Gul, Truong, and Veeraraghavan, 2012),

auditor-provided tax services (De Simone, Ege, and Stomberg, 2014), recent auditor and management changes (Rice and Weber, 2012), and managerial overconfidence (Chen, Lai, Liu, and McVay, 2014; Lee, 2016).

2.2.3. Earnings Conference Calls

While many studies focus on quantitative data, others emphasize the role of qualitative data in predicting certain financial events. Corporate conference calls are large-scale telephone conference calls during which the management makes presentations on earnings and other relevant information and answers participants' questions (Frankel, Johnson, and Skinner, 1999). They are considered to be the main communication venue between companies and all interested parties, including investors and by- and sell-side analysts (Frankel, Johnson, and Skinner, 1999; Skinner, 2003; Corbin Perception, 2015; Bushee, Matsumoto, and Miller, 2003; Chen, Demers, and Lev, 2016). Conference calls are usually conducted immediately after the quarterly earnings press release. In a quarterly earnings conference call, the chairman, CEO, CFO, or other senior executives provide an overview of all major issues that affect the company's performance, highlight the business successes, and answer informed questions from analysts and investors. The speech of management conveys and reinforces their opinion on current business situation and its implications for future performance. It has been shown by Allee and DeAngelis (2015) and Sedor (2002) that conference calls contain incremental information beyond mandated disclosures such as financial report and earnings announcement regarding the current and future situation of the company.

Conference calls are important supplementary disclosures especially when earnings contain unusual or extraordinary items as the management will explain the implications of those items to analysts. Compared to other written financial disclosures such as press releases, conference calls are less formal and more flexible, and the management is typically unsure of what exactly the investors and the analysts will ask (Frankel, Johnson, and Skinner, 1999). Therefore, presentations and answers in conference calls are more informative than other formal documents. The PCAOB (2010a) recommends that auditors should refer to earnings conference call narratives for better understanding of material misstatement risk. Extant literature considers the sentiment feature of conference calls as a new factor in addition to the traditional firm-level fundamentals for the study of a certain event. For example, research has investigated the corresponding market reaction (Henry, 2006; Henry and Leone, 2009; Matsumoto, Pronk and Roelofsen, 2011; Price, Doran, Peterson, and Bliss, 2012; Allee and Deangelis, 2015; Davis, Ge, and Matsumoto, 2015). In particular, there is evidence for that conference calls are related to increased stock trading volume and return variance (Frankel, Johnson, and Skinner, 1999; Price, Doran, Peterson, and Bliss, 2012; Bushee, Matsumoto, and Miller, 2003). Furthermore, managerial tone of conference calls is found to be related to future performance as well as analyst responses and uncertainty (i.e., Mayew and Venkatachalam, 2012; Druz, Wagner, and Zeckhauser, 2015; Davis, Ge, Matsumoto, and Zhang, 2015).

Another line of research links the linguistic cue of conference calls to financial reporting quality (and future events). For example, Hobson, Mayew, and Venkatachalam (2012) document that cognitive dissonance in CEO speech is diagnostic of adverse misreporting. Similarly, Larker and Zakolyukina (2012) claim that a series of linguistic

characteristics from conference calls are related to the management's deceptive discussions that are linked to subsequent financial restatements. Burgoon et al. (2016) identify a set of linguistic signs of deception derived from conference calls and find pitch and voice quality, vocal intensity, and other signs are associated with financial frauds.

The words are the gateway to the mind (Schafer, 2011). Analyzing the words that one chooses when he or she speaks provides insights into his or her thought process. If a company's ICFR has material weaknesses (it is called "bad news" hereafter), the speech of the management who has the knowledge of the "bad news" may contain linguistic clues revealing the different cognitive process of the speaker as compared to those who do not have such "bad news". The existence of ICMW indicates a reasonable possibility that a material misstatement of the company's financial statements cannot be prevented or detected by the internal control system (PCAOB, 2007). Therefore, the "bad news" is a critical concern of the management, with which the management may inadvertently provide some word clues (for the effect of the "bad news' on his/her sentiment or emotion) in a conference call (Druz, Wagner, and Zeckhauser, 2015). Especially, if the management tries to intentionally cover the "bad news", such behavior may be discovered by identifying the linguistic clues as signs of deceit in management speeches.

This view is supported by the accumulating evidence from experiments, case studies, and meta-analyses on the perpetration and detection of deceptive behaviors in social psychology research (e.g., Zuckerman and Driver, 1985 and DePaulo et al., 2003). According to the leakage hypothesis (Ekman and Friesen, 1969), the act of deception will make a single person feel guilty, stressful, and fear of detection. Furthermore, DePaulo, Rosenthal, Rosenkrantz, and Green (1982) and Kraut (1980) suggest that a person may

experience relatively heightened cognitive processing when telling a lie than telling the truth. Thus, it is possible to distinguish liars from truthtellers by examining the word clue that reveals their sentiment (including the emotion). Specifically, for example, feeling guilty, stressful, and fear, liars may try to dissociate themselves from their own responses by "making more neutral statement …, or … speaking in the third person" (DePaulo, Rosenthal, Rosenkrantz, and Green, 1982). As a result, the language clue (if it can be detected successfully) is capable of predicting the internal control weakness. Accordingly, this essay develops hypotheses as follows:

H1: The sentiment features of conference calls are significantly associated with the likelihood of internal control material weaknesses.

H2: The explanatory and predictive ability of the model that incorporates sentiment features of conference calls along with major financial determinants is superior to that of the model that merely uses the financial determinants.

2.3. Sentiment Analysis Method

2.3.1. Deep Learning for Sentiment Analysis

Deep learning was firstly proposed by G.E. Hinton and his coworkers in 2006 (Hinton, Osindero, and Teh, 2006). Inspired by the biological neural network in human brains, it contains layers of artificial neurons which allow the machine to learn representations of data with multiple levels of abstraction (LeCun, Bengio, and Hinton, 2015). Recent advances in deep learning have dramatically improved the state-of-the-art in image identification, speech recognition, text understanding, and many other domains.

For example, a deep neural network (e.g., Google's AlphaGo) plays games at better than human-level performance and on a scale much larger than the availability of human will allow in the given time frame (Heaton, Polson, Witte, 2016). Amazon Go, as another example of deep learning application, is a new type of store providing checkout-free shopping experience. It is powered by computer vision, sensor fusion[6], and deep learning technology. The technology automatically detects when products are taken from or returned to the shelves and keeps track of them in a virtual cart. Customers can just leave the store when they are done shopping without checkout and be charged shortly.

Deep learning works effectively in sentiment analysis. In Met Gala 2016, supermodel Karolina Kurkova wore a "cognitive" gown including 150 LED lights which change color in reaction to the sentiments and emotions of Kurkova's Twitter followers. The dress is empowered by Watson Tone Analyzer technology of deep learning, which is able to identify joy, passion, curiosity, excitement and encouragement[7]. While a traditional bag of words approach typically measures the sentiment by counting the number of words associated with a particular sentiment word list scaled by the total number of words in the document, a deep learning model "learns" from large-scale examples by developing a deep neural network (DNN) with multiple layers of numerous neurons to transform input data and identify the pattern underlying the data. A simplified

[6] Sensor fusion combines multiple data from different sensors to increase the reliability and accuracy of the results. For example, when an item is picked but then placed back to the inventory location, the image will be combined with the weight received from a pressure sensor located at the inventory location to determine the identity of the item. Specifically, "the image analysis may be able to reduce the list of potentially matching items down to a small list. The weight of the placed item may be compared to a stored weight for each of the potentially matching items to identify the item that was actually placed in the inventory location. By combining multiple inputs, a higher confidence score can be generated increasing the probability that the identified item matches the item actually picked from the inventory location and/or placed at the inventory location" (Bishop, 2016).

[7] https://www.telegraph.co.uk/fashion/events/met-gala-the-most-impressive-tech-looks-on-the-red-carpet/

structure of a DNN is shown in Figure 2.1. It has one input layer to receive raw data (e.g., the transcripts of conference calls), multiple hidden layers to process data and extract features, and one output layer to provide results for identified data features (e.g., sentiment). Each layer applies a nonlinear transformation on its preceding layer and provides a representation. In other words, the output representation of each layer is provided as input to its successor layer. As the input data goes deeper, the constructed nonlinear transformation becomes more complex and the representation becomes more abstract. The output of the final layer is the final representation of the raw data, which provides features extracted from the data that are useful for further classification, association, and other tasks (Najafabadi, et al., 2015).

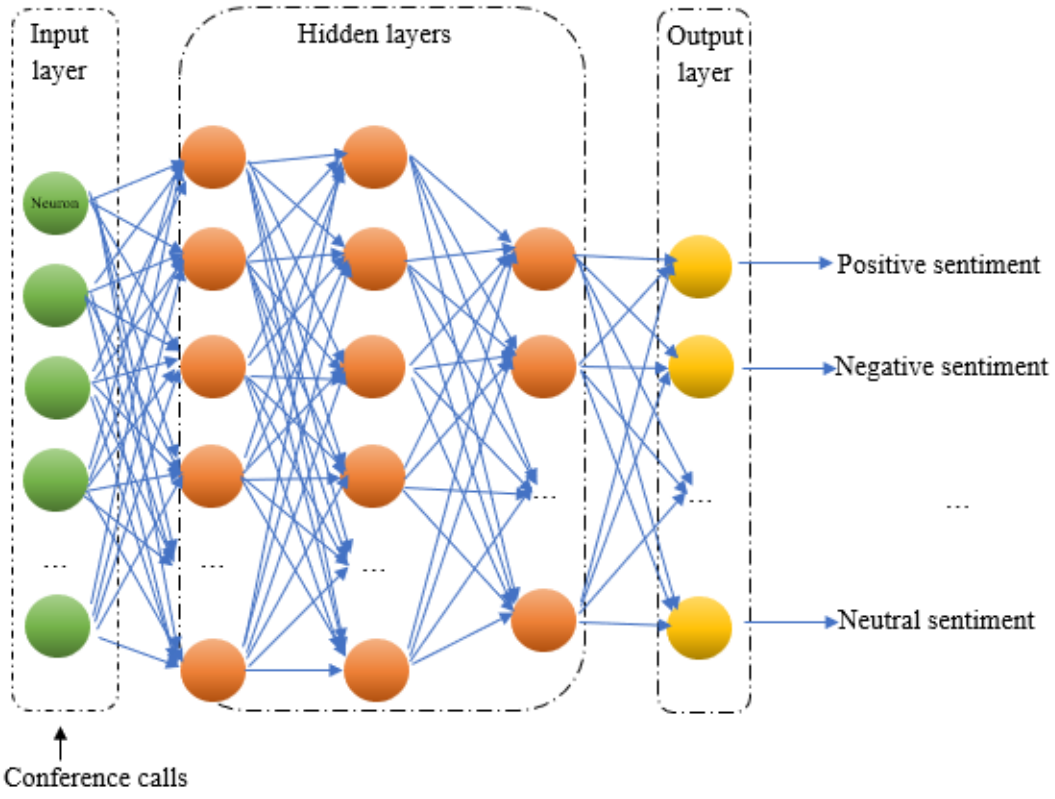

Figure 2.1 A Deep Neural Network for Conference Call Sentiment Analysis

2.3.2. Alchemy Language API

The tool used for sentiment analysis is Alchemy Language API, a deep learning-based text analysis cloud services provided by IBM Watson[8]. Unlike conventional text mining methods that requires laborious and time-consuming data preprocessing (e.g., removing HTML tags, transferring HTML characters to text characters, converting words from upper case to lower case, deleting punctuation and stop words), Alchemy Language API requires zero data preprocessing: the user only needs to provide the text/html raw file or even the URL. The tool removes irrelevant content such as links and advertisements and returns results[9]. This makes deep learning-based text analysis more efficient than traditional text mining approach. While the sentiment analysis provided by Watson identifies attitude, opinions, or feelings in the content that is being analyzed, the emotion analysis detects joy, anger, disgust, fear, and sadness implied in the text[10]. The sentiment analysis applies Watson's deep learning algorithms trained with text from billions of webpages[11] and the output predictions are based on the data pattern learned by the algorithms.

2.3.3. Sentiment Features

The sentiment features acquired from Alchemy Language API include the overall sentiment score and the joy score. The returned sentiment score measures the overall sentiment strength of the document, ranged from -1 to 1, where negative score represents

[8] https://www.ibm.com/watson/developercloud/alchemy-language.html
[9] https://watson-api-explorer.mybluemix.net/apis/natural-language-understanding-v1
[10] https://console.bluemix.net/docs/services/alchemy-language/index.html#index
[11] https://alchemy-language-demo.mybluemix.net/

negative sentiment, positive score represents positive sentiment, 0 represents neutral sentiment, and 1 indicates that the sentiment is both positive and negative[12]. The score of emotion joy values ranges from 0 to 1, which represents the confidence level indicating the probability that the emotion of joy is implied by any part of the sample text[13]. It is noteworthy that joy is not a component of sentiment but a specific type of emotion. Furthermore, during a conference call, the management tries to avoid expressing some "extreme" personal emotion such as anger, sadness or fear. Therefore, despite Watson provides us with emotion analysis for anger, disgust, sadness, and fear, those emotions rarely exist in the transcripts of conference calls. On the other hand, since joy is a common emotion that can be easily found in conference calls and the "extreme" emotions will make one feel less joyful, this paper focuses on the emotion of joy and excludes other emotions.

2.4. Research Design

The goal of this research is to (1) examine the relationship between sentiment features of management from earnings conference calls and the likelihood of ICMW; (2) demonstrate that the sentiment features contain incremental information for the prediction of ICMW by providing empirical evidence for the significant improvement of the explanatory and predictive power of the model with sentiment predictors as compared to the model with financial fundamentals only. Therefore, following models are developed.

2.4.1. Model Development

[12] https://www.ibm.com/watson/developercloud/alchemy-language/api/v1/#targeted_sentiment
[13] https://www.ibm.com/watson/developercloud/alchemy-language/api/v1/#emotion_analysis

Following existing literature (e.g., Doyle, Ge, and McVay, 2007a and Ashbaugh-Skaife, Collins, and Kinney, 2007), this essay employs logistic regressions to examine the relation between the extracted sentiment features and the existence of ICMW. It begins with the **baseline model A** below:

$$ICMW = \beta_0 + \beta_1 Marketvalue + \beta_2 Aggregateloss + \beta_3 Zscore + \beta_4 Segments + \beta_5 Foreign + \beta_6 Inventroy + \beta_7 Restructure + \beta_8 Acquisition + \beta_9 Resign + \beta_{10} Big4 + \beta_{11} Litigation + \sum IndustryFE + \varepsilon$$

where:

$ICMW$ = indicator equals 1 if there is at least one internal control material weakness identified under SOX 404, and 0 otherwise;

$Marketvalue$ = logarithm of share price multiplied by number of shares outstanding[14];

$Aggregateloss$ = indicator equals 1 if earnings before extraordinary items in year t and t–1 sum to less than zero, and 0 otherwise;

$Zscore$ = Z-score (Altman, 1968), which measures financial distress of the company;

$Segments$ = logarithm of the sum of the number of operating and geographic segments reported by the Compustat Segments database for the firm in year t;

$Foreign$ = indicator equals 1 if the company has a non-zero foreign currency transaction in year t, and 0 otherwise. This variable is reported by Compustat Segment database;

[14] The number of shares outstanding is presented in millions

$Inventory$= inventory scaled by total assets;

$Restructure$ = indicator equals 1 if the company was involved in a restructuring in the last three years, and o otherwise;

$Acquisition$=indicator equals 1 if the company engages in acquisitions in the last three years, and 0 otherwise.

$Resign$ = indicator equals 1 if the auditor resigned in the year prior to an ICW disclosure; and 0 otherwise;

$Big4$ = indicator equals 1 if the firm is audited by a Big 4 audit firm, and 0 otherwise;

$Litigation$ = indicator equals 1 if the company is in a litigious industry[15], and 0 otherwise.

$IndustryFE$ = industry fixed effects

The dependent variable equals 1 if ICMW exists in the company, and 0 otherwise. Most of the financial determinants for ICMW are consistent with those used in the studies of Doyle, Ge, and McVay, (2007a) and Ashbaugh-Skaife, Collins, and Kinney (2007). It controls for market value ($Marketvalue$), a measure of firm size. Although evidence on the association between firm size and control quality is mixed (Krishnan, 2005), intuitively larger firms have more complete and effective financial reporting procedure ensuring proper segregation of duties (Doyle, Ge, and McVay, 2007a).

[15] The definition of litigious industry follows Francis, Philbrick, and Schipper (1994). Companies in litigious industries are with SIC codes of 2833-2836 (biotechnology);3570-3577 (computer equipment);3600-3674 (electronics);5200-5961 (retailing); and 7370-7374 (computer services).

A second important predictor is related to the financial performance of the company as it is believed that a company with poor financial performance may not be able to maintain sufficiently effective internal control environment. For example, Defond and Jiambalvo (1991) find firms with weaker financial performance tend to have more accounting errors. Therefore, this paper examines two financial health related variables, including $Aggregateloss$, which indicates whether the sum of earnings before extraordinary items for the past two years is negative, and $Zscore$, refers to Altman z-score (Altman, 1968) of distress risk[16].

The complexity of operations is another important determinant of ICMW since internal control breaches are more likely to occur in firms with more diverse and multifaceted operations (Ashbaugh-Skaife, Collins, and Kinney, 2007). $Segments$ and $Foreign$ are used to control for the effect of operational complexity on internal control systems.

This study also takes into consideration of the influence of operating characteristics, including inventory status and sales growth, on the internal control. High level of inventory makes it difficult to accurately measure, record, and report. As a result, baseline model A uses $Inventory$ (defined as inventory scaled by total assets) to proxy for such operating characteristics (Kinney and McDaniel, 1989). Sales growth is controlled in other models which will be discussed later.

[16] Zones of discrimination:
Z > 2.99: "Safe" Zone
1.81 < Z < 2.99: "Gray" Zone
Z < 1.81: "Distress" Zone

Furthermore, firms recently undergo structural changes have higher chance of experiencing internal control difficulties due to the possible personnel and organization issues. Therefore, another control variable is $Restructure$, which is equal to 1 if the company was involved in a restructuring in the last three years, and 0 otherwise. Similarly, internal control problems are related to acquisitions as firms that recently engaged in acquisition have to integrate different internal control systems (Zhang, Zhou, and Zhou, 2007). This paper uses an indicator variable, *ACQUISITION* with a value of 1 if the company engages in acquisitions in the last three years, and 0 otherwise.

This study follows Krishnan (2005) to include *RESIGN*, a dummy coded 1 for an auditor resignation in the past one year, and 0 otherwise. A possible reason is that a recent auditor resignation occurs when the auditor realizes that the expected cost of audit will exceed the revenue the auditor charges, implying the internal control of the client is too weak to rely on (Krishnan, 2005; Ashbaugh-Skaife, Collins, and Kinney, 2007).

$Big4$, and $Litigation$ are two variables considered to be related to incentives to ICMW detection (Ashbaugh-Skaife, Collins, and Kinney, 2007). ICMWs in firms audited by Big 4 auditors are more likely to be discovered because Big 4 auditors are seen as providers for higher audit quality with more systematic examination and investigation procedures and more advanced data analytics techniques. $Litigation$ proxies for companies in litigious industries, including biotechnology, computer equipment, electronics, retailing, and computer services. This variable is used in the model because managers in litigious industries have greater incentive to reveal ICWs to reduce litigation risk (Collins, and Kinney, 2007). Finally, the fix effect of industry is included[17].

[17] Industry classifications are compiled using the following SIC codes: Agriculture: 0100–0999; Mining & Construction: 1000–1299, 1400–1999; Food & Tobacco: 2000–2141; Textiles and Apparel: 2200–2399;

Sentiment model A

Sentiment model A is developed by adding two sentiment features, *Sentiment* and *Joy*, to the baseline model A. Sentiment model A is defined as:

$$ICMW = \beta_0 + \alpha_1 Sentiment + \alpha_2 Joy + \beta_1 Marketvalue + \beta_2 Aggregateloss + \beta_3 Zscore + \beta_4 Segments + \beta_5 Foreign + \beta_6 Inventory + \beta_7 Restructure + \beta_8 Acquisition + \beta_9 Resign + \beta_{10} Big4 + \beta_{11} Litigation + \sum IndustryFE + \mathcal{E}$$

where:

Sentiment=sentiment score of the overall transcript.

Joy= joy score, ranged from 0 to 1.

Other variables are defined the same as in baseline model A.

Baseline model B

As mentioned earlier, models in group B control for the effect of sales growth on internal control since there is evidence supporting that rapid growth of sales are likely to lead to internal control problems (Doyle, Ge, and McVay, 2007a; Ashbaugh-Skaife, Collins, and Kinney, 2007). The variable measuring sales growth is defined as the average percentage change of sales in the last three years, which has limited availability.

Lumber, Furniture, & Printing: 2400–2796; Chemicals: 2800–2824, 2840–2899; Refining & Extractive: 1300–1399, 2900–2999; Durable Manufacturers: 3000–3569, 3580–3669, 3680–3999; Computers: 3570–3579, 3670–3679, 7370–7379; Transportation: 4000–4899; Utilities: 4900–4999; Retail: 5000–5999; Services: 7000–7369, 7380–9999; Banks & Insurance: 6000–6999; Pharmaceuticals: 2830–2836, 3829–3851.

A different model, **baseline model B**, is developed to fit this portion of data. The model is as follows:

$$ICMW = \beta_0 + \beta_1 Marketvalue + \beta_2 Aggregateloss + \beta_3 Zscore + \beta_4 Segments + \beta_5 Foreign + \beta_6 Inventroy + \beta_7 Growth + \beta_8 Restructure + \beta_9 Acquisition + \beta_{10} Resign + \beta_{11} Big4 + \beta_{12} Litigation + \sum IndustryFE + \mathcal{E}$$

where:

$Growth$ =Average growth rate (percentage) in sales for the last three years

Other variables are defined the same as in baseline model A.

Sentiment model B

To examine the relationship between sentiment features and ICMW, sentiment model B includes the same sentiment features, Sentiment and Joy, as does sentiment model A. **Sentiment model B** is defined as follows:

$$ICMW = \beta_0 + \alpha_1 Sentiment + \alpha_2 Joy + \beta_1 Marketvalue + \beta_2 Aggregateloss + \beta_3 Zscore + \beta_4 Segments + \beta_5 Foreign + \beta_6 Inventroy + \beta_7 Growth + \beta_8 Restructure + \beta_9 Acquisition + \beta_{10} Resign + \beta_{11} Big4 + \beta_{12} Litigation + \sum IndustryFE + \mathcal{E}$$

2.4.2. Data

The sample starts with 6379 transcripts of conference calls from Seek iNF filed from 2005 to 2014. The textual data is matched with *Compustat* by Central Index Key (CIK) and fiscal year (determined by the announcement date). Among the 6379 transcripts,1582 records miss CIK or fiscal year information. So, they are excluded from the sample. The corresponding fiscal year of the remaining sample is from 2004 to 2014. Each transcript is feed to Alchemy Language API to obtain the sentiment features. For those companies that have multiple conference calls in one year, the paper uses the conference call of the company in last quarter of this year. After this step, the aggregated sentiment features involve 2408 firm-years.

Next, the data of sentiment features are linked to *Audit Analytics* for information about material internal control weakness (to fulfill SOX 404). A record is identified as containing ICMWs if the count of ICMWs is more than 0 as provided by *Audit Analytics*. 15 observations missing internal control weakness information are removed.

To examine the control variables, the data is merged with the financial fundamentals in *Compustat*. It removes 731 firm-years with missing values of key financial variables used in the logistic model. Since the controls include some auditing-related variables, it further merges the sample with the data from *Audit Analytics* and eliminates 11 records of master data that do not have matching auditing-related control variables from *Audit Analytics*. The final sample contains 1651 firm-years. The sample selection procedure is reported in Table 2.1.

Table 2.1 Sample Selection Procedure

Initial conference call transcript samples from Seek iNF	6379
Less: Missing fiscal year or CIK information	(1582)
use the conference call in the last quarter if a company has multiple conference calls	
Remaining:	2408
Less:	
Missing internal control information	(15)
Missing Compustat data	(731)
Missing Audit Analytics data	(11)
Final sample	1651

The final sample contains 189 firm-years with ICMWs and 1462 observations as the control sample without any ICMWs. The distribution of firms with ICMWs and control firms over fiscal years and industries is summarized in table 2.2 and table 2.3, respectively. As shown in Table 2.2, ICMW occurs most frequently in fiscal years from 2005 to 2007, which is prior to the financial crisis period. It is noticed that the fiscal year of 2004 has the lowest number of ICMW. This is because the earliest filing year of conference call transcripts collected is 2005, which limits the availability of sample for fiscal year of 2004. From Table 2.3, it is found that firms in durable manufacturers and computers industries have disclosed much more material internal control weaknesses than firms in other industries. This may be caused by the complexity of operation.

Table 2.2 Sample Distribution over Fiscal Years

Fiscal year	ICMW sample	Control sample
2004	8	30
2005	26	103
2006	31	131
2007	27	151
2008	17	186
2009	12	150
2010	12	142
2011	17	137
2012	11	139
2013	14	164
2014	14	129
Total	189	1462

Table 2.3 Sample Distribution over Industries

Industry	ICMW sample	Control sample	Total
Agriculture	0	3	3
Mining & Construction	4	57	61
Food & Tobacco	4	25	29
Textiles & Apparel	1	6	7
Lumber, Furniture, & Printing	7	33	40
Chemicals	7	31	38
Refining & Extractive	10	49	59
Durable Manufacturers	45	280	325
Computers	45	285	330
Transportation	12	101	113
Utilities	3	35	38
Retail	7	105	112
Services	18	165	183
Banks & Insurance	2	18	20
Pharmaceuticals	24	269	293
Total	189	1462	1651

2.5. Results

2.5.1. Univariate Analysis and Descriptive Statistics

Table 2.4 presents Pearson product-moment correlations. It shows that many variables are correlated with one another. For instance, *Joy* and *Sentiment* are correlated with the largest correlation of 0.4529, which is normal since the emotion joy is a positive sentiment. *Sentiment* is also correlated to *Marketvalue* with a coefficient of 0.0749. *Litigation* and *Acquisition* are correlated to *Sentiment* and *Joy* at high significant level. Moreover, *Marketvalue, Aggregateloss, Segments,* and *Big 4* are correlated with each other or with other variables like *Acquisition* and *Restructure,* and so forth. Table 2.5 reports descriptive statistics and the results of univariate tests that statistically assess the differences between the ICMW sample and control sample. For numerical variables, it presents the summary statistics including the mean, standard deviation (std. dev.), first quartile (25% percentile), median, and third quartile (75 percentile), while for categorical variables, it shows the mean values and how significantly the means of two groups differs with each other.

The results of descriptive statistics in Table 2.5 show that managers of firms with ICMW tend to convey fewer joy emotion during conference calls. Additionally, larger firms are less likely to report internal control weaknesses. While there is no significant difference of the average z-score measuring financial distress between ICMW samples and the controlling sample, significantly lower median values for this variable for the treating sample are observed as compared to the controlling sample. It suggests that companies in treating sample is more likely to experience financial distress. Consistent

with prior literature (e.g., Doyle, Ge, and McVay, 2007a), the average value of *Growth* for companies with ICMW is higher than that for companies without ICMW.

2.5.2. Multivariate Analysis

Table 2.6 reports the results of multivariate analysis for two groups of models: group A includes baseline model A and sentiment model A; group B consists of baseline model B and sentiment model B. Though not tabulated, industry indicator variables are also included to capture the tendency of material weakness firms to cluster by industry.

Models in group A have 1651 records. The results of baseline model show that, *Marketvalue, Segment, and Resign* significantly affect the likelihood of ICMW. The coefficient of *Marketvalue* is -0.2551. This variable is negatively associated with the probability of ICMW at p-values (not tabulated) less than 0.01 under one-tailed tests. Variable *Segment* is positively related to the dependent variable at 0.01 level. Additionally, *Resign* has a positive effect on the predicted probability of a material weakness at p-value less than 0.01. All of these variables are significantly associated with the dependent variable in the expected direction. A p-value that less than 0.0001 for the likelihood ratio, χ^2, and a Pseudo R^2 of 0.0757 suggest that the overall explanatory ability of the model is economically significant.

Table 2.4 Pearson Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Sentiment	1.0000	0.4529 (0.0000)	0.0749 (0.0005)	0.0215 (0.3208)	-0.0154 (0.4888)	0.0348 (0.1090)	0.0126 (0.5612)	0.0145 (0.5015)	-0.0345 (0.1106)	0.1097 (0.0000)	0.0305 (0.1585)	-0.0363 (0.0956)	0.0088 (0.7970)	0.0718 (0.0009)
2 Joy		1.0000	0.0362 (0.0942)	-0.0413 (0.0560)	0.0443 (0.0460)	0.0331 (0.1273)	0.0193 (0.3717)	0.0139 (0.5208)	0.0051 (0.8132)	0.1041 (0.0000)	-0.0009 (0.9653)	0.0119 (0.5853)	0.0142 (0.6766)	0.0732 (0.0007)
3 Marketvalue			1.0000	-0.0619 (0.0042)	0.0290 (0.1924)	0.2321 (0.0000)	0.0108 (0.6163)	0.0070 (0.7471)	0.0972 (0.0000)	-0.0624 (0.0039)	-0.0083 (0.7007)	-0.0423 (0.0523)	0.0241 (0.4787)	0.2179 (0.0000)
4 Aggregateloss				1.0000	-0.1669 (0.0000)	-0.0964 (0.0000)	0.0184 (0.3948)	0.1025 (0.0000)	-0.0769 (0.0004)	0.0633 (0.0034)	0.0066 (0.7601)	-0.0024 (0.9138)	0.0657 0.0534	-0.1195 (0.0000)
5 Zscore					1.0000	-0.0074 (0.7412)	-0.0395 (0.0754)	-0.1357 (0.0000)	0.0100 (0.6513)	-0.0245 (0.2708)	0.0136 (0.5398)	0.0298 (0.1834)	-0.0387 (0.2657)	-0.0171 (0.4430)
6 Segments						1.0000	0.0418 (0.0537)	0.2250 (0.0000)	0.1518 (0.0000)	-0.2133 (0.0000)	-0.0093 (0.6696)	0.1133 (0.0000)	-0.0659 0.0531	0.2370 (0.0000)
7 Foreign							1.0000	0.0204 (0.3445)	0.0512 (0.0178)	-0.0094 (0.6629)	0.0147 (0.4953)	0.0419 (0.0542)	0.0093 0.7845	0.0576 (0.0078)
8 Restructure								1.0000	0.1306 (0.0000)	0.0053 (0.8061)	-0.0433 (0.0452)	0.0495 (0.0230)	-0.0084 (0.8403)	0.1135 (0.0000)
9 Big 4									1.0000	-0.0563 (0.0092)	-0.1225 (0.0000)	-0.0840 (0.0001)	-0.0348 (0.3066)	0.1006 (0.0000)
10 Litigation										1.0000	0.0110 (0.6112)	-0.1273 (0.0000)	0.0792 (0.0199)	-0.0536 (0.0134)
11 Resign											1.0000	0.0017 (0.9369)	0.0253 (0.4584)	-0.0057 (0.7935)
12 Inventory												1.0000	-0.0308 (0.3703)	-0.0944 (0.0000)
13 Growth													1.0000	-0.0254 (0.4555)
14 Acquisition														1.0000

All continuous variables that do not take log are winsorized at the 1% and 99% to mitigate outliers.

Table 2.5 Descriptive Statistics

	Mean	Std.dev.	25%	Median	75%
Sentiment:					
MW group	0.1544	0.1110	0.0787	0.1617	0.2292
Control group	0.1527	0.1292	0.0725	0.1611	0.2370
Joy:					
MW group	0.1882	0.1854	0.0654	0.0838	0.3027
Control group	0.2254***	0.2008	0.0700	0.0887	0.4443
Marketvalue:					
MW group	5.6820	1.5909	4.7246	5.7173	6.4711
Control group	6.3925***	2.0056	5.0263	6.3254***	7.6036
Zscore:					
MW group	2.5524	13.4819	1.4528	2.4460	4.0049
Control group	3.3752	11.6202	1.4580	2.9782***	5.3032
Segments:					
MW group	1.4099	0.6707	1.0986	1.3863	1.9459
Control group	1.3573	0.7850	0.6931	1.3863	1.9459
Inventory:					
MW group	0.0971	0.1242	0.0015	0.0567	0.1409
Control group	0.0880	0.1233	0.0000	0.0424	0.1309
Growth:					
MW group	0.5288	2.4284	0.0046	0.0884	0.3344
Control group	0.2676**	1.5891	0.0034	0.0925	0.2187
Aggregateloss:					
MW sample	0.1542	0.3656	0	0	0
Control sample	0.1651	0.3809	0	0	0
Foreign:					
MW sample	0.9900	0.1023	1	1	1
Control sample	0.9872	0.1176	1	1	1
Restructure:					
MW sample	0.3184	0.4639	0	0	1
Control sample	0.3404	0.4757	0	0	1
Acquisition:					
MW sample	0.3846	0.4901	0	0	1
Control sample	0.3814	0.4902	0	0	1
Resign:					
MW sample	0.0995***	0.3008	0	0	0
Control sample	0.0109	0.1053	0	0	0
Big4:					
MW sample	0.6070***	0.4876	0	1	1
Control sample	0.7566	0.4316	1	1	1
Litigation:					
MW sample	0.4080	0.4912	0	0	1
Control sample	0.4110	0.4882	0	0	1

***, **, * significant different from MW group at a one tailed p-value ≤ 0.01,0.05, and 0.10, respectively, under a t-test on the equality of means or nonparametric test on the equality of medians.

While other variables remain the same, sentiment model A includes two additional sentiment features, *Sentiment* and *Joy*. It shows that *Joy* is negatively related to the existence of ICMW with a p-value less than 0.01. This result suggests that the higher the joy score is, the less likely that the company has ICMW. The overall sentiment score is insignificant. A possible reason is that CEOs and CFOs tend to emphasize the success of the business to make the overall tone of their speech as positive as possible. Differently, the joy score is not designed to measure the overall joyfulness of the conference call document. It indicates the possibility that an emotion of joy is suggested by any part of a document. For a company with ICMW, the manager is less likely to be joyful even when he/she is talking about some good news of the company.

Same as that in the baseline model, *Marketvalue*, *Segments*, and *Resign* significantly affect the likelihood of ICMW in the expected direction. The likelihood ratio (Pseudo R^2) of the sentiment model rises to 98.17 (0.0827). It indicates that, by adding these two sentiment features, the explanatory ability of the model is improved.

Next, a likelihood ratio test (LR test) is used to compare the goodness of fit of the two models. The resulting likelihood ratio is 8.32 and statistically significant at 0.05, suggesting that the sentiment model A has significantly better explanatory ability.

Models in group B incorporate the sales growth variable, *Growth*. This reduces the sample size to 1228. In both baseline and sentiment model, *Growth* does not have a significantly influence on the likelihood of ICMW. Results in this specification are similar to that of models in group A, except *Segments* becomes insignificant. In the sentiment model, the coefficient of Joy changes from -1.3762 to -1.5264, representing a

stronger influence on the predicted value of the target variable. This pattern holds for *Marketvalue* and *Resign.* Pseudo R^2 for baseline model B (sentiment model B) reaches to 0.0785 (0.0872). The improvement of the explanatory power of the sentiment model B as compared to the baseline model B is significant, as shown by the Likelihood Ratio of 7.02 and its p-value of 0.0300.

From the perspective of the explanatory power of the model, both the univariate and multivariate findings support the hypotheses. Though sentiment score is not significantly related to the target variable, it has been found that the higher the joy score is, the less likely that the company will report ICMW under the requirement of SOX 404. Therefore, the confidence score of joy of conference calls extracted by deep learning technique is a useful predictor of ICMW and the incremental informativeness of management tone for internal control weakness prediction is supported.

Table 2.6 Logistic Regression of the Probability of ICMW

	Predicted sign	Estimate coefficients of group A		Estimate coefficients of group B	
		Baseline model A (1)	Sentiment model A (2)	Baseline model B (3)	Sentiment model B (4)
Intercept	+/-	-1.6784**	-1.6211*	-2.1248*	-2.0097*
$Sentiment$	-		0.6243		0.2979
Joy	-		-1.3762***		-1.5264**
$Marketvalue$	-	-0.2551***	-0.2495***	-0.2591***	-0.2537***
$Aggregateloss$	+	-0.3105	-0.3137	-0.1360	-0.1379
$Zscore$	-	-0.0040	-0.0008	-0.0047	-0.0035
$Segments$	+	0.3424***	0.3547***	0.2512	0.2559
$Foreign$	+	0.3927	0.4047	0.5328	0.5575
$Inventory$	+	0.1535	0.1585	0.5008	0.5555
$Growth$	+			-0.0193	-0.0286
$Restructure$	+	-0.1366	-0.1330	-0.1187	-0.1420
$Acquisition$	+	0.0601	0.0935	0.2901	0.3429
$Resign$	+	2.2631***	2.2322***	2.3476***	2.3188***
$Big4$	-	-0.1079	-0.0984	-0.0007	0.0260
$Litigation$	+	0.1908	0.2211	0.2760	0.3119
Industry indicator variables		Included	Included	Included	Included
Number of total observations		1651	1651	1228	1228
Likelihood ratio, χ^2 (p-value)		89.85 (0.0001)	98.17 (0.0001)	63.42 (0.0001)	70.44 (0.0001)
Pseudo R^2		0.0757	0.0827	0.0785	0.0872
Likelihood-ratio test: Likelihood ratio (p-value)		8.32** (0.0156)		7.02** (0.0300)	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

2.5.3. The Prediction Performance of the Model

Besides the explanatory ability of the sentiment attributes in the model, this study examines whether the sentiment features help improve the prediction accuracy. Table 2.7 displays the 10-fold cross validation results of a list of evaluation metric with Logistic Regression, Random Forest, and Artificial Neural Network algorithms. For Logistic Regression, the use of sentiment features makes the False Negative Rate decreases from 0.4621 to 0.3690. Since False Positive Rate increases from 0.2426 to 0.294 when the sentiment features are incorporated into the model, the overall accuracy decreases from 0.7288 to 0.6915. Since the sample is unbalanced, the number of companies without ICMW is much greater than that of companies with ICMW, the overall accuracy is biased. An unbiased metric is AUC, the area under the ROC curve. It measures the overall predictive performance of the model. It is found that the AUC for the sentiment model (0.6955) is slightly higher than that of the baseline model (0.6931). This is because the sentiment model performs more effectively than the baseline model in terms of detecting ICMW rather than identifying companies without ICMW.

Compared to the case of Logistic Regression, the models perform better when Random Forest is employed as the classifier. The AUC reaches to 0.7274 for the sentiment model, higher than that of the baseline model, which is 0.7228. Relative to the baseline model, the sentiment model has an increased overall accuracy of 0.8357 and a decreased False Positive Rate which is as low as 0.1033, but the False Negative Rate increased from 0.4069 to 0.5724.

In addition, an Artificial Neural Network with one hidden layer of 100 neurons is developed to predict ICMW. In the sentiment model, the False Positive Rate is 0.014 higher than that in the baseline model. However, the False Negative Rate is 0.0207 lower, compared to the results in the baseline model. Again, due to the sample imbalance, the overall accuracy of the sentiment model is lower than that of the baseline model. But the improved AUC shows that the overall predictive performance of the sentiment model is better than that of the baseline model.

Moreover, the unreported result of predictor importance shows that, each of these two sentiment features functions as one of the top 3 important predictors in all sentiment models. To summarize, the best classification algorithm is Random Forest. Overall, the sentiment features improve the predictive performance of the model for ICMW.

Table 2.7 10-Fold Cross Validation Result

		AUC	Overall Accuracy	False Positive Rate	False Negative Rate
Logistic Regression	Baseline model	0.6931	0.7288	0.2426	0.4621
	Sentiment model	0.6955	0.6915	0.2994	0.3690
Random Forest	Baseline Model	0.7228	0.7256	0.2545	0.4069
	Sentiment Model	0.7274	0.8357	0.1033	0.5724
ANN	Baseline model	0.6726	0.7171	0.2530	0.4828
	Sentiment model	0.6838	0.7081	0.2664	0.4621

2.6. Additional Analysis

2.6.1. The Number of ICMW

Next, this study examines the number of ICMWs. Table 2.8 provides the frequencies of the number ICMWs (*Countweak*). While 88.37% of the sample has zero weakness,

5.15% has only one material weakness. The rest has more than one weakness. Thus, a multinomial logit model is presented in Table 2.9. The dependent variable, *Countlevel*, is coded 2 for more than one material weakness, 1 for one material weakness, and 0 for no weakness. A priori, one would expect a negative association between the sentiment features and the level of the number of the material weakness. Columns (1)-(4) present the coefficients for the effect of the variables on the likelihood of *Oneweak* (columns (1)-(2)) and *Moreweak* (columns (3)-(4)) relative to the likelihood of there not being a weakness, where *Oneweak* equals 1 if the company reports one material weakness, and 0 if there is no material weakness; *Moreweak* equals 1 if the company reports more than one material weaknesses, and 0 if there is no material weakness. Column 5 presents the difference in the coefficients to test whether the coefficient is significantly different across the two circumstances. Among the 1651 observations, 107 firm-years have more than one material weakness and 85 firm-years report only one weakness.

Results in columns (1) – (4) indicate that, the results of the primary estimation in Table 2.6 generally carry over to both circumstances of internal control material weaknesses. *Marketvalue*, *Segments*, and *Resign* are significant to *Oneweak* and *Moreweak*. *Marketvalue* is more significant (at 0.01) to *Moreweak* than to *Oneweak* (at 0.5). Similarly, Segment is significant at 0.05 for the prediction of *Moreweak*, while it is significant at 0.1 for *Oneweak* prediction. *Resign* is equally significant (at 0.01 level) for both two cases. Sentiment is positive and significant at 0.1 level for firms with more than one weakness but not significant for firms with one material weakness. *Joy* is significant at 0.01 and negatively associated with *Moreweak* but insignificant for *Oneweak*.

Column (5) indicates that there are significant differences at various level between coefficients of *Sentiment*, *Joy*, and *Resign* for the two kinds of problems[18]. Thus, the number of material weakness is greater in situations of higher sentiment score, lower joy score, and auditor resignations. Other factors, such as market value and number of segments, do not affect the relative likelihood of one material weakness versus more material weakness.

Table 2.8 The Number of ICMW

Countweak	Frequency	Percentage	Cumulative percentage
0	1459	88.37	88.37
1	85	5.15	93.52
2	50	3.03	96.55
3	29	1.75	98.30
4	9	0.55	98.85
5	9	0.55	99.40
6	2	0.12	99.52
7	4	0.24	99.76
8	1	0.06	99.82
9	1	0.06	99.88
18	1	0.06	99.94
20	1	0.06	100.00
Total	1651	100	

[18] The result of Z score is not discussed as it is insignificant to both cases of ICMW

Table 2.9 Multinomial Logistic Regression

		Oneweak vs. Noweak		Moreweak vs.Noweak		(3)-(1)[19]
Independent Variable	Expected Sign	Coefficient	P-value	Coefficient	P-value	
		(1)	(2)	(3)	(4)	(5)
Intercept	+/-	-17.3622	0.996	-31.5101	0.986	-14.1479
Sentiment	-	-0.8239	0.421	1.8955*	0.061	2.7194*
Joy	-	-0.2783	0.679	-2.4116***	0.001	-2.1333**
Marketvalue	-	-0.1967**	0.016	-0.2944***	0.001	-0.0977
Aggregateloss	+	-0.5376	0.118	-0.1183	0.695	0.4193
Zscore	-	-0.0116	0.122	0.0158	0.111	0.0274**
Segments	+	0.3123*	0.097	0.3994**	0.023	0.0871
Foreign	+	-0.4587	0.551	14.7832	0.991	15.2419
Inventory	+	0.1788	0.883	0.3791	0.746	0.2003
Restructure	+	0.0677	0.790	-0.2939	0.240	-0.3616
Acquisition	+	0.0342	0.893	0.1367	0.559	0.1025
Resign	+	1.6018***	0.004	2.6901***	0.001	1.0883*
Big4	-	-0.2373	0.397	0.0069	0.979	0.2442
Litigation	+	-0.1963	0.620	0.5673	0.141	0.7636
Industry indicator variables		Included		Included		
Number of total observations		1651 (107 Moreweak, 85 oneweak, and 1459 Noweak)				
Likelihood ratio, χ^2 (p-value)		149.07*** (0.0001)				
Pseudo R^2		0.1027				

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

2.6.2. The Persistency of ICMW

To examine the effect of the sentiment features on the persistency of ICMW, companies with ICMW are divided into two categories. The first category includes companies with first year ICMW (which means this is the first year for the past three years that the company has ICMW), while the second category consists of companies with ICMW persistence (which means the company has other ICMW for the past three

[19] Models the probability of *Moreweak* relative to the probability of *Oneweak*..

years). In Table 2.10, it re-estimates the logistic regressions with the alternate dependent variables, *Firstyrmw* and *Persistmw*, where *Firstyrmw* equals 1 if it is the first year the company has ICMW, and 0 if the company does not disclose ICMW; *Persistmw* equals 1 if, besides the ICMW in this year, the company has other ICMW for the past three years, and 0 if the company does not have ICMW in this year. Note that the regression now compares the firms in each group to the original *Compustat* control group (companies with no ICMW) and not to the other groups. This analysis excludes *Growth* as this variable limits the sample size.

The first estimation has *Firstyrmw* as the dependent variable, and its result is shown in column (1) and (2) of Table 2.10. The number of observations with first year ICMW is 37, while the number of firms that do not report ICMW in this year is 1462. While *Joy* is not significantly related to *Firstyrmw*, *Marketvalue* is negatively associated with the dependent variable at 0.01. *Acquisition* is positive and significantly related to *Firstyrmw.* The likelihood ratio χ^2 is 29.24 with a P value of 0.1726, indicating that the overall model is not statistically significant. This result suggests that, with the current variables, the likelihood of first year material misstatements cannot be properly predicted.

The second model uses *Persistmw* as the dependent variable, and the results are reported in column (3) and (4). As shown in table 2.10, 92 companies persistently have material weakness. *Joy, Marketvalue, Segment, and Resign* are all significant in the hypothesized directions at p-values less than 0.01 (or 0.05) under one-tailed tests in this estimation. This suggests that companies that persistently report ICMW are larger, more complex and diversified, and more likely to have disagreement with the auditors. The coefficient of *Joy* is -1.7097, stronger than its counterpart (-1.3762) in the primary

estimation reported in table 2.6, suggesting that *Joy* performs more effectively when it is used to determine persistent ICMW than to determine general ICMW. Table 2.10 also shows that companies with historical ICMW do not appear to have lower overall sentiment score, which is consistent with the result of the primary estimation. In addition, compared to the primary estimation, this model has higher Pseudo R^2, which is 0.1327 (as opposed to 0.0827).

2.7. Conclusion, Limitation, and Future Research

2.7.1. Conclusion

This chapter examines the incremental informativeness of sentiment features of conference calls in identifying existing ICMW disclosed under SOX 404. The transcripts of conference calls from 2004 to 2014 are analyzed with Alchemy Language API, a textual analysis tool powered by deep learning, an emerging AI method that is built with a large number of training data to learn the underlying data pattern. Since the model is continuously trained and tested by new data, the classification errors are continuously decreasing and the performance is improving. As a result, with deep neural network a computer can perform more effectively and efficiently than a human expert and this technology has been widely applied for big data analysis (Sun and Vasarhelyi, 2017).

Table 2.10 Logistic Regression of the Probability of ICMW by the Persistency

		Dependent variable: Firstyrmw[20]		Dependent variable: Persistmw[21]	
Independent variable	Expected Sign	Coefficient	P-value	Coefficient	P-value
		(1)	(2)	(3)	(4)
Intercept	+/-	-2.2647***	0.007	-2.6044**	0.027
Sentiment	-	-0.5544	0.722	0.3843	0.719
Joy	-	-0.6643	0.530	-1.7097***	0.009
Marketvalue	-	-0.3090***	0.009	-0.2614***	0.001
Aggregateloss	+	0.6380	0.123	-0.5148	0.150
Zscore	-	0.0086	0.575	0.0004	0.964
Segments	+	0.2556	0.375	0.3628**	0.050
Foreign	+	0.1887	0.441	0.3161	0.765
Inventory	+	-0.3979	0.835	0.9401	0.395
Restructure	+	0.0986	0.802	-0.1742	0.512
Acquisition	+	0.6709*	0.082	0.0506	0.842
Resign	+	0.4467	0.705	2.6591***	0.001
Big4	-	-0.3437	0.398	-0.2630	0.339
Litigation	+	0.5390	0.348	0.5584	0.176
Industry indicator variables		Included		Included	
Number of total observations		1499		1554	
Number of observations with no MW		1462		1462	
Number of MW observations[22]		37		92	
Likelihood ratio,χ^2		29.24		92.11***	
(p-value)		0.1726		0.0001	
Pseudo R^2		0.0849		0.1327	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

[20] Firstyrmw=1 if it is the first year in the past three years that the company has material weakness(es), and =0 if the company has no material weakness in the current year.

[21] Persistmw=1 if the company, besides the current year's material weakness(es), had material weakness(es) in the past three years, and =0 if the company has no material weakness in the current year.

[22] It refers to the number of firm-years when Firstmw=1 and Persistmw=1, respectively

This research extends the application of deep learning to auditing domain. It obtains the overall sentiment score of the document as well as the confidence score of the emotion joy and use them as additional predictors to identify the probability of ICMW in companies. The results of the primary analysis indicate that, after taking the sentiment features into consideration, the explanatory ability of the model improves significantly, as opposed to the baseline model that merely utilizes the major ICW determinants suggested by prior literature (i.e., Doyle, Ge, and McVay, 2007a; Ashbaugh-Skaife, Collins, and Kinney, 2007). To examine the predictive performance of the model with these sentiment features, it applies three classification algorithms, Logistic Regression, Random Forest, and traditional Artificial Neural Network with one hidden layer consisting of 100 neurons, to develop the prediction models. The prediction results show that, while Random Forest achieves the best performance, model with sentiment features under all three machine learning algorithms generally outperform their baseline models with higher AUCs. In sum, the sentiment features, especially the joy score, improves both explanatory and predictive ability of the model for ICMW identification.

Additionally, using the sentiment features and other determinants of internal control weakness, this study investigates the relationship between the sentiment features and the number of material weakness and the persistency of material weakness, respectively. For the problem of material weakness count, a multinomial logistic regression is employed. The results show that as compared to the situation of one material weakness, companies with more than one material weakness have higher sentiment score, lower joy score, and more likely to have auditor resignation. To investigate the informativeness of sentiment features for the persistency of internal control weakness, this research considers two

circumstances: first year weakness and persistent weakness. The first circumstance refers to the fact that this is the first year in the past three years that the company reports ICMW, whereas the second one is that the company has other ICMW in the past three years. The finding is that the model with the current independent variables does not explain the first circumstance properly. However, for the second circumstance, results suggest that companies that persistently report ICMW appear to be larger, more complex and diversified, and more likely to have auditor resignation. Furthermore, the effect of *Joy* on the dependent variable is stronger than it is in the primary estimation, indicating that joy score determinates persistent ICMW situation more effectively than it determinates general ICMW condition. The analysis also finds that companies with historical ICMW are not likely to have lower overall sentiment score, which is consistent with the result of the primary estimation. In addition, compared to the primary estimation, this model has higher explanatory ability.

2.7.2. Limitation and Future Research

This study is subject to limitations. In particular, the deep learning algorithm applied in this study is not exclusively trained with finance-specific data, which may decrease the prediction accuracy. In the future, more finance-specific data should be collected, labeled, and used as the training set for a finance-specific deep learning model to support related decision making.

Secondly, this paper does not separate managers' discussion and analysts' questions in conference calls. To isolate the managers' sentiment and its effect, future work is needed to exact managers' speech out of the transcripts of conference calls.

Thirdly, while the presentation section of the conference calls can be scripted in advance, the Q&A section is hard to be prepared. This is because the management is unsure of the exact information needs of the participants during the conference call. As a result, it would benefit from separating the Q&A part from the presentation part of the conference call, which could be another direction for future research.

Chapter 3 The Performance of Sentiment Features of MD&As for Financial Misstatements Prediction: A Comparison of Deep Learning and Bag of Words Approaches

3.1. Introduction

This essay explores whether the sentiment features elicited from the transcripts of Management's Discussion and Analysis (MD&A) sections of 10-K filings are useful for predicting financial misstatements. Financial misstatement is of considerable interest to financial statement users. Prior literature examines a variety of finance and non-finance quantitative factors as financial misstatement predictors (e.g., Beneish, 1999; Dechow et al., 2011; Huang, Rose-Green, and lee, 2012; Cecchini et al., 2010, Perols et al., 2017). However, researchers argue that since most of the quantitative attributes are disclosed by financial statements, they may contain misleading information that does not fairly present the financial position and the performance of the company. This is because the management has the incentive to distort the information to present the company more favorably (Öğ˘üt et al., 2009).

As the research in social psychology suggests, emotions and cognitive processes of the speaker could result in linguistic cues that can help identifying deceptions (Zuckerman and Driver, 1985, DePaulo et al., 2003). Managers with the knowledge of the existence of fraud or errors has the intention to uncover the truth, and the cognitive processes could be revealed by some sentiment features from their language. Consistently, prior literature emphasizes the importance of text documents and points out that words and phrases in conferences calls, MD&As, audit reports, SEC comment letters, press release, and other business communication documents provide incremental

qualitative evidence of sentiment and other linguistic features that can be used to uncover financial misstatements (e.g., Larcker and Zakolyukina, 2012; Lee, Lusk, and Halperin, 2014; Czerney, Schmidt, and Thompson, 2014). Burgoon et al., (2016) document that certain linguistic and vocalic features (e.g., pitch and voice quality, vocal intensity, hedging-uncertainty, and immediacy-nonimmediacy) from earnings conference calls are related to future financial misstatements. Since financial misstatements are caused by unintentional errors or intentional fraud (AICPA 2011; AICPA, 2014), prior research on this topic is usually in conjunction with fraud detection. Hobson, Mayew, and Venkatachalam (2012) examine whether vocal markers of cognitive dissonance are useful for detecting financial misreporting. Using speech samples of CEOs during earnings conference calls, they find vocal dissonance markers are positively associated with the likelihood of fraudulent statements. However, such relationship is not supported for error-caused financial misreporting. Over the past few decades, besides investigating the statistical relationship between the sentiment features of text and financial misstatement, researchers have intensified their efforts to predict financial misstatements by developing various machine learning models. Examples include Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN) (Tsai and Chiou, 2009; Ögˇüt et al., 2009).

To extract linguistic characteristics of business communication documents, “bag of words” approach has been widely used in prior research. In particular, some psychosocial dictionaries such as General Inquirer (GI) or Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) are used to calculate the sentiment and other features. Tetlock (2007) captures investor sentiment from the Wall Street Journal by measuring the

pessimism index consisting of mostly negative and weak words from GI. But Loughran and McDonald (2011a) argue that dictionary designed to extract sentiments in ordinary speech may not apply properly to business document. They develop an alternative word lists that better reflect sentiment in financial text and find that their word lists are associated with 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings. The "bag of words" approach represents a text as the bag of its words and counts the frequency of particular words based on the predefined word lists. Despite its simplicity, replicability, and relevance for specific content, this approach does not consider the sequence, grammar, and the structure of the sentence (Larcker and Zakolyukina, 2012). Since different combination of the same words can imply different meanings, researchers believe "bag of words" approach is too simplistic to obtain the accurate meaning of the text (Salton and McGill, 1983).

An alternative approach for text analysis is deep learning, a new frontier in machine learning based on deep neural networks. It has the capability of automatically extracting features from data, especially the unstructured or semi-structured data such as videos, audios, and text. Trained with big volume of data and relies on its deep hierarchical structure of neural network layers with powerful computational capability, a textual analysis model based on deep learning is able to "learn" the patterns underling the text, "understand" its meaning, and output abstract features extracted from the text (Issa, Sun, and Vasarhelyi, 2016). This provides an opportunity to apply deep learning to analyze text to predict financial misstatements. Despite its success in Natural Language Processing and other related areas, limited research in accounting and auditing has

applied the deep learning technology, and this study aims to bridge this gap and extend this line of research by examining the effectiveness of deep learning extracted sentiment features on misstatement caused by frauds and errors. IBM Watson provides a textual analysis tool, Alchemy Language API[23], powered by deep neural network, linguistic, and statistical algorithms. It can read and understand text from various topics as it is trained with more than 200 billion texts of English website and news and continuously processes over 3 billion API requests per month from different industries (Turian, 2015). One obvious superiority of this tool is that it is able to directly read the webpage or HTML/text document with an input of an URL or HTML/text document. No text preprocessing actions are needed. However, unlike "bag of words" approach which utilizes finance-specific dictionary, the deep learning text analysis model is not exclusively trained with finance-specific content. Therefore, it is unclear whether it could provide sentiment features that are appropriate for financial misstatement prediction task.

The objective of the study is to provide insight into three questions: (1) Do sentiment features add information for financial misreporting prediction? (2) If the answer is yes, are they effective for fraud prediction only or for both fraud and error? (3) How effective the model using sentiment features obtained with deep learning technique performs as compared to the model using sentiment feature obtained by "bag of words" approach? This research analyzes 31,466 MD&As from 10-K filings corresponding to fiscal year 2006 to 2015 with "bag of words" and deep learning approaches and obtained three sentiment features, *Sentiment_TM*, *Sentiment_DL*, and JOY, where *Sentiment_TM* is calculated based on the frequency of positive and negative words in L&M word list

[23] https://www.ibm.com/watson/developercloud/alchemy-language.html

(Loughran and McDonald, 2011a), and *Sentiment_DL* and *JOY* are the sentiment and joy emotion extracted by Alchemy Language API. Among the 31,466 MD&As, 321 documents are related to misstated financial statements as identified by AuditAnalytics^{TM}. Besides the sentiment-related attributes, this research uses 82 factors related to financial misreporting based on previous research (Beneish, 1999; Dechow et al., 2011; Huang et al, 2012; Cecchini et al 2012, Perol et al., 2017).

Five machine learning algorithms, including the Random Forest, Logistic Regression, Naïve Bayes, Deep Neural Network, and Traditional Neural Network, are employed to analyze the same dataset. With each algorithm, three prediction tasks are conducted, which are detecting misstatements, predicting frauds, and identifying errors. For each task, it establishes three classification models. The first model is called baseline model. It uses solely the 82 financial misstatement-related factors provided by prior studies, without considering any sentiment measures of MD&A. The other two models have the identical structure with the exception of the sentiment measures. While model 1 uses *Sentiment_DL* as the sentiment measure and *JOY* as the emotion feature, model 2 employs *Sentiment_TM* as the sentiment measure. In total, 45 models are established.

The prediction results show that Random Forest algorithm outperforms other machine learning algorithms in terms of all evaluation metrics especially for fraud classification. In addition, generally models with sentiment related variables perform more effectively than the baseline models and prior models built by existing research, as evidenced by better AUC values. However, the predictive ability of sentiment features of MD&As is observed only for fraud detections. In other words, the sentiment features are informative only when managers have the intention to misreport.

The remainder of the essay consists of seven sections. Section 2 reviews prior literature. Section 3 discusses the two sentiment extracting approaches. Section 4 describes the research design. The results and discussion are presented in Section 5 and Section 6, respectively. Finally, conclusion and limitations are provided in Section 7.

3.2. Prior Literature

3.2.1. Financial Misstatement Detection

Research on financial misstatement prediction proposes financial and nonfinancial factors that can be used to predict fraud. Dechow et al. (2011) analyze financial characteristics connected to misstating companies involved in Accounting and Auditing Enforcement Releases (AAER) and find that several measures of accrual quality, gross profit, “soft” assets, and other financial factors are highly associated with misstatements. Beneish (1999) finds that firms with earnings overstatements that violate GAAP are likely to trade the holdings of stock and exercise stock appreciation rights, and the sales occur at inflated prices. Consistently, Summers and Sweeney (1998) report that the management tends to reduce the holdings through high levels of selling activity, in terms of the number of transactions, the number of shares sold, or the dollar amount of shares sold. Beasley (1996) conduct an empirical analysis of the relation between the board of director composition and financial statement fraud. Huang, Rose-Green, and Lee (2012) provide evidence that CEO age is positively associated with financial reporting quality.

Another stream of literature applies a variety of misstatement detection methods on a sample of fraudulent and nonfraudulent financial statements. Cecchini et al. (2010) provide a methodology based on Support Vector Machines (SVM) to detect frauds with

financial data. In particular, with a financial kernel, the power of the learning machine is increased to be able to correctly labeled 80% of the fraudulent companies. Goel et al. (2010) focus on the analysis of text in annual reports to detect fraud. They examine both the verbal content and the presentation style of the qualitative portion of the annual reports using natural language processing tools and find that linguistic features like tone, voice, readability index, etc. can improve the prediction accuracy of their fraud detection model. Specifically, they use two versions of 10-K forms: one consisted of 1027 documents (405 fraudulent 10-Ks vs. 622 non-fraudulent 10-Ks); the other consisted of 1,375 documents (405 fraudulent 10-Ks vs. 970 non-fraudulent 10-Ks). SVM is used to build the classification model and reached an accuracy of 89.51% for version 1 dataset (the accuracy of the classification model for Version 2 dataset is 89.04%).

3.2.2. Sentiment features of MD&A and Financial Misstatements

SEC requires public companies to disclose annual reports on Form 10-K. The annual report on Form 10-K provides a comprehensive overview of the company's business and financial condition[24]. The Management's Discussion and Analysis (MD&A) section in the 10-K form is considered a vital conduit of information to investors as it is the management’s narrative explanation of a company's financial statements, containing the management’s perspective on the current status of the company in the industry and future prospects of the company (Wheeler and Cereola, 2015; Humpherys et al., 2011). As it contains more inclusive information than does the audited financial statement, auditing standards such as SAS No. 118 (AICPA, 2010) and its predecessor, SAS No. 8 (AICPA,

[24] See https://www.sec.gov/answers/form10k.htm

1975), encourage auditors to examine MD&As for information indicating possible financial irregularities. The importance of the MD&A is supported by the prevalent notion that the qualitative contents provide incremental information to the quantitative contents (Kothari, Li and Short 2009; Bochkay and Levine, 2014). MD&As are a valuable source of clues for financial misstatement detection. Combining the qualitative contents with the traditional quantitative information assist decision makers to obtain a holistic view of the firm's situation. In a company with financial misstatements, the deceptive management will to use different linguistic cues as compared to the honest management in a company that is free of misstatement (Humpherys et al., 2011). This provides an opportunity to discriminate misreporting from non-misreporting for financial statements.

By analyzing the linguistic features of the MD&A, extant research finds that there are numerous clues (e.g., the sentiment, the complexity, and the readability) underlying the language expressed by the management, which could be used as predictors of financial misstatement. Churyk, Lee, and Clinton (2009) find significant differences of linguistic coding used in the MD&A of 10-K forms between fraudulent firms and non-fraudulent firms. The results show that compared to ones not filling restatements disclosed by AAER (Accounting and Auditing Enforcement Release), firms with restatements tend to use more words, less terms with positive emotions like optimism and energy, and more terms with negative emotion like anxiety. Similarly, Humpherys et al. (2011), by showing that MD&As of fraudulent firms are significantly more likely to contain active language than those of non-fraudulent firms, demonstrate that "*linguistic models of deception were potentially useful in discriminating deception and managerial*

fraud in financial Statements". Loughran and McDonald (2011b) conclude that the appearance of a list of 13 problematic phrases in 10-Ks are significantly related to fraudulent financial statements.

However, not all financial misstatements are fraudulent restatements. Financial misstatement can arise from either fraud or error, depending on whether the misstatement is caused by intentional or unintentional actions (Humpherys et al., 2011). Although errors are unintentional, they are usually caused by deficiencies or weaknesses in internal control. Management of companies with deficiencies or weaknesses has the incentive to not disclose the true situation of their internal controls, and cues for this fact can be observed from the text of MD&As. Thus, the linguistic features of MD&A sections could be associated with both intentional and unintentional financial misstatement. There are studies exploring the relationship between linguistic features of text documents (such as MD&As and conference calls) and financial misstatements including both intentional and unintentional misreporting. Lee, Lusk, and Halperin (2014) point out that the data used by Churyk et al. (2009) and Humpherys et al. (2011) is for the time period prior to SOX-era. They use data in the SOX-era and assert that the language cues of MD&As are still powerful to signal financial misstatement in the SOX-era.

3.3. Approaches of Textual Analysis

3.3.1. "Bag of Words" Approach

Extant papers in accounting and finance have extracted linguistic features from business communication documents, such as corporate disclosures (e.g., Loughran and McDonald, 2011a), press release (e.g., Davis, Piger, and Sedor, 2012), earnings

conference calls (Larcker and Zakolyukina, 2012), and media news (e.g., Tetlock, 2007). A prevalent and simple approach to obtain these linguistic features (e.g., sentiment) from text documents is called “bag of words”, which represents a text as the bag of its words. To extract features from the text, the frequency of particular words is counted based on the predefined general and finance-specific dictionary. Two of the most popular general dictionaries are “General Inquirer” (GI) and “Linguistic Inquiry and Word Count” (LIWC). Tetlock (2007) examines the “Abreast of the Market” column in Wall Street Journal and measures the investor’s pessimism sentiment by counting negative and weak words listed in GI dictionary. Similarly, the negative and positive word lists of GI dictionary are employed by Kothari, Li, and Short (2009), who construct firm-specific disclosure measures from more than 100,000 disclosure reports by management, analysts, and news reporters. The examples of research relying on LIWC include Churyk et al. (2009) and Lee, Lusk, and Halperin (2014). Since general dictionaries are designed for the generic English language, they do not contain certain words that are considered positive or negative in financial documents only and include some generally negative words like “liability” that are not negative in the financial context (Henry and Leone, 2009; Loughran and McDonald, 2011a). Therefore, researchers argue that finance-specific word lists are more appropriate for business communication analysis (Henry and Leone, 2009; Loughran and McDonald, 2011a). Li (2006) develops a risk sentiment words list for 10-Ks, containing words related to “risk” (including “risk”, “risks”, and “risky”) and “uncertainty” (including “uncertain”, “uncertainty”, and “uncertainties”). Core, Guay, and Larcker (2008) manually select a list of keywords and phrases with negative tone by reading approximately 200 press articles about CEO compensation.

Loughran and McDonald (2011a) create a financial dictionary (L&M List) consisting of all words that occurred in at least 5% of the 10-Ks from 1994 to 2008. The finance-specific dictionary has been proved to be significantly associated with 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings. Because it is designed for 10-K text analysis and the selected words are precise and well-targeted, the L&M list has been widely used by subsequent studies (e.g., Garcia, 2013; Jegadeesh and Wu, 2012; Chen et al. ,2013; Loughran and McDonald, 2013). Consequently, this study uses L&M word list to extract sentiment feature from MD&As from 10-Ks. Despite its simplicity and replicability, "bag of words" approach ignores language grammar, word sequence, as well as the various combinations of same words or phrases conveying different meanings (Manning and Schutze, 1999; Larcker and Zakolyukina, 2012). Another issue of this technique is that the manually selected word list is subjective as it mainly relies on the researcher's personal judgment and rules of extraction.

3.3.2. Deep Learning Approach

As an emerging AI technique, deep learning learns the pattern of the data from examples, and the learning process requires no human intervention. A deep learning model builds deep hierarchical layers consisting of numerous neurons to transform input data and identifies the pattern underlying the data. In a deep neural network, each layer applies a nonlinear transformation on its input layer and provides a representation. That is, the output representation of each input layer is provided as input to its next layer. As the input data goes deeper, the nonlinear transformation constructed becomes more complicated and the representation becomes more abstract. The output of the final layer

is the final representation of the input raw data, which provides features extracted from the data that are useful for further classification, association, and other tasks (Najafabadi, et al., 2015). Figure 3.1 presents a simplified example of a deep neural network identifying the sentiments from MD&As.

Deep learning performs effectively for the analysis of big data such as image, audio, video, and text. With its complex computation, a deep learning based textual analyzer such as Watson Alchemy Language API is able to "understand" the meaning of the document by extracting abstract features automatically. These features involve sentiment, emotion, keywords, concepts, relationship among concepts, involved entities, and etc. Realizing the great value that deep learning could add to audit profession, the Big 4 accounting firms are investing hundreds of millions of dollars into such cutting-edge technologies (Kokina and Davenport, 2017). KPMG forms an alliance with IBM Watson artificial-intelligence unit to develop AI tools for bank loan analysis. Deloitte works with Kira, a contract analysis system to develop deep learning models examining complex auditing-related documents (Deloitte, 2016). However, limited research in auditing academia applies this technology to audit procedures. This study aims to bridge the gap by exploring the application of deep learning in textual analysis for MD&A sections of 10-K filings and compares the power of the sentiment features extracted by deep learning text mining approach for financial misstatement detection.

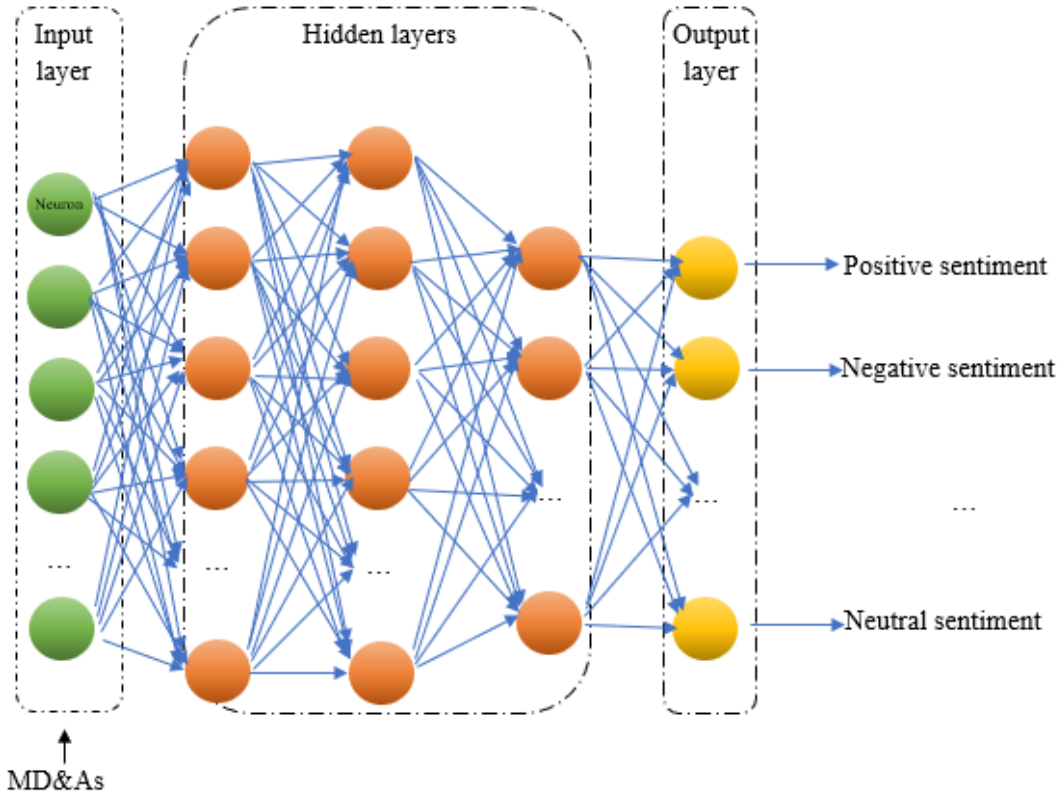

Figure 3.1 A Deep Neural Network for MD&A Sentiment Analysis

This essay employs IBM Watson Alchemy Language API, which is able to read and understand text and provides sentiment scores as the output. Alchemy API is continuously trained with more than 200 billion words from English websites such as tweets, blog posts, and Facebook comments as well as news articles. The training data set comes from dozens of industries, providing Alchemy API the ability to analyze a wide range of topics (IBM Watson, 2015).

This tool can automatically analyze both HTML/text document and webpage. Unlike the “bag of words” method which requires time-consuming data preprocessing steps, it

does not need a human user to preprocess the data as the model directly reads the webpage or HTML/text document with an input of an URL or HTML/text document. Alchemy Language API can automatically remove advertisements, navigation links, and other irrelevant content and perform analysis[25]. The returned value of the textual analysis service used by this work is the sentiment score, the overall attitude of the given document, as well as the emotion score of joy, measuring the how confident that the DNN model "believes" that text expresses joyful emotion. Table 3.1 compares the two textual analysis approaches.

Table 3.1 Deep Learning and "Bag of Words" Approaches

	Deep learning approach	Bag of words approach
Description of the technique	Emerging technique employing deep hierarchical neural network and trained with a large amount of text files	Prevalent technique using various word lists (dictionary), with each one representing a particular sentiment feature
Rationale	"understand" the meaning of a text file: extract high-level and abstract features from raw data by building complex concepts out of simpler concepts	count the frequency of the words originated from a specific dictionary
Output sentiment feature	Sentiment scores	Output sentiment feature
Tool	Alchemy language API	Loughran and McDonald (2011a)
Is it a finance-specific tool	No	Yes
Required text document	HTML/text document and webpage	HTML/text document
Does it need data preprocessing	No	Yes

[25] More information is available at http://www.alchemyapi.com/api/combined/htmlc.html

3.4. Research Design

3.4.1. MD&A Data

The management discussion and analysis (MD&A) sections of 10-Ks are obtained from SeekiNF[26]. The initial sample involves 61,686 MD&As filed from 2007 to 2015. This study eliminates 28,515 MD&As missing matched information of CIK or fiscal year specified in Compustat^{TM}. For those with multiple 10-Ks or with 10-K/A[27]s, only the latest one is kept, and the deleted ones are called "duplicated" MD&As in this research. This process further removes 1,705 MD&As. The remaining MD&A sample has 31,466 records corresponding to annual reports of fiscal year from 2006 to 2015. The sample selection process is reported in Table 3.2.

Table 3.2 Sample Selection of MD&As

MD&As filed from 2007 to 2015 as provided by SeekiNF	61686
Less: MD&As without matched information of CIK or fiscal year as specified in *Compustat*	(28515)
Less: "duplicated" MD&As	(1705)
Remaining sample	31,466

3.4.2. Misreporting Data

This study uses *AuditAnalytics* restatements database available via Wharton Research Data Services (WRDS) to identify misreporting sample. This database is queried in December 2016 to identify any restatements caused by financial misreporting[28]

[26] https://www.seekedgar.com/seekinf.html
[27] 10-K/As are amended filings for previously issued 10-Ks
[28] The original restatement sample includes 4351 firm-years. This study excludes 4,030 restatement observations caused by accounting rule (GAAP/FASB) applications failure.

for the fiscal years in the MD&A sample. To differentiate intentional (i.e., irregularities) from unintentional (i.e., errors) misstatements, this study refers to the data field RES_FRAUD and RES_CLER_ERR in *AuditAnalytics*, which provide information regarding the sample of misstatement in terms of whether they are intentional fraudulent reports or unintentional errors made by accounting clerks[29]. It finds that 321 of the sample firm-years have financial misstatements. Among the 321 records, 104 observations contain frauds and 218 firm-years have errors. In addition, this paper obtains the information of the auditor for the financial statement from AuditAnalytics^{TM}. The composition of the misreporting data across fiscal years is depicted in Table 3.3. It is found that more financial misstatements are in 2007 and 2008 than other years. This could be attributed to the global financial crisis of 2007–2008. In addition, there are fewer misstatements in recent years as some misstatements may not be identified until they will eventually be restated in the future.

Table 3.3 Distribution of Misstatements across Fiscal Years

Fiscal Year	Misstatements	Frauds	Errors
2006	43	7	36
2007	64	14	50
2008	65	13	52
2009	47	15	32
2010	39	15	24
2011	12	8	4
2012	14	9	5
2013	23	14	9
2014	12	7	5
2015	2	2	0
Total	321	104	217

[29] where, RES_FRAUD is equal to 1 if the misstated financial statement is fraudulent, and 0 otherwise; RES_CLER_ERR is equal to 1 if the misstatement is caused by errors of the accounting clerk.

The distribution of financial misstatements across industries is shown in Table 3.4. Manufacturers and computers are more susceptible to misstatements than others. Industries with fewer misstatement cases reported include agriculture, utilities, textiles & apparel, chemicals, and so on.

Table 3.4 Distribution of Misstatements across Industries

Industry[30]	Misstatements	Frauds	Errors
Agriculture	5	3	2
Mining & Construction	15	2	13
Food & Tobacco	7	0	7
Textiles & Apparel	6	2	4
Lumber, Furniture, & Printing	15	11	4
Chemicals	6	3	3
Refining & Extractive	9	1	8
Durable Manufacturers	73	28	45
Computers	71	18	53
Transportation	15	2	13
Utilities	2	1	1
Retail	25	9	16
Services	33	12	21
Banks & Insurance	12	2	10
Pharmaceuticals	27	10	17
Total	321	104	217

3.4.3. Sentiment Measures

This essay uses two approaches to obtain sentiment scores: deep learning approach and "bag of words" approach. With deep learning approach, the Alchemy Language API returns the overall sentiment score of the MD&A text, measuring the overall attitude in

[30] Industry classifications are compiled using the following SIC codes: Agriculture: 0100–0999; Mining & Construction: 1000–1299, 1400–1999; Food & Tobacco: 2000–2141;Textiles and Apparel: 2200–2399; Lumber, Furniture, & Printing: 2400–2796; Chemicals: 2800–2824, 2840–2899; Refining & Extractive: 1300–1399, 2900–2999; Durable Manufacturers: 3000–3569, 3580–3669, 3680–3999; Computers: 3570–3579, 3670–3679, 7370–7379; Transportation: 4000–4899; Utilities: 4900–4999; Retail: 5000–5999; Services: 7000–7369, 7380–9999; Banks & Insurance: 6000–6999; Pharmaceuticals: 2830–2836, 3829–3851.

the content of the document that is being analyzed. The score ranges from -1 to 1, where a positive value represents positive sentiment ("1" represents both positive and negative sentiment), a zero means neural, and a negative value stands for negative tones. The score measures the strength of the sentiment within the document as predicted by the deep neural network. This study calls the sentiment score obtained from Alchemy Language API "Sentiment_DL". Furthermore, it employs the confidence score for joy, called JOY, to indicates the probability that the emotion of joy is implied by the sample text. JOY is derived from a stacked generalization-based ensemble framework powered by a combination of machine learning algorithms (including deep learning) and language features such as words, phrases, punctuation, and the overall sentiments[31]. The joy index is provided by Alchemy API as well and is ranged from 0 to 1, with 0 representing no joy at all and 1 indicating the maximum of joy.

With "bag of words" approach, it refers to the positive and the negative word list of Loughran and McDonald (2011a[32]) to obtain the sentiment score, "Sentiment_TM". This method requires data preprocessing steps as follows:

(1) for each conference call, a text file of the transcript is processed by removing tags from HTML documents (MD&As), transferring HTML characters to text characters, dropping non-linguistic marks such as "^", "$", and replacing all the blank lines and duplicate spaces with a single space

[31] https://console.bluemix.net/docs/services/tone-analyzer/science.html#the-science-behind-the-service
[32] This essay uses the March 2015 version of LM word lists from http://www3.nd.edu/~mcdonald/Word_Lists.html)

(2) words are converted into lower case so that words like “fraud” and “Fraud” will not be identified as two different words

(3) punctuation, stop words such as “the”, “for”, “of”, and numbers are removed

(4) each word of a conference call transcript is identified and then categorized on the basis of whether it is included on the positive or the negative word list of Loughran and McDonald (2011a). This step generates raw word counts of positive words, negative words and a total word count, which are used to compute the positive score and the negative score:

Positive score = the word count of positive words/the total word count

Negative score = the word count of negative words/the total word count

Moreover, this study follows Druz et al (2015) to correct for negation. It excludes a positive word from the count when a negation word (no, not, none, neither, never, nobody, *n' t) presents among the three words preceding the positive word (except when there is a comma, a period, a semicolon, or a question mark in that range).

(5) the sentiment score is constructed by computing the difference between positive scores and negative scores. This ratio is bounded between –1 and +1 and provides a metric of the relative positivity of the conference call.

Table 3.5 provides the descriptive statistics of sentiment features.

Table 3.5 Descriptive Statistics of the Sentiment Features

	Obs.	Mean	Min.	P25	Median	P75	Max.
Sentiment_DL	31466	0.0194	-0.5606	-0.0292	0.0194	0.0661	0.7487
Sentiment_TM	31466	-0.0109	-0.0895	-0.0158	-0.0062	-0.0051	0.0419
JOY	31466	0.0593	0.0000	0.0460	0.0501	0.0541	1.0000

3.4.4. Other Variables

In addition to the sentiment features, this research uses 82 predictors for financial frauds and misstatements based on prior research (e.g., Perols, Bowen, Zimmermann, and Samba, 2017; Dechow, et al., 2011; Perols, 2011; Cecchini et al., 2010; Beneish, 1999; Huang, Rose-Green, and Lee, 2012; Churyk, Lee, and Clinton, 2009). The variables are described in Appendix A. It includes all variables from Perols (2011) and all variables[33] from the final model of Dechow, et al. (2011) that can be calculated using Compustat data. This paper also selects six representative variables from the research of Cecchini, et al. (2010), Beneish (1999), Huang, Rose-Green, and Lee (2012), and Churyk, Lee, and Clinton (2009). Those variables are described in Panel D in Appendix A. For instance, the SGAI (Selling, General, and Administrative Expenses Index) is the ratio of sales, general, and administrative expenses to sales in year t relative to the corresponding measure in year t - 1. This variable measures the portion of the SGA expenses in sales, where SGA expenses are used to 1) promote, sell, and deliver a company's products and services, and 2) manage the overall company. The use of this variable follows the recommendation of Beneish (1999) that analysts interpret a disproportionate increase in

[33] The variable "Declining cash sales dummy" is not included as it is similar to variable "Percentage change in cash sales" in the research of Dechow et al. (2011).

sales as a negative signal about a company's future prospects. In addition, Beneish (1999) suggests the use of other two indexes, DEPI and AQI. The DEPI (Depreciation Index), is the ratio of the rate of depreciation in year t - 1 to the corresponding rate in year t. A DEPI greater than 1 indicates that the rate at which assets are being depreciated has slowed, suggesting that the company has revised upward the estimates of assets' useful lives or adopted a new method to increase income. The AQ (Asset Quality Index) is the ratio of asset quality in year t to asset quality in year t- 1, where asset quality is the ratio of noncurrent assets other than property, plant, and equipment (PP&E) to total assets. The AQI measures the change in asset realization risk, which is the propensity to capitalize, and thus defer, costs. All three indexes recommended by Beneish (1999) are related to earnings manipulation. Churyk, Lee, and Clinton (2009) provide empirical evidence that the MD&A for companies that restate their financial statements will contain more words. Thus, FILESIZE, which is the number of words of MD&As, is included as one of the misstatement predictors.

3.4.5. Classification models

This study employs three target variables: MISSTATEMENT, FRAUD, and ERROR[34]. The target variable is also called “class label attribute”, containing values (“0” or “1”) indicating the predefined class (i.e., “misstatement” or “normal statement”) to which each observation belongs.

[34] MISSTATEMENT equals 1 if the financial statement contains any material misstatements; 0 otherwise. FRAUD (ERROR) identifies if the misstatement is caused by fraud (error). It equals 1 if it is a fraud (error) and 0 otherwise.

Identifying financial misstatements can be regarded as a typical two-step classification problem. In the first step, a model is trained with a training dataset. This step is called supervised learning (Russell and Norvig, 2010). In the second step, a testing dataset which does not belong to the training set is used to validate the model. Once the training is successful, the model is expected to successfully classify unlabeled samples in the testing dataset as misstatement or normal financial statement. Subsequently, the accuracy of misstatement predictions is evaluated against the actual misstatement class of the testing dataset. To build and validate the classification models, it utilizes 10-fold cross validation technique (Geisser, 2017). The dataset is highly imbalanced. For example, the ratio of frauds to non-fraudulent statements is 104:31362. To tackle the data imbalance, it employs "over-sampling" method to increase the number of instances from the under-represented classes in the training dataset (Drummond and Holte, 2003).

Five machine learning algorithms are applied to build the model, including the Random Forest, Logistic Regression, Naïve Bayes, Deep Neural Network, and Traditional Neural Network. With each algorithm, three prediction tasks are conducted, including detecting misstatements, predicting fraud, and identifying errors. For each task, it establishes three classification models. The first model is called baseline model. The baseline model uses solely the 82 predictors, without considering any sentiment measures of MD&A. The other two models have identical structure with the exception of the sentiment measures. While model 1 uses Sentiment_DL as the sentiment measure and JOY as the emotion feature, model 2 employs Sentiment_TM as the sentiment measure. Therefore, there are totally 45 models established.

The design of model structure is described in Table 3.6 and the same design applies to all machine learning algorithms. Panel A shows the models for the task of financial misstatement prediction. The target variable is MISSTATEMENT. It includes three models. The first model is the baseline model. The second model uses the sentiment features (the overall sentiment score and the joy score) extracted with deep learning approach, while the third model utilizes the sentiment score calculated based on bag of words approach. Panel B and C present the structure of the other two tasks, which is similar to that of the models in panel A, except that models in panel B use FRAUD as the target variable while models in panel C use ERROR as the target variable.

Table 3.6 The Structure of Models

Panel A: Misstatement Prediction				
		Baseline model	Model 1 (deep learning)	Model 2 (bag of words)
Dependent variable		MISSTATEMENT	MISSTATEMENT	MISSTATEMENT
Independent variables	Sentiment measures	N/A	SENTIMENT_DL JOY	SENTIMENT_TM
	Other predictors	82 variables related to misstatement	82 variables related to misstatement	82 variables related to misstatement
Panel B: Fraud Prediction				
		Baseline model	Model 1 (deep learning)	Model 2 (bag of words)
Dependent variable		FRAUD	FRAUD	FRAUD
Independent variables	Sentiment measures	N/A	SENTIMENT_DL JOY	SENTIMENT_TM
	Other predictors	82 variables related to misstatement	82 variables related to misstatement	82 variables related to misstatement
Panel C: Error Prediction				
		Baseline model	Model 1 (deep learning)	Model 2(bag of words)
Dependent variable		ERROR	ERROR	ERROR
Independent variables	Sentiment measures	N/A	SENTIMENT_DL JOY	SENTIMENT_TM
	Other predictors	82 variables related to misstatement	82 variables related to misstatement	82 variables related to misstatement

3.5. Results

3.5.1. Model Evaluation

Table 3.7 through table 3.11 summarize the predictive performance of the 10-fold cross validation of the five algorithms for these three prediction tasks. It uses the overall accuracy (Accuracy), the false positive rate (FPR, or type I error rate), the false negative rate (FNR, or type II error rate), and the area under the receiver operator characteristic curve (AUC) to evaluate the model performance. The overall accuracy measures the proportion of the accurate classifications in all observations. The FPR is related to the false positive finding, which is the incorrect rejection of a true null hypothesis. In this analysis, false positive means that an observation is identified as a misstatement/fraud/error one when in fact it is normal. Thus, the FPR is the proportion of false positives in all normal observations. Oppositely, the FNR is related to the false negative finding, which is incorrectly retaining a false null hypothesis. In this research, false negative means that an observation is identified as a normal one when in fact it has misstatement/fraud/error. So, the FNR refers to the proportion of the false negatives in all misstatement/frauds/error observations[35]. AUC measures the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots false positive rate (FPR) and Ture Positive Rate[36] (TPR) on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. AUC is a summary of the overall diagnostic accuracy of the model, with values of 0.5 representing a random model without discriminative power and 1 representing a perfectly accurate prediction model. AUC is not affected by the imbalanced distribution of positive and negative observations in the sample.

---

[35] The performance metrics used in this paper is computed as follows:
Accuracy = (TP+TN)/(TP+TN+FP+FN)
False positive rate (type 1 error rate) = FP/(FP+TN)
False negative rate (type 2 error rate) = FN/(FN+TP)

[36] True positive rate (also called sensitivity, hit rate, and recall) is defined as TP/(TP+FN).

- Random Forest

Table 3.7 exhibits the results of the 9 models using Distributed Random Forest (DRF) (Ho, 1995). DRF is a powerful classification and regression technique, which generates a forest of classification (or regression) trees, rather than a single classification (or regression) tree. Each of these trees is a weak learner built on a subset of rows and columns. To reduce the variance, 50 trees are built on the dataset in this study. The final prediction is made by taking the average prediction over all of the trees.

In the case of fraud detection, the baseline model performs worse than the other two models which use the sentiment features, in terms of the overall accuracy (75.36%), the false positive rate (24.64%), the false negative rate (24.04%), and the AUC (0.8288). Furthermore, model 1 that employs sentiment features with deep learning approach outperforms model 2 that uses sentiment features with bag of words approach, as model 1 has higher AUC (0.8524) and accuracy value (77.28%) and lower false positive rate (22.72%) and false negative rate (22.12%).

For error prediction, the performance of all three models is much worse than in the case of fraud detection. The highest AUC is 0.6786, which is achieved by model 2. Models with sentiment features do not exhibit stronger predictive power than the baseline model. Specifically, model 1 performs less effective than the baseline model in terms of a lower value of overall accuracy (61.77%) and a higher false positive rate (38.23%). Model 2 is better than the baseline model, with a slightly higher accuracy (63.73%) and AUC (0.6786) as well as lower false positive rate (36.27%) and false negative rate (36.70%).

For misstatement identification, model 1 and model 2 are not as effectively as they are for misstatement prediction. Model 1 outperforms the rest models in terms of the all four metrics (except the false negative rate, which is the same as that of the baseline model). Furthermore, the superiority of the model 1 is less significant as compared to the case of fraud detection.

Table 3.7 The Results of 10-Fold Cross Validation with Random Forest

		Accuracy	Type one error rate	Type two error rate	AUC
baseline	MIS	66.69%	33.31%	33.33%	0.7232
	FRAUD	75.36%	24.64%	24.04%	0.8288
	ERROR	61.91%	38.06%	38.53%	0.6673
Model 1 (deep learning)	MIS	66.88%	33.11%	33.33%	0.7325
	FRAUD	77.28%	22.72%	22.12%	0.8524
	ERROR	61.77%	38.23%	37.61%	0.6683
Model 2 (bag of words)	MIS	65.23%	34.77%	34.89%	0.7224
	FRAUD	75.76%	24.24%	24.02%	0.8506
	ERROR	63.73%	36.27%	36.70%	0.6786

- Logistic Regression

Logistic Regression is one of the most popular classification technique (Freedman, 2009). The results for Logistic Regression is shown in Table 3.8. For fraud detection, while model 2 has the lowest false negative rate of 31.73%, model 1 outperforms the baseline model and the model 2, in terms of the accuracy (67.93%), the false positive rate (32.07%), and the AUC (0.7473). But the superiority of the models with sentiment features over the baseline model is not observed for the other two prediction tasks. For instance, model 2 outperforms the other models for error prediction, with slightly higher accuracy of 61.37%, lower false positive rate of 38.62%, and lower false negative rate of

38.99%. For the task of misstatement identification, Model 1 performs the best in terms of all metrics.

Table 3.8 The Results of 10-Fold Cross Validation with Logistic Regression

		Accuracy	Type one error rate	Type two error rate	AUC
baseline	MIS	60.51%	39.49%	38.94%	0.6525
	FRAUD	65.70%	34.30%	34.62%	0.7125
	ERROR	60.61%	39.39%	39.45%	0.6331
Model 1 (deep learning)	MIS	62.33%	37.68%	37.07%	0.6695
	FRAUD	67.93%	32.07%	33.65%	0.7473
	ERROR	61.14%	38.85%	39.91%	0.6539
Model 2 (bag of words)	MIS	60.55%	39.45%	38.94%	0.6474
	FRAUD	66.59%	33.42%	31.73%	0.7448
	ERROR	61.37%	38.62%	38.99%	0.6345

- Traditional Neural Network

This research also uses traditional neural network algorithm (VanGerven and Bohte, 2017). It develops a traditional "shallow" Neural Network with one hidden layer consists of 100 nodes to conduct the same three prediction tasks. Table 3.9 summarizes the result. The fraud detection result shows that the baseline model performs better than model 1 and 2. The AUC is 0.7743, slightly higher than that of model 1 and model 2. Similarly, the accuracy of the baseline model is 73.50%, the highest value among the three models. The FPR and the FNR of the baseline model is lower than those of the other 2 models. The similar result holds for error prediction and misstatement identification. Models considering the sentiment features do not perform as well as the baseline model, as suggested by the evaluation metrics. A comparison of performance between model 1 and

2 shows that model 1 is slightly better than model 2 for all three prediction tasks, but the AUC of model 1 is lower that of model 2 for fraud detection.

Table 3.9 The Results of 10-Fold Cross Validation with Traditional ANN

		Accuracy	Type one error rate	Type two error rate	AUC
baseline	MIS	61.02%	38.98%	38.94%	0.6535
	FRAUD	73.50%	26.49%	27.88%	0.7743
	ERROR	59.87%	40.12%	40.83%	0.6359
Model 1 (deep learning)	MIS	59.34%	40.66%	40.81%	0.6400
	FRAUD	70.07%	29.93%	29.81%	0.7622
	ERROR	59.95%	40.05%	40.83%	0.6327
Model 2 (bag of words)	MIS	57.81%	42.20%	42.06%	0.6264
	FRAUD	69.81%	30.19%	30.77%	0.7631
	ERROR	58.79%	41.21%	41.74%	0.6243

- Deep Neural Network

The established deep neural network has three hidden layers. The first hidden layer has 175 nodes, the second has 350 nodes, and the third has 150 nodes[37]. Table 3.10 shows that, for fraud classification, model 1 has the best prediction performance with a high AUC of 0.7837. It also has the lowest FNR, which is 22.12%. The overall accuracy of model 1 (69.26%) is slightly lower than that of the baseline model (73.8%) but higher than that of model 2 (68.85%). The overall accuracy of the baseline model is better as the model tends to identity the observations as negative. So, it has lower FPR (26.17%) but higher FNR (33.65%). Model 1 is more effective in identifying fraud than model 2,

[37] This research uses a prevalent hyperparameter optimization technique, Grid Search, to select key hyperparameters and other settings in deep learning, such as the number of hidden layers and neurons as well as the activation function. The basic idea of Grid Search is that, the user selects several grid points of the hyperparameter and train the neural network using every combination of those hyperparameters. The combination of hyperparameters that produces the lowest validation error is selected.

evidenced by a lower FNR, 22.12%, as opposed to 28.85% for model 2. In addition, the FPR of model 1 (30.77%) is slightly lower than that of model 2 (31.16%).

For error prediction, the performance of all three models worsens, as indicated by the decreased AUC of baseline model (0.6150), model 1 (0.6159), and model 2 (0.6009). Other metrics also show the similar result. Among the three models, there is no big differences in the predictive performance, suggesting that the sentiment features do not help identifying errors, which are perhaps not surprising because not all misstatements are caused by frauds (Hennes, Leone, and Miller, 2008). Similar pattern is also observed for the models of misstatement prediction. Although model 1 has the highest AUC (0.6314), its accuracy, FPR, and FNR is worse than the baseline model.

Table 3.10 The Results of 10-Fold Cross Validation with DNN

		Accuracy	Type one error rate	Type two error rate	AUC
baseline	MIS	59.31%	40.68%	40.81%	0.6145
	FRAUD	73.8%	26.17%	33.65%	0.7477
	ERROR	58.97%	41.02%	42.66%	0.6150
Model 1 (deep learning)	MIS	58.82%	41.18%	42.06%	0.6314
	FRAUD	69.26%	30.77%	22.12%	0.7837
	ERROR	57.67%	42.36%	38.99%	0.6159
Model 2 (bag of words)	MIS	57.94%	42.04%	43.93%	0.6017
	FRAUD	68.85%	31.16%	28.85%	0.7804
	ERROR	54.13%	45.92%	38.99%	0.6009

- Naïve Bayes

As depicted in Table 3.11, with Naïve Bayes algorithm, the AUC for all models are low, with the highest score of 0.5888 for model 1 of frauds detection. The FNR is

extremely high. Model 1 has a 69.23% of FNR, suggesting that the Naïve Bayes models do not perform well for misstatement prediction with our dataset.

Table 3.11 The Results of 10-Fold Cross Validation with Naïve Bayes

		Accuracy	Type one error rate	Type two error rate	AUC
baseline	MIS	80.03%	19.73%	75.08%	0.5260
	FRAUD	85.78%	14.02%	73.08%	0.5653
	ERROR	91.48%	7.91%	93.94%	0.4874
Model 1 (deep learning)	MIS	80.31%	19.10%	77.26%	0.5174
	FRAUD	86.91%	12.90%	69.23%	0.5888
	ERROR	90.91%	8.50%	93.91%	0.4876
Model 2 (bag of words)	MIS	81.96%	17.40%	79.44%	0.5151
	FRAUD	89.11%	10.68%	73.08%	0.5770
	ERROR	91.27%	8.15%	95.24%	0.4808

3.5.2. Predictor Importance

Table 3.12 list the top ten important predictors of fraud detection models with Random Forest algorithms[38]. The sentiment features obtained with both bag-of-words and deep learning approach are listed as one of the top 10 important predictors. SENTIMENT_DL ranks fifth in model 1, with a scaled importance of 0.4160, while SENTIMENT_TX ranks fourth in model 2, with a scaled importance of 0.4223. JOY, which is not reported in table X, ranks 18th. Other important factors include, Soft assets, accounts receivable to total assets, Property, plant, and equipment to total assets, market value of equity, and etc.

[38] This section only reports the predictor importance of fraud detection models developed with the most effective algorithm, Random Forest.

Table 3.12 Top 10 Important Predictors of Fraud Detection Models: Random Forest

Baseline Model		Model 1		Model 2	
Predictor	Scaled Importance	Predictor	Scaled Importance	Predictor	Scaled Importance
SOFT	1	SOFT	1	SOFT	1
RECAT	0.7189	RECAT	0.9222	RECAT	0.6731
PPENTAT	0.4805	PPENTAT	0.5454	PPENTAT	0.5061
PENSION	0.3816	MVE	0.4480	SENTIMENT_TM	0.4223
MVE	0.3004	SENTIMENT_DL	0.4160	MVE	0.3474
LEASE	0.2775	FAAT	0.3788	PENSION	0.3313
AT	0.2557	PENSION	0.3613	AT	0.2906
FAAT	0.2512	AT	0.3042	LEASE	0.2458
LTXINT	0.2198	SALEAT	0.2372	FAAT	0.2447
CLEASE	0.1795	LTXINT	0.2152	SALEAT	0.2089

3.6. Discussion

In this section, the results of the 45 models are discussed from the perspective of the prediction tasks, the classification algorithms, and the structure of the model, respectively. Table 3.13 compares the results (AUC) of 10-fold cross-validation for all 45 models.

First of all, from the perspective of prediction tasks, it is found that the fraud detection models perform much better than the error and misstatement prediction models. It suggests that the predictive ability of sentiment features of MD&As is observed only for intentional misstatements and not for errors. In other words, the sentiment features are informative only when managers have the intention to misreport. A possible reason is that the management does not realize that there are unintentional errors, so there is no significant difference of the sentiment and emotion between the positive observations and

the negative ones. The misstatement models perform better than the error detections models but worse than the fraud detection model. This is because misstatements involve both errors and frauds. The ineffectiveness of detecting errors affect the ability of the models to detect the misstatements.

Secondly, considering fraud detection alone, Random Forest models achieve the best result of the five algorithms. The AUC score of all models exceeds 0.8, and the overall accuracy is higher than 75%. Random Forest also perform well in terms of FPR and FNR, which are no higher than 25%. Deep Neural Networks are less effective than Random Forest models, with a highest AUC of 0.7837. With this algorithm, the two models with sentiment features are more likely to incorrectly identify observations as frauds, with higher FPR than that of the baseline model. The traditional Neural Networks perform less effectively than the Deep Neural Networks. The best AUC is 0.7743, achieved by the baseline model. The AUC for model 1 and 2 are slightly lower, which are 0.7622 and 0.7631. Additionally, the baseline model outperforms the others in terms of other metrics. The Logistic Regression has its best AUC of 0.7473 lower than that of the Deep Neural Networks. Furthermore, the model 1 outperforms the other two models. The Naïve Bayes the least effective algorithm, which is caused by the collinearity of many predictors in our sample.

Lastly, models with sentiment variables generally have higher predictive accuracy than the baseline models. Furthermore, model 1 performs better than model 2 for the prediction of fraudulent financial statement as evidenced by higher overall accuracy and AUC and lower FPR and FNR, especially with those more effective algorithms such as Random Forest and Deep Neural Network. This result suggests that the sentiment

features extracted with deep learning approach have better prediction power than those calculated with “bag of words” approach.

Table 3.13 A Comparison Table of Prediction Performance for All 45 Models

	Misstatement		Fraud		Error	
Random Forest	Baseline	0.7232	Baseline	0.8288	Baseline	0.6673
	Deep Learning	0.7325	Deep Learning	0.8524	Deep Learning	0.6683
	Bag of words	0.7224	Bag of words	0.8506	Bag of words	0.6786
Logistic Regression	Baseline	0.6525	Baseline	0.7125	Baseline	0.6331
	Deep Learning	0.6695	Deep Learning	0.7473	Deep Learning	0.6539
	Bag of words	0.6474	Bag of words	0.7448	Bag of words	0.6345
Traditional Neural Network	Baseline	0.6535	Baseline	0.7743	Baseline	0.6359
	Deep Learning	0.6400	Deep Learning	0.7622	Deep Learning	0.6327
	Bag of words	0.6264	Bag of words	0.7631	Bag of words	0.6243
Deep Neural Network	Baseline	0.6145	Baseline	0.7477	Baseline	0.6150
	Deep Learning	0.6314	Deep Learning	0.7837	Deep Learning	0.6159
	Bag of words	0.6017	Bag of words	0.7804	Bag of words	0.6009
Naïve Bayes	Baseline	0.5260	Baseline	0.5653	Baseline	0.4874
	Deep Learning	0.5174	Deep Learning	0.5888	Deep Learning	0.4876
	Bag of words	0.5151	Bag of words	0.5770	Bag of words	0.4808

3.7. Conclusion, Limitation, and Future Research

3.7.1. Conclusion

Unlike bag of words approach that ignores the meaning of the words and phrases (Salton and McGill, 1983), deep learning approach is able to read and understand the

meanings of various combinations of words and phrases with the same appearance in the text. This essay applies deep learning technique to analyze 31,466 MD&As containing 321 firm-years with financial misstatement (among which 104 are caused by frauds, 218 are caused by errors) from 2006 to 2015. It employs a deep learning text analyzer, Alchemy Language API, which returns the sentiment score and the emotion index of joy for each MD&A text document. This essay uses the sentiment features as supplementary predictors in conjunction with 82 quantitative predictors provided by previous work (Perols, Bowen, Zimmermann, and Samba, 2017; Dechow et al., 2011; Perols, 2011; Cecchini et al, 2010; Beneish, 1999; Huang, Rose-Green, and Lee, 2012; Churyk, Lee, and Clinton, 2009).

This essay establishes 45 classification models under 3 types of model structure to conduct 3 predictions tasks (including predict frauds, errors, and misstatements) using 5 algorithms. It is found that all sentiment features are considered as important predictors by its model. The results also show that the deep learning-based sentiment features generally perform better than those based on bag of words approach. However, the models are only effective for fraud detection. Furthermore, among the 5 algorithms, Random Forest achieves the best performance. The AUC of the model with deep learning-based sentiment features reaches 0.8524. Therefore, the answers to the research questions are (1) the sentiment features obtained by both deep learning approach and bag of words approach provide essential information for financial misstatement prediction; (2) however, they are effective for fraud prediction only; (3) the deep learning approach generally performs better than the “bag of words” approach in this research. As a result,

considering its effectiveness of fraud detection and efficiency for text processing, deep learning based textual analysis is a promising technique for audit analytics.

3.7.2. Limitation and Future Research

This study has limitations. Since the restatement data was collected in December 2016, and *Compustat* is the only database used in this paper, the sample size of the restatements is limited. Specifically, there are fewer misstatements in 2014 and 2015, as some misstatements may not be identified until they will eventually be restated in the future. Accordingly, this study can be developed further by extending the data collection period and referring to AAER and other sources to obtain more restatements. Second, this study uses the sentiment and the emotion of joy to capture the linguistic cue of the MD&A. More characteristics, such as length, level of detail, complexity, hedging and uncertainty language, and immediacy (Burgoon et al, 2016) can analyzed in future research. Third, the deep learning tool used by this study is not trained exclusively with finance-specific text. Furthermore, Watson does not provide detailed information regarding the model development. For example, the model structure as well as the selection of other hyperparameters such as the activation function is unclear. In the future, finance-specific data (e.g., MD&As, Conference Calls, Articles in business journals, and Press Releases) can be collected to train deep neural networks for sentiment classification of audit-related documents.

Chapter 4 Predicting Audit Fee with Twitter: Do the 140 Characters reveal a firm's audit risk?

4.1. Introduction

Social media plays an increasingly important role in information sharing and social networking (Asur and Huberman, 2010). Due to its ease of use, high speed and wide reach, social media is gradually changing the nature of communication among users (Cong and Du, 2007; Kaplan and Haenlein 2010; Du and Jiang, 2015). In the business area, social media platforms, including Facebook, Twitter, Pinterest, LinkedIn, Tumblr, Google+, as well as their competitors, allow stakeholders to create, bookmark, share, and comment on content, creating enormous, various, and valuable data. This presents an excellent opportunity to investigate the social media data to obtain insights for research in accounting and auditing.

This paper investigates the association between the activity of Twitter's users and the audit fee to answer the research question "Do Tweets provide audit risk information that influences audit pricing"? In other words, this research aims to investigate whether Tweets reflect the audit risk of a company, which is consistent with the auditor's judgment on the company's risk, as measured by the audit fee. Auditors' professional judgment based on knowledge gained through a variety of information sources regarding the business of companies and frequent interactions with these companies provides a unique setting to assess whether Twitter can be used as a non-financial information source indicating the audit risk.

Previous literature provides a portrayal of the nature and effect of the adoption of social media by companies and claims that social media plays an important role in accounting information environment (e.g., Debreceny, 2015). However, it is still unclear whether and how social media could assist auditors to make decisions. Specifically, the vast majority of extant studies focus on companies' usage of social media and the market effects of this usage. For example, Du and Jiang (2015) observe significant association between social media adoption by S&P 1500 firms and firm performance measured by stock price and return on assets. The social media platforms examined include Facebook, Twitter, YouTube, blogs, discussion forums, RSS, and LinkedIn. Furthermore, the authors find that the significant association is mainly attributed to the usage of Facebook and Twitter. Focused on Facebook and Twitter, Zhou et al. (2015) track all messages posted in these two platforms from 2009 to 2013 for nearly 10,000 listed firms since their first adoption and find that the use of Twitter is prevalent early in the study period, while the adoption of Facebook is prevalent in the later period. Prokofieva (2015) documents a significant and negative association between "abnormal bid-ask spread" and Twitter dissemination of mandated continuous disclosures by Australian publicly listed companies. However, there is little understanding on the information content of social media for auditors' risk assessment.

Audit engagements are performed with a risk-oriented approach. Greater audit efforts will be applied on client firms with higher misstatement risk (Hoitash, Hoitash, and Bedard, 2008). Generally Accepted Auditing Standards (AU-C Section 300: Planning

an Audit[39]) state that, as an important pre-engagement activity, the auditor should perform investigation procedures regarding the acceptance or continuance of the client. The auditor devotes substantial time to understand as much as possible about the company and its management to mitigate the audit risk of engaging with or continuing to serve a client whose material misstatement risk exceeds the acceptable level of the accounting firm. The auditor investigates the company's industry, regulatory, economic conditions, impact of competition and other external factors affecting the company's risk of material misstatement in financial reporting. Audit standards also requires auditor to obtain an overall understanding of the financial performance, litigation status, nature of the business, control environment, management's integrity and reputation, and so forth. For existing clients, the auditor typically conducts similar retention reviews annually or when necessary (Louwers, et al., 2015). Based on their understanding of the company, auditors make fee decisions by evaluating the client's engagement risk related to material misstatement, which determines the nature and amount of audit evidence they need to gather (ISA 310: Knowledge of the Business). Consequently, the auditor needs various information from multiple sources (Louwers, et al., 2015). Besides inquiring the management, employees, banks, vendors and others within and outside the company and examining related documents, the auditor usually refers to industry trade publications, news articles, lawsuits, bankruptcy court outcomes, or even hires private investigators to conduct additional searches on the occurrence of unusual events (i.e., the company is accused of fraud or under the investigation of SEC).

[39] AU-C Section 300: Planning an Audit
https://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00300.pdf

The efficiency of social media makes more information regarding the company's operational and financial situation available to us. Such information could indicate the company's potential litigations, deteriorating reputation, increased business risk, internal control deficiency, unethical behaviors of management, inappropriate business strategy, and etc., which are red flags of the company's engagement risk. For instance, customer complaints on a product's quality or poor customer services can predict a downward sales revenue or profitability, which creates incentives for the company to commit financial fraud (Kreutzfeldt and Wallace, 1986). Therefore, social media could provide a wealth of useful information for the auditor to establish the "frame of reference" for the decision of audit pricing. Recognized the value of social media, the Internal Revenue Service (IRS) is reported using social media activity trackers to monitor the untrustworthy taxpayers' behavior (such as showing off their recent purchase of cars, houses, designer bags, watches while working at low-paying jobs) and look for clues of tax cheats (Bea, 2013).

This research focuses on Twitter, a social networking and microblogging platform providing multi-way communication among users via short text messages, so called tweets. With hashtags (e.g., #DeepLearning) and StockTwits (e.g., $APPL), the user creates tweets regarding topics and public companies that interest them. Other users can read, bookmark, respond to, or retweet them in real-time. Unlike other social media like Facebook which allows 63,206 characters, Twitter limits its message to 140 characters, which makes tweets usually more straightforward and concise.

With the expectation that there is a significant relationship between the content of tweets and the audit fee, this essay conducts tests on a sample of U.S. publicly listed firms in 2015. Following existing studies (e.g., Zhou et al, 2015; Debreceny, Rahman,

and Wang, 2016; Rui, Liu, and Whinston, 2013), this study employs the sentiment of tweets sentiment, an important metric indicating the attitude of the message generator, as well as the volume of tweets and retweets as characteristics of information in Twitter. Due to the complexity and variance of natural language expressions, the biggest problem with traditional textual analysis techniques (i.e., bag of words approach) is the difficulty in effectively identifying the sentiment within the document (Godarzi, 2011). Furthermore, compared to text in traditional media such as news articles and blogs, tweets are even harder to analyze as people like using slangs and abbreviation (Synthesio, 2011). As a result, this study applies deep learning, an Artificial Intelligence technique that excels at "understanding" the meaning of words, to acquire the sentiment of tweets. Specifically, this paper uses *Twitter Insight* to use the deep learning algorithm trained by IBM Watson.

This essay hypothesizes that the more negative tweets are posted discussing the client company the higher will the audit fee be. Furthermore, as the number of retweets measures the popularity of certain topics about the company in Twitter. The second hypothesis is that the association between negative tweets and audit fees is stronger for companies with more retweets. This study first uses the full sample to test the hypotheses and does not observe a significant relationship between the negative sentiment of tweets and audit fees. However, the results show that audit fees are sensitive to the negative sentiment of tweets when there are more retweets responding to the tweets about the company. Furthermore, the effect of companies' risk conditions on the association between characteristics of Twitter information and audit fees is examined. The results of empirical analysis show that Tweets reflect the company's audit risk when their clients

are free of going-concern issues and with median level of restatement risk. In particular, negative sentiment of tweets will positively affect audit fees, especially when more retweets are received. These results hold when one-year lagged value of the control variables are used to avoid errors in auditor's anticipation of clients' risk. This research suggests that Twitter can be used as an additional source of information for auditors to evaluate companies' engagement risk. This is meaningful not only for the pre-engagement process but also the entire planning process.

The remainder of this study is organized as follows: The next section provides the background, literature review, and hypotheses development. The third section introduces deep learning-based sentiment analysis. Section 4 discusses the research design, followed with section 5 discussing the results. Robustness test is conducted in section 6. Finally, section 7 draws conclusion and discusses limitations and directions for future research.

4.2. Background, prior Literature, and Hypotheses Development

4.2.1. Audit Fees

Previous literature provides evidence that risky clients are likely to pay high audit fees (O'Keefe et al. 1994; Lyon and Maher 2005; Venkataraman, Weber, and Willenborg, 2008). The audit fee model developed by Simunic (1980) suggests that audit fee is a function of the size and complexity of the company, as well as the audit risk assessed by the auditor (Gul, 2007). To enable the auditor to estimate the audit hours and the hourly rate, and consequently propose the audit fee for the new client, an overall assessment of

the client's engagement risk (primarily material misstatement risk) occurs within the client acceptance process. The audit fee is negotiated and determined in the engagement letter. Once engaged, the negotiated fee will not change except in response to unexpected significant changes (Hackenbrack, Jenkins, and Pevzner, 2014). Prior to fee negotiation, the auditor spends a great deal of time searching for information regarding the client's operations, industry, and the economy, such as regulatory requirement, industrial condition, economic environment, governance profile, funding structure, and special issues, to assess the inherent risk (Castro, Peleias, and Silva, 2015). For the existing client, the auditor conducts a retention review annually or when necessary to decide if it is appropriate to continue serving the client (Louwers, et al, 2015). The auditor considers both inherent risk and control risk as there is prior knowledge with respect to the client's nature of business and internal control system. Therefore, the audit fee typically reflects the auditor's judgment of the potential client's material misstatement risk, based on a wealth of information from various sources, such as news articles (Redmayne, Bradbury, and Cahan, 2010), analysts' forecast (Foo et al., 2016), past SEC filings and announcements, and industry trade publications.

4.2.2. Twitter

As a public available information source, Twitter is a social networking and microblogging platform providing multi-way communication among users via short text messages, tweets. With hashtags (e.g., #DeepLearning) and StockTwits (e.g., $APPL), the user creates tweets regarding topics and public companies that interest them. Other users can read, bookmark, respond to, or retweet them in real-time. Thanks to its instantaneous latency and ubiquitous accessibility, Twitter has speedily gained its

popularity since its birth in 2006, and the number of its monthly active users had reached 319 million as of the fourth quarter of 2016. Unlike other social media like Facebook which allows 63,206 characters, Twitter limits its message to 140 characters, which makes tweets usually more straightforward and concise. Consequently, Twitter is an ideal communications channel for stakeholders and provides information revealing the stakeholder's interest and attitude in terms of the volume of tweets and retweet, the sentiments of the tweeters, as well as the topic of their conversations (Debreceny, Rahman, and Wang, 2016).

Prior literature primarily investigates the company's adoption of Twitter and its market effect. For example, using a sample of technology firms, Blankespoor, Miller, and White (2013) document that the market liquidity is enhanced by the additional dissemination of firm-initiated news via Twitter. Prokofieva (2015) studies the Australian public companies and demonstrates that using Twitter as an information dissemination channel is negatively associated to abnormal bid-ask spread, especially for firms that have lower levels of analyst coverage and/or lesser visibility in traditional media. In contrast, Lee, Hutton and Shu (2015) provide evidence for the negative market effect of the usage of Twitter. They study the short-window price reaction around 177 product recalls from 2008-2012 and find that the increased tweeting exacerbates the negative price reaction. Besides Twitter, Du and Jiang (2015) examine other six types of social media including Facebook, YouTube, blogs, discussion forums, RSS, and LinkedIn and find that half of the S&P 1500 firms use one or both of these platforms and their use of social media is related to firm performance, measured by stock price and return on assets. Similarly, Yu, Duan and Cao (2013) investigate the impact of social media (blogs,

forums, and Twitter) and conventional media (major newspapers, television broadcasting companies, and business magazines) on capital markets. They claim that blogs and Twitter have a positive association with short term stock performance.

While these studies focus on company-initiated tweets, Debreceny, Rahman, and Wang (2016) center on user-initiated tweets posted around 8-K disclosures of S&P 1500 companies and examine whether tweeting is associated with stock market reactions to corporate disclosures. They argue that existing literature using company-initiated tweets usually regard social media as an additional way of information dissemination rather than "a tool for measuring information recognition". Eschenbrenner, Nah, and Telaprolu (2015) address the issue of social media (Facebook and Twitter) usage by Big 4 and second-tier accounting firms other than public companies. They categorize content of information from social media into several classes and observe that "knowledge sharing", "socialization and onboarding,'' and ''branding and marketing'' are the most common class of communication.

Despite the well-documented finding that Twitter is an important information dissemination channel for companies and affects the capital market, fewer studies (i.e., Debreceny, Rahman, and Wang, 2016) recognize that it contains important information reflecting stakeholders' recognition and attitude (i.e., sentiment). Little research explores the powerful insights provided by social media and how the auditor could leverage them to support risk assessment in planning (Debreceny, 2015). This research mainly examines the negative sentiment of tweets, while controlling for the frequency of tweets as it captures the volume effect of tweeting on audit fee (Rui, Liu and Whinston, 2013).The sentiment of a text shows the author's perception, attitude, or opinion and is regarded by

extant research as a critical feature influencing decision makers' affective states and, hence, their decision (Bonner, 2008; Mian and Sankaraguruswamy, 2012; Prokofieva, 2015; Debreceny, Rahman, and Wang, 2016). Therefore, more and more studies make an effort to use the sentiment feature. For instance, Baker and Wurgler (2006) employ Principal Component Analysis (PCA) to measure the investor's sentiment. Pak and Paroubek (2010) develop a sentiment classifier and categorize the sentiment of tweets to positive, negative, and neutral sentiments for opinion mining. Debreceny, Rahman, and Wang (2016) utilize abnormal sentiment measure to proxy the investor's perception of the corporate disclosure. Mian and Sankaraguruswamy (2012) find that investor sentiment influences the stock price sensitivity to earnings news.

It is noteworthy that positive and negative sentiments perform differently with respect to their influence on decision making as one tends to be more sensitive to bad news as opposed to good news. Yu, Duan, and Cao (2013) observe that, compared to the effect of positive sentiment on positive market returns, negative sentiment of tweets is more strongly associated with negative returns. Similarly, Shiller (2005) finds larger stock price shocks for negative sentiment rather than for positive sentiment. Negative discussion on Twitter timely identifies potential issues in the operation of a company and its financial situation. To illustrate, information posted on Twitter revealing the CEO's questionable behavior such as insider trading, résumé fraud, and sexual misconduct uncover his or her ethical problem, which increases the risk of the financial misstatement, and consequently the risk of audit engagement.

Accordingly, it hypothesizes:

H1: The audit fee of a company is positively associated with the negativity of the Tweets mentioned the company.

4.2.3. Retweets

When a Twitter user is interested in a tweet he or she will repost or forward this tweet. This behavior is called retweeting (Cha, et al., 2010) and the tweets received by their followers are called retweets (Wu and Shen, 2015). The number of retweets is an important metric of popularity for a tweet, as other users found it interesting enough to share with their audience. The popularity of the tweets for a company strengthens the association between the negative tweets and the risk of material misstatement, as it suggests that Twitter users are interested in specific topics about the company. For a company receiving negative tweets, the more retweets the company has, the more likely that the company is involved in controversial events or issues. In other words, a large number of retweets for companies receiving negative tweets is a signal of increased risk of the company. As the number of retweet for a Tweet varies from zero to millions, this paper uses the maximum number of retweets for all tweets about the company in the research time period to measure the popularity of the company.

As a result, it hypothesizes:

H2: The association between the audit fee and the negativity of the Tweetsis stronger for companies with more Retweets.

4.3. Sentiment Analysis Method

The sentiment features of tweets in this essay are extracted by deep learning technique (also called deep neural network) (LeCun, Bengio, and Hinton, 2015). It is an Artificial Intelligence method that has been frequently adopted and effectively performed for big data analysis over the last decade (Najafabadi et al., 2015). Deep learning employs deep artificial neural network to abstract data representations and generalize to future data. As computers become more and more powerful, the architecture of Artificial Neural Network becomes deeply hierarchical and consists of multiple hidden layers. Due to its "depth", Deep Neural Networks have produced state-of-the art achievement in computer vision, speech recognition, natural language processing, and other tasks. Figure 4.1 shows a simplified example of a DNN for sentiment analysis of tweets. The DNN is trained with large number of tweets and the output is the identified sentiment. The input layer is used to receive the raw tweets, the multiple hidden layers process the data, and the output layer classifies the data. Each layer is composed of neurons, in which complex data processing takes place to conduct linear and nonlinear transformation of the received data /features from last layer and form a new version (more abstract) of feature (Goodfellow, Bengio, and Courville, 2016). The parameter of the complex computation is "learned" through training with massive amount of data without human intervention.

This essay uses IBM *Twitter Insights* to surface sentiment and other enrichments from tweets. With Deep natural language processing algorithms from IBM Social Media Analytics, this tool includes APIs that allow searches for Twitter content based on keywords, timeframes, and other query parameters, provides real-time analysis of Tweets, and returns Tweets with related properties, such as the number of retweet and the

overall sentiment (e.g., positive, negative, ambivalent, or neutral) [40]. The returned values used in this essay is the count of tweets, negative tweets, and retweets for each tweet.

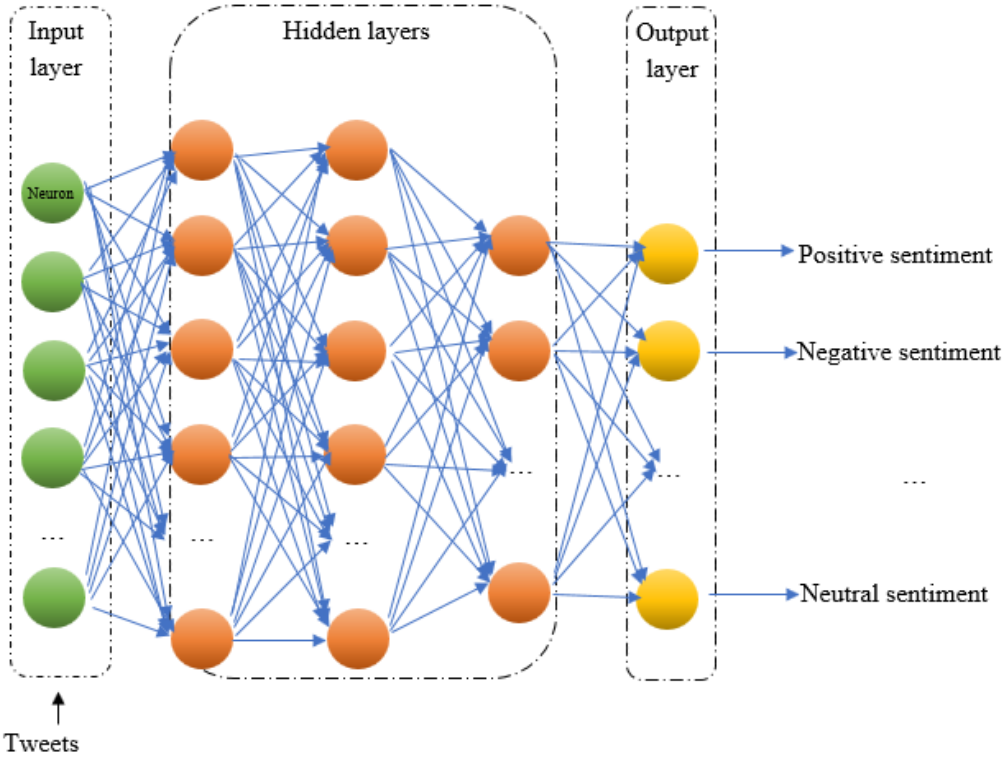

Figure 4.1 A Deep Neural Network for Tweets Sentiment Analysis

4.4. Research Design

4.4.1. Sample

Table 4.1 describes the sample selection procedure, which begins with a list of 6130 U.S. public companies for fiscal year 2015. Following prior research (e.g., Simunic 1980;

[40] https://console.bluemix.net/docs/services/Twitter/twitter_overview.html#about_twitter

Francis, 1984; DeFond, Francis, and Wong, 2000), 235 financial services companies (SIC codes 6000-6900) are eliminated[41], leaving 5895 companies for analyses. For those companies, the financial data is drawn from Compustat and Compustat Segments databases and the audit data is from AuditAnalytics. After merging with these datasets, 3084 observations missing related financial or audit information are excluded. This process leaves 2811 companies. For each of the 2811 companies, related twitter data is collected using *Twitter Insights*[42]. The collected data includes information about all Tweets and Retweets containing the company's StockTwit (for example, $APPL) posted from 12 months prior to the first day of the fiscal year to the first day of the fiscal year[43]. Figure 4.2 shows the timeline for tweets collecting process. 479 observations missing related twitter data are deleted. Finally, the sample collection process yields 2332 companies involving 326,659,114 tweets. Table 4.2 provides the distribution of sample across industries.

[41] The finance industry is excluded from the study because many of the financial ratios used to estimate audit fees, such as leverage, are not relevant to financial institutions (DeFond, Francis, and Wong, 2000).
[42] Twiter Insights has been retired since April 2017. Alternatively, researchers can use Twitter Gnip APIs to retrieve the Twitter data. More information is discussed in "Limitations and Future Research" section of this chapter.
[43] Auditing Standard No. 16 (PCAOB 2010b) requires auditors to document their understanding of the terms of an engagement in an engagement letter. The engagement letter also documents the negotiated audit fee in the first quarter of the year under audit. The negotiated audit fee is sticky and usually will not be changed unless there are "significant unexpected changes in the amount of the auditor and the client" (Hackenbrack and Hogan 2005; Hackenbrack, Jenkins, and Pevzner, 2014).

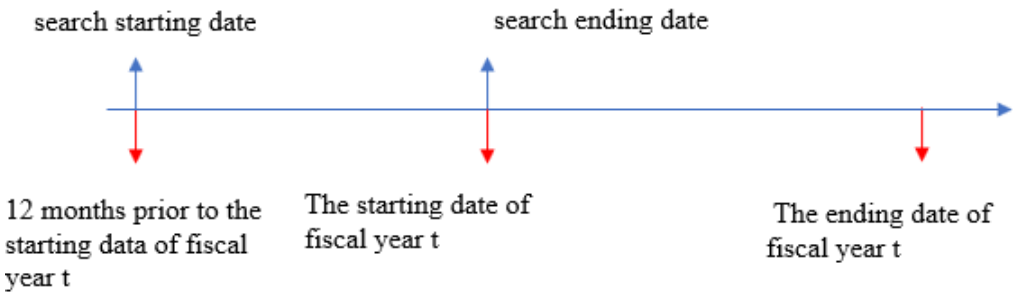

Figure 4.2 Timeline for Tweets Collection

Table 4.1 Sample Selection Procedure

	Number of Observations
U.S. listed companies in 2015	6130
Less: financial, insurance, and real estate firms	(235)
Less: observations with financial variable data missing in Compustat or Compustat Segments	(1,869)
Less: observations with audit data missing in AuditAnalytics	(1,215)
Less: observations with missing Twitter data	(479)
Final sample	2332

Table 4.2 Sample Distribution across Industries

Industry		Distribution	
		Count	Percentage
1	Agriculture	6	0.26
2	Mining & Construction	71	3.04
3	Food & Tobacco	56	2.40
4	Textiles & Apparel	16	0.69
5	Lumber, Furniture, & Printing	58	2.49
6	Chemicals	76	3.26
7	Refining & Extractive	165	7.08
8	Durable Manufacturers	382	16.38
9	Computers	432	18.52
10	Transportation	178	7.63
11	Utilities	42	1.80
12	Retail	167	7.16
13	Services	243	10.42
14	Banks & Insurance	39	1.67
15	Pharmaceuticals	401	17.20
Total		2332	100

4.4.2. Audit Fee Model

To analyze the relationship between audit fees and sentiment factors of tweets, the analysis starts with a traditional audit fee model designed to capture a company's financial and operational situation, auditor choice, audit complexity, audit risk, and other variables reflecting the demand for audit services (Francis and Wang 2005; Krishnan, Sami, and Zhang, 2005; Ghosh and Pawlewicz 2009; Choi et al. 2010; Stanley 2011). It then incorporates the twitter variables to test the hypotheses using the following model (variable definitions are provided in Appendix B):

$$Lnauditfee = \beta_0 + \beta_1 Negative + \beta_2 Retweets + \beta_3 Negativity * Retweets + \beta_4 Tweets + \beta_5 Roaearnings + \beta_6 Size + \beta_7 Invrec + \beta_8 Leverage + \beta_9 Currentratio + \beta_{10} BTM + \beta_{11} Growth + \beta_{12} Loss + \beta_{13} Segments + \beta_{14} Foreign + \beta_{15} Merger + \beta_{16} Special + \beta_{17} Firstyear + \beta_{18} Big4 + \beta_{19} IC + \beta_{20} GC + \sum IndustryFE + \mathcal{E}_1$$

(1)

where:

Lnauditfee: natural log of audit fees;

Negative: the percentage of tweets with negative sentiment among all tweets mentioned the client company minus the percentage of tweets with positive sentiment among in all tweets mentioned the client company;

Tweets: the count of all tweets mentioned the client company

Retweets: the maximum number of retweets for each single tweet mentioned the client company.

Roaearnings: earnings, calculated as operating income after depreciation (OIADP) divided by total asset (AT)

Size: natural log of total assets (AT);

Invrec: inventory (INVT) plus accounts receivable (RECT) divided by total assets (AT);

Leverage: the difference between total liabilities (LT) and current liabilities (LCT) divided by total assets (AT);

Currentratio: current assets (ACT) divided by current liabilities(LCT);

BTM: the difference between total assets (AT) and total liabilities (LT) divided by market value of common equity (PRCC_F × CSHO);

Growth: the percentage of change in sales (SALE) from year t-1 to year t;

Loss: equals 1 if the client firm reports a net loss (NI<0), and 0 otherwise

Segments: the number of business segments

Foreign: equals 1 if the client firm has foreign operations (TXFO), and 0 otherwise;

Merger: equals 1 if the client firm reports the item related to acquisition and merger (AQP), and 0 otherwise;

Special: equals 1 if the client firm reports special items (SPI), and 0 otherwise;

Firstyear: equals 1 if initial year of audit, and 0 otherwise;

Big4: equals 1 if Big 4 auditor, and 0 otherwise;

IC: equals 1 if the current auditor indicates internal control weakness, and 0 otherwise;

GC: equals 1 if the current auditor issues a going-concern opinion, and 0 otherwise;

To measure the effect of the sentiment of tweeting on audit fees, variable *Negative* is used to measure the strength of negativity for tweets mentioned the company. Consistent with Rui, Liu, and Winston (2013), this approach avoids the multicollinearity issue that will arise if the absolute numbers of negative/positive tweets are used in the model as they are significantly correlated with the total count of tweets. The control variable, TWEETS, is the total count of tweets mentioned the company (containing the StockTwit of the company), which is used to control the volume effect of tweeting on audit fee. It is consistent with previous research in the context of marketing (for instance, Rui, Liu, and Winston, 2013; Chintagunta, Gopinath, and Venkataraman, 2010). To test

the second hypothesis, *Retweets* is used to measure the popularity of a specific topic about the company. The interaction term examines the simultaneous influence of Negative, Retweet, as well as Negative and Retweet on the audit fee. All other control variables are measured as of the end of the fiscal year, and all continuous variables that do not take log are winsorized at the 1st and 99th percentiles. Finally, the audit fee model includes industry fixed effects to isolate the effects of industry in the determination of audit fees.

4.5. Results

4.5.1. Descriptive Statistics

Table 4.3 provides descriptive statistics on the variables included in the audit fee model. The average of *Lnauditfee* is 13.6714 (the average of actual audit fee is $865,792.0226). *Negative* is negatively skewed, with a mean of -0.1339, suggesting that there are overall more positive tweets than negative tweets. The mean *Tweets* for the 12-month period is 30940.69, while the mean of *Retweets* is 5725.934. The average *ROA* is 5.88 percent, more than 12 % of firms received a going concern, and more than 22 % of the observations reported a loss. In addition, 9.52% of our sample changed auditors (*Firstyear*), and 66.25 % of the observations are audited by the Big four audit firms.

Table 4.4 presents the Pearson correlation matrix among individual variables and the dependent variable of the audit fee model in Equation (1). As anticipated, the correlations between *Lnauditfee* and all three twitter variables are positive, but *Negative* is insignificantly associated with *Lnauditfee*. In addition, significant correlations are observed between *Lnauditfee* and all other control variables, which is consistent with

prior research. The significant correlations between *Tweets* and both *Size* and *Big 4*, and *Tweets* and *Retweets* support that bigger companies or companies audited by big four audit firms tend to have more tweets and retweets. The variance inflation factor (VIF) is examined for each independent variable in multiple regression analyses for estimation and inference concerns arising from multicollinearity.

Table 4.3 Descriptive Statistics

Variable	Mean	Std.	P25	Median	P75
lnauditfee	13.6714	1.4779	12.7281	13.8117	14.6671
Negative	-0.1339	0.1436	-0.1951	-0.1237	-0.0475
Tweets	30940.6900	157825.2000	13.5000	100.5000	955.5000
Retweets	5725.9340	21870.9200	2.0000	23.5000	434.0000
roa_earnings	0.0588	0.1317	0.1614	0.0676	0.1196
size	5.9939	2.6908	4.4417	6.2309	7.8274
invrec	0.2043	0.1754	0.0588	0.1622	0.3065
leverage	0.3304	0.3389	0.0659	0.2700	0.4883
current_ratio	2.8842	3.8402	1.1009	1.8703	3.0977
BTM	0.4312	1.4926	0.1540	0.3648	0.7250
growth	0.2259	1.1407	-0.1056	0.0245	0.1667
Segments	1.6700	1.0900	1.0000	1.0000	2.0000
Loss	0.2247	0.4996	0.0000	0.0000	1.0000
Foreign	0.5279	0.4993	0.0000	1.0000	1.0000
Merger	0.1934	0.3950	0.0000	0.0000	0.0000
Special	0.1535	0.3605	0.0000	0.0000	1.0000
Firstyear	0.0952	0.2936	0.0000	0.0000	0.0000
Big4	0.6625	0.4730	0.0000	1.0000	1.0000
IC	0.0472	0.2121	0.0000	0.0000	0.0000
GC	0.1239	0.3296	0.0000	0.0000	0.0000

All continuous variables that do not take log are winsorized at the 1% and 99% to mitigate outliers.

Table 4.4 Pearson Correlation Matrix

Panel A: Part 1										
	1	2	3	4	5	6	7	8	9	10
1 Auditfee	1.0000									
2 Tweets	0.0473*	1.0000								
3 Negative	0.0241	-0.0146	1.0000							
4 Retweets	0.1048*	0.3752*	-0.0539*	1.0000						
5 Roaearnings	0.0789*	0.0023	0.0102	0.0071	1.0000					
6 Size	0.8547*	0.0508*	0.0246	0.1065*	0.1358*	1.0000				
7 Invrec	-0.0134*	0.0023	0.0099	0.0120	0.0292*	0.1215*	1.0000			
8 Leverage	-0.0323*	0.0035	-0.0146	-0.0066	-0.0970*	-0.0615*	-0.0117	1.0000		
9 Currentratio	-0.0761*	-0.0069	0.0066	-0.0150	0.0052	-0.0356*	-0.0376*	-0.0028	1.0000	
10 BTM	0.0160*	-0.0014	0.0055	0.0071	0.0005	0.0231*	-0.0084	-0.0003	-0.0006	1.0000
11 Growth	-0.0256*	-0.0018	-0.0949*	-0.0052	0.0003	-0.0245*	-0.0102	-0.0000	-0.0005	0.0001
12 Loss	-0.2315*	-0.0130	-0.0174	-0.0680*	-0.0495*	-0.5156*	-0.2580*	0.0124	0.0219*	-0.0106
13 Segments	0.2015*	-0.0026	0.0174	-0.0058	0.0116*	0.1253*	-0.0041	-0.0057	-0.0141*	-0.0027
14 Foreign	0.5804*	0.0324	0.0375	0.0604*	0.0411*	0.5087*	0.2054*	-0.0176*	-0.0566*	0.0007
15 Merger	0.3293*	0.0270	0.0262	0.0290	0.0190*	0.2437*	0.0302*	-0.0099	-0.0271*	-0.0008
16 Special	0.4326*	0.0315	0.0141	0.0452*	0.0059	0.2467*	-0.0001	-0.0150*	-0.0509*	-0.0128*
17 Resignation	-0.1448*	-0.0084	-0.0130	-0.0239	-0.0701*	-0.1157*	-0.0007	0.0006	0.0008	-0.0009
18 Dismissal	-0.0764*	-0.0101	0.0275	-0.0214	-0.0080	-0.0932*	0.0082	-0.0035	-0.0082	-0.0008
19 GC	-0.4269*	-0.0233	-0.0383*	-0.0435*	-0.0974*	-0.6007*	-0.1340*	0.0456*	-0.0385*	-0.0200*
20 Big4	0.5314*	0.0389*	0.0148	0.0704*	0.0542*	0.5850*	-0.1618*	-0.0241*	-0.0319*	0.0166*
21 IC	-0.2641*	-0.0133	-0.0008	-0.0251	-0.0594*	-0.3301*	-0.0457*	0.0145*	-0.0116	-0.0010

Panel B: Part 2											
	11	12	13	14	15	16	17	18	19	20	21
11	1.0000										
12	0.0099	1.0000									
13	-0.0005	-0.0532*	1.0000								
14	-0.0177*	-0.3353*	0.2125*	1.0000							
15	-0.0018	-0.0879*	0.0926*	0.3312*	1.0000						
16	0.0016	0.0243*	0.1234*	0.3042*	0.4615*	1.0000					
17	0.0014	0.0464*	-0.0072	-0.0697*	-0.0449*	-0.0367*	1.0000				
18	-0.0016	0.0745*	0.0087	-0.0450*	0.0085	0.0181*	-0.0277*	1.0000			
19	0.0183*	0.4119*	-0.0758*	-0.3382*	-0.1576*	-0.0656*	0.1131*	0.1190*	1.0000		
20	-0.0156*	-0.2661*	0.0917*	0.3919*	0.1314*	0.1112*	-0.1171*	-0.1434*	-0.3943*	1.0000	
21	0.0057	0.2234*	-0.0410*	-0.1685*	-0.0737*	-0.0245*	0.1150*	0.1312*	0.3817*	-0.2423*	1.0000

* Denotes two-tailed significance levels at 0.05. All continuous variables that do not take log are winsorized at the 1% and 99% to mitigate outliers.

4.5.2. Main Multivariate Results

Model 1 and 2 in Table 4.5 present multivariate results on the association between audit fees and the twitter variables with the purpose of testing the first hypothesis. Columns (1) and (2) display the coefficients and p-values for independent variables in model 1. The results show non-significant relationship between each of the twitter variables and the audit fee (p=0.573 for Negative, p=0.834 for *Tweets*, and p=0.365 for *Retweets*), indicating that audit fees are not significantly associated with more negative tweeting for the full sample.

Prior research documents that companies with predecessor auditor resignations are likely to have higher audit fees, as auditor resignations signal higher audit risk (Yoon, 2016). Therefore, model 2 uses *Resignation* and *Dismissal* to compare the effect of *auditor resignation* and *auditor dismissal* on audit fee decision. Similar results as that of model 1 are observed, regarding the relationship between audit fees and the twitter variables as well as the control variables. Coefficients of other control variables for both model 1 and 2 are consistent with predictions based on prior research, except that BTM is non-significant (p=0.461 for model 1 and p=0.433 for model 2). Thus, H1 for the full sample is not supported.

To test H2, model 3 and 4 incorporate the interaction between Negative and Retweet. Similar to the setting of model 1 (2), model 3 (4) uses *Firstyear* (*Resignation* and *Dismissal*) to capture the effect of auditor change (the reason of auditor change) on the audit fee. The results of Model 3 (Column (5) and (6)) show that, while *Negative* remains insignificant, *Retweets* (β=1.39e^{-6}, p=0.036) and the interaction *Negative* × *Retweets* (β=7.14e^{-6}, p=0.011) become significant. These results indicate that the auditor prices a company with

more negative tweets higher when there are specific topics about the company receiving more retweets. Similar results are shown for model 4. H2 is supported.

The explanatory power of each model in Table 4.5 is consistent with prior research, explaining approximately 86% of the variation in audit fees. the variance inflation factor (VIF) of each explanatory variable is reviewed and there is no indication that multicollinearity draws concerns about these inferences.

4.5.3. The Effect of Risk Conditions

This research also seeks to provide evidence of associations between audit fees and tweeting sentiment when companies' risk conditions are likely to affect the tweeting activities (which may or may not accurately reflect the company's risk) as well as other audit fee determinants. For this purpose, two types of risks are considered: going-concern risk and financial restatement risk.

- Going-concern risk

A going-concern opinion is issued when an auditor perceives a heightened threat to a company's ability to continue largely in its present form for an indefinite future (AICPA,1988; Blay, Geiger, and North, 2011). To examine whether the going-concern risk affects the weight that the auditor put on information in Twitter when pricing the client, the full sample is divided into two groups: companies that receive going-concern opinion in the current fiscal year and companies that does not receive going-concern opinion, namely "GC companies" and "Non-GC companies", respectively.

Table 4.5 Regression of Tweets Sentiment on Audit fees

		Model 1		Model 2		Model 3		Model 4	
Variable	Expected Sign	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	+/-	9.8690***	0.001	9.8707***	0.001	9.8701***	0.001	9.8577***	0.001
Negative	+	0.0411	0.573	0.0453	0.534	0.0745	0.314	0.0790	0.285
Retweets	+	$8.85e^{-8}$	0.834	$8.61e^{-8}$	0.839	$1.39e^{-6}$**	0.036	$1.41e^{-6}$**	0.033
Negative×Retweets	+					$7.14e^{-6}$**	0.011	$7.26e^{-6}$***	0.010
Tweets	+/-	$-7.24e^{-9}$	0.365	$-7.14e^{-9}$	0.371	$-7.96e^{-9}$	0.318	$-7.89e^{-9}$	0.323
Roaearnings	-	-0.0005***	0.001	-0.0005***	0.001	-0.0005***	0.001	-0.0005***	0.001
Size	+	0.4433***	0.001	0.4428***	0.001	0.4430***	0.001	0.4431***	0.001
Invrec	+	0.3891***	0.001	0.3887***	0.001	0.3894***	0.001	0.3855***	0.001
Leverage	+	0.1010***	0.001	0.1013**	0.001	0.1017***	0.001	0.1013***	0.001
Currentratio	-	-0.0012**	0.014	-0.0012**	0.015	-0.0012**	0.014	-0.0012**	0.015
BTM	-	-0.0005	0.461	-0.0006	0.433	-0.0005	0.469	-0.0006	0.439
Growth	-	-0.0009**	0.019	-0.0009**	0.020	-0.0009**	0.017	-0.0008**	0.021
Loss	+	0.1179***	0.001	0.1182***	0.001	0.1188***	0.000	0.1153***	0.001
Foreign	+	0.3588***	0.001	0.3593***	0.001	0.3591***	0.001	0.3586***	0.001
Merger	+	0.1272***	0.001	0.1280***	0.001	0.1272***	0.001	0.1267***	0.001
Special	+	0.1596***	0.001	0.1592***	0.001	0.1596***	0.001	0.1542***	0.001
Firstyear	-	-0.0782*	0.061			-0.0771*	0.064		
Resignation	+			-0.1769*	0.065			-0.1767*	0.065
Dismissal	-			0.0136	0.791			0.0203	0.694
GC	+	0.3067***	0.001	0.2900***	0.001	0.2892***	0.001	0.2840***	0.001
Big4	+	0.4669***	0.001	0.4669***	0.001	0.4665***	0.001	0.4745***	0.001
IC	+	0.0634***	0.001	0.0635***	0.001	0.0634***	0.001	0.0633***	0.001
Industry effect		Included		Included		Included		Included	
Observations		2332		2332		2332		2332	
Adjusted R^2		0.8612		0.8611		0.8615		0.8615	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively

Table 4.6 report the results of estimation for these two groups of companies. As shown in column (1) and (2), while *Negative* is not significant (β = -0.1755, p=0.387), *Retweets* is significant at 0.1 level but the sign is negative (β = -7.28e^{-6}). The interaction is insignificantly associated with the audit fee. Among control variables, only *Roaearnings, Size, Foreign, Resignation,* and *Big4* are significantly related to the audit fee. This is different from the results of full sample that almost all coefficients of control variables are consistent with predictions based on prior research. This model explains 74.12% of the variation in audit fees, which is lower than the full sample models shown in Table 4.6. Column (3) and (4) show the regression results of the non-GC group. It is found that all twitter variables are positive and significantly associated with *Lnauditfee.* Specifically, *Negative* is significantly associated with the audit fee at 0.05 level (β = 0.1703), and the interaction *Negative* × *Retweets* significantly strengthens this association (β = 8.90e^{-6}, P<0.001). It suggests that Tweets are less likely to reflect a company's audit risk when it faces going-concern issues. A possible reason is that companies with going-concern threats tend to adopt strategies to cope with financial distress, creating many positive Tweets. In addition, the number of maximum retweets strengthens the positive association between *Negative* and *Lnauditfee*. So, both hypotheses are supported for non-GC companies. All other quantitative control variables, with the exception of *Resignation,* become significant, and the model explains 84.56% of the variation in audit fees.

Table 4.6 Regression of Tweets Sentiment on Audit fees by the Existence of GC Opinions

		GC companies		Non-GC companies	
Variable	Expected Sign	Coefficient	p-value	Coefficient	p-value
		(1)	(2)	(3)	(4)
Intercept	+/-	11.8946***	0.001	9.2993***	0.001
Negative	+	-0.1755	0.387	0.1703**	0.032
Retweets	+	$-7.28e^{-6}$*	0.095	$1.88e^{-6}$***	0.004
Negative× Retweets	+	$-3.17e^{-5}$	0.201	$8.90e^{-6}$***	0.001
Tweets	+/-	$2.88e^{-9}$	0.994	$-1.18e^{-8}$	0.121
Roaearnings	-	-0.0005***	0.002	0.0159***	0.001
Size	+	0.3330***	0.001	0.4680***	0.001
Invrec	+	0.1294	0.587	0.5596***	0.001
Leverage	+	0.0446	0.177	0.1813***	0.001
Currentratio	-	0.0021	0.906	-0.0008*	0.093
BTM	-	-0.0012	0.357	-0.0129***	0.000
Growth	-	-0.0006	0.180	-0.0045*	0.425
Loss	+	-0.3048	0.219	0.1511***	0.001
Foreign	+	0.4588***	0.009	0.3256***	0.001
Merger	+	0.0814	0.610	0.1104***	0.001
Special	+	0.0758	0.446	0.1502***	0.001
Resignation	+	-0.4833*	0.053	-0.1539	0.137
Dismissal	+	0.0022	0.986	0.0468	0.408
Big4	+	0.6886***	0.001	0.4318***	0.001
IC	+	0.0058	0.849	0.0960***	0.001
Industry effect		Included		Included	
Observations		289		2043	
Adjusted R^2		0.7412		0.8456	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

- Restatement risk

To examine the influence of financial restatement risk on the association between Twitter feeds and audit pricing, a restatement risk indicator is needed. This study follows Lobo and Zhao (2013) and Liu et al. (2018) to obtain the restatement risk indicator, namely *Pscore*, by calculating the predicted probability of restatement using the model proposed by Dechow et al. (2011), as shown in Equation (2).

$$Restatement = \beta_0 + \beta_1 Totalaccural + \beta_2 \Delta Rec + \beta_3 \Delta Inv + \beta_4 Softassets + \beta_5 \Delta Csale + \beta_6 \Delta Roa + \beta_7 Issuance + \beta_8 \Delta Emp + \beta_9 Lease + \beta_{10} Abret + \beta_{11} Lagabret + \varepsilon \quad (2)$$

Variable definitions in equation (2) are summarized in Appendix B. This model utilizes a list of financial variables to predict the probability of financial restatement before the audit is conducted. Since the predicted value of restatement probability (*Pscore*) is used before the audit is conducted, it is also called the indicator of pre-audit restatement risk. Based on the value of *Pscore*, the full sample is classified into three groups. The first group includes companies with *Pscore* greater than 0.1185, which is the top quintile of *Pscore* of the whole sample. Those companies are considered as observations with high pre-audit restatement risk. The second group consists of observations with low pre-audit restatement risk. The *Pscore* values of those companies are lower than 0.0551, which is the bottom quintile of *Pscore* of the whole sample. Companies in the third group are with median level of restatement risk as the *Pscore* values are between 0.051 and 0.1185.

Column (1) and (2) of Table 4.7 depicts the coefficients and p-value for high risk observations, respectively. It is shown that all Twitter variables as well as many control variables, such as *Growth*, *Foreign*, and *Merger,* become insignificant. This is consistent with the case of GC companies, and the same pattern holds for the group of low risk companies. In contrast, for companies with median level of pre-audit restatement risk, there are positive and significant associations between Twitter variables and the audit fee. Specifically, *Negative* is significant at 0.05 level (p=0.034) with a coefficient of 0.1976, and *Negative* × *Retweets* is significantly associated with the dependent variable (β =

$6.13e^{-6}$, p=0.047). In addition, the vast majority of the control variables become significant with their expected sign, but *Leverage, Growth, and Firstyear* remain non-significant. These results indicate that the audit fee is not sensitive to the negativity of tweets when a company has an exceptional high or low risk of pre-audit restatement, and the positive association between audit fee and negative tweets is stronger for companies with more retweets. H1 and H2 are supported for companies with median risk. The results are consistent with those for companies in Non-GC group as reported in Table 4.6. For the explanatory power, the first two models explain no less than 84% of the variation in audit fees, while the last model explains approximately 75% of the variation.

In sum, the results provide evidence that, Tweets can be used as an additional information source to help auditors make fee decisions when their clients have no issues affecting the going-concern assumption. In addition, Tweets are unreliable when the company has an extremely high/low risk of restatement.

Table 4.7 Regression of Tweets Sentiment on Audit fees by the Level of Restatement Risk

		High risk (Top Quintile)		Median risk (Middle Quintile)		Low risk (Bottom Quintile)	
Variable	Expected Sign	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
		(1)	(2)	(3)	(4)	(5)	(6)
Intercept	+/-	12.3178***	0.001	9.7059***	0.001	11.0437***	0.001
Negative	+	0.3887	0.303	0.1976**	0.034	-0.1123	0.787
Retweets	+	$-6.10e^{-6}$	0.593	$1.36e^{-6}$*	0.058	$-3.13e^{-6}$	0.636
Negative × Retweets	+	$-6.10e^{-5}$	0.391	$6.13e^{-6}$**	0.047	$5.50e^{-6}$	0.866
Tweets	+/-	$4.49e^{-7}$	0.385	$-1.23e^{-8}$	0.114	$4.56e^{-8}$	0.741
Roaearnings	-	-0.2972**	0.050	-0.3312***	0.001	-0.1006	0.727
Size	+	0.4234***	0.001	0.4814***	0.001	0.3070***	0.001
Invrec	+	0.9707**	0.050	0.5887***	0.001	1.3936*	0.058
Leverage	+	0.4758*	0.074	-0.0026	0.963	-0.0989	0.709
Currentratio	-	-0.0602**	0.038	-0.0160***	0.001	-0.0143	0.365
BTM	-	-0.0079	0.891	-0.0467***	0.001	-0.2754**	0.045
Growth	-	-0.1223	0.187	-0.0037	0.682	-0.0385*	0.088
Loss	+	0.2706*	0.051	0.0995***	0.002	0.0386	0.863
Foreign	+	0.2077	0.241	0.2909***	0.001	0.1768	0.278
Merger	+	0.0658	0.663	0.0582**	0.044	0.0560	0.794
Special	+	0.0820	0.738	0.1490***	0.001	0.2699*	0.094
Pscore	+	-10.8065**	0.035	-0.8217	0.384	0.8056	0.932
Firstyear	-	0.0310	0.873	-0.0732	0.158	-0.0509	0.843
GC	+	0.1367	0.648	0.2059**	0.012	0.7562	0.134
Big4	+	0.7411***	0.001	0.4124***	0.001	0.8107***	0.001
IC	+	0.1234**	0.046	0.1122***	0.001	Omitted	
Industry effect		Included		Included		Included	
Observations		90		2172		70	
Adjusted R^2		0.8620		0.8400		0.7496	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

4.5.4. Prediction Performance of the Prediction Model

Furthermore, this study examines the extent to which the twitter variables are able to improve the predictive performance of the audit fee model. It develops a Linear Regression, a Random Forest consisting of 50 Regression Trees, and a traditional Artificial Neural Network with one hidden layer of 100 nodes, with all determinant

variables that used in previous analysis. 10-fold cross validation is applied to validate the constructed models. Table 4.8 presents the prediction results of a baseline model (that uses all control variables but does not consider Twitter variables in equation (1)) and a sentiment model (that incorporate all Twitter variables and control variables) for each algorithm. The predictive accuracy is measured by two of the most commonly used metrics for regression problems: MAE (Mean of Absolute Error) and RMSE (Square Root of Mean of the Squared Errors). MAE is the average over the test sample of the absolute differences between prediction $\hat{y}_j$ and actual observation y_j . RMSE is the square root of the average of squared differences between prediction $\hat{y}_j$and actual observations y_j. The formulas for MAE and RMSE are listed below.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

The results indicate that the sentiment model using Linear Regression algorithm outperforms all other models for audit fee prediction (RMSE=0.4720 and MAE=0.3671). The predictive accuracy generally increases as the Twitter variables are considered as additional predictors for audit fees, especially for the Linear Regression model. Specifically, RMSE reduces 0.1264 (0.5984-0.4720), and MAE decreases 0.0626 (0.4297-0.3671). With Random Forest and ANN, the RMSE of the sentiment model

reduces to 0.6873 and 0.6248, respectively, but the MAEs for both models increase slightly.

Table 4.8 The Results of 10-Fold Cross Validation

	Linear Regression		RF		ANN	
	Baseline model	Sentiment Model	Baseline model	Sentiment model	Baseline model	Sentiment model
	(1)	(2)	(3)	(4)	(5)	(6)
RMSE	0.5984	0.4720	0.6902	0.6879	0.6261	0.6248
MAE	0.4297	0.3671	0.4170	0.4269	0.4617	0.4619

4.7. Robustness Tests

Prior research of audit fees primarily uses contemporaneous financial variables as it is assumed that auditors are able to accurately anticipate the risks of their prospective clients (Yoon, 2016). However, this assumption is questioned as auditors, especially the successor auditors for the initial year of audit, may not fully predict the financial ratios and other factors related to audit pricing (Hackenbrack et al. 2014; Yoon, 2016). To explore the robustness of our results, we estimate Equation (1) with one-year lagged values for all control variables except *Big 4* and *Firstyear*. Table 4.9 and Table 4.10 illustrate the new estimation using the modified audit fee model using one-year lagged values for these control variables. Column (1) and (2) of Table X show an insignificant coefficient of *Negative* (β=0.0718, p=0.361), but the coefficients of *Retweets (*β=2.54e^{-6}, p<0.001*)* and *Negative* × *Retweets (*β=1.08e^{-5}, p<0.001*)* are positive and significant at 0.01 level. Column (3) and (4) present the results for GC companies. Similar to the main results, the coefficients of Negative, Retweets, and the interaction term are insignificant. The last two columns of this table show positive and significant coefficients of *Negative*

*(*β=0.2140, p=0.013*), Retweets* (β=2.29e^{-6}, p<0.001, and the interaction (β=9.57e^{-6}, p<0.001).

The robustness test for groups of companies with different levels of restatement risk is performed and the results are presented in Table 4.10. Consistent with the main test, the Twitter variables are positive and significantly associated with the audit fee for companies with median restatement risk. Similar results are not observed for groups with extremely high or low risk. To summarize, the robustness check supports the main results that audit pricing is sensitive to the sentiment of tweets discussing the company especially when there are a great number of retweets.

Table 4.9 Regression of Tweets Sentiment on Audit Fees for a Robustness Test: Full Sample and Groups by the Existence of GC opinions

		Full sample		GC companies		Non-GC companies	
Variable	Expected Sign	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
		(1)	(2)	(3)	(4)	(5)	(6)
Intercept	+/-	9.6759***	0.001	11.2758***	0.001	9.6981***	0.001
Negative	+	0.0718	0.361	0.3220	0.121	0.2140**	0.013
Retweets	+	$2.54e^{-6}$***	0.001	$3.51e^{-6}$	0.769	$2.29e^{-6}$***	0.001
Negative × Retweets	+	$1.08e^{-5}$***	0.001	$-3.52e^{-5}$	0.580	$9.57e^{-6}$***	0.001
Tweets	+/-	$-1.08e^{-8}$	0.173	$-1.92e^{-6}$	0.124	$-1.1e^{-8}$	0.143
Roaearnings_{t-1}	-	0.0043*	0.068	-0.0417	0.025	0.0078***	0.001
Size_{t-1}	+	0.4540***	0.001	0.4072***	0.001	0.4789***	0.001
$Invrec_{t-1}$	+	0.4087***	0.001	0.5394**	0.033	0.4717***	0.001
$Leverage_{t-1}$	+	0.0612***	0.001	0.0452	0.105	0.0865*	0.064
$Currentratio_{t-1}$	-	-0.0109***	0.001	-0.0157	0.347	-0.0073***	0.008
BTM_{t-1}	-	-0.0025***	0.003	-0.0021*	0.056	-0.0874***	0.001
$Growth_{t-1}$	-	$-8.11e^{-5}$	0.150	$-6.60e^{-5}$	0.360	0.0002	0.426
$Loss_{t-1}$	+	0.1767***	0.001	0.4367**	0.023	0.1895***	0.001
$Segments_{t-1}$	+	0.0177***	0.001	0.0100	0.724	0.0172***	0.001
$Foreign_{t-1}$	+	0.2582***	0.001	0.1696	0.322	0.2374***	0.001
$Merger_{t-1}$	+	0.1394***	0.001	0.1837	0.343	0.1255***	0.001
$Special_{t-1}$	+	0.0995***	0.001	-0.0117	0.905	0.1145***	0.001
Firstyear	-	-0.0990**	0.020	-0.0605	0.604	-0.0925**	0.042
GC_{t-1}	+	0.1519***	0.005				
Big4	+	0.4307***	0.001	0.7095***	0.001	0.3660***	0.001
IC_{t-1}	+	0.0598***	0.001	0.0397	0.256	0.0608***	0.001
Industry effect		Included		Included		Included	
Observations		2332		289		2043	
Adjusted R^2		0.8662		0.7695		0.8535	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

Table 4.10 Regression of Tweets Sentiment on Audit Fees for a Robustness Test: Groups by the Risk of Financial Restatements

		High risk (Top Quintile)		Median risk (Middle Quintile)		Low risk (Bottom Quintile)	
Variable	Expected Sign	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
		(1)	(2)	(3)	(4)	(5)	(6)
Intercept	+/-	10.0545***	0.001	9.3131***	0.001	10.0124***	0.001
Negative	+	0.3164	0.560	0.2350**	0.015	-0.3043	0.710
Retweets	+	$-8.13e^{-6}$	0.556	$2.19e^{-6}$***	0.005	$9.47e^{-7}$	0.901
Negative × Retweets	+	$-9.20e^{-5}$	0.321	$7.88e^{-6}$**	0.012	$-1.31e^{-5}$	0.734
Tweets	+/-	$-2.10e^{-6}$	0.747	$-1.10e^{-8}$	0.155	$-9.96e^{-8}$	0.495
Roaearnings	-	-0.4788	0.380	-0.3404***	0.001	-0.1076	0.575
Size	+	0.5041***	0.001	0.4925***	0.001	0.3842***	0.001
Invrec	+	0.8338	0.102	0.5072***	0.001	1.9899**	0.018
Leverage	+	0.0855	0.756	-0.0783	0.160	0.1666	0.621
Currentratio	-	0.0157	0.721	-0.0143***	0.007	0.0052	0.802
BTM	-	-0.2351	0.143	-0.0967***	0.001	-0.0940	0.300
Growth	-	0.1075	0.511	0.0003	0.894	-0.0100	0.275
Loss	+	0.0075	0.973	0.1272***	0.001	0.2282	0.394
Foreign	+	0.3961*	0.069	0.2796***	0.001	0.3682*	0.087
Merger	+	-0.0754	0.701	0.0819***	0.006	-0.1348	0.598
Special	+	-0.0233	0.915	0.0905**	0.011	0.1319	0.525
Pscore	+	-2.0639	0.308	1.3284	0.168	2.5919	0.867
Firstyear	-	0.2502	0.365	-0.0602	0.252	-0.1266	0.721
GC	+	-0.1743	0.689	0.2444**	0.021	1.1520*	0.066
Big4	+	0.6331***	0.005	0.3716***	0.001	0.6691**	0.027
IC	+	0.1779**	0.031	0.0630***	0.001	0.1894	0.677
Industry effect		Included		Included		Included	
Observations		90		2172		70	
Adjusted R^2		0.8219		0.8433		0.7485	

***, **, * indicates significance at the 0.01, 0.05, and 0.10 level or better, respectively.

4.8. Conclusion

While recent research provides evidence that qualitative factors, such as 10-K filings, 8-K filings, CEO letters, MD&As, and earnings press, influence auditor pricing (e.g., Yoon, 2016, Liu, Vasarhelyi, and Yoon, 2018; Dikolli et al. 2016; Liu, 2015), researchers primarily focus on disclosures from management. Limited research uses the qualitative information provided by both management and other stakeholders (e.g., investors and customers). This study extends prior research with an examination of whether certain characteristics of utterances in social media provide companies' audit

risk information affecting audit fees. Specifically, it investigates the association between the sentiment of user-generated tweets and the audit fee as well as how retweets affect this association. It uses *Negative, Retweets,* and *Tweets* to measure three characteristics of information in Twitter, which are the strength of the negative sentiment of tweets, the popularity of specific topics about the company, and the volume of tweets, respectively. Furthermore, this study uses an Artificial Intelligence technique, deep learning, to obtain the sentiment of tweets, enriching existing textual analysis approaches in accounting and auditing research that rely on "bag of words" approach, which neglects the semantic content of the textual data.

This research argues that the information in Twitter reveals the client's risk related to audit engagement. However, for the full sample analysis, higher audit fees are not significantly associated with more negative tweets. It further seeks to partition the sample to investigate whether risk factors influence the relationship between these tweeting characteristics and audit fee decisions. The results indicate that, for clients without going-concern audit opinion or companies with median level of financial restatement risk, the more negative tweets the company receives, the higher the auditor charges, while controlling for the total volume of tweets and other factors. The popularity of tweets about the company is measured by the maximum number of retweets for each tweet about the company and it is hypothesized that the relationship between audit fee and the negative sentiment of tweets becomes stronger if the tweets are popular among the users. This hypothesis is supported only for companies without going-concern opinion and with median restatement risk.

To test whether the Twitter variables help improve the prediction accuracy of the audit fee model, three algorithms are employed, including Linear Regression, Random Forest, and Artificial Neural Network. It is found that the predictive ability of the model, measured by RMSE and MAE, increases as these Twitter attributes are considered. The robustness test uses one-year lagged value of control variables other than *Tweets*, *Big 4*, and *Firstyear* and still shows an audit fee premium for clients with median level of restatement risk and without auditor-perceived going-concern issues.

This research offers suggestions for accounting research and practice. It documents that for companies with certain characteristics, auditors will incorporate information in Twitter for additional evidence of audit pricing. It suggests that as social media such as Twitter provides qualitative information regarding the risk of the prospective client, it can be used as a technology shortcut to improve the quality of audit decision making (Western Intergovernmental Audit Forum, 2013). AI technology like deep learning can be applied to identify the sentiment of textual data offer efficient and effective evidence with limited human bias to support audit judgment.

4.9. Limitations and Future Research

This study is subject to limitations. First, this study directly uses the result of sentiment analysis of Twitter Insights provided by IBM Watson. It remains a black-box regarding how sentiments are calculated. Second, the observations are all from one single fiscal year, due to the limitation of the availability of twitter data. Watson Twitter Insight has been expired since April 2017. Currently, Twitter Gnip API service provides two APIs to allow users to collect twitter data, REST API and Streaming API. While the

former provides historical tweets (but with limited accessibility for the free version[44]), the latter streams unlimited real-time tweets generated in past seven days. Future study can cumulatively retrieve twitter data using Streaming API. There are a number of tools to access the Twitter API with varied capabilities and required levels of technical skills. These tools include software libraries (e.g., Tweepy for Python and retweet for R), command line tools (e.g., Twarc), web applications (e.g., DMI_TCAT), and plugins for popular analytic packages (e.g., NVIVO, NodeXL for Excel, and TAG for Google Sheets) (Littman, 2017[45]). With those tools, future research can apply DNN to extract more features other than the sentiment as additional evidence to support other types of audit judgment, e.g., client acceptance and continuance, internal control risk evaluation, and audit plan design. A potential obstacle is the scarcity of labelled data for machine training purpose as it is costly and extremely time-consuming to obtain labels of abstract characteristics generated by human experts. A possible solution is using Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), one of the most important new development in deep learning (LeCun, 2016), to generate artificial labelled data. In addition, to enrich the qualitative database of potential audit evidence, it is necessary to explore more data sources such as news articles and analysts reports.

[44] The free version is called Sandbox endpoint, which is designed for testing and developing a proof of concept. It provides 100 tweets per data request with a maximum of 3200 most recent tweets. https://developer.twitter.com/en/docs/tweets/search/overview/premium
[45] https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data

Chapter 5 Conclusions

5.1. Summary

This dissertation attempts to contribute to the auditing field by demonstrating how deep learning technology can be implemented to analyze textual data to support auditor decision making. It seeks to answer: (1) whether the sentiment features of earnings conference calls extracted by deep learning technique provide incremental information regarding the existence of internal control material weaknesses; (2) whether deep learning-based sentiment analysis perform more effectively and efficiently than "bag of words" approach for financial misstatement prediction; (3) whether Twitter information obtained by using deep learning provides insights for the assessment of the prospective client's risk, and consequently helps the auditor determine the audit fee. Table 5.1 summarizes the results of the three essays in this dissertation. It lists the audit-related risk of interest for each essay as well as how the deep learning-based sentiment features improve the explanatory and predictive ability of the models.

Table 5.1 A Summary of Results for the Three Essays

	Risk of interest	Explanatory ability	Predictive ability
Essay 1	ICMW	Improved	Improved
Essay 2	Financial misstatement	N/A	Improved for fraud prediction
Essay 3	Audit engagement risk	Improved for most of the companies	Improved

Essay 1 uses the deep learning technique to measure the overall sentiment and the strength of the "joy" emotion in earnings conference calls. These sentiment measures are

used as additional variables to predict Internal Control Material Weakness disclosed under SOX404. The tool used in this essay is IBM Watson Alchemy Language API, which allows users to call web service provided by a deep learning-based textual analysis to analyze the sample of conference calls transcripts from 2004 to 2014. The results of the primary analysis indicate that, with the sentiment score and the joy score, the explanatory ability of the model improves significantly compared to the baseline model that merely utilizes the major ICMW determinants suggested by prior literature (i.e., Doyle, Ge, and McVay, 2007a; Ashbaugh-Skaife, Collins, and Kinney, 2007). To examine the prediction accuracy, Logistic Regression, Random Forest, and Artificial Neural Network algorithms are employed to construct models to predict the ICMW. The 10-fold cross validation results report that Random Forest outperforms other machine learning models, in terms of a list of model evaluation metrics. In addition, an improvement in prediction accuracy of the model is observed after incorporating the sentiment measures. This study further tests the number and the persistency of material weakness and finds that sentiment features are more effective in predicting companies with more than one ICMW and companies that persistently have ICMW.

Essay 2 applies deep neural network to analyze the sentiment features from a sample of 31,145 MD&As of 10-K filings from 2006 to 2015. The objective of the study is to investigate the ability of the sentiment features for predicting financial misstatements. Similar to essay 1, this essay uses the sentiment score and joy score as a supplementary predictor in conjunction with 82 quantitative predictors provided by previous work (Perols, Bowen, Zimmermann, and Samba, 2017; Dechow et al., 2011; Perols, 2011; Cecchini et al, 2010; Beneish, 1999; Huang, Rose-Green, and Lee, 2012; Churyk, Lee,

and Clinton, 2009). Since financial misstatements includes both frauds and errors, this research aims to use these qualitative and quantitative attributes to conduct three tasks: predicting misstatements involving both frauds and errors, detecting frauds, and identifying errors. To demonstrate the superiority of the deep learning-based sentiment analysis, it compares the predictive performance of two models: one using sentiment features extracted with deep learning in addition to 82 factors suggested by prior literature; the other using sentiment features calculated with the "bag of words" and the same 82 factors. In addition, a model that only considers the 82 variables is developed as a baseline. Furthermore, this study employs Logistic regression, Random Forest, Naïve Bayes, traditional Artificial Neural Network, as well as Deep Neural Network to build the final classification model. The results show that, while all sentiment features are important predictors in the models, deep learning-based sentiment features exhibit the best performance in predicting frauds. However, similar results are not observed for the task of predicting errors and misstatements. Consequently, it concludes that (1) the sentiment features obtained by both deep learning approach and bag of words approach provide essential information for financial misstatement prediction; (2) however, they are effective for fraud prediction only; (3) the deep learning approach generally performs better than the "bag of words" approach in this research.

The last essay examines the information in Twitter and attempts to use the sentiment and other characteristics of tweeting activities of a client company to predict the audit fee. Furthermore, it investigates the effect of risk conditions of the client firm on the association between the characteristics of tweets and the audit fee. With IBM *Twitter Insights*, this research uses their deep natural language processing tool for tweets and

constructs three Twitter variables: *Negative*, *Tweets*, and *Retweets*, where *Negative* is the difference between the percentage of negative tweets and the percentage of positive tweets in all tweets mentioned the client company; *Tweets* is the count of all tweets mentioned the company; and *Retweets* refers to the maximum count of retweets for all tweets about the company, which measures the popularity of the tweets. The three Twitter variables are incorporated in the audit fee model based on prior literature (Francis and Wang 2005; Krishnan et al. 2005; Ghosh and Pawlewicz 2009; Choi et al. 2010; Stanley 2011). Although it does not find a significant coefficient of *Negative* in a full sample test, the interaction *Negative*×*Retweets* is found to be positive and significantly related to audit fee in this test. Furthermore, this study partitions the sample into different groups based on risk conditions regarding the existence of going-concern opinion and the probability of financial restatement. The results show that, for clients without going-concern audit opinion and companies with median level of financial restatement risk, the more negative tweets the company receives, the higher the auditor charges. The relationship between audit fee and the negative sentiment of tweets becomes stronger if the tweets are popular among the Twitter users, measured by *Retweets*. The results also show that the Twitter variables improve the prediction accuracy of audit fee models developed with three algorithms: Linear Regression, Random Forest, and Artificial Neural Network. Finally, the results of robustness test using one-year lagged value of all control variables other than *Tweets*, *Big 4*, and *Firstyear* reinforce the conclusion that there is an audit fee premium for clients with median level of restatement risk and without auditor-perceived going-concern issues.

5.2. Contributions

The main contributions of this dissertation are threefold. First, it is among the first studies to apply deep learning technology to support audit decision making and demonstrates that deep learning is an effective and efficient audit data analytics tool. Second, it explores the incremental informativeness of textual documents for audit risk assessment. Specifically, three types of textual documents are examined, including conference calls, MD&As, and Tweets, to evaluate the risk of internal control material weakness, financial misstatement, and audit engagement, respectively. Third, it offers useful insights to both audit practice and academia in terms of demonstrating the usefulness of sentiment, emotion, and other linguistic characteristics from the business communication documents for the improvement of audit quality.

5.3. Limitations

"Artificial Intelligence isn't coming. It's already here" (Ovaska-Few, 2017). Leading professionals like the big four companies are leveraging this technology to automate mundane and inefficient audit processes (e.g., checking inventory at client's warehouses and reading business contract or confirmations) (Kokina and Davenport, 2017). This dissertation is only an initial attempt to demonstrate the effectiveness and efficiency of deep learning in audit. The major limitation of this dissertation is that the deep learning models developed by IBM Watson is not trained exclusively with finance-specific text, which may produce biased results. Moreover, the underlying mechanism of data processing and calculation of these deep neural networks is a black-box. For example, it

is unclear how the sentiment and emotion scores are extracted from the raw data and how the hyper-parameters of the deep neural networks are configured.

Due to the availability issue of historical tweets, the current sample used in the last essay is restricted to tweets in the year of 2015. Similarly, the misstatement samples in the second essay are obtained solely from Compustat. For example, there are fewer misstatements in 2014 and 2015, as some misstatements may not be identified until they will eventually be restated in the future.

In addition, this dissertation does not discuss how the auditor without programming skills can use the deep learning tools in practice. While some open-source deep learning programming libraries such as TensorFlow and Theano need relatively high level of programming and data analytics skills, many pre-developed deep learning tools such as Watson Analytics and H2O flow have a minimum requirement of these skills (Sun and Vasarhelyi, 2018).

5.4. Future Research

To prompt the application of this technique to more audit procedures, more work needs be done. Future research can be conducted to provide a framework to guide auditors to apply deep learning to different audit stages and procedures. More applications, such as client acceptance and continuance, internal control risk evaluation, and audit plan design, can be conducted with the help of deep learning. Besides textual understanding, other two capabilities of deep learning in big data analytics, speech recognition and visual identification, need to be explored to obtain more sources of audit evidence. It would be interesting to discuss how the interaction between human and

machine can help facilitating the application of deep learning and improving the prediction accuracy.

Besides the three data sources in this dissertation, future research can explore more data sources (e.g., example news articles, press releases, CEO letters, analyst reports, customer reviews, and other social media platforms like Facebook and LinkedIn) and apply deep neural network to extract features other than the emotion of joy. Other types of emotions such as anger, disgust, fear, and sadness may also be explored.

Another direction for future research is to combine sentiment features obtained with deep learning approach and other linguistic characteristics extracted with traditional text mining approach. Examples of those linguistic characteristics include the level of detail, the complexity, the use of hedging and uncertainty language, and immediacy (Burgoon et al, 2016).

Moreover, in the future, with the availability of audit-specific data, researchers can develop their own deep neural networks to support audit judgment. While a potential obstacle is the scarcity of labelled data for machine training purpose as it is costly and extremely time-consuming to obtain labels of abstract characteristics generated by human experts, a possible solution is using Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), one of the most important new development in deep learning (LeCun, 2016), to generate artificial labelled data.

The comparison of deep learning with traditional data analysis technique is conducted only in the second essay. However, similar comparison can be performed for all three essays or other audit applications. Furthermore, a comparison between AI and

human auditors can be conducted as a behavior research to analyze the differences of prediction performance and thinking process between human and machine.

Bibliography

Abbott, L. J., Parker, S., and Peters, G. F. 2004. Audit committee characteristics and restatements. Auditing: A Journal of Practice & Theory, 23(1), 69-87.

Abbott, L. J., Parker, S., and Presley, T. J. 2012. Female board presence and the likelihood of financial restatement. Accounting Horizons, 26(4), 607-629.

Allee, K. D., and DeAngelis, M. D. 2015. The structure of voluntary disclosure narratives Evidence from tone dispersion. Journal of Accounting Research, 53(2), 241-274.

Alles, Gray, and Takagi, 2016. Auditing as an effective technology: the medium is the message. working paper.

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

American Institute of Certified Public Accountants (AICPA). 1975. Other Information in Documents Containing Audited Financial Statements. Statement on Auditing Standards No. 8, AU Section 550. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA). 1988. The Auditor's Consideration of an Entity's Ability to Continue as a Going Concern. Statement on Auditing Standards No. 59. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA). 2010. Other Information in Documents Containing Audited Financial Statements. Statement on Auditing Standards No. 118, AU-C Section 720. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA). 2011. Statements on Auditing Standards: Clarification and Recodification. Statement on Auditing Standards No. 122, AU-C Section 240. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA). 2014. Statement on Auditing Standards: Using the Work of Internal Auditors. Statement on Auditing Standards No. 128, AU-C Section 240. New York, NY: AICPA.

Asare, S.K., Fitzgerald, B.C., Graham, L.E., Joe, J.R., Negangard, E.M. and Wolfe, C.J., 2013. Auditors' internal control over financial reporting decisions: Analysis, synthesis, and research directions. Auditing: A Journal of Practice & Theory, 32(sp1), 131-166.

Ashbaugh‐Skaife, H., R. LaFond, and B. Mayhew. 2003. Do nonaudit services compromise auditor independence? Further evidence. The Accounting Review 78 (3): 611–639

Ashbaugh‐Skaife, H., Collins, D.W. and Lafond, R., 2009. The effect of SOX internal control deficiencies on firm risk and cost of equity. Journal of Accounting Research, 47(1), 1-43.

Ashbaugh-Skaife, H., Collins, D. W., and Kinney, W. R. 2007. The discovery and reporting of internal control deficiencies prior to SOX-mandated audits. Journal of Accounting and Economics, 44(1), 166-192.

Ashbaugh-Skaife, H., Collins, D.W., Kinney Jr, W.R. and LaFond, R., 2008. The effect of SOX internal control deficiencies and their remediation on accrual quality. The Accounting Review, 83(1), 217-250.

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, 492-499). IEEE.

Baker, M., and J. Wurgler. 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61: 1645–1680

Barr-Pulliam, D., Brown-Liburd, and Sanderson, K.A., 2017. The Effects of the Internal Control Opinion and Use of Audit Data Analytics on Perceptions of Audit Quality, Assurance, and Auditor Negligence. Working paper.

Bea, F. 2013. The IRS may be snooping through Facebook and Twitter to nab tax evaders. Digital Trends. https://www.digitaltrends.com/social-media/irs-could-audit-you-for-social-media-posts/

Beasley, M. 1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. The Accounting Review 71 (4): 443–465.

Beneish, M. D. 1999. The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24-36.

Besch, D. 2009. Remarks before the 2009 AICPA National Conference on Current SEC and PCAOB Developments. https://www.sec.gov/news/speech/2009/spch120709db.htm

Bishop, T. 2016. How ‘Amazon Go’ works: The technology behind the online retailer’s groundbreaking new grocery store. GeekWire. https://www.geekwire.com/2016/amazon-go-works-technology-behind-online-retailers-groundbreaking-new-grocery-store/

Blankespoor, E., Miller, G. S., and White, H. D. 2013. The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. The Accounting Review, 89(1), 79-112.

Blankley, A. I., Hurtt, D. N., and MacGregor, J. E. 2012. Abnormal audit fees and restatements. Auditing: A Journal of Practice & Theory, 31(1), 79-96.

Blay, A.D., Geiger, M.A. and North, D.S., 2011. The auditor's going-concern opinion as a communication of risk. Auditing: A Journal of Practice & Theory, 30(2), 77-102.

Bochkay, K., and C. B. Levine. 2014. Using MD&A to improve earnings forecasts. Working paper. Rutgers University.

Bonner, S. E. 2008. Judgment and Decision Making in Accounting. Upper Saddle River, NJ: Prentice Hall

Burgoon, J. et al., 2016. Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. Journal of Language and Social Psychology, 35(2), 123-157.

Bushee, B. J., D. A. Matsumoto, and G. S. Miller. 2003. Open versus closed conference calls: the determinants and effects of broadening access to disclosure. Journal of Accounting & Economics 34 (1-3):149-180.

Castro, W. B. D. L., Peleias, I. R., and Silva, G. P. D. 2015. Determinants of Audit Fees: a Study in the Companies Listed on the BM&FBOVESPA, Brazil. Revista Contabilidade & Finanças, 26(69), 261-273.

Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P., 2010. Detecting management fraud in public companies. Management Science, 56(7), 1146-1160.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K., 2010. Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.

Chen, H., De, P., Hu, Y., and Hwang, B. 2013. Customers as advisors: The role of social media in financial markets. Available at SSRN. http://ssrn.com/abstract=1807265

Chen, J., Demers, E., and Lev, B. 2016. Oh What a Beautiful Morning! Diurnal Variations in Executives' and Analysts' Behavior: Evidence from Conference Calls. Working Paper. http://www.darden.virginia.edu/uploadedFiles/Darden_Web/Content/Faculty_Research/Seminars_and_Conferences/CDL_March_2016.pdf

Chen, Y., Gul, F. A., Truong, C., and Veeraraghavan, M. 2012. Audit Quality and Internal Control Weakness: Evidence from SOX 404 Disclosures. Working paper. Available at SSRN 1979323.

Chen, S. S., Lai, S. M., Liu, C. L., and McVay, S. E. 2014. Overconfident managers and internal controls. Working paper.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. Marketing Science, 29(5), 944-957.

Choi, J.H., Kim, C., Kim, J.B., and Zang, Y. 2010. Audit Office Size, Audit Quality, and Audit Pricing. Auditing: A Journal of Practice & Theory 29 (1):73-97.

Choi, J.H., Kim, J.B., Liu, X. H., and Simunic, D. A. 2008. Audit pricing, legal liability regimes, and Big 4 premiums: Theory and cross-country evidence. Contemporary Accounting Research 25 (1): 55–99.

Churyk, N.T., Lee, C.C., and Clinton, D.B. 2009. Early detection of fraud: evidence from restatements. Advances is Accounting Behavioral Research, 12, 25-40.

Coats, P.K. and Fant, L.F., 1993. Recognizing financial distress patterns using a neural network tool. Financial management, 22(3), 142-155.

Collins, D., A. Masli, A. L. Reitenga, and J. M. Sanchez. 2009. Earnings restatements, the Sarbanes-Oxley Act and the disciplining of chief financial officers. Journal of Accounting Auditing and Finance 24 (1): 1–34.

Corbin Perception. 2015. Earnings Calls as a Competitive Tool. http://www.corbinperception.com/research-portal/leadership-white-papers/earnings-callscompetitive-tool/

Core, J. E., Guay, W., and Larcker, D. F. 2008. The power of the pen and executive compensation. Journal of Financial Economics, 88(1), 1-25.

Cong, Y. and Du, H., 2007. Welcome to the World of Web 2.0. The CPA Journal, 77(5), 6-10.

COSO, 2004. Enterprise Risk Management—Integrated Framework.

Czerney, K., Schmidt, J. J., and Thompson, A. M. 2014. Does auditor explanatory language in unqualified audit reports indicate increased financial misstatement risk?. The Accounting Review, 89(6), 2115-2149.

Davis, A. K., Ge, W., Matsumoto, D., and Zhang, J. L. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies, 20(2), 639-673.

Davis, A.K., Piger, J.M. and Sedor, L.M., 2012. Beyond the numbers: Measuring the information content of earnings press release language. Contemporary Accounting Research, 29(3), 845-868.

Debreceny, R. S. 2015. Social media, social networks, and accounting. Journal of Information Systems, 29(2), 1-4.

Debreceny, R., Rahman, A., and Wang, T. 2016. Is Twittersphere activity associated with stock market reactions to corporate announcements. Working paper.

Deeplearning4j. 2017. Introduction to Word2Vec. https://deeplearning4j.org/word2vec.html#intro

Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. 2011. Predicting material accounting misstatements. Contemporary accounting research, 28(1), 17-82.

DeFond, M., Jiambalvo, J. 1991. Incidence and circumstances of accounting errors. The Accounting Review 66, 643-655.

DeFond, M.L., Francis, J.R. and Wong, T.J., 2000. Auditor industry specialization and market segmentation: Evidence from Hong Kong. Auditing: A Journal of Practice & Theory, 19(1), 49-66.

Deloitte. 2016. Press release. Deloitte forms alliance with Kira Systems to drive the adoption of artificial intelligence in the workplace. New York.

Depaulo, B. M., et al. 2003. Cues to deception. Psychological Bulletin 129: 74–118.

DePaulo, B. M., Rosenthal, R., Rosenkrantz, J., and Green, C. R. 1982. Actual and perceived cues to deception: A closer look at speech. Basic and Applied Social Psychology, 3(4), 291-312.

De Simone, L., Ege, M. S., and Stomberg, B. 2014. Internal control quality: The role of auditor-provided tax services. The Accounting Review, 90(4), 1469-1496.

Dikolli, S. S. and Keusch, T. Mayew, W. J. and Steffen, T. D. 2017. CEO Behavioral Integrity, Auditor Responses, and Firm Outcomes. Working paper. https://ssrn.com/abstract=2131476 or http://dx.doi.org/10.2139/ssrn.2131476

Doyle, J., Ge, W., and McVay, S. 2007a. Determinants of weaknesses in internal control over financial reporting. Journal of accounting and Economics, 44(1), 193-223.

Doyle, J.T., Ge, W. and McVay, S., 2007b. Accruals quality and internal control over financial reporting. The Accounting Review, 82(5), 1141-1170.

Drummond, C. and Holte, R.C. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In Workshop on learning from imbalanced datasets II. August, Vol. 11:1-8. Washington DC: Citeseer.

Druz, M., Wagner, A. F., and Zeckhauser, R. J. 2015. Tips and Tells from Managers: How Analysts and the Market Read Between the Lines of Conference Calls (No. w20991). National Bureau of Economic Research.

Du, H., and W. Jiang. 2015. Do social media matter? Initial empirical evidence. Journal of Information Systems 29 (2):51-70.Ekman, P., and Friesen, W. V. 1969. Nonverbal leakage and cues to deception. Psychiatry, 32, 88-106.

Eschenbrenner, B., Nah, F. and Telaprolu, V. R. 2015. Efficacy of social media utilization by public accounting firms: Findings and directions for future research. Journal of Information Systems 29 (2).

Feldman, B. 1996. It's your call. Investor Relations (December): 35-37.

Ferguson, A., Francis, J. R., and Stokes, D. J. 2003. The effects of firm-wide and office-level industry expertise on audit pricing. The accounting review, 78(2), 429-448.

Foo, W., Bliss, M. A., Gul, F. A., and Lai, K. 2016. Auditors' Response to Analysts' Forecast Properties: Some Evidence from Audit Fee Pricing. Working paper.

Francis, J. R. 1984. The effect of audit firm size on audit prices: A study of the Australian market. Journal of accounting and economics, 6(2), 133-151.

Francis, J.R. and Ke, B., 2006. Disclosure of fees paid to auditors and the market valuation of earnings surprises. Review of Accounting Studies, 11(4), 495-523.

Francis, J., Philbrick, D., Schipper, K. 1994. Shareholder litigation and corporate disclosures. Journal of Accounting Research 32, 137–164.

Francis, J. R., and Wang, D. 2005. Impact of the SEC's public fee disclosure requirement on subsequent period fees and implications for market efficiency. Auditing: A Journal of Practice & Theory 24 (s-1):145-160.

Frankel, R., Johnson, M., and Skinner, D. 1999. An Empirical Examination of Conference Calls as a Voluntary Disclosure Medium. Journal of Accounting Research 37 (1):133-150.

Franzel, M. J. 2015. Current Issues, Trends, and Open Questions in Audits of Internal Control over Financial Reporting. A speech made in American Accounting

Association Annual Meeting. Chicago. Aug. 8, 2015. https://pcaobus.org/News/Speech/Pages/08102015_Franzel.aspx

Freedman, D. A. 2009. Statistical Models: Theory and Practice. Cambridge University Press.

Galant, D.1994. The Technology Trap. Institutional Investor (May): 141-45.

Garcia, D. 2013. Sentiment during recessions. The Journal of Finance, 68(3), 1267-1300.

Geisser, S., 2017. Predictive inference. Routledge.

Ghosh, A., and R. Pawlewicz. 2009. The impact of regulation on auditor fees: Evidence from the Sarbanes-Oxley Act. Auditing: A Journal of Practice & Theory 28 (2):171-197.

Goel, S., Gangolly, J., Faerman, S. R., and Uzuner, O. 2010. Can linguistic predictors detect fraudulent financial filings? Journal of Emerging Technologies in Accounting, 7(1), 25-46.

Godarzi, T. 2011. IBM's Watson could be used for sentiment analysis in social media. The blogherald. http://www.blogherald.com/2011/02/21/ibms-watson-could-be-used-for-sentiment-analysis-in-social-media/

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. Deep learning (Vol. 1). Cambridge: MIT press. http://www.deeplearningbook.org

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In Advances in neural information processing systems: 2672-2680.

Google speech team. 2015. Google voice search: faster and more accurate. http://googleresearch.blogspot.com/2015/09/google-voice-search-faster-and-more.html.

Gul, F.A. 2007. Hong Kong Auditing: Economic Theory and Practice. City University of Hong Kong Press, Hong Kong

Hackenbrack, K. E., and Hogan, C. E. 2005. Client retention and engagement-level pricing. Auditing: A Journal of Practice & Theory 24 (1):7-20

Hackenbrack, K. E., Jenkins, N. T., and Pevzner, M. 2014. Relevant but delayed information in negotiated audit fees. Auditing: A Journal of Practice & Theory 33 (4):95-117.

Hammersley, J.S., Myers, L.A. and Shakespeare, C., 2008. Market reactions to the disclosure of internal control weaknesses and to the characteristics of those weaknesses under section 302 of the Sarbanes Oxley Act of 2002. Review of Accounting Studies, 13(1), 141-165.

Heaton, J.B., Polson, N., and Witte. J.H. 2016. Deep Learning for Finance: Deep Portfolios. Working paper. https://ssrn.com/abstract=2838013

Hennes, K., Leone, A., and Miller, B. 2008. The Importance of Distinguishing Errors from Irregularities in Restatement Research: The Case of Restatements and CEO/CFO Turnover. The Accounting Review 83: 1487–1519.

Henry, E. 2006. Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. Journal of Emerging Technologies in Accounting, 3, 1–19.

Henry, E., and Leone, A. J. 2016. Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. The Accounting Review, 91(1), 153-178.

Hinton, G.E. et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6), .82-97.

Hinton, G.E., Osindero, S. and Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

Hobson, J. L., Mayew, W. J., and Venkatachalam, M. 2012. Analyzing speech to detect financial misreporting. Journal of Accounting Research, 50(2), 349-392.

Huang, H. W., Rose-Green, E., and Lee, C. C. 2012. CEO age and financial reporting quality. *Accounting Horizons*, *26*(4), 725-740.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3), 585-594.

Ho, T.K. 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. 278–282.

Hoitash, R., Hoitash, U. and Bedard, J.C., 2008. Internal control quality and audit pricing under the Sarbanes-Oxley Act. Auditing: A Journal of Practice & Theory, 27(1), .105-126.

Hoitash, R., Hoitash, U. and Johnstone, K.M., 2012. Internal control material weaknesses and CFO compensation. Contemporary Accounting Research, 29(3), pp.768-803.

IBM Watson. 2015. Thought Leadership Whitepaper. Sentiment analysis with AlchemyAPI: A hybrid approach.

Issa, H, Sun, T., and Vasarhelyi, M.A. 2016. Research ideas for Artificial Intelligence in auditing: the formalization of audit and workforce supplementation. Working paper. Rutgers Business School.

Jegadeesh, N., and Wu, A.D. 2012. Word power: A new approach for content analysis. AFA 2012 Chicago Meetings Paper. http://ssrn.com/abstract=1787273

Jennings, M.M., Pany, K. and Reckers, P.M., 2008. Internal control audits: Judges' perceptions of the credibility of the financial reporting process and likely auditor liability. Advances in Accounting, 24(2), 182-190.

Jin, X., Wah, B. W., Cheng, X., and Wang, Y. 2015. Significance and challenges of big data research. Big Data Research, 2(2), 59–64.

Johnstone, K., Li, C. and Rupley, K.H., 2011. Changes in corporate governance associated with the revelation of internal control material weaknesses and their subsequent remediation. Contemporary Accounting Research, 28(1), 331-383.

Kaplan, A. M., and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of social media. Business Horizons 53: 59–68.

Kraut, R.E. and Poe, D.B., 1980. Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. Journal of Personality and Social Psychology, 39(5), p.784.

Kim, J. B., Liu, X., and Zheng, L. 2012. The impact of mandatory IFRS adoption on audit fees: Theory and evidence. The Accounting Review, 87(6), 2061-2094.

Kinney, W., McDaniel, L. 1989. Characteristics of firms correcting previously reported quarterly earnings. Journal of Accounting and Economics 11, 71–93.

Kokina, J. and Davenport, T.H. 2017. The Emergence of Artificial Intelligence: How Automation is Changing Auditing. Journal of Emerging Technologies in Accounting: 14(1), 115-122.

Kothari, S. P., Li, X., and Short, J. E. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. The Accounting Review, 84(5), 1639-1670.

Kreutzfeldt, R.W. and Wallace, W.A. 1986. Error characteristics in audit populations: Their profile and relationship to environmental factors. Auditing: A Journal of Practice and Theory 6 (Fall): 20-43.

Krishnan, J. 2005. Audit committee quality and internal control: An empirical analysis. The accounting review, 80(2), 649-675.

Krishnan, J., Sami, H., and Zhang, Y. 2005. Does the provision of nonaudit services affect investor perceptions of auditor independence? Auditing: A Journal of Practice & Theory 24 (2):111-135.

Lam, B. 2015. The hidden messages in corporate conference calls. The Atlantic. https://www.theatlantic.com/business/archive/2015/03/the-hidden-messages-in-corporate-conference-calls/387100/

Larker, D., and Zakolyukina, A. 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495–540.

LeCun, Y. 2016. Answer to "What are some recent and potentially upcoming breakthroughs in deep learning?" Quora. https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), 436.

Lee, J. E. 2016. CEO overconfidence and the effectiveness of internal control over financial reporting. Journal of Applied Business Research, 32(1), 81.

Lee, L. F., Hutton, A. P., and Shu, S. 2015. The role of social media in the capital market: evidence from consumer product recalls. Journal of Accounting Research, 53(2), 367-404.

Lee, C. H., Lusk, E., and Halperin, M. 2014. Content analysis for detection of reporting irregularities: Evidence from restatements during the SOX Era. Journal of Forensic and Investigative Accounting, 6, 99-122.

Li, F. 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan. http://ssrn.com/paper=898181.

Liu, Z., 2015. MD & A Disclosure Tone and Audit Pricing. Doctoral Dissertation. Drexel University.

Liu, Y., Vasarhelyi, M.A., and Yoon, K. 2018. Are auditors professionally skeptical? Evidence from audit efforts and voluntary disclosure tone. Working Paper.

Lobo, G.J. and Zhao, Y., 2013. Relation between audit effort and financial report misstatements: Evidence from quarterly and annual restatements. The Accounting Review, 88(4), 1385-1412.

Lopez, T.J., Vandervelde, S.D. and Wu, Y.J., 2009. Investor perceptions of an auditor's adverse internal control opinion. Journal of Accounting and Public Policy, 28(3), 231-250.

Loughran, T., and McDonald, B. 2011a. When is a liability not a liability? Textual Analysis, Dictionaries, and 10‐Ks. Journal of Finance, 66, 35–65.

Loughran, T., and McDonald, B. 2011b. Barron's red flags: Do they actually work? Journal of Behavioral Finance, 12(2), 90–97.

Loughran, T., and McDonald, B. 2013. Measuring readability in financial disclosures. Available at SSRN. http://ssrn.com/abstract=1920411

Louwers, T. J., Ramsay, R. J., Sinason, D. H., Strawser, J. R., and Thibodeau, J. C. 2015. Auditing and assurance services. New York, NY: McGraw-Hill/Irwin. 6th Revised ed. Edition.

Lyon, J.D. and Maher, M.W., 2005. The importance of business risk in setting audit fees: Evidence from cases of client misconduct. Journal of Accounting Research, 43(1), .133-151.

Manning, C., and Schutze, H. 1999. Foundations of statistical natural language processing. Cambridge, MA: The MIT Press.

Matsumoto, Dawn, Maarten Pronk, and Erik Roelofsen. 2011.What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions. Accounting Review, vol. 86, no. 4 (July): 1383–1414.

Mayew, W. J., and Venkatachalam, M. 2012. The power of voice: Managerial affective states and future firm performance. The Journal of Finance, 67(1), 1-43.

McConnell Jr, D.K. and Banks, G.Y., 2003. How Sarbanes-Oxley will change the audit process. Journal of Accountancy, 196(3), 49.

Mian, G. M., and Sankaraguruswamy, S. 2012. Investor Sentiment and Stock Market Response to Earnings News. The Accounting Review 87 (4):1357-1384

Munsif, V., Raghunandan, K., Rama, D. V., and Singhvi, M. 2011. Audit fees after remediation of internal control weaknesses. Accounting Horizons, 25(1), 87-105.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. 2015. Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21.

National Research Council 2013 Frontiers in Massive Data Analysis. The National Academies Press, Washington, DC. http://www.nap.edu/openbook.php?record_id=18374

Nielsen, M.A. 2015. Neuralnetworks and Deep learning. Determination Press.

O'Keefe, T.B., Simunic, D.A. and Stein, M.T., 1994. The production of audit services: Evidence from a major public accounting firm. Journal of Accounting Research, .241-261.

Öğüt, H., Aktaş, R., Alp, A., and Doğanay, M. M. 2009. Prediction of financial information manipulation by using support vector machine and probabilistic neural network. Expert Systems with Applications, 36(3), 5419-5423.

Ovaska-Few, S. 2017. How artificial intelligence is changing accounting. Journal of Accountancy. https://www.journalofaccountancy.com/newsletters/2017/oct/artificial-intelligence-changing-accounting.html

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. 2010. Press, J. 2008. London

Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K., 2015. The development and psychometric properties of LIWC2015. http://hdl.handle.net/2152/31333

PCAOB. 2004. AU Section 325: Communications About Control Deficiencies in an Audit of Financial Statements.

PCAOB. 2007. Auditing Standard No. 5: An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements.

PCAOB. 2010a. Auditing Standard No.12: Auditing Standard Related to the Auditor's Assessment of and Response to Risk and Related Amendments to PCAOB Standards.

PCAOB. 2010b. Auditing Standard No. 16: Communications with Audit Committees. https://pcaobus.org/Standards/Auditing/pages/auditing_standard_16.aspx

Perols, J.L., 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 30(2), 19-50.

Perols, J.L., Bowen, R.M., Zimmermann, C. and Samba, B., 2017. Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 92(2), 221-245.

Price, S. M., Doran, J. S., Peterson, D. R., and Bliss, B. A. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance, 36(4), 992-1011.

Prokofieva, M., 2015. Twitter-based dissemination of corporate disclosure and the intervening effects of firms' visibility: Evidence from Australian-listed companies. Journal of Information Systems, 29(2), 107-136.

Raphael. 2015. How Artificial Intelligence Can Boost Audit Quality. http://ww2.cfo.com/auditing/2015/06/artificial-intelligence-can-boost-audit-quality/

Rapoport ,2016. Auditing Firms Count on Technology for Backup. The Wall Street Journal. http://www.wsj.com/articles/auditing-firms-count-on-technology-for-backup-1457398380.

Redmayne, N. B., Bradbury, M. E., and Cahan, S. F. 2010. The effect of political visibility on audit effort and audit pricing. Accounting & Finance, 50(4), 921-939.

Rice, S. C., and Weber, D. P. 2012. How effective is internal control reporting under SOX 404? Determinants of the (non‐) disclosure of existing material weaknesses. Journal of Accounting Research, 50(3), 811-843.

Romanus, R. N., Maher, J. J., and Fleming, D. M. 2008. Auditor industry specialization, auditor changes, and accounting restatements. Accounting Horizons, 22(4), 389-413.

Russell, S.J. and Norvig, P. 2010. Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall (ISBN 9780136042594).

Rui, H., Liu, Y., and Whinston, A. 2013. Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems, 55(4), 863-870.

Salton, G and McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY, USA.

Schafer J. 2011. Reading People by the Words They Speak. Psychology Today. https://www.psychologytoday.com/blog/let-their-words-do-the-talking/201106/reading-people-the-words-they-speak

SEC. 2010. Internal control over financial reporting in exchange act periodic reports of non-accelerated filers. SEC.gov. https://www.sec.gov/rules/final/2010/33-9142.pdf

Sedor, L. M. 2002. An explanation for unintentional optimism in analysts' earnings forecasts. The Accounting Review, 77(4), 731-753.

Shiller, R. J. 2005. Irrational Exuberance. Princeton, NJ: Princeton University Press.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

Simunic, D. A. 1980. The pricing of audit services: Theory and evidence. Journal of accounting research, 161-190

Sinnett, W.M., 2007. CFO skillsets changing... again: with CFO turnover still near record levels, Financial Executives Research Foundation (FERF) asked some in the executive search business to discuss the trends and what they are looking for in senior financial executive recruits. Financial Executive, 23(5), 35-38.

Skinner, D. J. 2003. Should firms disclose everything to everybody? A discussion of "Open vs. closed conference calls: the determinants and effects of broadening access to disclosure". Journal of Accounting & Economics, 181-187.

Stanley, J. D. 2011. Is the Audit Fee Disclosure a Leading Indicator of Clients' Business Risk? Auditing: A Journal of Practice & Theory 30 (3):157-179.

Sun, T, and Vasarhelyi, M. A. 2017. Deep Learning and the Future of Auditing: How an Evolving Technology Could Transform Analysis and Improve Judgment. CPA Journal. Jun, p24-29.

Sun, T. and Vasarhelyi, M.A. 2018. Embracing Textual Data Analytics in Auditing with Deep Learning. International Journal of Digital Accounting Research. Forthcoming.

Sun, T., Liu, Y., and Vasarhelyi, M.A. 2017. The Performance of Sentiment Feature of MD&As for Financial Misstatements Prediction: A Comparison of Deep Learning and Bag of Words Approach. Working paper.

Synthesio. 2011. The truth about natural language processing. https://www.synthesio.com/wp-content/uploads/2010/11/SYNTHESIO-NLP.pdf

Tetlock, P. C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62: 1139–68.

Tsai, C. F., and Chiou, Y. J. 2009. Earnings management prediction: A pilot study of combining neural networks and decision trees. Expert systems with applications, 36(3), 7183-7191.

Turian, J. 2015. Using AlchemyAPI for enterprise-Grade text analysis. AlchemyAPI.

vanGerven, M. and Bohte, S.M., 2017. Artificial neural networks as models of neural information processing: Editorial on the Research Topic Artificial Neural Networks as Models of Neural Information Processing. https://doi.org/10.3389/fncom.2017.00114

Venkataraman, R., Weber, J.P. and Willenborg, M., 2008. Litigation risk, audit quality, and audit fees: Evidence from initial public offerings. The Accounting Review, 83(5), .1315-1345.

Waroff, D. 1994.The Well-Connected IR Officer. Institutional Investor (May): 134-38.

Western Intergovernmental Audit Forum. 2013. Using Technology to Enhance Audit Work and Communication. https://classic.regonline.com/custImages/300000/300628/TATE_WIAFTempe2013-TatePresentation.pdf

Wheeler, S., and Cereola, S. J. 2015. Auditor scrutiny of unaudited client disclosure outlets: Recognized vs. disclosed financial statement items also appearing in the MD&A. Advances in Accounting, 31(1), 91-95.

Wu, B. and Shen, H., 2015. Analyzing and predicting news popularity on Twitter. International Journal of Information Management, 35(6), 702-711.

Wu, Y.J. and Tuttle, B., 2014. The interactive effects of internal control audits and manager legal liability on managers' internal controls decisions, investor confidence, and market prices. Contemporary Accounting Research, 31(2), 444-468.

Yoon, K., Hoogduin, L., and Zhang, L. 2015. Big Data as Complementary Audit Evidence. Accounting Horizons. June, Vol. 29, No. 2, 431-438.

Yoon, K. 2016. Three essays on unorthodox audit evidence. Doctoral Dissertation. Rutgers, The State University of New Jersey. https://doi.org/doi:10.7282/T3DJ5HVN

Young, C. S., Tsai, L. C., and Hsu, H. W. 2008. The effect of controlling shareholders' excess board seats control on financial restatements: evidence from Taiwan. Review of Quantitative Finance and Accounting, 30(3), 297-314.

Yu, Y., Duan, W., and Cao, Q. 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems, 55(4), 919-926.

Zhang, X., Zhao, J., and LeCun, Y. 2015. Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems. 649-657.

Zhang, Y., Zhou, J., and Zhou, N. 2007. Audit committee quality, auditor independence, and internal control weaknesses. Journal of accounting and public policy, 26(3), 300-327.

Zhou, M., Lei, L., Wang, J., Fan, W., and Wang, A. G. 2015. Social media adoption and corporate disclosure. Journal of Information Systems, 29(2), 23-50.

Zuckerman, M., and Driver R. 1985. Telling Lies: Verbal and Nonverbal Communication of Deception, in Multichannel Integrations of Nonverbal Behavior. Hillsdale, NJ: Lawrence Erlbaum: 129–147.

Appendices

Appendix A: Definitions of Explanatory Variables for Chapter 3

Panel A: 22 Variables from Dechow et al. (2011)		
Variable	Explanation	Definition
ACOB	Abnormal change in order backlog	(OB -OB_{t-1})/OB_{t-1} - (SALE - $SALE_{t-1}$)/$SALE_{t-1}$
ISSUE	Actual issuance	if SSTK > 0 or DLTIS > 0 then 1, else 0
BM	Book-to-market	CEQ/ (CSHO * PRCC_F)
CPENSION	Change in expected return on pension plan assets	PPROR -$PPROR_{t-1}$
CFCF	Change in free cash flows	(IB - RSST Accruals)/Average total assets - (IB_{t-1}- RSST $Accruals_{t-1}$)/Average total $assets_{t-1}$
CINV	Change in inventory	(INVT - $INVT_{t-1}$)/Average total assets
CLEASE	Change in operating lease activity	((MRC1/1.1 + MRC2/1.1^2 + MRC3/1.1^3 + MRC4/1.1^4 + MRC5/1.1^5) -($MRC1_{t-1}$/1.1 + $MRC2_{t-1}$/1.1^2 + $MRC3_{t-1}$/1.1^3 + $MRC4_{t-1}$/1.1^4 + $MRC5_{t-1}$/1.1^5))/Average total assets
CRECV	Change in receivables	(RECT - $RECT_{t-1}$)/Average total assets
CROA	Change in return on assets	IB/Average total assets – IB_{t-1})/ Average total $assets_{t-1}$
DTE	Deferred tax expense	TXDI/AT_{t-1}
DFF	Demand for financing (ex ante)	if (OANCF - ($CAPX_{t-3}$+ $CAPX_{t-2}$+ $CAPX_{t-1}$)/3)/ACT < -0.5, then 1, else 0
EP	Earnings to price	IB/ (CSHO x PRCC_F)
LEASE	Existence of operating leases	if MRC1 > 0, or MRC2 > 0, or MRC3 > 0, or MRC4 > 0, or MRC5 > 0, then 1, else 0
PENSION	Expected return on pension plan assets	PPROR
FR	Level of finance raised	FINCF/Average total assets
LEV	Leverage	DLTT/AT
PCCM	Percentage change in cash margin	((1 - (COGS + (INVT – $INVT_{t-1}$))/ (SALE - (RECT – $RECT_{t-1}$))) - (1 - ($COGS_{t-1}$ + ($INVT_{t-1}$ – $INVT_{t-2}$))/ ($SALE_{t-1}$ – ($RECT_{t-1}$ – $RECT_{t-2}$))))/ (1

		- $(COGS_{t-1} + (INVT_{t-1} - INVT_{t-2}))/(SALE_{t-1} - (RECT_{t-1} - RECT_{t-2}))$
PCCS	Percentage change in cash sales	$((SALE - (RECT - RECT_{t-1})) - (SALE_{t-1} - (RECT_{t-1} - RECT_{t-2})))/(SALE_{t-1} - (RECT_{t-1} - RECT_{t-2}))$
RSSTAC	RSST accruals	RSST Accruals = (△WC + △NCO + △FIN)/Average total assets, where: WC = (ACT - CHE) - (LCT - DLC) NCO = (AT - ACT - IVAO) - (LT - LCT - DLTT) FIN = (IVST + IVAO) - (DLTT + DLC + PSTK)
SOFT	Soft assets	(AT - PPENT - CHE)/Average total assets
UEMP	Unexpected employee productivity[46]	$(SALE/EMP - SALE_{t-1}/EMP_{t-1})/(SALE_{t-1}/EMP_{t-1}) - INDUSTRY((SALE/EMP - SALE_{t-1}/EMP_{t-1})/(SALE_{t-1}/EMP_{t-1}))$
WCA	WC accruals	$(((ACT - ACT_{t-1}) - (CHE - CHE_{t-1})) - ((LCT - LCT_{t-1}) - (DLC - DLC_{t-1}) - (TXP - TCP_{t-1})) - DP)$/Average total assets
Panel B: 34 Variables from Perols (2011)		
Variable	Explanation	Definition
RECSALE	Accounts receivable to sales	RECT/SALE
RECAT	Accounts receivable to total assets	RECT/AT
RECD	Allowance for doubtful accounts	RECD
RECDREC	Allowance for doubtful accounts to accounts receivable	RECD/RECT
RECDSALE	Allowance for doubtful accounts to net sales	RECD/SALE
ZSCORE	Altman Z-score	3.3 * (IB + XINT + TXT)/AT + 0.999 * SALE/AT + 0.6 * CSHO * PRCC_F/LT + 1.2 * WCAP/AT + 1.4 * RE/AT
BIGFOUR	Big Four auditor	if 0 < AU < 9, then 1, else 0
CINVTSALE	Current minus prior year inventory to sales	$INVT/SALE - INVT_{t-1}/SALE_{t-1}$

[46] Similar variable used in both Dechow et al. (2011) (abnormal change in employees) and Perols (2011) (unexpected employee productivity)

DAYS	Days in receivables index	(RECT/SALE)/ ($RECT_{t-1}$/$SALE_{t-1}$)
DE	Debt-to-equity	LT/CEQ
DCS	Declining cash sales dummy[47]	if SALE - (RECT – $RECT_{t-1}$) < $SALE_{t-1}$– ($RECT_{t-1}$– $RECT_{t-2}$) then 1, else 0
FAAT	Fixed assets to total assets	PPEGT/AT
GROWTH	Four-year geometric sales growth rate	(SALE/$SALE_{t-3}$) ^ (1/4) - 1
GM	Gross margin	(SALE - COGS)/SALE
PCPRCCF	Holding period return in the violation period	(PRCC_F- $PRCC\_F_{t-1}$)/ $PRCC\_F_{t-1}$
ROEDIFF	Industry ROE minus firm ROE	$NI_{industry}$/$CEQ_{industry}$ – NI/CEQ
INVSALE	Inventory to sales	INVT/SALE
SALE	Net sales	SALE
PA	Positive accruals dummy	if (IB - OANCF) > 0 and (IB_{t-1} - $OANCF_{t-1}$) > 0, then 1, else 0
PROAAT	Prior-year ROA to total assets current year	(NI_{t-1}/AT_{t-1})/AT
PPENTAT	Property, plant, and equipment to total assets	PPENT/AT
SALEAT	Sales to total assets	SALE/AT
TURNOVERS	The number of auditor turnovers	if AU <> AU_{t-1}, then 1, else 0 + if AU_{t-1}<> AU_{t-2}, then 1, else 0 + if AU_{t-2} <> AU_{t-3} then 1, else 0
INTEREST	Times interest earned	(IB + XINT + TXT)/XINT
TATA	Total accruals to total assets	(IB - OANCF)/AT
LTAT	Total debt to total assets	LT/AT
DA	Total discretionary accrual	RSST $Accruals_{t-1}$ + RSST $Accruals_{t-2}$ + RSST $Accruals_{t-3}$
ISMV	Value of issued securities to market value	if CSHI > 0, then CSHI * PRCC_F/ (CSHO * PRCC_F) else if (CSHO – $CSHO_{t-1}$) > 0, then ((CSHO – $CSHO_{t-1}$) * PRCC_F)/ (CSHO * PRCC_F), else 0

[47] As this variable is similar to "Percentage change in cash sales" (Dechow et al., 2011). It is not used in this paper.

RECV1.1	Whether accounts receivable > 1.1 of last year's	if RECT/$RECT_{t-1}$> 1.1, then 1, else 0
AMEX	Whether firm was listed on AMEX	if EXCHG = 5, 15, 16, 17, 18, then 1, else 0
GM1.1	Whether gross margin percent >1.1 of last year's	if ((SALE - COGS)/SALE)/ (($SALE_{t-1}$- $COGS_{t-1}$)/ $SALE_{t-1}$) > 1.1, then 1, else 0
LIFO	Whether LIFO	if INVVAL = 2, then 1, else 0
NEWSEC	Whether new securities were issued	if (CSHO – $CSHO_{t-1}$) > 0 or CSHI > 0, then 1, else 0
MANUF	Whether SIC code between 2999 and 4000	if 2999 < SIC < 4000, then 1, else 0
Panel C: 20 Variables based on Cecchini et al. (2010)		
Variable	Explanation	definition
CSALEAT	Change in sales to assets	SALE/AT - $SALE_{t-1}/AT_{t-1}$
SALEEMP	Sales to employees	SALE/EMP
CSALEEMP	change in Sales to employees	SALE/EMP - $SALE_{t-1}/EMP_{t-1}$
SALEXOPR	Sales to operating expenses	SALE/XOPR
CSALEXOPR	change in Sales to operating expenses	SALE/XOPR - $SALE_{t-1}/XOPR_{t-1}$
ROE	Return on equity	NI/CEQ
ROA	Return on assets	NI/AT
CROE	Change in return on equity	NI/CEQ - NI_{t-1}/CEQ_{t-1}
ROS	Return on sales	NI/SALE
CROS	Change in return on sales	NI/SALE - $NI_{t-1}/SALE_{t-1}$
APINVT	Accounts payable to inventory	AP/INVT
CAPINVT	Change in accounts payable to inventory	AP/INVT - $AP_{t-1}/INVT_{t-1}$
LTXINT	Liabilities to interest expenses	LT/XINT
CLTXINT	Change in liabilities to interest expenses	LT/XINT - $LT_{t-1}/XINT_{t-1}$
XOPR	Expenses	XOPR
CXOPR	Change in expenses	XOPR-$XOPR_{t-1}$

AT	Total assets	AT	
CAT	change in total assets	AT-AT_{t-1}	
LT	Liabilities	LT	
CLT	Change in liabilities	LT-LT_{t-1}	
Panel D: other variables			
SGAI	Selling, general, and administrative expenses index	($XSGA_t$/$Sales_t$)/ ($XSGA_{t-1}$ /$Sales_{t-1}$)	Beneish, 1999
MVE	Market value of equity	Ln(PRCC_F*CSHO)	Huang et al, 2012
DEPI	Depreciation index	(DP_{t-1}/(DP_{t-1}+ $PPENT_{t-1}$))/(DP_t/(DP_t+$PPENT_t$))	Beneish, 1999
AQI	Assets quality index	((1-ACT_t+$PPENT_t$)/ AT_t)/ ((1-ACT_{t-1}+$PPENT_{t-1}$)/ AT_{t-1})	Beneish, 1999
BIGFOUR	Big four audit firm	Equals 1 if the financial statement for year t is audited by Big 4 audit firms (AU=1,2,3, or 4), and 0 otherwise	Huang et al., 2012
FILESIZE	MD&A file size	The number of words of the MD&A document	Churyk et al., 2009

Appendix B: Variable Definitions for Chapter 4

Variable	**Definition**
Lnauditfee	natural log of audit fees
Tweets	count of all tweets mentioned the client company
Negative	(Count of negative tweets– count of positive tweets)/ Tweets
Retweets	maximum number of retweets for each tweet mentioned the client company
Roaearnings	OIADP/AT
Size	natural log of AT
Invrec	(INVT+RECT)/AT
Leverage	(LT-LCT)/AT
Currentratio	ACT/LCT
BTM	(AT-LT)/ PRCC_F × CSHO
Growth	$(Sale_t - Sale_{t-1})/Sale_{t-1}$
Loss	if NI<0, then 1, else 0
Segments	the number of business segments
Foreign	if the client firm has foreign operations (TXFO), then 1, else 0
Merger	if the client firm reports the item related to acquisition and merger (AQP), then 1, else 0
Special	if the client firm reports special items (SPI), then 1, else 0
Firstyear	if initial year of audit, then 1, else 0
Big4	if Big 4 auditor, then 1, else 0
IC	if the current auditor indicates internal control weakness, then 1, else 0
GC	if the current auditor issues a going-concern opinion, then 1, else 0
Restatement	if the annual report is restated, then 1, else 0
Totalaccural	$[(AT - CH - LT - PSTK)_t - (AT - CH - LT - PSTK)_{t-1}] / (AT_t + AT_{t-1})/2$
ΔRec	$RECT_t/AT_t$ - $RECT_{t-1}/AT_{t-1}$
ΔInv	$INVT_t/AT_t$ - $INVT_{t-1}/AT_{t-1}$
Softassets	(AT - PPENT - CHE)/AT
$\Delta Csale$	$(CSALE_t - CSALE_{t-1})/CSALE_{t-1}$ where CSALE = SALE - ΔREC
ΔRoa	(IB_t/AT_t)- (IB_{t-1}/AT_{t-1})
Issuance	if DLTIS > 0 or SSTK>0, then 1, else 0
ΔEmp	$\left[\frac{EMP_t - EMP_{t-1}}{EMP_{t-1}}\right] - \left[\frac{AT_t - AT_{t-1}}{AT_{t-1}}\right]$
Lease	if MRCT > 0, then 1, else 0

Abret	the difference between annual buy-and-hold stock return and annual buy-and-hold value weighted index return
Lagabret	ABRET lagged by 1 year
Pscore	predicted probability of restatement using equation (2)