Predicting Insider Trading from Financial Text: An Interpretable Deep Learning Approach

Rong Liu
School of Business
Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030, USA
Email: rong.liu@stevens.edu
Phone: +1 (201)-216-5028

Feng Mai
School of Business
Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030, USA
Email: feng.mai@stevens.edu
Phone: +1 (201)-216-3815

Zhe Shan
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Oxford, OH 45056, USA
Email: jayshan@miamioh.edu
Phone: +1 (513)-529-4940

Ying Wu*
School of Business
Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030, USA
Email: ying.wu@stevens.edu
Phone: +1 (201)-216-3510

* Corresponding author.
Abstract

The detrimental effects of insider trading on the financial markets and the economy are well documented. However, resource-constrained regulators face a great challenge in detecting insider trading and enforcing insider trading laws. Studies have shown that publicly-listed companies’ textual disclosures contain significant information about economic activities and intangibles related to unethical corporate behaviors such as insider trading. We propose a regtech framework that uses machine learning to predict insider trading from text data. Distinct from typical black-box neural network models, which have difficulty tracing a prediction back to key features, our approach combines the predictive power of deep learning with attention mechanisms to provide interpretability to the model. Attention mechanisms provide visibility to focal words and paragraphs that are deemed important by the model. Other novel features of our model include learning representations from a business proximity network and incorporating the temporal variations of a firm’s financial disclosures for prediction. We train the models using over 90,000 U.S. firms’ SEC filings to predict whether a firm will be involved in an insider trading related to class action lawsuit. We evaluate the performance of the models by comparing them with benchmark models with both numeric and text inputs. We demonstrate how text-based models can offer new insights about insider trading and discuss the practical implications of predictive models. Our study contributes to the emerging fintech literature by showing that deep learning is a promising method for regulators to examine a large amount of text in order to monitor and predict financial misconduct. We also contribute to information systems literature by showcasing how predictive analytics can reconcile performance and interpretability.

**Keywords:** Insider Trading, Design Science, Predictive Analytics, Text Mining, Deep Learning, Attention Models, Graph Mining, RegTech
1 INTRODUCTION

Insider trading refers to the buying or selling of a public company’s securities based on material, nonpublic information. Corporate executives have access to private information about public companies and thus possess a significant advantage over the majority of investors. By making perfectly timed trades based on nonpublic information, insiders can yield millions in profit. Insider trading based on private information clearly compromises the fairness and integrity of the capital markets. If left unchecked, it has dire and long-term consequences such as damaging investor confidence, raising the cost of capital [33], impeding innovation [64], and hurting financial market stability [57]. Because of the high economic and social costs, in the United States and many other countries, it is illegal to trade while in possession of material nonpublic information about a security.

Yet, despite the laws and regulations in place, their efficacy in restraining insider trading is limited. Prosecution and conviction of insider trading cases are difficult. In the U.S., after a series of legal precedents in favor of the alleged offenders, it is now considerably harder to win convictions for insider trading violations [59]. In the U.K., two principal suspects of a large insider trading case were recently acquitted after a nine-year Financial Conduct Authority investigation that cost $20 million [92]. According to the study by [13], insider trading laws exist in more than eighty percent of countries that have a stock market, but prosecutions have taken place in only about forty percent of them. Making things worse, the duty of policing insider trading violations largely falls on the shoulders of market regulators such as the Securities and Exchange Commission (SEC) in the U.S. These agencies are severely resource-constrained and must rely on staff discretion to select cases to pursue [27].

Given the challenges faced by regulators and policymakers, an increasing body of empirical research has investigated what factors give rise to informed insider trading. Research in finance
has demonstrated the effects of various internal and external mechanisms on insider trading intensity, such as firm size, stock return volatility, and regulatory change [6,55]. Although these studies throw considerable light on the roles of different economic and behavioral factors, little is known about whether, and to what extent, these factors together can guide future regulatory practices. Part of the reason is that the existing empirical literature focuses on the effects of individual mechanisms and does not address a crucial question: How best can regulators detect future illegal insider trading?

We answer the question by framing it as a novel predictive analytics problem. We use machine learning models to learn from historical data and predict the likelihood of insider trading violations in the future. Predictive analytics can hold great practical value while also generating new theoretical insights [85]. Analytical technology for regulatory considerations (a.k.a. regtech) is especially relevant for information systems (IS) research in the age of fintech, as policymakers and regulators are in need of modern tools and frameworks to keep up with technology-driven changes in the field [41]. Currently, regulators must wait passively for a whistleblower or a large suspicious trade to make a case [4]. With a prediction model, regulators could actively target future investigation and enforcement efforts. In addition, the ability to detect potential illegal insider trading could provide beneficial deterrent effects.

However, building a prediction model for insider trading is a challenging task for several reasons. First, there are many economic, behavioral, and contextual factors that influence the probability of financial misconduct. It is hard to capture these factors by quantitative information from the company’s financial statement. As such, it is an open question whether informative and opportunistic insider trades that violate regulations can be predicted. Second, the very definition of illegal insider trading is vague [4]. Although U.S. law requires corporate insiders to report
transactions to the SEC, not all trades made by insiders, however profitable, are illegal. Moreover, it does not require an insider to trade on private information directly to be considered illegal—the SEC has targeted other actions such as tipping friends and relatives with confidential information. Yet using SEC enforcement actions as the target variable is prone to Type II errors because of the SEC’s limited resources and selection bias [28]. Third, the antecedents of corporate financial misconducts not only include a firm’s own current conditions, but also its competitive environment [47] and changes in historical financial reports [24]. Incorporating these elements in a prediction model is difficult as they entail high dimensional features. In addition, because the environment is “in a state of constant change” [80], we need to capture the dynamic competitive relationship between firms. Fourth, for regulatory agencies to justify their actions, the prediction model needs to be more than a black box. Many machine learning methods are not well received among criminal justice agencies because they lack clear logic or explanation for their decisions [16]. High interpretability is crucial for machine learning models to reach their full potential in assisting regulators and policymakers.

To address these challenges, we compile a comprehensive dataset that contains all the U.S. public firms that were the subject of insider trading related class action lawsuits between 1996 and 2015. Figure 1 provides several excerpts from the lawsuit filings. We provide theoretical arguments for why insider trading can be predictable, and why firms’ annual reports (10-K’s) are useful. We exploit financial text data to predict whether a firm will be a defendant in a case. Specifically, we extract two sections, Business Description and Managerial Discussion & Analysis (MD&A), from the 10-K filings to the SEC and match them with our insider trading dataset. Our combined dataset contains a large sample of 90,152 U.S. firms’ filings with more than 717M words.

[Insert Figure 1 about here]
We propose an interpretable deep learning framework to automatically extract patterns for prediction from this large corpus. Deep learning is a machine learning paradigm that combines multiple layers of neural networks to learn representations of data with multiple levels of abstraction [63]. Despite its ubiquity, effective integration of text data in financial models remains a challenging mission due to the difficulty in both obtaining and quantifying textual data [62]. We show that the deep learning model is suitable for such tasks by condensing the sparsely-encoded information in the financial text. We compare the predictive performance of the deep learning models to several benchmark models and show that deep learning models have superior out-of-sample AUCs.

We map our model design to the logic of the problem by making three adaptations to a typical deep text classification model. First, we add hierarchical attention mechanisms to the deep learning model. The attention mechanisms automatically assign context-dependent weights to words and paragraphs during the model training process, thus allowing us to decode which part of the data the neural network is paying attention to. This is important for regulatory considerations because the trained neural network is directly interpretable. Second, we use a graph-based machine learning approach to represent firms’ dynamic context in a network. Specifically, we extract business description sections from the 10-K’s and construct a business proximity network [83] using a novel text similarity measure. The network supports the intuition that two competing firms usually have similar product descriptions. We then use a neural network to represent each firm’s position on the network with a low-dimensional vector. This graph-based definition of the environmental context can capture competition across industry boundaries; it is also inherently dynamic because the competitive relationships are updated along with the annual filings. We encode this environmental context learned from textual data using Gated Recurrent Units (GRUs),
a type of recurrent neural network. Third and finally, we measure changes in historical financial reports and encode the sequence of temporal variations of a firm’s reports with GRUs to extract patterns indicative of insider trading.

Overall, this study contributes to the IS literature in several ways. First, we add to the emerging fintech literature by offering a new set of predictive analytics tools for market overseers to regulate insider trading. We show that textual disclosure is relevant for its ability to reflect information cues that are missing in quantitative variables. In particular, our textual model can identify a coherent set of meaningful words and topics that are resonant with relevant economic and behavioral theories. In addition, our empirical findings can shed light on theoretical tensions in prior studies regarding the effects of competitive environment and temporal differences of filing documents on risks of insider trading. The prediction framework can potentially be used for detecting other financial misconduct and corporate wrongdoing.

Second, we contribute to IS methodology by designing an attention-based deep learning approach that integrates multiple inputs from textual, numerical, and relational data. Recently, deep learning has shown promising results in many areas of IS research including user-generated content [44,84,86,102,103], process analytics [35], and healthcare analytics [68,99], thanks to its ability to extract features from high dimensional, unstructured data. We add to this growing stream of analytics literature by (1) showing how attention-based models can strike the balance between the better performance of neural networks and the higher interpretability of linear models, and (2) extending the work of [83] by proposing a deep learning model to construct and to learn features from a business proximity network, thereby capturing information about firms’ competitive environment to enhance prediction.

Third, we show that predictive analytics for insider trading is an important and viable research
topic for IS researchers. To the best of our knowledge, only a few studies looked at insider trading from a computational perspective and existing works are mostly exploratory [94]. We demonstrate, theoretically and empirically, that insider trading cases are predictable from publicly available archival data. With the integrity of the modern financial market at stake, this is an area where IS researchers can make a significant contribution.

2 BACKGROUND AND RELATED WORK

We review three strands of related literature: (1) insider trading and its costs, (2) predictive analytics for regulators, and (3) deep learning methods in IS research. We also discuss how our work fills gaps in the prior literature.

2.1 Insider Trading and Its Costs

The trades of corporate insiders are among the most widely scrutinized activities in the stock market [25]. The social costs of insider trading are well documented. Easley et al. [32] show that a change of ten percent in the probability of insider trades increases the cost of equity by 2.5 percent per year. Easley and O’Hara argue that information asymmetry associated with insider trading is a systematic risk that cannot be diversified away [33]. Although laws and regulations governing insider trading are implemented in many countries, they are not always effective. Using data from 52 countries, Bris reports a surprising finding: the intensity of insider trading and profitability increases after new laws are enforced [17]. U.S. law requires corporate insiders — defined as directors, officers, or owners of more than 10% of the company stock—to report transactions to the SEC.¹ Still, based on the abnormal returns of these reported transactions, it is clear that insiders take advantage of their information to make profitable trades [55].

¹ For details please see https://www.sec.gov/answers/form345.htm.
While many scholars studied the factors behind the variation of insider trading activities [6,55,57], few have looked at the predictability of illegal insider trading. Cohen et al. show that only routine trades with zero abnormal returns are predictable, which has limited value for regulators’ enforcement actions [25]. Tamersoy et al. conduct a large-scale analysis of insiders’ trades using the Form 4 filings [94]. They extract distinctive temporal patterns in insiders’ trades that may be explained by regulations, policies and other factors. However, the study is exploratory and does not address the predictability of the trades. Our study fills a critical gap in the literature by providing evidence on the predictability of illegal insider trades.

2.2 Predictive Analytics for Regulatory Enforcement and Fraud Detection

In many domains, predictive analytics can be as, if not more, valuable than explanation [5]. Law enforcement is one of these domains. Predictive analytics allows enforcement agencies to stay ahead of the game using targeted intervention and more efficient resource allocation [38]. Researchers have also applied computational methods to detect financial fraud. We refer readers to [14,76,98] for reviews of these techniques. Our work is inspired by recent IS studies that use machine learning to detect financial statement fraud. Abbasi et al. propose a business intelligence framework that detects fraud from publicly available financial information [1]. Text mining methods have been used to identify SEC investigations [22,40]. Siering et al. use content-based and linguistic cues for online crowdfunding fraud detection [87]. Dong et al. use financial ratios and language-based features from social media data for corporate fraud prediction [29].

Our study adds to this stream of literature by focusing on the prediction of insider trading cases using text. We design our model to work with a much larger sample size (more than 90,000 observations, compared with from 122 to 652 in prior studies [22,29,40,87]). Further, although several studies consider textual features, they rely on different features engineering methods. We
propose an end-to-end deep learning model [75], in which the learning algorithm goes directly from the raw textual input to the prediction. As will be discussed later, the training of such end-to-end model can be more efficient given a large amount of input data. Lastly, prior approaches mainly utilize case-specific features and do not consider relational data (i.e., firms’ relationship to its peers and its own past). Inclusion of a temporal component for individual firms would be an important extension to these methods [22]. Leveraging new methods to learn from relational data is also much needed in a networked business environment [83].

2.3 Deep Learning for Textual Data

Firms’ text disclosure plays an important role in how financial information is conveyed to the public. Studies have demonstrated that qualitative corporate filings contain valuable information about credit risk [69] and financial statement frauds [40]. However, most studies rely on simple text summarization techniques such as word count, sentiment, and readability. The information in financial text goes well beyond these measures [15]. To leverage the full value of the textual disclosures, there is a need for more efficient algorithms to extract information from textual data.

In the past decade, artificial neural network has been used as an important tool in IS research (see for example [65,78,96,97]). Built on classical neural network theories, deep learning models [63] add additional processing layers which can transform raw inputs into higher-level representations. For textual data, a deep learning model maps discrete words or phrases into continuous representations in a vector space, which allows the model to operate on the semantics of raw inputs. Several recent IS studies successfully leverage deep learning models such as recurrent neural network and word embeddings for prediction tasks in healthcare domains and online reviews [68,84,99,103].

Despite their remarkable predictive capabilities, deep learning models have been criticized for
being black boxes with low interpretability [67]. As machine learning models penetrate critical areas such as financial markets and medicine, there is increased anxiety about whether they can be trusted if they cannot be understood. New regulations from the European Union propose that individuals affected by algorithmic decisions have a right to explanation [42]. Law enforcement agencies especially require high interpretability of machine learning models since they must defend their decisions [16]. To address the issue of interpretability, we introduce attention mechanisms. Attention mechanism [95] is a novel technique in deep learning that allows a neural network model to focus dynamically on a subset of an input based on a given context to enhance pattern recognition. We show that hierarchical attention mechanisms allow our model to achieve both high prediction accuracy and interpretability.

3 THEORETICAL FOUNDATION

3.1 Predictability of Insider Trading

Before discussing the details of the prediction model, we need to answer the question: why do we believe future insider trades are predictable in the first place? To justify the prediction models, we note that findings from both behavioral psychology and corporate finance lend support to our approach. To begin with, behavioral experiments have shown that unethical behaviors such as informed insider trading are not random. Various economic and psychological causes—such as external and internal reward [8], the probability of punishment [74], and saliency of dishonesty [39]—can trigger or curtail acts of cheating [71]. In other words, having knowledge of environmental cues makes future dishonest actions predictable.

At the firm level, the financial reports and market performance often disclose the environmental factors that elicit insider trading. Several specific channels can contribute to the predictability of insider trading behavior. First, financial documents contain cues about managers’ personal
incentives that can often foretell their actions. The quality of managerial disclosure can reveal subsequent insider trading decisions [81]. A lower quality disclosure is associated with future insider purchasing, as managers wish to maintain their information advantage. Second, financial reports may disclose weakness in internal control, such that certain actions and policies of the firm management can lead to an unethical work environment [88]. Moreover, certain business cultures are more likely to weaken the honesty norm and promote dishonest behavior [26], and a firm’s ethical culture can be reflected in its financial report [93]. Lastly, corporate insiders, for their own incentives, have the propensity to engage in trading prior to material events such as bankruptcy [45]; these material events are known to be predictable using machine learning models. Taken together, just as how lab experiments and empirical studies can estimate the causal effects of an isolated factor, a machine learning model should be able to derive patterns from historical data prior to opportunistic insider trading behavior.

3.2 Utility of Financial Text

We propose that financial text, corporate annual filings in particular, can be used as an important source of information for a machine learning model. Theoretically, the Information Manipulation Theory [72] argues that insiders who tend to conduct opportunistic trading have incentives to withhold relevant information so that investors will make incorrect inferences. Meanwhile, once the insider has chosen to commit opportunistic insider trading, he or she has incentives to engage in a type of word "shell game" which grandstands certain aspects of performance to deflect attention away from the economic events that precipitated the insider trading [51]. Given such subtle motives, highly aggregated variables such as document tone are limited by their low dimensionality. For example, an opportunistic insider might have incentives to be abnormally positive on one topic, but abnormally negative on another, washing away any signal in aggregate
tone. A deep learning model has the potential to learn new representations from the raw text—a high dimensional data structure—to identify both effects.

Specifically, the Management Discussions and Analysis (MD&A) section of the 10-Ks can contribute significant prediction power for three reasons. First, while the numbers in the financial statements are summaries of historical performance, the purpose of MD&A is to provide a management perspective not only on their firms' past performance and current financial positions, but also their future prospects [36]. Second, public text information provides a setting that is more consistent with the argument of limited investor rationality and investor attention developed in the recent literature [49]. In contrast, numerical data extracted from financial statements or the stock market are unlikely to be subject to simple limited investor attention. Finally, managers are likely to have more freedom in writing the texts of the annual report than of the numbers, since the latter are subject to Generally Accepted Accounting Principles. Therefore, MD&As can shed light on more nuanced managerial behaviors and strategic intent.

Other than the MD&A section, two additional sources of information from 10-Ks can add predictive power. The business description section can capture firms’ competitive environment [53], which may either encourage or restrain insider trading. On the one hand, because insiders’ potential profit depends on the information asymmetry, they are less likely to extract rents in highly competitive environments closely monitored by analysts and peers [54,79]. On the other hand, a complex environment—partly embodied by competitive pressures from peers [100]—creates opportunities for insiders to exploit their information advantage. This can be due to the lapse of ex-ante preventive measures [47]. Or, as shown in [52], firms with similar descriptions are more likely to engage in mergers and acquisition (M&A) deals because of product market synergies. Informed insider trading is prevalent ahead of these M&A announcements [9].
Further, firms’ changes in the filings also carry valuable information. Cohen et al. [24] present the “lazy prices” phenomenon: changes to the 10-Ks predict future earnings, news announcement, and most importantly, insider selling activities. This is because the public digest these changes with a lag, but the insiders hold material information on why these changes took place. Therefore, we contend that the temporal differences of filing documents for individual firms can also add value to a prediction model.

4 DATA AND PREPROCESSING

We construct our insider trading database by merging four data sources: class action lawsuit data from Stanford Securities Class Action Clearinghouse (SCAC), textual disclosure data from 10-K annual filings to the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, accounting data from Compustat North America, and equity trading data from Center for Research in Security Prices (CRSP). Figure 2 summarizes our analysis procedure, including the database construction, data preprocessing, feature extraction, and model training and evaluation. Next, we describe each dataset and analysis procedure in detail.

4.1 Insider Trading Indicator

Our sample begins with the firms that were the subject of a class action lawsuit identified through the Stanford Securities Class Action Clearinghouse (SCAC). The SCAC provides detailed information regarding the filing date, class period (i.e., the period over which the alleged fraudulent behavior occurred), nature of the complaint, and settlement terms. We examine security

---

2 Another litigation database is the SEC litigation database. SEC litigation cases are not suitable for our study as most of the cases are related to fraud or embezzlement. The majority of the insider trading court cases were brought to the court by the stakeholders, not the SEC, due to funding constraints.
class-action suits occurring after passage of the Private Securities Litigation Reform Act (PSLRA) with filing dates between 1996 and 2015.

Since not all the cases in SCAC are insider trading, we identify cases violating any law or rule related to insider trading. The Securities Exchange Act of 1934 was the first law against insider trading. The Act grants the SEC the authority to set the rules. Section 20A of the Act provides a private right of action against persons engaged in insider trading. SEC has also enacted Rule 10b5-1 to deter insiders from trading on private information. Thus, within all case summaries, we search for keywords mentioning this law or this rule using regular expressions. If either Section 20A or Rule 10b5-1 is mentioned in a case summary, then the case is annotated as positive. From our original sample of 4,903 security-related cases, we are able to identify 1,151 cases that are related to insider trading. For each identified positive case, we further assign it to the corresponding fiscal year of the firm based on its filing date. Note that, typically, a legal case may last for many years and some may be dismissed at the end. To make the problem more tractable, the classification target is identified based only on the initial filing status of cases.

4.2 Numerical Financial Predictors

For numerical inputs, we compile eight firm-level predictor variables based on the literature on insider trading. When predicting financial misbehavior, it is common to consider the accounting information and up-to-date market information that may reflect the company’s liability, liquidity, and profitability status. Table 1 lists these variables and the rationale for including them. In our study, all variables are obtained by merging annual accounting data from Compustat North America with equity data from CRSP. To avoid recording errors or outliers, we further winsorize all the numerical predictor variables at 1% and 99% by replacing values that are lower than 1% with the variable’s first percentile and higher than the 99% with its ninety-ninth percentile. In
addition, we include the Fama-French 12-industry classification as a categorical predictor. Table 2 provides the summary statistics of the numerical variables.

[Insert Tables 1 and 2 about here]

4.3 Textual Predictors

A key innovation of our study is that we consider an untapped textual data source to predict insider trading. The SEC requires all public firms to file a 10-K form at the end of each fiscal year. A complete 10-K consists of fourteen items that provide a comprehensive yearly summary of a company’s business. Common items include business (description), financial performance, organization structure, executive compensation, equity, among others.

Our prediction model focuses on the Business (Item 1) and the MD&A (Item 7) sections of 10-K. These two sections serve three different purposes in our prediction model. First, according to the theory of insider trading, the main predictive power shall come from the MD&A section. Second, the Business sections are used to build a business proximity network each year (akin to [83]). The network captures firms’ dynamic relatedness with each other in terms of product, market, and technology. Our model then uses a firm’s structural position on the business proximity network—which represents the organizational environment of the firm—for prediction. Finally, we encode the temporal differences of the MD&A section to capture the recent changes in the firm.

We download all 10-K forms and its variants, 10-K405, 10KSB, and 10KSB40, from the SEC EDGAR system. We link 10-K forms to CRSP and the Compustat database using the Central Index Key (CIK). For each linked 10-K filing, we transform the file to plain text by removing the HTML tags, tables, and exhibits. We then extract the Business and MD&A sections using regular
expressions. These two sections usually appear as Item 1 and Item 7 respectively. The raw corpus is preprocessed in four steps. (1) We tokenize documents into individual words using spaCy, an open-source library for natural language processing (NLP). (2) We also use spaCy to lemmatize words to remove the inflectional forms of words and return them to basic forms. (3) We remove the common stop words, numeric, and punctuation marks. (4) Since some phrases have meanings that are not captured by simply summing up the individual words, we use the method recommended by [73] to find common phrases in the text and treat them as single words.

We then merge numerical predictors with all the textual predictors based on Compustat identifier (GVKEY) and fiscal year to get our primary samples, which include all the firms from 1996 to 2015. For each firm-year combination, we identify whether the firm has at least one lawsuit related to insider trading in the following year as a binary indicator. This binary indicator becomes our prediction target. We lag the textual and numeric predictors by one year and use them to predict the insider trading cases. For the sake of efficiency, we remove words with few frequencies in the entire corpus and include only 45,000 most frequent words. Such filtering procedure is a common practice in natural language processing. In total, our data includes 11,612 firms and 90,152 firm-years with no missing observations. Table 3 provides a yearly distribution of insider trading cases in the sample.

Using these numerical and textual predictors, we create deep learning models with attention mechanisms to predict insider trading. We also compare deep learning models with baseline models for benchmarking purposes (see Figure 2). We describe these models in the next section.

[Insert Table 3 about here]

---

3 We exclude Item 1A (Risk Factors) and Item 1B (Unresolved Staff Comments) subsections because they are not directly related to firms’ main products. We include the cases where MD&A appearing in the annual shareholder letters at the end of the 10-K.
5  METHOD

Figure 3 presents the deep learning model architecture. Modules for processing different inputs are shown using different colors. These modules, annotated as A-F in the figure, are introduced one by one as follows.

[Insert Figure 3 about here]

5.1  Word Embedding (Module A)

In the first layer of the model, we learn vectorized representations of words from the documents. The goal is to represent each word as a vector of $p$ dimensions ($p = 100$ in our experiments). This layer addresses the “curse of dimensionality” problem in the commonly-used bag-of-words model. The bag-of-words model treats each unique word as a feature, which can increase the dimensionality of input features to over tens of thousands (the size of vocabulary). As the dimensionality increases, the amount of data required for the model to have acceptable variance increases rapidly [90]. Therefore, dimension reduction on the inputs usually improves the out-of-sample performance of the model. Moreover, with a dense, vectorized presentation of the words, the prediction model makes use of the semantics rather than the syntax of the text.

There are two ways to obtain word vectors. Their first steps are the same. We start by observing that each unique word $w$ can be naturally expressed using a one-hot binary vector of size $n$, where $n$ is the size of the vocabulary. Formally, let $C$ denote the set of $n$ unique words in total in the corpus, $C = [w_1, w_2, ..., w_n]$. The one-hot encoding of word $w = [x_1, x_2, ..., x_n]$, where $x_j = \begin{cases} 1, & \text{if } w = w_j \\ 0, & \text{otherwise} \end{cases}$. For example, the fourth word in the vocabulary can be expressed as $[0, 0, 0, 1, 0, ..., 0]$. The one-hot vector is a very sparse representation because there are $n - 1$ zeros in the vector. Then the neural network uses the corpus to learn an embedding matrix $W$ with dimension $n \times p$. The matrix $W$ projects the one-hot vector of the $j$-th word to a vector $e_j$ in $\mathbb{R}^p$, where $e_j$ is
the corresponding row in $W$. We use $e_j$ as the vector representation of the word. The vector $e_j$ is a dense vector with a much lower dimension.

The difference between these two ways is how to complete the feed-forward neural network for training. The first way is using the skip-gram model proposed by [73] to pre-train word vectors. It uses an unsupervised approach to learning word embedding vectors from a large corpus. The skip-gram model summarizes the contextual information of each word by predicting its surrounding words using a neural network. The network adds an output *softmax* layer after $e_j$. The layer uses the $e_j$ as the input to predict the probability of observing each context word surrounding the $j$-th word. The second way is optimizing the parameter matrix $W$ directly in a supervised model. In this approach, word vectors are initialized randomly and then learned as a part of the prediction. As a result, word vectors are tuned toward the learning target of the model. We experimented with both approaches and found that a hybrid approach renders better prediction performance. In this approach, we first train word vectors using the skip-gram model with our 10-K corpus, these word vectors are then used to initialize the embedding matrix $W$ in our deep learning model, and $W$ is further tuned during training the deep learning model.

5.2 Word Attention (Module B) and Paragraph Representations

Motivated by the human visual attention system, attention mechanisms have been actively used in image recognition models to find the focal points in images. This technique has recently been introduced to NLP and achieved excellent performance in many tasks [95,101]. Traditional NLP models (such as bag-of-words and deep learning models built solely on word embeddings) learn a fixed weight for the same input feature. For example, a word is assumed to be equally informative throughout a document. The essence of attention is that the feature weight can change according to context, so the model can dynamically focus on the most relevant input features for prediction.
Moreover, a significant advantage of attention is that it provides interpretability to the model based on the context-dependent weights of words and paragraphs.

We adopt the hierarchical attention network structure [101] and include two layers of attention mechanisms in our model. The first layer generates paragraph representations based on words, and the second layer generates document representations based on paragraphs. Such design allows us to interpret the model both at the local (paragraph) level and the global (document) level. We introduce the first layer in this subsection and the second layer later in subsection 5.4.

In the first layer, attention is allocated to words within each paragraph. Previous studies show that specific words or topics in financial documents can help predict misreporting or litigation risks [18]. Yet, a word can have different meanings depending on the paragraph context. For instance, the word “extraordinary” can be used to describe the excellent performance of the firm; it can also mean extraordinary circumstances in which an abnormal course of action is warranted. Attention mechanisms can learn important words based on the local context and aggregate these informative words to form a representation of the paragraph.

Formally, a document with \( m \) paragraphs is denoted as \( d = [s_1, s_2, ..., s_m] \), where the \( j \)-th paragraph \( s_j = [x_{j,1}, x_{j,2}, ..., x_{j,h}] \), \( x_{j,i} (i = 1, 2, ..., h) \) is the one-hot encoding vector of the \( i \)-th word in paragraph \( s_j \), and \( h \) is the number of words in a paragraph. We set \( m = 40 \) by merging surplus paragraphs and padding documents with less than 40 paragraphs. Similarly, we truncate long paragraphs and pad shorter ones so that each paragraph has a uniform length \( h = 200 \). Let \( W \) denote the \( n \times p \) embedding matrix for all word, where \( p \) is the embedding dimension. The first attention layer generates a representation for paragraph \( s_j \) through Equations (1) - (4):

\[
\begin{align*}
    e_j &= s_j W \\
    u_j &= \tanh (e_j W_a + b) \\
    a_j &= \frac{\exp(u_j u_a^T)}{\sum\exp(u_j u_a^T)}
\end{align*}
\]
\[ z_j = a_j \circ e_j \quad (4) \]

Equation (1) returns a word embedding matrix \( e_j \ (e_j \in \mathbb{R}^{h \times p}) \) for all words in paragraph \( s_j \). Equations (2)-(4) describe the attention mechanism. In Equation (2), through a \( q \)-unit hidden layer with weights \( W_a \) and bias \( b \), we get a hidden representation of \( e_j \), denoted as \( u_j \ (u_j \in \mathbb{R}^{h \times q}) \), where \( W_a \in \mathbb{R}^{p \times q}, \ b \in \mathbb{R}^{1 \times q} \). Moreover, the attention mechanism introduces another parameter, \( U_a \in \mathbb{R}^{1 \times q} \). We can consider \( U_a \) as a context vector representing a query “should the model pay attention to a vector of \( u_j \)?” Then, in Equation (3), the cosine similarity between \( u_j \) and \( U_a \) is calculated as the attention weight. The weight \( (a_j \in \mathbb{R}^{h}) \) is further normalized over all words in \( s_j \). Finally, as shown in Equation (4), the dot product of attention weight \( a_j \) and \( e_j \), i.e., the weighted sum of the word embedding vectors in \( s_j \), produces the paragraph presentation \( z_j \in \mathbb{R}^{1 \times p} \). In this attention mechanism, \( W_a, b, \) and \( U_a \) are the parameters to be learned.

### 5.3 Business Proximity Network and Node Embedding (Module C)

As argued earlier, a firm’s competitive environment provides important contextual information for predicting insider behaviors. We adopt a social network approach to define a firm’s environment. In each year, we use the business description sections in the 10-K filings to construct a business proximity network [53,83]. Each firm is a node in the network. Two firms are connected if they have high product, market, or technology overlap, as measured by document similarity of their business descriptions. Then, we use a node embedding model to represent each firm as a 100-dimensional vector that summarizes the firm’s structural position in the network [43].

We use a document similarity algorithm built upon word embeddings, called Word Mover’s Distance (WMD), to calculate the dyadic proximity between firms [61]. Intuitively, when measuring the similarity of two documents \( d \) and \( d' \), WMD captures the “total costs” of mapping.
each word in \( d \) to its closest synonym in \( d' \). The mapping is less costly if two words are similar in meanings; the cost is zero if the same word appears in both documents. Formally, for a pair of documents \( d \) and \( d' \) both with length \( l \), we use embedding vectors to represent all the nouns, i.e., \( d = [e_1, e_2, \ldots, e_l] \) and \( d' = [e'_1, e'_2, \ldots, e'_l] \). Let \( c(i,j) = \| e_i - e'_j \|_2 \) be the Euclidean distance of word vectors \( e_i \) and \( e'_j \). Then the WMD between \( d \) and \( d' \) can be calculated as \( r = \sum_{i,j=1}^{l} T_{ij} c(i,j) \), where \( T_{ij} \) is the optimal solution to the optimization problem: \( \min \sum_{i,j=1}^{l} T_{ij} c(i,j) \), subject to \( \sum_{j=1}^{l} T_{ij} = e_i \), \( \sum_{i=1}^{l} T_{ij} = e'_j \), and \( T_{ij} \geq 0 \), for \( \forall i, j \in [1, 2, \ldots, l] \). This problem is essentially a transportation problem: the words in \( d \) are the sources and the words in \( d' \) are the destinations; the costs are the semantic distances between pairs of words.

After we calculate the pairwise proximity between firms using WMD, we follow [53] and define the network by adding edges between the top 2.05% of the most similar firm pairs. The threshold resembles the granularity of the Standard Industrial Classification (SIC) code at the three-digit level (2.05% firm pairs belong to the same SIC-3 industry). Because we use the word embedding technique to measure the similarity between firms, our approach is a deep learning alternative to the text-based business proximity networks proposed by recent literature [53, 83]. Compared with the cosine similarity approach in [53], our method is robust to synonymy—firms sometimes use different words to describe the same product. Compared with the topic modeling approach used in [83], our approach naturally fits into the deep learning framework and does not entail estimating a Bayesian model.

Given the business proximity network, we need to find useful features in the network to represent firms’ environments. Traditional social network analysis relies on hand-picked features

\(^4\) We only include the nouns because they are used to describe the product market that a firm operates in.
such as node degrees, centralities, and local communities to represent the nodes. By contrast, node2vec can automatically learn low-dimensional vector representations of nodes based on their network positions [43]. The idea behind the node2vec algorithm is conceptually similar to that of skip-gram model, i.e., we can learn a word’s meaning from its neighbors. In a network, a node’s neighbors determine its role and position. The algorithm thus takes two steps. First, it learns each node’s neighbors by simulating \( r \) random walks of fixed length \( l \) starting from each node.\(^5\) In the second step, it applies the skip-gram model on the traces of random walks (sequences of nodes) as if they were a text corpus. The nodes are treated as words, and each walk is treated as a sentence. In essence, node2vec uses a neural network to predict a node’s neighborhood defined by the random walks. Finally, because we construct one network each year, we align the yearly dimensions of node vectors using the orthogonal Procrustes method [46].

5.4 **Paragraph Attention (Module D) and Document Representation**

Now we turn to the second layer in the hierarchical attention model. This layer generates a representation for each document. It learns the meaning of a paragraph based on the global context, which consists of the preceding paragraphs and the firm’s competitive environment. In the prior subsection, we describe how a node vector encodes a firm’s structural position in a business network. Following the sequence-to-sequence learning techniques [91], we use GRU to generate a document presentation from paragraph representations concatenated with the node vector at the beginning.\(^6\) Such representation emulates how humans typically process the information in an MD&A document: starting with the firm’s overall business context as an initial memory, in a

---

\(^5\) The random walks are governed by two additional parameters \( p \) and \( q \). Together they decide whether the random walks should explore more distant regions in the network or focus on the local neighbors. We use the Python library *stellargraph* to train the node2vec model. We use the default parameters \( r = 10, l = 100, p = 1, q = 1 \).

\(^6\) GRU is a simplified variant of the popular Long-Short Term Memory (LSTM) network that combines the forget and add gates of LSTM into a single update gate (see [68] for an introduction of LSTM). We find the LSTM encoder not as effective and takes much longer to train in our task.
sequential manner, utilizing the memory to interpret each paragraph in the document, updating the memory with the most important information extracted from the paragraph, and finally generating a new representation of the paragraph. As shown in Figure 3, final representations, denoted as $Z' = [nv', z'_1, z'_2, ..., z'_m]$, are generated for the node vector and the paragraphs after the GRU.

While the first word-level attention layer identifies informative words in a paragraph, the second paragraph-level attention layer identifies important paragraphs when constructing a document representation. With a different set of parameters $W_a, b$, and $U_a$, we treat $Z'$ as $s_j$ when applying the same attention mechanism described by Equations (2)-(4). This paragraph attention layer assigns an attention weight to each paragraph (Equation (3)), and generates a document representation as the weighted sum of the node vector and the paragraph representations (Equation (4)).

5.5 Temporal Variation of MD&A (Module E)

Temporal variation of a firm’s MD&A documents is another textual input of our model. The temporal variation of firm $i$’s MD&A document at period $t$ is represented as a sequence \( R^{(i)}_t = [r_{(t-k, t-k+1)}, r_{(t-k+1, t-k+2)}, ..., r_{(t-1, t)}] \), where \( r_{(t_1, t_2)} \) is the WMD of a pair of MD&A documents at consecutive periods $t_1$ and $t_2$. We set $k = 5$ in our experiment, i.e., the model considers the extent of changes in a five-year window. In order to generate a final representation of $R^{(i)}_t$ that can capture patterns of temporal variations, we pass the sequence through a bidirectional GRU so that each element is linked with both its preceding and succeeding elements in the sequence.

So far, we have processed all textual predictors. As shown in Figure 3, the final presentations of node vectors, MD&A documents, and temporal variations are concatenated and sent to another dense layer to allow interactions among them.
5.6 Numerical Financial Indicator Predictors (Module F)

As the final input of our model, the numerical predictors first enter a dense layer to allow interactions among the predictors. Then the output of this hidden layer is merged with the textual predictors in another dense hidden layer to generate the final features for prediction. These features enter a single neuron with a sigmoid activation function to predict inside trading probabilities in accordance with prior neural network literature [56,97].

5.7 Comparison with Baselines Models

We now compare the attention-based deep learning model with a set of standard baseline models that use the bag-of-words representation of text documents. The bag-of-words model has been used in many other IS studies that use text as predictors (e.g., [40,87]). While the bag-of-words model is conceptually simple with good interpretability, our deep learning framework has two advantages. First, the word embedding layer learns the semantics of the words and represent them in a low-dimension space. Although singular value decomposition (SVD) on a document-term matrix can achieve the same goal (e.g., [103]), SVD is computationally expensive ($O(\min\{mn^2, m^2n\})$) for an $m \times n$ matrix. Moreover, SVD requires the entire document-term matrix to be fit into the memory. These hurdles make SVD impractical for a large corpus. A deep learning model, on the other hand, can leverage the parallel computing power of Graphics Processing Units (GPUs). It also supports mini-batch training, whereby the model can be trained using a stream of small subsets of the data. Therefore, our model is easily scalable.

The second advantage of our framework is that models using bag-of-words representations allocate the same weight to a word within all samples, while an attention mechanism assigns context-dependent weights locally to features within a sample. The context can be general semantics, as determined by surrounding words and sentences. The context can also be specific to
the problem itself, such as the position of the firm in the business proximity network. Moreover, as shown in our hierarchical attention network, attention mechanisms can be flexibly applied to any layer in the deep learning architecture. This allows the model to be interpretable at different levels of granularity.

Empirically, we consider two common algorithms to compare with our deep learning model using textual predictors: support vector machine (SVM) and Naïve Bayes. In a text classification problem, common baseline algorithms include Naïve Bayes, k-nearest neighbors (kNN), SVM, logistic regression, and decision tree [70]. We exclude kNN, decision tree, and logistic regression because they yield poor performance in our experiment. Distance-based algorithms such as kNN fail because the MD&A documents from the same firm are usually more similar to each other than to those from other firms. As a result, the nearest neighbors are usually historic documents of the same firm. Also, in our end-to-end approach (in contrast to the feature engineering approaches as in [50,60]), the models learn directly from the bag-of-words representations. Without careful regularization, logistic regression and decision tree tend to overfit the bag-of-words inputs, as it is possible to find a subset of words that fits every training sample perfectly (complete separation). Therefore, we follow prior work and use Naïve Bayes and SVM as baseline algorithms because of their ability to generalize well in large feature spaces [1,29,70].

Finally, we consider three baseline models with only numerical predictors: a logistic regression, a decision tree, and a neural network model. The neural network model is similar to Module F in Figure 3, except that the output of the hidden layer is not merged with the textual features but enters a single neuron with a sigmoid activation function to produce predictions.
6 EMPIRICAL RESULTS

6.1 Model Implementation and Evaluation

In this section, we present the results of our experiments with attention models and compare their performance with baseline models. All models are trained on an Ubuntu 14.04 server with a Xeon E5-2620 v4 CPU, 64 GB of memory, and a GTX 1080 GPU.

To measure model performance, we apply stratified five-fold cross-validation and report the average AUC obtained from test sets. Because insider trading cases are rare, using the classification accuracy to measure a model’s performance can be misleading. This is because the accuracy score assumes that Type I and Type II errors are equally costly. For regulators, however, prevention and deterrence are less costly than enforcement and punishment [89]. In other words, the cost of false negatives is higher than that of false positives. Although it is possible to assign a higher cost to the false negatives in the model’s objective function, such assignment would be arbitrary. Moreover, regulators are interested in more than a dichotomous prediction. The probability of insider trading can provide guidance for resource allocation.

Hence, we use AUC as a more flexible performance measure. AUC calculates the trade-off between the false positive rate and the true positive rate as the decision criterion (cutoff probability) varies. It can be used to evaluate a model’s overall ability without assuming a relative cost structure. The AUC score ranges from 0.5 to 1, with 0.5 indicating a baseline of random assignment of class labels, and 1 suggesting a perfect classification. We also report the averaged decile-ranking table on the test sets. We rank the company’s predicted insider trading probabilities into deciles, where

---

7 The stratified 5-fold cross validation (CV) splits the dataset into five equal folds (20% each). In each of the five iterations, 80% of data are used for training and 20% for testing. Each fold has the same proportion of positive cases and is used as the test set exactly once. Stratified CV is a recommended approach for estimating the out-of-sample performance of the model [58]. Because our dataset contains limited number of positive cases (less than 1%), $k$ cannot be very large as we want to make sure each fold includes enough positive samples. To further validate the choice of $k = 5$, we conducted an experiment following [58] with $k$ varying from 2 to 20. $k=5$ gives the best trade-off between the running speed and model performance.
the top decile contains the companies with the highest risk and the bottom decile contains firms with the lowest risk. The decile table is constructed by tabulating the cumulative percentage of actual insider trading firms in each decile. A high percentage in the high deciles implies better out-of-sample prediction accuracy.

We implement our baseline models using scikit-learn, an open-source library for Python. All neural network models are implemented using Keras, a Python API built on top of Google’s TensorFlow library. As discussed earlier, for the word-embedding layer, we initialize the word-embedding matrix \( W \) with pre-trained word vectors and then train it as a part of the models. The embedding dimension \( p \) is set to 100. We obtain pre-training word vectors with the entire text corpus using the skip-gram model [73]. The context vector \( U_a \) used in the attention layers is also set to 100 dimensions \( q \). The GRU for encoding paragraphs is configured with 100 cells. The bidirectional GRU for temporal variations has 3 cells to produce an output with 6 numbers. All hidden dense layers have 64 units and use ReLU activation function. In addition, to prevent overfitting, we apply a combination of L1 and L2 regularizations in all layers. Also, we apply batch normalization to each hidden layer to increase the stability of our model. Since our dataset is imbalanced, we incorporate class weights into the binary cross entropy cost function.

6.2 Model Performance

Table 4 reports the model performance for all the baseline models. When using numerical predictors, the best-performing model is the neural network model (denoted as Model C in table 5) with an AUC of 78.95%. As shown in Figure 3, the hidden layer with 64 units can allow the 8 numerical variables to have sufficient interactions with each other. As a result, it outperforms the logistic regression, which can be considered as a single neuron. When using text data as bag-of-words features, the Naïve Bayes and support vector machine models perform equivalently with an
AUC around 70%. Now we turn to the performance of our deep learning models as shown in Table 5. First, with only the text (word embeddings) as the input, the model achieves AUC of 75.75%, 5% higher than Naïve Bayes and support vector machine baselines. Then we add other predictors to this model one at a time to examine the incremental effectiveness of each predictor. With the temporal variations added, the model performance increases to 76.19%. Similarly, incorporating the node vectors also improves the AUC. With all the textual predictors, the model (denoted as Model I in table 5) achieves a decent AUC score of 76.81%. Overall, we find 10-K textual disclosures have high predictive power for insider trading.

Finally, when both textual and numerical predictors are included in the model, the AUC is further increased by more than 4% to reach 81.22%. This model also outperforms the best baseline model (Model C) by more than 2%. When comparing the decile table, this model has the highest 1st decile performance (45.63%). This means that if the regulators target the top 10% of the firms with the highest predicted insider trading probability, 45.63% of the positive cases will be included. Also, the sum of 1st and 2nd deciles in this model reaches 67.5%, exceeding the best baseline model (Model C) by 10%.

[Insert Table 5 about here]

6.3 Insights from Error Analysis

We further analyze the prediction errors made by the best model with numerical predictors (Model C) and the deep learning model with all textual predictors (Model I) to determine if numerical and textual predictors are complementary or substitutable. Using a decision threshold of 0.5, we plot the recall rate of each class in Figure 4, i.e. the percentage of positive or negative samples that are successfully recovered. The left chart shows that about 46% of positive samples
can be correctly identified by both models. The numerical model can recognize 25.65% of insider trading cases that are missed by the textual model, whereas the textual model recovers 11.82% of cases that cannot be identified by the numerical model. This seems to indicate that the numerical variables alone are more effective in recognizing positive cases. Yet it comes at the expense of a high false positive rate, as shown by the right pie chart. The numerical model misidentifies 19.15% negative samples as positives, but these samples can be correctly identified by the textual model. In comparison, the textual model has about 8% fewer false positives. Overall, we conclude that textual and numerical predictors are complementary because 1) the model with both inputs (Model J) outperforms all other models, and 2) each input can help identify a portion of cases which otherwise would be missed by the other.

[Insert Figure 4 about here]

6.4 Interpreting Attentions on Text

A key feature of attention models is their interpretability. After the attention models have been trained, we retrieved all attention weights of Model I on the testing dataset to understand how a prediction of a specific sample is reached. These weights offer a new way to analyze the context-specific attention on individual words. To illustrate, we use the industry that a firm is part of as the context. We first categorize the firms using the Fama-French 12 industry classification. Then, we select the top 25% cases based on their predicted probability, among which Finance, Healthcare, and Wholesale & Retail industries contain a large number of cases. For these three industry sectors, we collected the top 25% high-attention paragraphs and then top 25% high-attention words. There are 157 unique words. As shown in Table 7, these three sectors share 49 common high-attention words.

[Insert Table 7 about here]
First, the attention model quantifies a coherent set of words that captures the underlying incentives when managers plan to trade on their private information. Theory offers two hypotheses regarding managers’ disclosure [77,81]. On the one hand, managers may provide informative disclosure in advance of their transactions to reduce litigation risk. On the other hand, to maintain their advantage, they may exercise considerable discretion over specific facts and their interpretation of these facts. Many of the words shown in Table 7 are associated with these two hypotheses. When discussing the outlook of firms, insider managers are more likely to use negative language (loss, reduction, discontinue, or decrease) and uncertainty terms (change, issue, or defer), which may reduce the risk of private lawsuits. In particular, more emphasis on risk factors in the financial operation (issuance, repurchase, maturity, fund, exposure, instrument) and accounting management (tax, amortization, defer, balance sheet, SFAS, and repayment) may temper investor expectation and provide less ammunition for them to file lawsuits. Second, there is evidence of manager exercising discretion over disclosure quality. By providing words such as recognize, realize, believe, estimate, managers may attempt to convince investors and regulators that they are obeying the regulations while using their discretion to protect their private information. Overall, Table 7 provides real-world examples of how insiders strive to balance the “revealing” or “concealing” choice.8

To further identify the amount of discretion that managers choose to exercise, we use the attention model to explore how it varies by context (industry sectors). As shown in Table 7, the high-attention words in the finance section suggest some risk factors are industry-specific, such as

8 To the best of our knowledge, our study provides the first large-sample empirical evidence on how insider trading alters disclosure, which has important practical implications for regulators and financial professionals. Our results suggest that extracting information from disclosure narratives is a rich approach for capturing high-risk financial misconduct. Brown et al. [18] find a set of topics that predicts financial misreporting, while we delve deeper into subtle linguistic features and provide empirical evidence that is more align with the theory of insider trading.
loan, trading, mortgage, portfolio, derivative, and pre-payment. For the healthcare industry, the focus is more related to innovation and commercialization. Firms in the wholesale and retail industry have higher attention on customer, operations and logistics, e.g. customer, open, closure, integration, rent, reduce, and delivery. The finding confirms the importance of identifying industry-specific factors related to insider trading.

As we add hierarchical attention to the model, we can study how the model zooms in and out of a document. The model offers regulators the convenience of focusing their limited capacity on those high-attention paragraphs. In Appendix A (in the companion file), we provide a case study in which the model highlighted two specific paragraphs in a positive case. We also use topic modeling to provide a bird’s-eye view of all the paragraphs that the model highlights. Prior research [10,83] has shown that latent Dirichlet allocation (LDA) is able to quantify the main themes in a firm’s textual disclosure. We first train a ten-topic LDA model using all the paragraphs in the entire sample. Appendix B (in the companion file) contains the word cloud of the high-probability words in each topic. We assign the following labels to the ten topics: Liability, Accounting, Production, Regulation, Operation, Hedging, Economy, Clinical, Proper, Plant and Equipment (PP&E), and Partnership. We then calculate the valence of topics as their mixture percentage in the entire sample. For example, overall the MD&A sections have 23% of the discussions on Liability, and 15% on Accounting-related issues. Finally, we use the LDA model trained from the full sample to infer the topic mixtures in the highlighted paragraphs (top ten high-attention paragraphs from each positive sample).

Figure 5 compares the topic proportions from the full sample with the proportions in the high-attention paragraphs. Consistent with the motive to avoid ex-post-facto shareholder litigation, insider managers are more likely to emphasize future liability risk and regulatory risks in the
MD&A section. In addition, we consider two hypotheses derived from communication and psychology literature. First, consistent with the Information Manipulation Theory, it is clear from Figure 5 that high-attention paragraphs do under-report details from the internal aspects, such as accounting, production, and PP&E. The second hypothesis is derived from the finding in [19] that insider managers avoid references to themselves because they wish to minimize personal responsibility if the fraud and illegal insider trading are discovered. We find consistent evidence in support of this hypothesis from Figure 5. These insider managers do reduce the discussion in the MD&A on how they have evolved in the firm partnership management, in particular, using fewer words like partnership, advisory, affiliated, membership, and advisor.

[Insert Figure 5 about here]

6.5 Interpreting Node Embedding and Temporal Variations

Recall that we presented two competing theories about the effect of competitive environment: one suggests that competitors distract preventive measures and create more insider trading opportunities around M&As; the other suggests that competition erodes insiders’ information advantage. Attentions on node embedding vectors may offer a resolution. We examine network statistics for those firm with both high attention weights on node vectors and high prediction probability (both in top decile). In Figure 6, we plot a probability distribution of these firms’ node degrees, i.e., the number of competitors with similar products. We compare the degree distribution of high attention firms with that of all the firms in our sample. Figure 6 suggests that the high attention firms with possible insider trading activities have greater node degrees than an average firm. In other words, firms with more competitors are worth paying attention to.

[Insert Figures 6 and 7 about here]
Next, we examine the effect of temporal variations on the prediction of insider trading, again with Model I. Our hypothesis is that if temporal variations contribute to the prediction of insider trading, their patterns should be different in the predicted positive cases than in the other cases. To verify this hypothesis, we divide our samples into two groups: cases with top one decile prediction probability (containing 41.50% true positive cases) and all the others. For each firm-year, the temporal variation is represented as a sequence of moving similarities of the firm’s MD&A in the past 5 years (see Section 5.5). Figure 7 shows the average similarities of the sequence over the samples in each group. The differences between these two groups are substantial: 0.08, 0.11, 0.10, 0.10, and 0.08 for the past 1-5 years respectively. All of the differences are statistically significant at the 1% level. Therefore, we conclude that our prediction model captures the impact of temporal variations on the possibility of illegal insider trading. Our experiment result supports the proposal that temporal variations would allow researchers to better understand how firm changes are manifest in financial texts [22].

7 DISCUSSION AND CONCLUSION

Before Equifax, Inc., announced its data breach that affected 145 million consumers to the public in September 2017, three senior executives unloaded $1.8 million of company shares [21]. Such insider trading instances led to public and media outcry, but financial regulatory agencies often failed to prosecute offenders due to limited resources. In this paper, we propose a machine learning framework to predict insider trading. By combining a unique dataset of shareholders’ class action lawsuits against insider trading with U.S. public firms’ financial reports, we conduct a comprehensive analysis to predict whether a firm will be the target of litigation. Our key methodological contribution is that we develop an interpretable deep learning model to extract information from the textual disclosures for both prediction and interpretation. We find that text is
a useful information source for insider trading prediction. We also find models with both numerical features and financial text can give stronger out-of-sample prediction accuracy than the former alone. Moreover, we find that as theory predicts, contextual information extracted from a business proximity network and temporal changes of the documents adds predict power. We interpret the model by identifying context-specific words and topics that are informative using the attention mechanisms.

7.1 Implications for Literature

Our study offers several implications for IS literature and knowledge base. First, from a design science perspective [48], we create an innovative artifact that is relevant for both practitioners and researchers. We design our framework specifically for regulatory considerations by achieving high predictability, interpretability, and scalability. We develop a deep learning method for text by considering a much larger corpus than is the case with existing methods. Overall, we demonstrate that deep learning is a valuable tool for text-based predictive analytics in IS research.

Second, both qualitative and quantitative IS researchers are increasingly using textual data to generate new theory [12]. For theory development, being able to “zoom in and out” of data is crucial because it enables a richer understanding of both the details and the broad patterns [37]. The hierarchical attention mechanism introduced in this study lends itself well to such tasks. By design, an attention-based model provides a principled approach to learn the broad predictive relationship between the text and dependent variable while also supporting zooming in and out of either informative words or paragraphs that support the predictions. In addition, the business proximity network constructed from text and the attention weights associated with node embedding vectors can provide fresh theoretical insights on the growing networked economy.

Third, we explore a new fintech research opportunity with profound societal implications.
Gomber et al. recognized the “one problem, one data set, one publication” problem in fintech research that may hinder the collaboration between academic researchers and practitioners [41]. To this end, our response variable and the predictor variables are publicly available. The availability of such “ground truth” and large data sources provides a common ground for a continuing conversation between the IS academia and regulators who have an interest in data-driven approaches.

7.2 Implications for Practice

The predictability of insider trading allows regulators to target future prevention, enforcement, and investigation efforts. Regulators are currently resigned to waiting for a trader to execute a trade of a size and timing precision that would warrant investigative work. Predictive models are much needed for resource-constrained regulatory agencies to flag and monitor highly probable firms. For example, the SEC took an initiative to develop software to examine language use in financial reports for signs of fraud [31]. Our work represents a novel development in this front. To the best of our knowledge, very few published studies use deep learning techniques to help financial regulators and policymakers streamline and automate the process of curbing illegal insider trading. Our framework takes an end-to-end approach, uses publicly available data without time-consuming feature engineering, and achieves superior performance. Moreover, attention mechanisms applied in our framework provide interpretability of predictions. The high-attention features have high face validity and corroborate recent high-profile insider trading cases. For example, industry-specific high-attention words reveals insider trading risks are associated with drug approval. The attention on node vectors suggests that insider trading is more likely in the highly competitive industries where M&A activities take place. The features highlighted by

10 See for example, US v. Kluger, 722 F.3d 549 (3d Cir. 2013).
 attentions allow regulators to reason and validate high-risk cases while proactively carrying out preventative interventions on targeted firms or individuals.

Apart from offering a decision support system for regulators, the mere existence of a prediction model could deter future opportunistic insider trading. According to Becker’s classical economic theory of crime [11], one would weigh the costs and benefits before making the choice of whether to commit an offense. It follows that there are two ways to deter illegal insider trading by swaying a would-be offender’s expected utility: raising the severity of punishment and increasing the certainty of detection. Results from behavioral experiments suggest that the certainty of detection has an even stronger effect than severe punishment [74]. On this basis, we believe that because a prediction model can result in swifter and more certain enforcement actions, the accompanying deterrence effect could indirectly benefit social welfare.

Predicting insider trades can also help expose other forms of corporate fraud. Research has shown that opportunistic insider trading goes hand-in-hand with other forms of corporate misbehavior such as earnings management [6]. The magnitude of insider trading often reflects the gravity of other problems—many insiders brazenly trade on a fraud that they or their colleagues committed. Together, these misbehaviors can pose even greater threats to shareholders and result in enormous welfare loss in the entire economy. Therefore, being able to predict opportunistic insider trading provides a glimpse of ongoing and future corporate fraud. It would help prevent the next Enron or WorldCom collapse that shattered investor confidence and crashed the market.

7.3 Limitations and Future Research

Our study has several limitations. First, we predicted only whether a firm faces insider trading litigation; we did not consider the severity of the case. Future research could consider alternative
targets such as the amount of abnormal return. Second, we conduct a firm-level analysis of insider trading. An individual-level analysis is also possible using executives’ individual characteristics from corporate governance databases. Third, our research used a single textual data source (10-K). Our methodology can be combined with other data sources such as social media (e.g., [29]) or earnings call transcripts to improve prediction accuracy. Our work can foster future research in this important area. Finally, although our model is trained with a comprehensive dataset, as new 10-K filings are added, the model may need to be fine-tuned to fit the changing corpus. How to create an adaptive deep learning framework (in the same vein as Metafraud [1]) is another interesting topic for future work.
REFERENCES

text as a predictor of financial events. *Decision Support Systems*, 50, 1 (December 2010), 164–175.


64. Levine, R., Lin, C., and Wei, L. Insider trading and innovation. *The Journal of Law and
Figures and Tables

The complaint further states defendants engaged in the alleged wrongdoing so that they could profit by selling their personally held ANSI shares at artificially inflated prices. During the Class Period, ANSI insiders, including certain of ANSI’s officers, sold a total of 700,759 shares of ANSI stock for gross proceeds of $28,617,666.


Lacking in a reasonable and therefore materially false and misleading. On April 8, 2002, prior to the opening of the market, Accredo shocked the market by announcing that it was reducing its previously issued earnings guidance and that it was examining the adequacy of reserves for accounts receivables that it acquired in a recent acquisition. In response to this announcement, the price of Accredo Health common stock declined precipitously falling from $25.40 per share to as low as $13.76 per share, on extremely heavy volume. During the Class Period, Accredo insiders sold more than $12 million worth of their personally-held Accredo stock while in possession of the true facts about the Company.

(b) Joan Ferrari, et al. v. Accredo Health, Inc., et al.

The original complaint alleges that Allaire and its principals knew or recklessly disregarded that, in its attempt to capture a broad customer base, the Company had spread itself too thin and consequently, lacked the resources to develop and market its core infrastructure product, ColdFusion, while simultaneously developing a market presence for its newer products, Spectra and Jrun. It also claims that company insiders sold 407,000 shares of Allaire stock in August, at artificially inflated prices for proceeds of $13.3 million. Only after they thus cashed out a substantial portion of their investment did the Company reveal, on September 18, 2000, that they expected Allaire to post a third-quarter net loss in the range of $(0.05) to $(0.20) per share, below both analysts’ third quarter consensus estimates and the previous two quarters’ results.


Figure 1: Excerpts from Insider Trading Class-Action Lawsuits

Figure 2: Research Framework
Figure 3: Deep Learning Model Architecture
Figure 4: Error Analysis of Predictive Models with Different Inputs

Figure 5: Comparison of Topic Mixture (All Documents vs. High-Attention Paragraphs)
Figure 6: Number of Competitors and Firm Attentions

Figure 7: Effect of Temporal Variations on Insider Trading Prediction
Table 1: Description of Numerical Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Rationale and References</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGAT</td>
<td>Log (Total Assets)</td>
<td>Size is an important determinant for abnormal returns from informed insider trading [82]. Insiders from smaller firms are more likely to have greater abnormal returns. A possible reason is that smaller firms have greater information asymmetry due to less closely monitored by institutional investors and analysts [66].</td>
</tr>
<tr>
<td>RSIZE</td>
<td>Relative market capitalization of the firm. Calculated using the log market capitalization of the firm divided by the market value of all securities</td>
<td>We follow [66] and consider the relative size (along with the absolute size) of the firm to the market.</td>
</tr>
<tr>
<td>SIGMA</td>
<td>Stock volatility. Calculated as the standard deviation of the daily stock return observed over the previous three months</td>
<td>Chakravarty et al. (2004) show that informed trading in options market is prevalent [23]. Acharya and Johnson (2007) show that insiders could also profit from the credit derivatives markets. Activities on these markets are associated with asset price volatility [2]. When the firm is in a more volatile condition or has higher default risk, the risk of insider trading also rises. The effect, however, could be the opposite for insiders with limited capital as they are unable to diversify away risk [3].</td>
</tr>
<tr>
<td>EXCESS_RETURN</td>
<td>The firm’s log excess return on its equity relative to that on the S&amp;P 500 index</td>
<td>A widespread form of insider trading is accompanied by earnings management [6]. For example, managers can inflate stock prices through misstating earnings before selling their stockholdings. Literature also documents that insiders exhibit “buy low and sell high” behavior [34]. Following negative excess return compared to the market, insiders purchase stocks and intend to sell high based on future bullish private information.</td>
</tr>
<tr>
<td>LTAT</td>
<td>Total Liabilities / Total Assets</td>
<td>Managers in highly leveraged firms are more likely to undertake risky projects, thereby raising the volatility of the stock price, which can induce insider trading on the derivative market [30]. Leverage could also be associated with leveraged buyouts and wider credit spread on the CDS markets, both providing lucrative trading opportunities for insiders [3].</td>
</tr>
<tr>
<td>LOGSALE</td>
<td>Log (Sale)</td>
<td>Similar rationale to LOGAT.</td>
</tr>
<tr>
<td>MB</td>
<td>Market-to-Book Ratio. Calculated using the ratio of the market equity to the adjusted book equity to which we add a 10% difference between the market equity and book equity</td>
<td>MB measures if a firm is over or undervalued by the stock market relative to its accounting value. Insider trading is usually positively related to the MB ratio. Such behavior reflects insiders’ recognition of market misvaluation based on past performance and insiders’ superior knowledge of the real potential of future performance [20].</td>
</tr>
<tr>
<td>INDUSTRY</td>
<td>Fama-French 12-industry classification (dummy coded)</td>
<td>Insider trading can be affected by industry-wide information flow, through channels such as suppliers sharing information from common customers [7].</td>
</tr>
</tbody>
</table>
### Table 2: Summary Statistics of Numerical Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGAT</td>
<td>5.89</td>
<td>2.14</td>
<td>-6.91</td>
<td>5.86</td>
<td>14.99</td>
</tr>
<tr>
<td>RSIZE</td>
<td>-10.88</td>
<td>1.97</td>
<td>-17.2</td>
<td>-10.97</td>
<td>-3.46</td>
</tr>
<tr>
<td>SIGMA</td>
<td>0.03</td>
<td>0.12</td>
<td>0</td>
<td>0.01</td>
<td>15.19</td>
</tr>
<tr>
<td>EXCESS_RETURN</td>
<td>-0.02</td>
<td>0.27</td>
<td>-0.82</td>
<td>-0.03</td>
<td>3.74</td>
</tr>
<tr>
<td>LTAT</td>
<td>0.57</td>
<td>4.73</td>
<td>0</td>
<td>0.54</td>
<td>1410.5</td>
</tr>
<tr>
<td>LOGSALE</td>
<td>5.21</td>
<td>2.29</td>
<td>-6.91</td>
<td>5.21</td>
<td>12.98</td>
</tr>
<tr>
<td>MB</td>
<td>2.03</td>
<td>43.17</td>
<td>-4907.93</td>
<td>1.59</td>
<td>8803.86</td>
</tr>
</tbody>
</table>

### Table 3: Distribution of Insider Trading Cases by Year (846 cases in total)

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Firms</th>
<th>Insider Trading Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>3112</td>
<td>6</td>
</tr>
<tr>
<td>1997</td>
<td>5688</td>
<td>28</td>
</tr>
<tr>
<td>1998</td>
<td>5850</td>
<td>33</td>
</tr>
<tr>
<td>1999</td>
<td>5743</td>
<td>49</td>
</tr>
<tr>
<td>2000</td>
<td>5595</td>
<td>56</td>
</tr>
<tr>
<td>2001</td>
<td>5362</td>
<td>167</td>
</tr>
<tr>
<td>2002</td>
<td>4969</td>
<td>62</td>
</tr>
<tr>
<td>2003</td>
<td>4696</td>
<td>58</td>
</tr>
<tr>
<td>2004</td>
<td>4451</td>
<td>64</td>
</tr>
<tr>
<td>2005</td>
<td>4413</td>
<td>60</td>
</tr>
<tr>
<td>2006</td>
<td>4365</td>
<td>37</td>
</tr>
<tr>
<td>2007</td>
<td>4322</td>
<td>43</td>
</tr>
<tr>
<td>2008</td>
<td>4289</td>
<td>51</td>
</tr>
<tr>
<td>2009</td>
<td>4108</td>
<td>20</td>
</tr>
<tr>
<td>2010</td>
<td>3921</td>
<td>21</td>
</tr>
<tr>
<td>2011</td>
<td>3839</td>
<td>42</td>
</tr>
<tr>
<td>2012</td>
<td>3755</td>
<td>21</td>
</tr>
<tr>
<td>2013</td>
<td>3718</td>
<td>15</td>
</tr>
<tr>
<td>2014</td>
<td>3795</td>
<td>5</td>
</tr>
<tr>
<td>2015</td>
<td>3937</td>
<td>8</td>
</tr>
</tbody>
</table>

### Table 4: Performance of Baseline Models

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Baseline Models</th>
<th>Textual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>Numerical</td>
<td></td>
</tr>
<tr>
<td>AUC (%)</td>
<td>76.84</td>
<td>76.63</td>
</tr>
<tr>
<td>Deciles (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>37.94</td>
<td>37.70</td>
</tr>
<tr>
<td>2</td>
<td>17.26</td>
<td>18.32</td>
</tr>
<tr>
<td>3</td>
<td>15.60</td>
<td>13.83</td>
</tr>
<tr>
<td>4</td>
<td>6.97</td>
<td>10.17</td>
</tr>
<tr>
<td>5</td>
<td>8.75</td>
<td>6.38</td>
</tr>
<tr>
<td>≥ 6</td>
<td>13.47</td>
<td>13.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

49
Table 5: Performance of Deep Learning Models

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Deep Learning Models with Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F) Text only</td>
</tr>
<tr>
<td></td>
<td>(G) Text + Temporal Variation</td>
</tr>
<tr>
<td></td>
<td>(H) Text + Node Vector</td>
</tr>
<tr>
<td></td>
<td>(I) Text + Temporal Variation + Node Vector</td>
</tr>
<tr>
<td></td>
<td>(J) Text + Temporal Variation + Node Vector + Numerical</td>
</tr>
<tr>
<td>AUC (%)</td>
<td>75.75</td>
</tr>
<tr>
<td></td>
<td>76.19</td>
</tr>
<tr>
<td></td>
<td>75.96</td>
</tr>
<tr>
<td></td>
<td>76.81</td>
</tr>
<tr>
<td></td>
<td><strong>81.22</strong></td>
</tr>
<tr>
<td>Deciles (%)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>40.31</td>
</tr>
<tr>
<td></td>
<td>41.61</td>
</tr>
<tr>
<td></td>
<td>39.25</td>
</tr>
<tr>
<td></td>
<td>41.50</td>
</tr>
<tr>
<td></td>
<td><strong>45.63</strong></td>
</tr>
<tr>
<td>2</td>
<td>16.32</td>
</tr>
<tr>
<td></td>
<td>16.31</td>
</tr>
<tr>
<td></td>
<td>16.31</td>
</tr>
<tr>
<td></td>
<td>14.77</td>
</tr>
<tr>
<td></td>
<td>21.87</td>
</tr>
<tr>
<td>3</td>
<td>11.70</td>
</tr>
<tr>
<td></td>
<td>9.69</td>
</tr>
<tr>
<td></td>
<td>12.29</td>
</tr>
<tr>
<td></td>
<td>12.65</td>
</tr>
<tr>
<td></td>
<td>10.16</td>
</tr>
<tr>
<td>4</td>
<td>7.68</td>
</tr>
<tr>
<td></td>
<td>8.15</td>
</tr>
<tr>
<td></td>
<td>7.44</td>
</tr>
<tr>
<td></td>
<td>8.75</td>
</tr>
<tr>
<td></td>
<td>6.38</td>
</tr>
<tr>
<td>5</td>
<td>6.15</td>
</tr>
<tr>
<td></td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>9.34</td>
</tr>
<tr>
<td></td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>5.67</td>
</tr>
<tr>
<td>≥ 6</td>
<td>17.84</td>
</tr>
<tr>
<td></td>
<td>18.68</td>
</tr>
<tr>
<td></td>
<td>15.37</td>
</tr>
<tr>
<td></td>
<td>15.01</td>
</tr>
<tr>
<td></td>
<td>10.28</td>
</tr>
</tbody>
</table>

Table 7: Industry-Dependent High-Attention Words

<table>
<thead>
<tr>
<th>Industry</th>
<th>Industry Specific Words</th>
<th>Common Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>loan, policy, trading, mortgage, gain, hold, shareholder, client, allowance, transaction, portfolio, derivative, origination, income, consider, collateral, unrealized, reserve, manage, specific, default, prepayment, summary, guarantee, classify, present, residual, collect, strategy, adjustment, economic</td>
<td>recognize, loss, reflect, change, purpose, condition, tax, reduction, amortization, issuance, impact, repurchase, balance sheet, position, adopt, recognition, contract, issue, discussion, sfas, realize, require, maturity, discontinue, estimate, settlement, write, decrease, contractual, implement, review, fund, extent, company, determine, exposure, repayment, record, accounting, previously, instrument, continue, recent, believe, number, defer, accordingly, result, repay, accordance</td>
</tr>
<tr>
<td>Healthcare, Medical Equipment, and Drugs</td>
<td>delay, potential, inception, payment, milestone, collaboration, announce, progress, commence, likely, private placement, research development, commercialization, limitation, hospital, eliminate, research, eitf, intend</td>
<td></td>
</tr>
<tr>
<td>Wholesale, Retail, and Some Services (Laundries, Repair Shops)</td>
<td>open, initiative, improve, fiscal, integration, goodwill intangible, rent, reduce, customer, closure, delivery, earning, statement, method, negatively impact</td>
<td></td>
</tr>
</tbody>
</table>