Introduction

Scholars were attempted to catch the code words of fraud from 10-K documents by varied methods, such as creating financial measures, studying the semantic of the documents or adopting computer science methodologies etc. Among all these scientific methodologies, Natural Language Processing (NLP) applications combined with Machine Learning (ML) and/ or Artifactual Intelligence (AI) is the novel and attractive one in Accounting, Auditing and Finance (Fisher et al., 2016).

We proposed the use of textual and linguistic features from the Management Discussion and Analytics (MD&A) section in the 10-K documents for the following reasons. First, U.S. Securities and Exchange Commission (SEC) requires all the publicly-traded company provide the 10-K document, and the MD&A is the obligated part, which minimize the selection bias. Second, the worthiest to read section in 10-K document is MD&A since the company can explain its business performance and result from previous fiscal year in its own words. MD&A provide an opportunity for the shareholders, auditors and researchers to observe the firms’ performance and prospects from the perspectives of management (Purda and Skillicorn, 2015). Third, accounting fraud occurs when the fraudulent manager intends to deceive shareholders by presenting misleading information on previous performance and the future expectation. Therefore, it is noted by researchers that MD&A is most relevant in addressing the mistakes in the financial statement (Hoberg and Lewis, 2017).

There are over 180,000 10-K documents in SEC’s EDGAR website since 1993. To comprehend, analyze and abstract such large number of documents is tedious and impossible for human, while NLP combined with ML has the capability to deal with “big data”. NLP build the bridge between the humen and the computer by processing unstructured language to structured data for the goal to achieve human-like language processing (Liddy, 2001). ML has wide applications, and ML-based classifiers facilitate researchers to determine the material fraud. Several researchers have applied this method on the prediction or detection in the fraudulent financial reporting in latest 10 years. Cecchini et al applied Support Vector Machine (SVM) with financial kernel (Cecchini et al., 2010), and Purda et al adopted bag of word methodologies and decision tree algorithm in 2015 (Purda and Skillicorn, 2015).

This research explores the application of one of Deep Learning architectures, Hierarchical Attention Network (HAN), in detecting linguistic pattern, form or signal in the misstatements. HAN has its own unique and competitive attribute comparing with other Deep Learning models. It has two levels of attention mechanisms for upper level and down level. Therefore, the structure of documents, such as paragraph level and sentence level, or sentence level and word level, varies in attention importance (Yang et al., 2016). With the attention importance of the Deep Learning model, we were aimed at excavating the patterns or rules in the syntax and semantics of the misstatements with the assistance of Deep Learning Model. Meanwhile, we also compared our model with other fraud detectors. For example, the classical classifiers, based on Support Vector Machine (SVM) and Naïve Bayes (NB), leveraged the financial text information; The F-score, as a signal of the likelihood of earnings misstatement (Dechow et al., 2011), and the MetaFraud, a novel meta-learning framework used the numerical information (Abbasi et al., 2012).

Therefore, we prepared the comprehensive data set of textual and numerical information, which was suitable for the detectors we mentioned above, although we had to face the loss of approximately 35% observations for the consistency of the data. We derived our exploration and the comparison by 56288 observations from 1995 to 2012. For each sample, we extracted the “Item 7. Management Discussion and Analytics” as the textual feature, and we compiled the numerical features from Compustat according to the financial features adopted in the researches of Dechow and Abbasi.

(Result of my experiment. To be continue)

Literature Review & Research Question

The timeline of financial fraud detection research lasts two decades. Some researchers used numerical financial or non-financial measures to track the proof of the fraudulent business activities and achieved good result (Abbasi et al., 2012; Dechow et al., 2011; Abbasi et al., 2012). ~~Most of studies selected 8-10 features, and the common classifications are logistic regression, neural networks, decision tree Bayesian networks and SVM~~. ~~There are several remarkable achievements. (TBC, need list some papers/examples).~~ The other research method is textual analysis which is widely applied in the academic research and business practice. Researchers have found out that the linguistic abnormalities in the fraudulent financial report, and studies examined the effectiveness and feasibility of the text mining and/ or machine learning in detection of misstatements (Humpherys et al., 2011; Purda and Skillicorn, 2015; Hoberg and Lewis, 2017; Brown et al., 2018). (TBC, need list some papers/ examples)

One of remarkable researches was conducted by (Dechow et al., 2011). From accrual quality, financial performance, nonfinancial performance, off-balance-sheet activities, and market-related variables, they investigated and discovered the numerical characteristics of misstating firms. For example, during the time of fraudulent behavior committed, performances of the company on accrual quality, financial performance and nonfinancial performance are deteriorating. With the logistic regression, they proposed a prediction model and the output of it is a scaled probability, called F-score. A F-score greater than one indicates the existence of misstatements, and a higher F-score it is, a higher probability of fraudulent occurred for this firm-year. After went through the Accounting and Auditing Enforcement Releases (AAERs), Dechow et al also concluded and compiled a database listing the fraudulent firm-years. Many financial fraud researches referred or used this dataset after Dechow’s work which is essentially meaningful and far-reaching (Dechow et al., 2011; Abbasi et al., 2012; Purda and Skillicorn, 2015). In our study, we also completed our work based on this dataset.

Meta-learning strategy achieved a good performance on the financial fraud detection. Compared with other machine learning algorithm, the advantage of meta-learning is it can learn from the learning process, that is the model can well adapt or genialize the new tasks in test data set (Brazdil et al., 2008). After went through the previous researches, Abbasi *et al* selected 12 financial ratios as features and construct the meta-learning model with SVM, Logit Regression, Naïve Bayes etc. for yearly and quarterly context-based data. The significant contribution of the research is to introduce a novel meta-learning framework which upgrades business intelligent methods into a meta-learning artifact, and to display the confidence scores generated by the model as a red flag for fraudulent detection (Abbasi et al., 2012).

Some researches which adopted text mining and natural language processing inspired us to do more work on explore the linguistic characteristics of financial misstatements. For MD&A of 10-K documents, with text mining technology and the deception theory from Communication and Psychology, scholars found that “those crafting fraudulent disclosures use more activation language, words, imagery, pleasantness, group references, and less lexical diversity than non-fraudulent ones” (Humpherys et al., 2011). The verbal abnormality of fraudulent MD&As was also be pointed out: fraudulent MD&A had less details about the sources of the performance while contained more positive aspect of performance (Hoberg and Lewis, 2017). Words also had the predictive or discriminative power to assess the likelihood of fraudulent financial statement after scholar processed the words by some machine learning algorithm, such as Random Forests and/ or SVM (Purda and Skillicorn, 2015). In additional, topics, generated by Latent Dirichlet Allocation (LDA) from MD&As, empowered the ability of the classifiers, which shown that the thematic content of the financial was useful for detections (Brown et al., 2018). Therefore, we believe that the linguistic differences between fraudulent and non-fraudulent ones are detective, no matter in semantics, lexicon or content.

The goal of our research is to explore and seek some linguistic characteristics of fraudulent financial statements with unique attribute of Hierarchical Attention Network (HAN) model. Meanwhile we want to validate the predict ability of this model in fraud detection task by comparing with other classical predictive models. It is common sense that not all parts of a document are equally important for answering a classification question. Since the documents have a hierarchical structure, such as sentences from paragraphs or paragraphs from documents, and there are two levels of attention mechanisms in HAN model, the HAN model therefore provides masterly assistance to find out units in the documents are differentially informative. In our research, the HAN model can build a presentative attention weights of input documents by aggregating important words into paragraph vectors and aggregating important paragraphs into document vectors (Yang et al., 2016). From the attention matrix, we can learn which words and paragraphs contribution the most for the classifier to decide the labels. In generally, the attention weights are human interpretable, and the attention mechanism works like a gating unit (Vashishth et al., 2019). In additional, we also planned to use some advanced interpretable technology to visualize the result of HAN model, such as Lime. From those words and paragraphs with high attention weights, we expected to find out some linguistic characters on content or semantics. At the same time, we were curious about the prediction ability of the HAN model compared with other classical model, such as the F-score and meta-learning based on numerical dataset, and the SVM and Naïve Bayesian model based on textual dataset.