Master's Thesis

Predicting Fraudulent Financial Statement using Textual Analysis and Machine-Learning Techniques

by

DIMAS Lagusto

52116610

September 2018

Master's Thesis Presented to

Ritsumeikan Asia Pacific University

In Partial Fulfillment of the Requirements for the Degree of

Master of Business Administration

Table of Contents

Certification Page

I, DIMAS Lagusto (Student ID 52116610) hereby declare that the contents of this Master's Thesis are original and true, and have not been submitted at any other university or educational institution for the award of degree or diploma.

All the information derived from other published or unpublished sources has been cited and acknowledged appropriately.

DIMAS Lagusto

2018 / 05 / 28

Acknowledgements

I would like to extend my gratitude for my seminar supervisor Prof. William B Claster for his continuous support and supervision throughout the period of my research in APU as well as during the development of this Master's thesis. I also would like to express my gratitude to my dear wife Nisa for her constant support during the development of this thesis.

Summary

Textual analysis and machine-learning algorithm can be applied to predict the existence of fraudulent financial statement. An automated prediction model that is implemented in this master's thesis managed to predict, with a relatively high probability, the existence of fraudulent financial statement. Measurement of the implemented prediction model showed that it achieved accuracy and sensitivity ratios of 79% and 93% respectively.

Detection of fraudulent financial statement is important for investors, regulators and auditing firms. Ever since the disclosure of a series of financial statement fraud in the late 90s and early 2000s, all relevant stakeholders have felt the adverse impact of the fraud, both socially and financially. Early detection of fraud may mitigate the adverse impact for these stakeholders. However with limited resource to conduct a manual detection of financial statement fraud, a more resource effective automated method is required. Utilizing textual analysis and machine learning algorithm an effective automated method of detection can be obtained.

Using financial statement in the form of annual 10-K and quarterly based 10-Q reports, this thesis implemented two textual analysis methods, i.e., distributed representative and bag of words methods, and a hybrid combination of both methods to extract textual features. Utilizing the extracted text features, the applied machine-learning training method created a classification model to predict fraudulent financial statement. The model that utilized the hybrid method managed to achieve the

aforementioned ratios.

1. Introduction

This master's thesis explores the possibility of implementing machine-learning technique to predict fraudulent financial fraud statement by analyzing the text content of publically available financial statements, i.e., 10-K annual report and 10-Q quarterly report, which are published by publically listed companies in the United States. Utilizing machine-learning techniques, this thesis found that prediction of fraudulent financial statement using textual analysis and machine-learning technique could be achieved with 79% accuracy and 93% sensitivity.

Implementing textual analysis and machine-learning technique to predict fraudulent financial statement has the potential to improve the effectiveness of the prediction process. Especially, when comparing machine learning technique with the traditional method of prediction that relies on the manual auditing process. This improvement should be an interesting subject for investors, regulators and auditing firms.

In the United States, all company with publically traded securities has the obligation to submit periodic reports to the US Securities and Exchange Commission (SEC) (The Securities Exchange Act of 1934, 2018). Two examples of the periodic reports that publically listed companies have to submit are the annual Form 10-K and the quarterly based Form 10-Q. The two reports contain disclosure of information regarding the nature of the business, current operational and financial condition of the company as well as any possible risk and legal proceedings faced by the company (US Securities and Exchange Commision, 2009). All of these submitted report can be accessed by the general public using the SEC's Electronic Data Gathering, Analysis and

Retrieval (EDGAR) system (US Securities and Exchange Commision, 2017).

The wealth of information that companies disclosed in their 10-K and 10-Q reports is an information source which investors could utilize to make informed decisions on their investment portfolio (You & Zhang, 2009). Encouraging operational and financial results may encourage investor to maintain or increase their investment in the company's securities. On the other hand, possible lengthy legal proceedings or increased market risk that the company faced may encourage investor to divest their investment or discourage them to invest in the company's securities. This made the accuracy of the information disclosed in both reports essential for investors in making accurate decisions about their investment.

However, the revelation of financial scandals that involved large multinational companies such as Enron, WorldCom and Tyco in the late 1990s and early 2000s (Hogan, Rezaee, Riley, Jr., & Velury, 2008) involving the disclosure of fraudulent information within their 10-K reports that were submitted to SEC caused damaging impact for both the investors and the general public. In addition to that the scandal also brought up questions about the accuracy of the Form 10-K submitted to the SEC (Goel & Gangolly, 2012).

The scandal and their subsequent effect increased the awareness on the importance of predicting the existence of fraudulent financial statement (Hogan, Rezaee, Riley, Jr., & Velury, 2008). In response to these scandals, SEC brought the companies that were suspected to publish fraudulent financial statement to court. All information about the litigation process, that includes both the individuals and the companies involved and the type of violation, is available to the public through the SEC website (US Securities and Exchange Commision, 2018). In addition to that, the US government

introduced a new regulation called the Sarbanes-Oxley Act in July 2002 to improve corporate governance (Hogan, Rezaee, Riley, Jr., & Velury, 2008).

In light of the existence of fraudulent financial statement, the biggest challenge facing the relevant stakeholder - such as regulators, investors as well as auditing firms - is how to effectively predict fraudulent financial statement that were published by the companies and subsequently submitted to the SEC. Effective prediction will bring many benefit to all the stakeholders. The benefits include improved focused in investigation for regulators, improved investment decision making for investors and improved decision-making in client engagements and during regular audits for auditing firms (Abbasi, Albrecht, Vance, & Hansen, 2012), (Albrecht, Albrecht, & Albrecht, 2008).

The traditional method to predict the existence fraudulent statement is by conducting a manual auditing process. This involves the presence of auditors to conduct an audit process on a company and their financial statement. However this method of prediction is neither efficient nor reliable (West & Bhattacharya, 2016). One reason for the inefficiency and unreliability is the volume and complexity of both the 10-K and 10-Q reports published by the companies in comparison with the amount of auditor that could be employed to conduct a comprehensive audit on the company and their reports. This problem of auditing resource deficiency is exacerbated with the diverse and complex business nature of many companies. This factor has increased the difficulties of the auditing activities and reduced their efficiency and reliability in predicting fraudulent financial statement. In summary, utilizing the auditing process to predict fraudulent financial statement is costly, time consuming and inaccurate (West & Bhattacharya, 2016). Therefore, an alternative method for prediction that is more efficient and reliable is required.

An alternative method for prediction is to have an automated process in place. The automated process will utilize computer algorithm to conduct quantitative analysis on the available data in an attempt to predict fraudulent financial statement (Hajek & Henriques, 2017). One indication of the adoption of the automated process is the inception of the Center for Risk and Quantitative Analytics by the SEC. This center was established to identify risk and threats for investors by utilizing data analytics (US Securities and Exchange Commision, 2013).

To explore the possibility of using machine learning technique for fraudulent financial statement prediction by analyzing the textual part of the statement, this master's thesis implemented a model that is depicted in Figure 1. The model utilized an automatic analysis and prediction process that is enabled by prevalent textual analysis and machine learning technique.

In general, the model could be divided into seven related processes. The first process is the process of identifying and obtaining both fraudulent and non-fraudulent 10-K and 10-Q reports. To identify fraudulent reports, this thesis analyzes SEC litigation releases to find litigation brought by the SEC involving fraudulent financial statements (US Securities and Exchange Commision, 2018). From the litigation releases, this thesis can identify the name of the companies that published the 10-K and/or 10-Q reports that were considered by the SEC to be fraudulent. In addition to that, based on the same litigation release, this thesis could also identify the specific publication period, i.e., the year or the quarter, of the fraudulent report.

The second process involves the manual extraction of textual information, i.e., texts, from a particular part of the report. The text extraction is implemented to exclusively obtain the textual portion of the report and remove the non-textual portion

such as tables, graph and document headers and footers. In this thesis, textual information is extracted from the Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) segment within the 10-K and 10-Q report. MD&A is a segment of the report where a company provides disclosures on the overall performance of the company that includes both financial and operational performance (US Securities and Exchange Commision, 2009). This information is important because it provides the reader with insights on not only the financial and operational accomplishments of the company but also on both internal and external circumstances that influenced the current and/or may influence future performance of the company. In essence, it also provides the reader with the future outlook of the company. Studies on textual analysis by (Cecchini M. , Aytug, Koehler, & Pathak, 2010) and (Humpherys , Moffitt, Burns, Burgoon, & Felix, 2011) also focused on the MD&A segment and the result suggested the value of conducting textual analysis on the MD&A segment.

The third process is the creation of two sets of text that will be used as the training and test sample by the prediction algorithm. Each set of text consists of both fraudulent and non-fraudulent MD&A text. In general both sample contains the same data. The basic difference between the two samples is that the training sample has a specific indicator or flag that indicates whether the particular text is fraudulent or not. While, on the other hand the test sample doesn't have that indicator. The training sample will be utilized to develop the classification model that will predict the existence of fraudulent reports. On the other hand, the test sample will be utilized afterwards at the end of the entire prediction process to measure the performance of the classification model in predicting fraudulent reports.

The fourth process includes the implementation of text pre-processing techniques to obtain useable text for the machine-learning techniques. The machine-learning technique then utilizes the text to develop a classification model. The text pre-processing steps used in this master thesis include lemmatization, tokenization and the removal of stop words. A research by (Manning, Raghavan, & Schütze, 2008) discussed the definition and significance of the pre-processing steps. The lemmatization process is the process of obtaining the dictionary form of words. The tokenization process will break down a text or sentence into individual words. The last pre-processing step, i.e., removal of stop words involves the removal of common words within sentences.

The fifth process in the overall prediction process is the extraction of features from the pre-processed text. The feature extraction process, that utilizes machine-learning techniques, is implemented to obtain unique characteristics or property of a text (Bishop, 2006). The unique feature of the text is essential to train the classification model used by the prediction process. To extract the features from the pre-processed text, this thesis will utilize textual analysis methods.

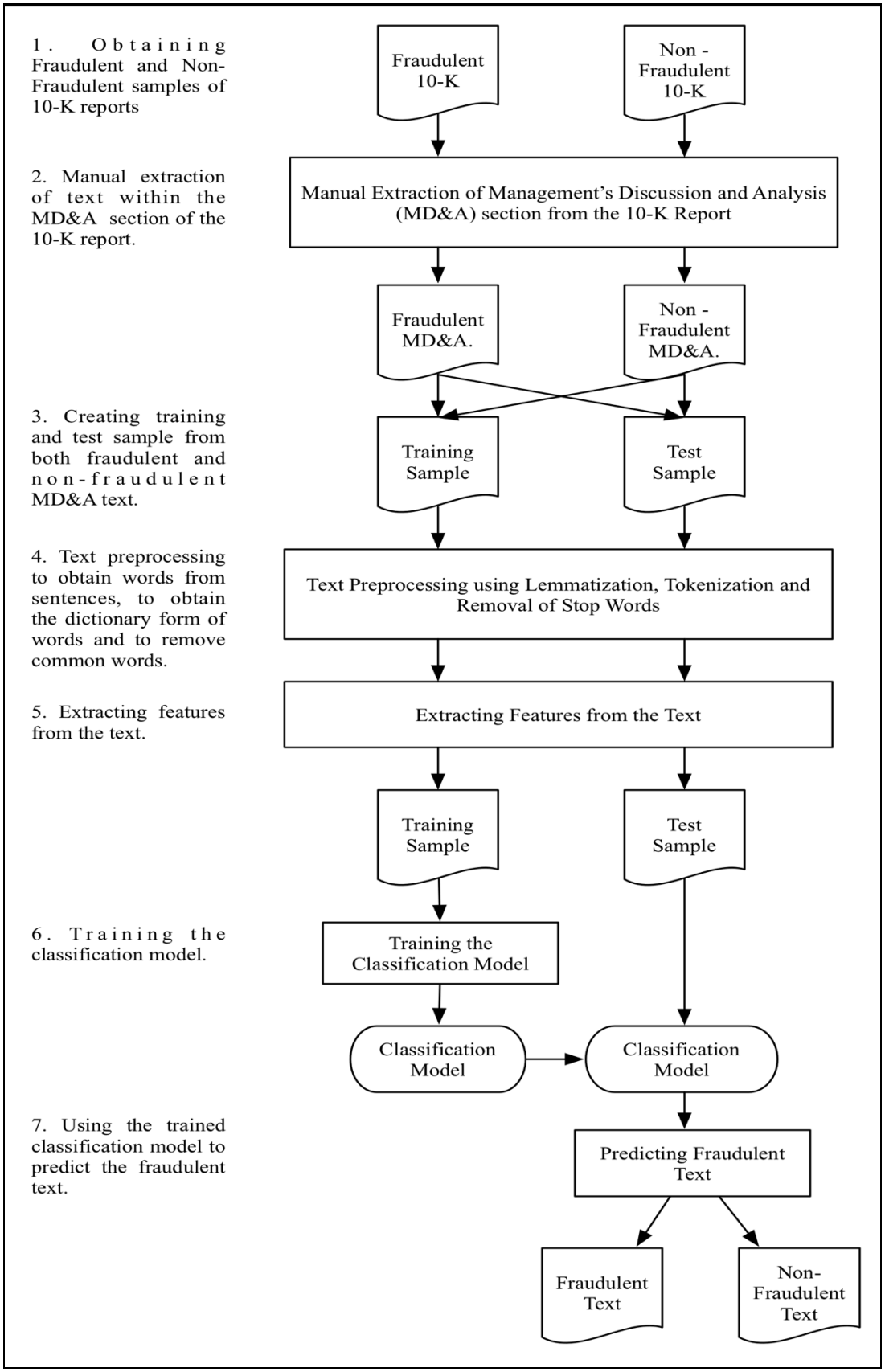
The sixth process is the training of the classification model using machine-learning techniques. The term training the model is the process of creating a mathematical model using the available training data (James, Witten, Hastie, & Tibshirani, 2017). Training of the model utilizes the features of the text within the training sample, obtained in the previous process, to develop the classification model. The trained classification model will be later used to predict fraudulent 10-K and 10-Q report. Throughout his thesis, the term classification model will be used to represent the prediction model. The reason is that the intended prediction process can also be considered as a classification process that classifies the reports into fraudulent and non-

fraudulent reports.

The seventh and last process is the process where the trained classification model predicts whether a particular text in the test sample is fraudulent or not. To measure the overall prediction process, this thesis will observe the accuracy of the classification model in their prediction.

In this chapter, the main argument of this master's thesis is outlined. In addition to that, this chapter also provides a brief description of the background and significance of research as well as the general framework used throughout the research.

Figure 1 General Process Flow of the Fraudulent Financial Statement Prediction Process



Two limitations of this master's thesis are as follows. First the text extraction process is implemented manually. The manual process is implemented due to the high complexity of implementing an automated text extraction process. The effect of the manual text extraction process is that only a relatively small sample of financial statement text could be extracted for further training and testing by the machine-learning algorithm.

The second limitation is that due to a relatively small sample of fraudulent and non-fraudulent financial statement text, the amount of features extracted by the textual analysis methods would also be limited. The limited features, from which the machine-learning training process utilized to create the classification model, may result in a less accurate classification model.

2. Literature Reviews

In the previous chapter, this master's thesis introduced the research that will be conducted within this thesis. The introduction includes the main argument of the thesis, brief description over the background and significance of the research. In addition to that the general framework of the research was also mentioned.

In this chapter, reviews of the literature relevant to the research are outlined. The reviewed literatures include literatures related to five main concepts. The first concept is related to financial statement as well as the 10-K and 10-Q report. The second concept is fraudulent financial statement. The third concept is the prediction of fraudulent financial statement. The fourth concept is textual analysis and its application to predict financial statement fraud. The fifth and last concept is the application of machine learning for fraudulent financial statement prediction.

2.1 Financial Statement and 10-K Report

The International Accounting Standard 1 (IAS 1) defined financial statement as a document that discloses an entity's financial position and performance within a certain time interval (International Financial Reporting Standards Foundation, 2007). According to the same standard, the financial statement was intended to provide their reader with the ability to make informed economic related decision using information about the financial and cash flow condition about the entity. Looking at the definition, it could be understood that the focus of the financial statement is more on financial information such as position of assets, liabilities and equities, revenue and expense, as

well as cash flow of the company within a certain period.

However the standard also requires that the financial statements include information on significant accounting policies and other relevant information that could provide the user of the statement with information about the overall financial position of the entity. The other relevant information previously mentioned, may include reviews by the management of the entity over the influencing factors or uncertainties faced, both internal or external, that have contributed or may contribute to the entity's financial performance as well as the plan and policy that the management takes as response.

In the United States, the SEC as the regulatory authority in the securities industry requires publically listed companies to submit periodic financial statement. Two examples of the periodic statements that have to be submitted in a standardized format are the annual 10-K Form (US Securities and Exchange Commision, 2017) and the quarterly 10-Q report (US Securities and Exchange Commision, 2011). In addition to financial information, both forms requires that companies, that are required to submit the two forms, to disclose a wide-range of information that are relevant to their business. The information includes information over the nature of the business; the risk faced by the business due to its operations; a review and analysis of the company's business and financial performance; accounting and financial policies and procedures in place as well as material relationships and transactions of the business (US Securities and Exchange Commision, 2009).

The development of a standardized form has an added benefit in improving the comparability of the financial statement. The improved comparability that is achieved through the standardization has been shown to reduce the cost of obtaining relevant information as well as to increase both the quantity and quality of the information that

can be used for further analysis (De Franco, Verdi, & Verdi, 2011). The benefit of standardization is the first reason for the selection of both the 10-K and 10-Q Forms as samples of financial statements that will be used for this research.

The second reason for using both forms to represent the financial statement that will be analyzed throughout this research is the ease of accessing the forms. All forms submitted to the SEC is considered public information and could be easily accessed using the SEC's Electronic Data Gathering, Analysis and Retrieval (EDGAR) system (US Securities and Exchange Commision, 2017).

The third reason for using the two forms is their ability to provide coherent, comprehensive and complete business and financial information about a company. These two forms, in most cases, contain notable information about both a company's financial position as well as its overall business performance that may not be available in other documents or press released published by the company (Griffin, 2003).

This section of the thesis has discussed the definition of financial statement as well as the 10-K and 10-Q forms as the corporate financial statement in the US. In addition to that, this section also discussed the reason for selecting both forms as the financial statement that is used and analyzed throughout the research.

2.2 Fraudulent Financial Statement

After discussing the definition of financial statement and the statements that are used in this research. This section will discuss the definition, motivation and effect of fraudulent financial statement.

A concise definition of financial statement fraud that is mentioned by (Arens, Elder, & Beasley, 2010) involves the provision of wrong or inaccurate information

within the statement. While (Rezaee, 2005) argues that financial statement fraud occurred when a company acted intentionally to deceive or mislead the readers of its financial statement through the inclusion of material misstatement in their published statements. The fraud may occur through the following four actions. The first action involves the falsification, alteration or manipulation of material records and information related to financial transactions that a company undertake. The second action includes the deliberate omission or misrepresentation of material records and information in the published financial statement. The third action happens when the company intentionally applied incorrect accounting principles, policies and procedures or applied the correct accounting principles, policies and procedures for the wrong purposes on material records and information published in the financial statement. The fourth action happens when certain accounting principles and policies are deliberately omitted or inadequately omitted from the report (Rezaee, 2002).

The motivation of committing fraud includes personal economic interest from executives in the company (Dechow, Ge, Larson, & Sloan, 2011), market pressure for continuous growth or improved revenue (Dechow, Ge, Larson, & Sloan, 2011) and lack of ethical values from the company executives (Soltani, 2014).

To understand the effect of financial statement fraud, this thesis looked into the staggering financial loss that materialized due to financial statement fraud. The research conducted by (Rezaee, 2002) suggested that the cost reach billions of US Dollars. The financial statement fraud that was committed by Enron, was estimated to cost investors and employee that possess Enron's shares as much as $70 billion in market capitalization (Rezaee, 2005).

2.3 Prediction of Fraudulent Financial Statement

The previous section discussed the definition, motivation and effect of financial statement fraud. In this section the different prediction methods available will be discussed.

Predicting the existence of fraudulent financial statement can be achieved manually through an audit process that is conducted either internally or by external independent parties (Rezaee, 2005). This manual auditing process normally relies on auditing internal corporate data (Abbasi, Albrecht, Vance, & Hansen, 2012) and the analysis of financial metrics such as liquidity and profitability to identify deviation based on historical financial data of the company or based on comparison with other company within the industry (Robinson, van Greuning, Henry, & Broihahn, 2008). Another method to assist the manual auditing process is the use of financial and process red flags that acts as indicator to help auditor to find areas to focus their audit (Kassem & Hegazy, 2010).

However the effectiveness and reliability of this audit process is not very high (West & Bhattacharya, 2016). The difficulties in predicting the occurrence of fraudulent financial statement could be explained by the insufficient level of knowledge on the subject, the small occurrence of the fraud and the high level of complexity due to the high level of knowledge of the perpetrator (Maes, Tuyls, Vanschoenwinkel, & Manderick, 2002). Previous research showed that the audit process that is conducted by SEC only contributed to only around 6 percent of all frauds that were correctly predicted (Dyck, Morse, & Zingales, 2010). While the audit process that were conducted by auditors contributed to around 11 % to 14 % of all correctly predicted fraud (Aliabadi, Dorestani, & Qadri, 2011), (Dyck, Morse, & Zingales, 2010).

In addition to that, since the manual auditing process is a process that can only be achieved by either an auditing firm with a contractual agreement with the company to carry out an audit or by auditors with the legal mandate to conduct an audit, e.g., auditors that works for or on behalf of regulatory bodies, many internal corporate data is inaccessible to investors and the general public (Cecchini M. , Aytug, Koehler, & Pathak, 2010). This means that investors and the general public has to rely on external corporate data that is available for the general public

This reality made the development of new methods to predict fraudulent financial statement a necessity for the improvement of efficiency and reliability of the prediction process. This improvement is important, especially for investors and the general public that can only depend on publically available corporate data, such as the corporate financial statement.

With the development of computer technology that improved the performance and accessibility of powerful computing power and algorithm, new automated methods that utilized computer technology has emerged. The new methods that enabled computers to analyze of large quantity of data and subsequently solve problems without having prior knowledge of the problems have shown to have the potential to improve fraudulent financial statement prediction process (Ravishankar, Ravi, Rao, & Bose, 2011), (Li, Yu, Zhang, & Ke, 2015). Several researchers defined this new method as machine-learning method (Li, Yu, Zhang, & Ke, 2015), (Goel, Gangolly, Faerman, & Uzuner, 2010), (Kim, Baik, & Cho, 2016), (Hajek & Henriques, 2017) and (Cecchini M. , Aytug, Koehler, & Pathak, 2010).

In general, fraudulent financial statement prediction using the machine learning method involves the analysis of data, collected from financial statement, using machine-

learning techniques to classify the financial statement into two categories namely fraudulent and non-fraudulent (Abbasi, Albrecht, Vance, & Hansen, 2012).

The analyzed data that represent the input of the machine-learning techniques can be classified into two types of data, i.e., financial variables and text within the financial statement (Hajek & Henriques, 2017). The first type of data is financial variables data that can indicate financial irregularities and subsequently financial statement fraud. The irregularities that can be identified include assets, liabilities, revenue and/or expense under or overstatement (Hajek & Henriques, 2017).

A research done by (Abbasi, Albrecht, Vance, & Hansen, 2012) provided the theoretical proof for the application of financial variables in predicting the existence of financial statement fraud. In general, the reason for using financial variable is based on the use of these variables to measure the financial performance of a company (Beaver, 1966). Hence unsatisfactory financial performance may encourage the management to commit fraudulent activities (Fanning & Cogger, 1998). In particular, if the performance of the company and their stock price is tied with the management compensation (Dechow, Ge, Larson, & Sloan, 2011).

Examples of the financial variables that were used includes asset turnover and leverage ratio that is used to by (Cecchini M. , Aytug, Koehler, & Pathak, 2010) and (Kirkos, Spathis, & Manolopoulos, 2007). The reasoning behind the use of asset turnover is that this variable may be indicative of revenue fraud. Revenue fraud is usually achieved by artificially increasing the revenue or net-sales obtained by a company. One indication of the artificial increment is when the rate of increment is abnormally high, i.e., the increase in net sales happens rapidly. In addition to asset turnover, leverage ratio is also used to identify whether a company is artificially

increasing their asset on the balance sheet. This could be identified, particularly when the increased of asset happened without any corresponding increase in their liability (Cecchini M. , Aytug, Koehler, & Pathak, 2010), (Kirkos, Spathis, & Manolopoulos, 2007).

Another example of financial variable that is used is the days in sales receivable that is used by (Kaminski, Wetzel, & Guan, 2004) and (Lin, Hwang, & Becker, 2003) to identify company committing revenue fraud scheme using fictitious receivables and the selling and general administrative expenses ratio used by (Dikmen & Küçükkocaoğlu, 2009) and (Cecchini M. , Aytug, Koehler, & Pathak, 2010) to identify firms committing revenue fraud by artificially reducing the selling and general administrative expenses.

In summary, the manipulation of these variables showed the company's efforts to artificially increase their revenue and asset as well as artificially reducing their liabilities and expenses to artificially make their financial performance seemed better than the actual one.

The second data type used as input for the machine-learning techniques is textual data that is contained within the financial statement. The argument behind the use of textual data is that the words used in the financial statement to describe the performance of the company have been shown to have correlation with fraudulent activities (Loughran & McDonald, 2016). Also (Goel, Gangolly, Faerman, & Uzuner, 2010) argues that the textual data contained in the financial statement contained indicators that may be used to identify fraudulent financial statement. The indicators may be identified by of placements of certain phrases and in the selective use of certain adjectives, sentence construction and adverbs (Goel, Gangolly, Faerman, & Uzuner, 2010). Another indicator may include the use of negative and uncertain words (Throckmorton,

Mayew, Venkatachalam, & Collins, 2015) that were based on the theory of non-verbal behavior of an individual when the commit deception (Vrij, 2008).

In this section, this thesis discussed the methods for predicting fraudulent financial statement that include manual audit process and the utilization of an automated machine-learning technique. The data used as input by the machine-learning techniques were also outlined. Next section, this thesis will discuss the available literature on textual analysis and its application in financial statement fraud prediction.

2.4 Textual Analysis and its Application

After discussing the methods for predicting fraudulent financial statement in the previous section. In this section, this thesis will first discuss textual analysis, which is a method to analyze the text portion of the financial statement. Afterwards this section will discuss the application of textual analysis to predict fraudulent financial statement. In this thesis, the textual analysis process is implemented in order to extract features of the text. Using the extracted features, the subsequent machine learning based prediction process creates a classification model that will predict the existence of fraudulent financial statement text.

Textual analysis is the study of analyzing text to identify patterns and then identify the meaning and sentiment that may be explicitly or implicitly as well as intentionally or unintentionally mentioned in the text (Loughran & McDonald, 2016). This meaning and sentiment will be then considered as the feature of the text that could be then used for further analysis, e.g., by the machine-learning technique. The study of textual analysis has a long history. In the early 1900s textual analysis was implemented in order to identify the actual writer of a literary work (Williams, 1975). During the

First World War, textual analysis was utilized in the analysis of political speeches in order to identify diplomatic signals that are contained within the carefully written speech (Burke, 1939).

Within the accounting and finance discipline, the application of textual analysis can be considered as an emerging field (Loughran & McDonald, 2016). The emergence of textual analysis is promoted by increased accessibility to large amount to electronic financial data that is in turn promoted by Internet technology and SEC requirement for public company to submit periodic financial statement (Cecchini M. , Aytug, Koehler, & Pathak, 2010).

Textual analysis was used in researches related stock market activity. In which pessimistic news in the media induces lower prices in the stock market and unusual level of pessimistic news, either high or low, will increase the trading in the stock market (Tetlock, 2007). In a related research, (Das & Chen, 2007) investigated the application of textual analysis to obtain investor's sentiment using data from stock message boards. The research concluded that market activity is related to the sentiment of the investors expressed in their comment in the message boards.

Another application of textual analysis was outlined by (Davis, Piger, & Sedor, 2012), in which textual analysis was applied to analyze corporate press releases for the purpose of predicting the future performance of a company. The argument behind the research was that texts contained in the press release contained subtle communications by the company's managers about the expected performance of the company. A different application of textual analysis is conducted by (Li F. , 2006), in which textual analysis on 10-K reports is used to predict future company's earnings. The analysis utilized the number of occurrence of particular words that are related to uncertainty or

risk within a company's 10-K reports. The research found a correlation between the frequency of risk related words within a company's 10-K reports and their future revenue. Companies with an increasing number of words that are related to uncertainty and risk within their year-over-year 10-K reports have a tendency to have decreasing revenue in their next year financial results.

One application of textual analysis that is relevant with this thesis is the application of textual analysis to predict fraudulent financial statement. Textual analysis has been a useful method to predict fraudulent financial statement due to the availability in large amount of publically accessible financial text (West & Bhattacharya, 2016). A research conducted by (Humpherys , Moffitt, Burns, Burgoon, & Felix, 2011) analyzed the Management's Discussion and Analysis (MD&A) segment within corporate 10-K reports to predict fraudulent financial statement. Their research, which resulted in a 67.3% accuracy rate, showed the potential of applying textual analysis to predict fraudulent financial statement. Another research conducted by (Cecchini M. , Aytug, Koehler, & Pathak, 2010) and (Glanchy & Yadav, 2011) showed a promising result. By utilizing text within the MD&A segment, their developed prediction models managed to achieve 82% and 84% accuracy respectively. One of the most promising results came from the research on fraudulent financial statement analysis using textual analysis was the model developed by (Goel, Gangolly, Faerman, & Uzuner, 2010) that achieved 89.51% average accuracy.

Analyzing the detail of textual analysis, this thesis found that the main component of the analysis is the representation of words that are contained within a particular text. This word representation can be then used to analyze the meaning or sentiment of the text.

After analyzing existing literature, this thesis identified two word representation methods. The first identified method is a method that relies on word counts to represent words within a text. The basic assumption of this method is that the order in which words occurred within a text is less important compared with the number of occurrence of the word itself (Manning, Raghavan, & Schütze, 2008). Therefore in this method the text "Alice likes Bob" and "Bob likes Alice" have the same representation due to the same number of word count between the two texts. This method is known as the bag of words method. This method has several implementation techniques to enhance its performance.

One of the simplest but quite powerful techniques is the targeted phrases technique (Loughran & McDonald, 2016). In this method, a calculation to find the frequency of occurrence of certain words or phrases - that is specifically selected in order to identify a certain event – within a text is used to represent the text. A research by (Loughran, McDonald, & Yun, 2009) specifically identifies the occurrence of the word "ethics" in combination with the occurrence of its variations, i.e., words that have similar meaning with "ethics", within corporate 10-K reports. In addition to that, their research also identified three phrases related to corporate social responsibility in their analysis of the report. The analysis was done in order to determine whether a company is more likely to become socially irresponsible. Their research finds that a strong correlation occurred between the frequency of the targeted words and phrases within a company's 10-K report and the likelihood of socially irresponsible activities by the company. In addition to that a strong correlation also appeared between the frequency of the targeted words and phrases with the probability of the company to be litigated in a class action lawsuit or to have poor corporate governance policies.

Another technique that improves the performance of the targeted phrases technique is the word list of dictionary technique (Loughran & McDonald, 2016). This technique utilized a compilation of words that shares a common meaning or sentiment, e.g., positive, negative, ethical or risky. Using the list, analysis can be done on a text by calculating the frequency of occurrence of words within the list. A research by (Li F. , 2006) showed a correlation between the occurrences of words that shared the same meaning, i.e., risk or uncertainty, in their 10-K reports, and the future revenue. The emergence of this technique prompted the creation dictionaries that were specifically compiled for analyzing of financial text. One of the first dictionaries compiled is the Henry dictionary (Henry, 2008). The Henry dictionary consisted of words with positive and negative sentiment that were compiled from annual press releases from companies that belong to the computer and telecommunications industry. A research by (Price, Doran, Peterson, & Bliss, 2012) also utilized the Henry dictionary to show correlation between the discussion within the question and answer (Q&A) section of corporate quarterly earnings conference calls and the stock prices of those companies.

Considering the favorable outcome that came with the utilization of the bag of words method, weaknesses can still be identified (Le & Mikolov, 2014). The first weakness is that the method didn't consider the ordering of words within the text. This could make two texts with two different meaning to have the same representation provided that these two texts contained the same set of words. This means the texts "Alice likes Bob" and "Bob likes Alice" would have the same representation although the actual meaning of the two text is different.

The second weakness is the inability of the bag of words method to capture the meanings of the words. The example is the following words "strong", "powerful" and

“Paris” would be considered by the bag of words method to have the same representation. Even though the meaning of the word “strong” and “powerful” is closer to each other than “Paris” (Le & Mikolov, 2014).

After discussing the bag of words method that utilizes word count to represent words meanings within a text. This thesis will discuss an alternative word representation method that addressed the weaknesses of the bag of words method.

The argument behind the alternative method is that the meaning of a particular word could be attributed to their co-occurrence with other word within a text as well as the order of the words within the same text (Church & Hanks, 1990).

An example for this concept is the word “boy” within the text “a boy is eating the banana” and the word “man” within the text “the man was eating a banana” will have meanings that are similar to each other, i.e., semantically closer compared with the word ”banana”. This is because both words have the same order within the respective text and both words co-occurred with the word “eating” and banana”.

With that in mind, the representation of the meaning of a word doesn’t depend on the frequency of occurrence of the word itself but also on its co-occurrence with other words within a text as well as the order of words within that same text. These dependencies established the context for that particular word. In essence, this representation method doesn’t merely represent the word but also able to represent the context of the word (Bengio, Ducharme, Vincent, & Jauvin, 2003). The term coined for the concept is called distributed representations (Hinton, 1986).

Analyzing the existing literature on textual analysis application for predicting financial fraud statement using machine learning method, this thesis found that most of the researches on this particular area, utilized the bag of words method that uses word

count to represent the meaning of words within a text (Hajek & Henriques, 2017). One example of a recent research that utilized the bag of word technique is the research done by (Goel & Uzuner, 2016) that counts the frequency of words with subjective, positive, negative and neutral sentiment for the analysis of the MD&A segment of corporate 10-K Forms. Their developed model managed to obtain accuracy rate between 63.6% and 81.8%.

Based on this lack of research on the application of the distributed representations method, this thesis will explore the use of the distributed representation method as the textual analysis method for textual feature extraction.

This thesis just reviewed the available literature on textual analysis and their application. In the next section this thesis will discuss machine-learning technique and their application for fraudulent fraud prediction.

2.5 Machine Learning Technique

This thesis has mentioned the term machine learning multiple times without explicitly describing what the term means. In this section, this thesis will discuss the term machine learning and its application to predict fraudulent financial statement.

In his publication (Samuel, 1959) describe machine learning as a process in which a machine, i.e., a computer, is programmed to learn a particular subject so that it would be able to solve a particular problem within the subject without any addition external assistance. The research done by (Samuel, 1959) investigated the possibility of a computer to play the game of checkers better than the person who programmed the computer to learn how to play the game. He found that it is possible for a computer to play checkers after it learned the rules of the game, the directional flow of the game and

some parameters related to the game.

In (Kohavi & Provost, 1998), machine learning is defined as a scientific field that study the application of induction algorithm or other algorithm that is used to learn.

Based on the two term, this thesis define the term machine learning as a study and application of algorithms that enable a computer systems to learn a particular subject and subsequently applies the knowledge learned to solve a problem related to the subject that was previously learned.

Machine learning is particularly useful when programming a computer to do a particular task cannot be implemented, either due to lack of human expertise or inability to access the knowledge that a human has. In this scenario, the ability to program a computer to learn a particular task and the letting the computer to learn to required knowledge to do the task is beneficial (Alpaydin, 2010).

In order for a computer to learn a particular subject, the computer required relevant data. The computer will then use the data to learn about the subject and subsequently develop a model that would be a representation about the knowledge on the particular subject (Alpaydin, 2010).

One famous example is the development of an automatic spam-filtering program for our emails. The knowledge and algorithm that can predict spam from legitimate email is something that we don't have. However, we have large number of samples of both spam and legitimate email for that purpose. The solution can be obtained by using a learning algorithm. From which the computer will first extract the features or characteristics of both spam and legitimate email and then afterwards develop a model that could classify or separate the two.

An extensive research conducted by (Hajek & Henriques, 2017) compared the use

of several machine learning technique for fraudulent financial prediction process. Their research used both financial variables and text data extracted from 10-K reports as input for the machine learning techniques. The machine learning techniques used include logistic regression, Bayesian classifier, decision trees, support vector machine, neural networks and ensemble methods. The result of their research showed the possibility of using machine learning technique to predict fraudulent financial statement by analyzing text data (Hajek & Henriques, 2017).

Using prediction accuracy as one of the measurement method, the results of (Hajek & Henriques, 2017) machine learning application on text data are as follows. Using logistic regression model, i.e., a regression method for analyzing binary valued dependent variables (Hosmer & Lemeshow, 2000), the accuracy is around 62%. The Bayesian classifier models produce an accuracy of around 58%. The decision tree models have an accuracy of around 59%. The support vector machine produces an accuracy percentage of around 62%. The neural network models produce an accuracy of 55%. Finally the ensemble models have an accuracy level of around 61%. As an important note, the research done by (Hajek & Henriques, 2017) utilized bag of words method for word representation.

Analyzing the relatively low accuracy of the results that were produced by the research conducted by (Hajek & Henriques, 2017), this thesis searches several literatures to find a technique that may have the potential to produce a better prediction results.

Analyzing the research done by (Chen & Guestrin, 2016), this thesis found a method called XGBoost that was widely used as the winning solutions in several international machine-learning competitions such as Kaggle challenges and KDDCup

2015. Examples of the solutions that utilized XGBoost are sales prediction, text classification, motion detection and malware-classification (Chen & Guestrin, 2016). Looking into the promising potential of this machine-learning method, this thesis will utilize the XGBoost method as the machine-learning prediction method.

3. Research Design

This master thesis seeks the possibility of using machine learning technique to predict fraudulent financial statement by analyzing the text data within the financial statement. The availability of standardized publically accessible financial statement, such as that 10-K reports (US Securities and Exchange Commision, 2017), promotes the utilization of textual analysis to predict fraudulent statement. In addition to that, the application of machine learning method allows the application of an automated prediction method to increase the efficiency of the auditing process (West & Bhattacharya, 2016). Furthermore the application of an automatic analysis and prediction process allows for a more efficient allocation of auditing resource to conduct a detailed investigation process to find and collect evidence of fraudulent activities from suspected company.

In this chapter, this master thesis will discuss the design of the classification model that will be implemented to explore the possibility of using machine learning technique for fraudulent financial statement prediction through the analysis of textual data. This thesis uses the term classification model because the developed model will classify a particular report into either fraudulent or non-fraudulent. The general process flow of the classification model that will be implemented is depicted in Figure 1. The discussion of each process will be outlined in sections within this chapter of this master's thesis. This thesis will utilize the Python version 3.6 programming language to implement the prediction model.

3.1 Data Identification and Retrieval

The first process in the model designed for this thesis is the data identification and retrieval process. The data will be utilized as the input for the classification model. The data is utilized for two specific purposes. The first one is, as training data for the creation of the classification model by the machine-learning algorithm. The second purpose of the data is, as test data to measure the performance of the classification model in predicting fraudulent and non-fraudulent financial statement. This thesis will use both the 10-K and 10-Q forms, i.e., the required regulatory annual and quarterly report for companies that are listed in securities exchanges throughout the United States, as representation of financial statements data (US Securities and Exchange Commision, 2009). Both reports will include samples of both fraudulent and non-fraudulent reports. As it has been mentioned in sub-chapter 2.1 of this thesis, the reports are selected because of its accessibility and standardized form (US Securities and Exchange Commision, 2017).

To identify whether a particular report is fraudulent of not, this thesis analyzes the US Securities and Exchange (SEC)'s litigation release (US Securities and Exchange Commision, 2018). The litigation release is a public announcement that is released by the SEC concerning civil lawsuit against a particular party, e.g., a company, a person or a group of person, that the SEC brought to the US Federal Court (US Securities and Exchange Commision, 2018). The analysis needs to identify two important variables, namely the company name and the specific period when a fraudulent report was submitted.

This thesis analyzed litigation releases from 1995 until 2018 to find information about lawsuits brought by the SEC that are related to financial statement fraud. A general indicator that is used to find financial statement fraud related litigation process

is when the litigation release contained statements by the SEC indicating the occurrence of material misstatement in a period specific report, i.e., annual or quarter, submitted by a particular company. Another indicator is when the litigation release mentioned that a period specific report contained materially false and/or misleading information. These two indicators enable this thesis to identify a period specific fraudulent report submitted by a particular company.

On the other hand if the analysis found no litigation release mentioning the occurrence of fraudulent activities within a company's period specific report. Then the report from that particular company in that particular period is considered as non-fraudulent.

After both the fraudulent and non-fraudulent reports have been identified, this thesis proceeds to download the report individually from the following SEC Edgar's webpage https://www.sec.gov/edgar/searchedgar/companysearch.html. That particular webpage enabled the general public to search both the 10-K and 10-Q reports, in addition to other reports that are required by the SEC, using the company's name as the search term. The reports will be downloaded in the text format, i.e., TXT.

3.2 Extraction of Textual Data

After both sets of data, i.e., fraudulent and non-fraudulent reports have been downloaded. The next process is the manual extraction of texts from a segment of the 10-K and 10-Q reports that contained specifically interesting content for textual analysis. The segment of the report that will be manually extracted is the Management Discussion and Analysis of Financial Condition and Results of Operations (MD&A) segment. Within that segment, the corporate management discusses the financial and

operational conditions and results within one financial period. The segment also provided the discussion of major plans that the company may have in addition to the company's financial and business operations future outlook (US Securities and Exchange Commision, 2009). In addition to that, as it has been mentioned in chapter 1 and sub-chapter 2.4 of this thesis, many textual analysis research have been achieved by utilizing text that is contained within the MD&A segment.

The manual extraction is done to obtain the only the textual content of the MD&A segment. Hence all non-textual parts such as, header, footer, tables and graphs, that may be contained in the MD&A segment is not extracted. The manual extraction process is selected because it is easier to implement compared with an automatic extraction. Even though an automated extraction may produce a faster process, the development of the script to automatically extract a text according to the requirement for extraction, i.e., obtaining exclusively textual content of the MD&A segment, is much more difficult to achieve and therefore much more time consuming.

After the manual extraction process is achieved, the extracted text is then saved in a comma-separated value format, i.e., CSV. Each extracted text will be considered as one entry, i.e., one row, in the file. In addition to the extracted text, each entry in the CSV file will also include a unique identifier and a flag that indicates whether the entry came from a fraudulent or non-fraudulent report. The unique identifier utilized the unique identifier that SEC's Edgar system provided to each reports, while the flag used the number "1" to indicate a fraudulent text and the number "0" to indicate a non-fraudulent one. In this step, the model will have two CSV files. One file is for the fraudulent text entry and another one for the non-fraudulent text entry.

3.3 Creation of Training and Test Sample

After the manual extraction process is completed and the process has obtained two CSV files containing two types of extracted texts, i.e., fraudulent and non-fraudulent, the next process is the creation of training and test sample that will be used in latter parts of the model.

The training sample is used to create or, in machine learning term, to train the machine-learning model that will be used to predict fraudulent report. While the test sample will be utilized for the purpose of measuring the accuracy of the classification model in predicting the fraudulent report. One important aspect about the training data is that whenever possible, a balanced proportion between the samples in the training data, e.g., between fraudulent and non-fraudulent text samples, has to be maintained. The reason for maintaining a balanced proportion is that a classification model created using an unbalanced training data may become less accurate (Kotsiantis, Kanellopoulos, & Pintelas, 2006). The reason is that during the creation or learning process of the classification model, an imbalanced training data may cause the learning process to have difficulties in learning the feature of the minority set, i.e., the data set with less number of samples. Thus subsequently, the established classification model may have difficulties in identifying or classifying sample from the minority set (Batista, Prati, & Monard, 2004).

Both training and test samples will have entry from both fraudulent and non-fraudulent report. The main difference between the two samples is that the test samples will have the indicator flag, i.e., the flag that identifies whether the financial statement is fraudulent or not, removed. Hence the test sample will be composed of only the text data and a unique identifier. The reason for the removal is because for the purpose of

measuring the overall performance of the classification model, this thesis needs to assess the performance of the model in correctly classifying financial statement without any prior information about its status, i.e., fraudulent or non-fraudulent.

The size ratio between the training and test data will be around 1 to 4. This means the training data will have around 4 times more text entry than the test data.

3.4 Text Pre-Processing Process

In the text pre-processing process, the textual data is modified before further textual analysis process can be implemented. The modification is implemented with the purpose of transforming the available textual data from an "unstructured" form into a "structured" form. The term "unstructured" and "structured" refer to the level of difficulties that is required for a computer to process the textual data. "Unstructured" textual data is more difficult, for a computer, to process (Manning, Raghavan, & Schütze, 2008).

Several preprocessing techniques are available for the transformation process. The following discussion will discuss three techniques that will be used in this thesis. All discussion is based on explanation outlined in (Manning, Raghavan, & Schütze, 2008). The first technique is tokenization where the textual data is broken down into individual words, i.e., tokens. Hence for the following sentence "Alice likes Bob", the tokenization technique will produce these three words or tokens "Alice", "likes", and "Bob".

The second preprocessing technique that is used in this thesis is the lemmatization process. This process reduces the inflectional form of a word by transforming a word into its dictionary form. Thus these following words "learning", "learned", "learnt" will

be transformed to its dictionary form “learn”.

The last preprocessing technique used is the removal of stop words. Stop words are common words with little value for the textual analysis process. Stop words usually appear in high frequency within the text. Common stop words include particle, pronoun and preposition such as “a”, “an”, “and”, “the”, “be”, “from” and “for”.

In this model developed for this thesis, the preprocessing process will be implemented automatically using an algorithm implemented using Python version 3.6 programming language.

3.5 Feature Extraction

The term feature in the machine-learning discipline can be understood as attribute or characteristic that describes a particular item including its value (Kohavi & Provost, 1998). An example of a feature and its corresponding value is the feature color with its value red, green and blue. A more detailed feature and its corresponding value is having the color blue as a feature with a corresponding numerical value to represent the full gradient of the color blue, from the lightest to the darkest. To extract the feature of the text contained in the 10-K and 10-Q reports, this thesis will utilize textual analysis methods.

As it had been mentioned in section 2.4 of this thesis, a relatively current textual analysis method, i.e., distributed representation method, will be implemented to extract the feature of the textual data. To implement the distributed representation method, this thesis will utilized the word2vec software library that is described in (Mikolov, Chen, Corrado, & Dean, 2013). The library takes the MD&A text data that has been processed as its input and produces word vectors as its output (Najafabadi, Villanustre,

Khoshgoftaar, Seliya, Wald, & Muharemagic, 2015). The calculated word vector value represents not only the occurrence of a single word but also its co-occurrence with other words within a text as well as the order of words within that same text. Hence, the word vectors produced by the word2vec library will represent the features of the text within the report. Subsequently, the classification model will utilize this feature for identifying fraudulent text.

However to better understand the performance of the distributed representation method, this thesis will also implement the bag of words method. The implementation of the bag of words method is done in order to have a comparative benchmark when comparing the performance of the two methods. To implement the bag of words method, this thesis will utilize the TFIDVectorizer software library available in the scikitlearn framework (Scikit Learn Community, 2017).

In addition to two methods, this thesis will also combined the features produced by both methods as the input for the classification model. The reason for this combination is to see whether the increased number of feature will improve the performance of the classification model in predicting fraudulent financial statement.

All implementation of the textual analysis method to extract the features from the 10-K and 10-Q reports will be done using the Python 3.6 programming language.

3.6 Training the Classification Model

The classification model is the machine-learning model that will be used to classify a particular financial statement's text into a certain category. In this thesis, the categories being discussed are fraudulent and non-fraudulent. In essence, the classification model is part of the prediction process that actually predicts whether a

particular financial report is fraudulent or not. This thesis used the term training when referring to the process of creating a mathematical model that will be used by the machine-learning algorithm to do a specific task (James, Witten, Hastie, & Tibshirani, 2017). The term mathematical model can be defined as a generalized (Witten & Frank, 2005) or summarized (Kohavi & Provost, 1998) pattern that is obtained from the training data. Hence, training the model is a process in which pattern is obtained from a given training data. Using this model, a prediction can be made on new test data.

In this thesis, the model is trained using a type of machine learning method called supervised learning method. This means that during the training process the actual results that can be attributed to the training data is provided (Witten & Frank, 2005). In the training data that this thesis used, the actual results that are attributed to the training data are the information about whether a particular text is fraudulent or not. This thesis utilized a numerical flag to provide that information, i.e., "1" for fraudulent and "0" for non-fraudulent. The developed model then tries to predict the existence of fraudulent report from test samples without any numerical flag to identify whether it is fraudulent or not.

To train the model, this master thesis will utilize the XGBoost algorithm, a relatively current algorithm that has delivered promising results in problem solving tasks (Chen & Guestrin, 2016). Training the model using the XGBoost algorithm will be implemented using the Python version 3.6 programming language.

3.7 Predicting the Test Data and Measuring the Prediction Results

The final process in the prediction model is predicting fraudulent 10-K report from the test data samples. In this process, the trained model will predict whether a

particular 10-K report is fraudulent or not. The prediction is done on the test data according to the process outlined in section 3.3 and 3.4 of this thesis. As it has been outlined before, the test data sample doesn't have numerical flags to identify whether the particular text sample is fraudulent or not. Hence the prediction is made solely based on the pattern that was obtained by the classification model that was trained utilizing the training data.

To analyze the performance of the classification model, this thesis will use five commonly used measurement metrics (Witten & Frank, 2005), (Hajek & Henriques, 2017). However before discussing the measurement metrics, this thesis needs to discuss a few notations that will be used for measurement. The definition is outlined in Table 1.

Table 1 Measurement Notation

Notation	Definition
True Positive (TP)	The number of fraudulent sample correctly predicted as fraudulent.
True Negative (TN)	The number of non-fraudulent report correctly predicted as non-fraudulent.
False Positive (FP)	The number of non-fraudulent report incorrectly predicted as fraudulent.
False Negative (FN)	The number of fraudulent report incorrectly predicted as non-fraudulent.
Positive	The number of fraudulent report
Negative	The number of non-fraudulent report

Based on the notations outlined in Table 1, the first metric that will be utilized in the classification model performance measurement process is accuracy. Accuracy is used to measure the percentage of correct predictions that were made by the model, i.e., a fraudulent report is predicted as fraudulent and vice versa. The accuracy is measured based on all report, both fraudulent and non-fraudulent. The formula for accuracy is outlined in Equation 1.

Equation 1 Accuracy

$$Accuracy = \frac{TP + TN}{P + N} \; x \; 100\%$$

The second measurement metric is sensitivity. Sensitivity refers to the ratio of fraudulent reports that were correctly classified as fraudulent, based on all fraudulent report. This metric measures the ability of the classification model to correctly predict the existence of fraudulent report from all available fraudulent report. The formula for sensitivity is defined in Equation 2.

Equation 2 Sensitivity

$$Sensitivity = \frac{TP}{P} \; x \; 100\%$$

The third measurement metric is specificity. Specificity refers to the percentage of non-fraudulent reports that were correctly classified as non-fraudulent, based on all non-fraudulent report. This metric measures the ability of the classification model in making a correct prediction over the existence of non-fraudulent report from all available non-fraudulent report. The formula for specificity is defined in Equation 3.

Equation 3 Specificity

$$Specificity = \frac{TN}{N} \; x \; 100\%$$

The next two measurement metrics are metrics used to measure the error

percentage of the prediction. The first error percentage measurement, i.e., the fourth measurement metric used in this thesis, is the type I error percentage or the false positive percentage. The type I error measure the percentage of non-fraudulent report wrongly classified as fraudulent based on all non-fraudulent report. The formula for the type I error is given in Equation 4.

Equation 4 Type I Error

$$Type\ I\ error = \frac{FP}{N}\ x\ 100\%$$

The second error percentage measurement as well as the fifth measurement metric used in this thesis is the type II error percentage or the false negative percentage. The type II error measure the percentage of fraudulent report wrongly classified as non-fraudulent based on all fraudulent report. The formula for the type II error is given in Equation 5.

Equation 5 Type II Error

$$Type\ II\ error = \frac{FN}{P}\ x\ 100\%$$

Although this thesis finds that all the five measurements are equally important, this thesis also believes that for practical application, in particular for regulators and auditing firms, the three most important measurements are accuracy, sensitivity and the type II error. Accuracy is important in measuring the quality of a classification model. Sensitivity is important to measure the effectiveness of the model in correctly

identifying fraudulent report.

As for the practical importance of type II error, this thesis believes that the risk that comes when a fraudulent report is misclassified as non-fraudulent is higher than when a non-fraudulent report is misclassified as fraudulent. The reason is when a fraudulent report is misclassified as non-fraudulent, this classification normally would not lead to further investigation or audit to find detailed evidence of fraud. Hence the fraudulent report would be less probable to be detected down the road and the mistake would be less probable to be mitigated. However, when non-fraudulent report is wrongly classified as fraudulent report, further investigation or audit should be able to clear the report and the company submitting the report. Hence any mistake can be mitigated down the road.

This chapter has discussed the design of the prediction model that is implemented in this master's thesis. In the next chapter, this master thesis will discuss the results of the implemented prediction model.

4. Analysis of the Implemented Prediction Model

In the previous chapter, this thesis has discussed the design of the prediction model. Next in this chapter, this thesis will analyze the results of the implemented prediction model.

4.1 The Data

After analyzing SEC's litigation releases (US Securities and Exchange Commision, 2018), this thesis managed to identify and downloaded 1,265 samples of fraudulent 10-K or 10-Q reports from 113 companies. From the downloaded fraudulent sample, this thesis randomly selected 127 samples for the prediction model. In addition to that, this thesis also downloaded 246 samples of non-fraudulent reports from 21 companies. From the downloaded non-fraudulent sample, this thesis randomly selected 121 samples for the prediction model. Thus in total, this thesis selected 248 10-K and 10-Q samples for further use by the prediction model. Table 2 outlined the distribution of the downloaded reports as well as the reports that are selected for the prediction model.

The number of selected samples was decided in order to have a relatively balanced proportion of training data, i.e., ratio between the fraudulent and non-fraudulent report. Another consideration was the difficulties of manually extracting the MD&A text from the downloaded 10-K and 10-Q reports.

Table 2 Number of Downloaded and Selected 10-K and 10-Q Data

	Downloaded	**Selected for Prediction Model**

	Downloaded	Selected for Prediction Model
Fraudulent	1,265	127
Non-Fraudulent	246	121

From the total 248 samples, the prediction model took 200 samples as the training data. The selection of the 200 samples for the training data was done in such a way to maintain a balanced proportion between fraudulent and non-fraudulent samples. One important reason for maintaining this balance is to ensure that the classification model would maintain its accuracy (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Thus, this thesis took 100 samples each for the training data. The remaining 48 samples will make the test data. Table 3 outlines the distributions between training and test data that are utilized by the prediction model.

Table 3 Distributions of Training and Test Data

	Training Data	Test Data
Fraudulent	100	27
Non-fraudulent	100	21

4.2 Results of the Prediction Model

After the prediction model is implemented and the measurement process produced measurement results according to the metrics discussed in section 3.7, this section of this thesis will present and discuss the results. In particular this section will discuss the three implemented textual analysis methods that are utilized for extracting the features of the texts contained in the financial statement. The discussion on the three textual analysis methods is outlined in section 3.5.

The first result that will be discussed is the results of the prediction model that

utilized the distributed representation method to extract features from the 10-K and 10-Q texts. The method is implemented using the word2vec library. From the measurement results, the model showed limited accuracy, sensitivity and specificity with ratios of 58%, 59% and 57% respectively. The error level is still relatively high with ratios of 43% and 41% for type I and type II error respectively. Results of the distributed representation method are outlined in Table 4.

Table 4 Results of the Distributed Representation Method

Measurement Metrics	Results of the Distributed Representation Method
Accuracy	58%
Sensitivity	59%
Specificity	57%
Type I Error	43%
Type II Error	41%

The results suggested the limited performance of the distributed representation method for predicting fraudulent financial statement.

The next result that will be discussed is the results of the prediction model that utilized the bag of words method to extract features from the 10-K and 10-Q texts. From the measurement results, the model showed an improvement on accuracy and specificity with ratios of 67% and 100% respectively. A specificity ratio of 100% means that the prediction method managed to correctly predict all non-fraudulent report. However, the sensitivity ratio of the bag of words method showed a reduced performance compared with distributed representation method with a ratio of 41% compared with 59% produced by the distributed representation method.

Type I error for the prediction model that is utilizing the bag of words method showed a dramatic improvement with no error made in predicting non-fraudulent report.

However, the type II error for the bag of words method is relatively higher when compared with the distributed representation method. The type II error for the bag of words method is 59% compared with 41% produced by the distributed representation method. Results of the bag of words method are outlined in Table 5.

Table 5 Results of the Bag of Words Method

Measurement Metrics	Results of the Bag of Words Method
Accuracy	67%
Sensitivity	41%
Specificity	100%
Type I Error	0%
Type II Error	59%

After looking into the results of both the distributed representations method and the bag of words method, this thesis will discuss the results of the prediction model that utilized a combination of features that are produced by the two previously discussed method.

The combined method produced a comparatively improved result. Especially in terms of accuracy and sensitivity when compared with both the bag of words and distributed representation method. The combined method produced an accuracy and sensitivity ratio of 79% and 93% respectively. The type II error produced by the combined method is also improved with an error ratio of only 7%. With regards to specificity, the results of using the combined method is better than the distributed representation method but still lower compared with using the features produced by the bag of words method. The combined method produces a specificity ratio of 62%. A similar results can be seen in the type I error measurement. With a type I error rate of 38% produced by the prediction model that utilized feature from both the feature

extraction method. Table 6 outlines the results of the combined method.

Table 6 Results of the Combined Method

Measurement Metrics	Results of the Combined Method
Accuracy	79%
Sensitivity	93%
Specificity	62%
Type I Error	38%
Type II Error	7%

4.3 Analysis of the Results

Analysis of the three textual analysis methods that were implemented for extracting textual features from the financial statement showed that the combined method provides the best performance, especially using the three measurement metrics - refer to section 3.7 - that this thesis believes to have important practical application, i.e., accuracy, sensitivity and type II error. Table 7 provides a comparison between the implemented textual analysis methods implemented for the fraudulent financial statement prediction process.

Table 7 Comparison between the Implemented Textual Analysis Methods

Measurement Metrics	Distributed Representation	Bag of Words	Combined
Accuracy	**58%**	**67%**	**79%**
Sensitivity	**59%**	**41%**	**93%**
Specificity	57%	100%	62%
Type I Error	43%	0%	38%
Type II Error	**41%**	**59%**	**7%**

The results showed that the combined method managed to correctly predict 93% of fraudulent financial statement within the test data sample with only 7% of error in predicting fraudulent financial statement. This result is important because prediction of

fraudulent financial statement should serve as a trigger for further analysis and investigation by the relevant stakeholder. Having a high sensitivity ratio should lead to a more productive further analysis and investigation process in confirming the existence of fraud since most of the fraudulent statement is correctly predicted by the prediction model.

However a relatively lower specificity rate may hinder the overall effectiveness of the prediction process due to a relatively high number of non-fraudulent statements that are wrongly classified as fraudulent. This may lead to unproductive results during further analysis and investigation process.

5. Conclusion and Future Works

This thesis explore the application machine-learning technique to predict fraudulent financial fraud statement by analyzing the text content of publically available financial statements, i.e., 10-K annual report and 10-Q annual report. Implementation of two textual analysis methods along with a third method that combined the two methods showed a promising result for the application of machine learning technique to predict fraudulent financial fraud statement through analysis of its textual content. The combined method produced the best results for accuracy, sensitivity and type II error with a ratio of 79%, 93% and 7% respectively.

Having a high sensitivity ratio should increase the confidence of relevant stakeholder in taking further actions such a more in-depth analysis and investigation to confirm the existence of fraud. The reason is that since the prediction method managed to identify a high proportion of fraudulent financial statement, further analysis and investigation should bring, with higher probability, evidence of existing fraud.

This thesis also found that the distributed representation method for textual analysis, implemented using word2vec library, produced a relatively low performance compared with the bag of words methods. However combining the two methods improved the some performance of each individual method.

Future works that could be taken from this thesis includes the implementation of an automated text extraction process to obtain particular texts from the 10-K and 10-Q reports. By having an automated process, the amount of text available for training and testing the classification model should increase. This in turn should improve the accuracy of the classification model.

Another future works that could be implemented is parameter adjustment. Both the textual analysis methods and classification training method have parameters that could be adjusted for the purpose of improving the accuracy of the classification model. Improvement to this parameter through adjustment may bring an improvement to the overall performance of the prediction model.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud A Meta-Learning Framefork for Detecting Financial Fraud. *MIS Quarterly , 36* (4), 1293-1327.

Albrecht, W., Albrecht, C., & Albrecht, C. C. (2008). Current Trends in Fraud and its Detection. *Information Security Journal: A Global Perspective , 17*, 2-12.

Aliabadi, S., Dorestani, A., & Qadri, M. (2011). United, Fraud Prevention and Detection in the United States: A Macro Perspective. *Journal of Forensic & Investigative Accounting , 3* (3), 150-165.

Alpaydin, E. (2010). *Introduction to Machine Learning 2nd Edition.* Cambridge, Massachusetts: The MIT Press.

Arens, A. A., Elder, R. J., & Beasley, M. S. (2010). *Auditing and Assurance Services: An Integrated Approach* (13th Edition ed.). New Jersey: Pearson Education Inc.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations , 6* (1), 20-29.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research , 4*, 71-111.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research , 3*, 1137-1155.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Berlin: Springer.

Burke, K. (1939). The Rhetoric of Hitler's 'Battle'. *The Southern Review , 5*, 1-21.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems ,*

50, 164-175.

Cecchini, M., Aytug, H., Koehler, G., & Pathak, P. (2010). Detecting Management Fraud in Public Companies. *Management Science , 56* (7), 1146-1160.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 785-794). San Francisco: Association for Computing Machinery.

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computer Linguistics , 16* (1), 22-29.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science , 53* (9), 1375-1388.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research , 29* (3), 845-868.

De Franco, G., Verdi, R. S., & Verdi, R. S. (2011). The Benefits of Financial Statement Comparability. *Journal of Accounting Research , 49* (4), 895-931.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatement. *Contemporary Accounting Research , 28* (1), 17-82.

Dikmen, B., & Küçükkocaoğlu, G. (2009). The Detection of Earnings Manipulation: The Three-Phase Cutting Plane Algorithm using Mathematical Programming. *Journal of Forecasting , 29* (5), 442-466.

Dyck, A., Morse, A., & Zingales, L. (2010). Who Blows the Whistle on Corporate Fraud. *The Journal of Finance , 65* (6), 2213-2253.

Fanning, K. M., & Cogger, K. O. (1998). Neural Network Detection of Management Fraud using Published Financial Data. *Intelegent Systems in Accounting, FInance*

and Management , 7 (1), 21-41.

Glanchy, F. H., & Yadav, S. B. (2011). A Computational Model for Financial Reporting Fraud Detection. *Decision Support Systems , 50*, 595-601.

Goel, S., & Gangolly, J. (2012). Beyond the Numbers: Mining the Annual Reports for Hidden Cues Indicative of Financial Statement Fraud. *Intelligent Systems in Accounting, Finance and Management , 19*, 75 -89.

Goel, S., & Uzuner, O. (2016). Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports. *Intelligent Systems in Accounting, Finance and Management , 23*, 215-239.

Goel, S., Gangolly, J., Faerman, S. R., & Uzuner, O. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings. *Journal of Emerging Technologies in Accounting , 7*, 25-46.

Griffin, P. A. (2003). Got Information? Investor Response to Form 10-K and Form 10-Q Edgar Filings. *Review of Accounting Studies , 8*, 443-460.

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science , 17*, 26-32.

Hajek, P., & Henriques, R. (2017). Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud - A Comparative Study of Machine Learning Methods. *Knowledge-Based Systems , 128*, 139-152.

Henry, E. (2008). Are Investors Influenced by How Earnings Press Releases are Written. *Journal of Business Communication , 45* (4), 363-407.

Hinton, G. E. (1986). Learning Distributed Representations of Concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* , (pp. 1-12). Erlbaum.

Hogan, C. E., Rezaee, Z., Riley, Jr., R. A., & Velury, U. K. (2008). Financial Statement Fraud: Insights from the Academic Literature . *Auditing: A Journal of Practice & Theory , 27* (2), 231-252.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Hoboken, NJ: John Wiley and Sons, Inc.

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

International Financial Reporting Standards Foundation. (2007). *International Financial Reporting Standards Foundation - IAS 1 Presentation of FInancial Statements.* Retrieved May 2, 2018, from International Financial Reporting Standards Foundation: https://www.ifrs.org/issued-standards/list-of-standards/ias-1-presentation-of-financial-statements/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning.* New York: Springer.

Kaminski, K. A., Wetzel, T., & Guan, L. (2004). Can Financial Ratios Detect Fraudulent Financial Reporting. *Managerial Auditing Journal , 19* (1), 15-28.

Kassem, R., & Hegazy, M. (2010). Fraudulent Financial Reporting: Do Red Flags Really Helps? *Journal of Economics and Engineering , 4*, 69-79.

Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting Financial Misstatements with Fraud Intention using Multi-Class Sensitive Learning. *Expert Systems with Applications , 62*, 32-43.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statement. *Expert Systems with Applications ,*

32, 995-1003.

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning , 30* (2/3), 271-274.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling Imbalanced Datasets: A Review. *International Transactions on Computer Science and Engineering , 30*.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning. 32.* Beijing: Journal of Machine Learning Research.

Li, B., Yu, J., Zhang, J., & Ke, B. (2015). Detecting Accounting Frauds in Publicly Traded U.S. FIrms: A Machine Learning Approach. *Proceedings of Machine Learning Research. 45*, pp. 173-188. Hong Kong: Journal of Machine Learning Research.

Li, F. (2006). *Do Stock Market Investors Undestand the Risk Sentiment of Corporate Annual Reports.* University of Michigan, Stephen M. Ross School of Business. Ann Arbor: Stephen M. Ross School of Business, University of Michigan.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. *Managerial Accounting Journal , 18* (8), 657-665.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research , 54* (4), 1187-1230.

Loughran, T., McDonald, B., & Yun, H. (2009). A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports. *Journal of Business Ethics , 89*, 39-49.

Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit Card Fraud Detection using Bayesian and Neural Network. *Proceedings of the 1st*

international naiso congress on neuro fuzzy technologies.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* New York: Cambrige University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Workshop Proceeding at the International Conference on Learning Representations.* Scottsdale, Arizona.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep Learning Applications and Challenges in Big Data Analytics. *Journal of Big Data , 2* (1), 1-21.

Price, S. M., Doran, J., Peterson, D. R., & Bliss, B. A. (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking and Finance , 36* (4), 992-1011.

Ravishankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of Financial Statement Fraud and Feature Selection using Data Mining Technique. *Decision Support Systems , 50*, 491-500.

Rezaee, Z. (2002). *Financial Statement Fraud Prevention and Detection.* New York: John Wiley and Sons, Inc.

Rezaee, Z. (2005). Causes, Consequences, and Deterence of Financial Statement Fraud. *Critical Perspectives on Accounting , 16*, 277-298.

Robinson, T. R., van Greuning, H., Henry, E., & Broihahn, M. A. (2008). Financial Analysis Techniques. In T. R. Robinson, H. van Greuning, E. Henry, & M. A. Broihahn, *International Financial Statement Analysis (CFA Institute Investment Series)* (pp. 259-314). John Wiley and Sons, Inc.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers.

IBM Journal , 3 (3), 535-554.

Scikit Learn Community. (2017). *Feature Extraction*. Retrieved May 5, 2018, from Scikit Learn: http://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

Soltani, B. (2014). The Anatomy of Corporate Fraud: A Comparative Analysis of High Profile American and European Corporate Scandals. *Journal of Business Ethics , 120*, 251-274.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance , LXII* (3), 1139-1168.

The Securities Exchange Act of 1934, 15 (U.S.C. § 13 or § 15(d) March 23, 2018).

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial Fraud Detection using Vocal, Linguistic and Financial Cues. *Decision Support Systems , 74*, 78-87.

US Securities and Exchange Commision. (2009). *US Securities and Exchange Commision Form 10-K*. Retrieved May 1, 2018, from US Securities and Exchange Commision: https://www.sec.gov/fast-answers/answers-form10khtm.html

US Securities and Exchange Commision. (2011). *US Securities and Exchange Commision Form 10-Q*. Retrieved May 5, 2018, from US Securities and Exchange Commision: https://www.sec.gov/fast-answers/answersform10qhtm.html

US Securities and Exchange Commision. (2013, July 2). *SEC Announces Enforcement Initiatives to Combat Financial Reporting and Microcap Fraud and Enhance Risk Analysis*. Retrieved May 01, 2018, from US Securities and Exchange

Commision: https://www.sec.gov/news/press-release/2013-2013-121htm

US Securities and Exchange Commision. (2017). *US Securities and Exchange Commision Fillings and Forms*. Retrieved May 01, 2018, from US Securities and Exchange Commision: https://www.sec.gov/edgar.shtml

US Securities and Exchange Commision. (2018). *US Securities and Exchange Commision Litigation Releases*. Retrieved May 1, 2018, from US Securities and Exchange Commision: https://www.sec.gov/litigation/litreleases.shtml

Vrij, A. (2008). *Detecting Lies and Deceit.* John Wiley and Sons.

West, J., & Bhattacharya, M. (2016). Intelligent Financial Fraud Detection: A Comprehensive Review. *Computers & Security , 57*, 47-66.

Williams, C. B. (1975). Mendenhall's Studies of Word-length Distribution in the Works of Shakespeare and Bacon. *Biometrika , 62*, 207-212.

Witten, I. H., & Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann.

You, H., & Zhang, X.-j. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies , 14*, 559-586.