7-2021

The First Sign: Detecting Future Financial Fraud from the IPO Prospectus

Lisa Spadaccini Anderson
University of Arkansas, Fayetteville

The First Sign: Detecting Future Financial Fraud from the IPO Prospectus

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy of Business Administration with a concentration in Accounting

by

Lisa Spadaccini Anderson
Clemson University
Bachelor of Science in Economics, 2010
Southern Methodist University
Master of Science in Accounting, 2012

July 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council:

_____
Kristian D. Allee, Ph.D.
Dissertation Director

_____
T.J. Atwood, Ph.D.
Committee Member

_____
Vernon J. Richardson, Ph.D.
Committee Member

Abstract

In this study, I examine whether it is possible to predict future financial statement fraud using disclosure content prior to the fraud. Specifically, I employ a machine learning algorithm to construct a unique measure based on the lexical cues embedded within a firm's first public disclosure, the Management's Discussion and Analysis section of the S-1 filing, during the Initial Public Offering process. I use this measure to predict whether a firm that is not already committing fraud will commit fraud within five years of the Initial Public Offering (IPO) that results in an Accounting or Enforcement Release (AAER). I find there is information within the S-1 filing that is useful in the prediction of out-of-sample fraud. Additionally, I find that the measure performs better than both benchmark measures from prior literature and a new measure using quantitative information, when using information available at the S-1 date. Furthermore, the lexical cues measure performs well in predicting fraud relative to the benchmark measures even after updating the benchmark measures with misstated annual filings to aid their (but not my measure's) fraud detection abilities. I find that my new measure is not limited to only predicting AAER based misconduct, but that the out-of-sample results hold when using an alternate sample based on 10(b)-5 filings as well as a comprehensive set of quantitative variables. Lastly, my measure identifies firms more likely to manage earnings to meet/beat analyst forecasts, firms who experience higher levels of information asymmetry around earnings announcements within the five years following the IPO, and has some predictive ability over future abnormal returns.

Acknowledgements

I thank my dissertation committee-Kris Allee (chair), TJ Atwood, and Vernon Richardson-for helpful direction and comments. I also thank workshop participants at the University of Arkansas and California Polytechnic University.

Table of Contents

1. Introduction

In this study I examine whether it is possible to identify fraud using the nuances of textual disclosure in corporate filings prior to the fraud. Financial statement fraud is a threat to the efficiency of capital markets as it impairs the trust between corporations and capital market participants (Amiram et al. 2018). The implications of fraud range from reducing the usefulness and value relevance of financial reporting to creating large scale losses of $74-$180 billion dollars in shareholder value if left unchecked (Healy and Wahlen 1999; Abarbanell and Lehavy 2003; American 2016). Therefore, the timely identification of fraud is beneficial to investors and regulators in that it likely preserves a significant amount of investor capital. A large body of literature identifies or predicts instances where financial statement fraud is presently occurring, as opposed to identifying instances in which fraud is likely to occur in the future (Cecchini et al. 2010; Dechow et al. 2011; Larcker and Zakolyukina 2012; Purda and Skillicorn 2015; Hoberg and Lewis 2017). Even the most recent studies focus on the actual manipulated financial statements (or other disclosures) during the SEC-identified manipulation period without considering the pre-manipulation financial statements as a standalone information source (Bao et al. 2020; Brown et al. 2020).[1] My study differs from prior research by examining whether there is information content in pre-fraud disclosure indicative of a higher likelihood of future fraud, enabling investors to make alternative investment decisions at an earlier point in time and providing regulators with an opportunity to increase monitoring on firms with the greatest likelihood of future misconduct.

[1] The identified manipulation period refers to the period named by the SEC in the applicable Accounting and Auditing Enforcement Release. While not all financial statement manipulation events are investigated and therefore some remain undetected, I follow prior literature and consider the periods identified as manipulation periods compared with non-manipulation-periods for the purposes of this study, consistent with prior literature (Dechow et al. 2011; Bao et al. 2020; Brown et al. 2020).

The ability to identify fraud-firms prior to their fraud is important for several reasons. First, as enforcement and regulatory monitors, such as the Securities and Exchange Commission (SEC), are subject to resource constraints, it enables them to maximize their limited resources and potentially prevent the occurrence of fraud (Bao et al. 2020). Furthermore, targeted monitoring has the potential to limit the extent of the impact of financial manipulation on investors and the period over which financial statements are manipulated. Second, providing internal monitors (i.e., boards) with a list of attributes associated with future manipulation likelihood enables them to watch for and potentially limit the extent of fraud within their own organization. Additionally, the timely detection or prevention of fraud preserves investor capital as firms suffer negative returns when financial statements are restated and investigated for financial misconduct (Palmrose et al. 2004; Karpoff et al. 2008; Gande and Lewis 2009).

In order to provide a timelier assessment of fraud risk, I examine the qualitative content in the earliest public firm disclosure of the Initial Public Offering (IPO) process: the initial S-1 filing. I use this setting for three reasons. First, the IPO period is a time when the firm is heavily monitored, potentially constraining firms who may otherwise manipulate their disclosures (Stoughton and Zechner 1998; Ritter and Welch 2002, Hanley and Hoberg 2012). Therefore, the IPO setting provides the opportunity to examine how a firm responds to pressure in a high-stakes firm disclosure environment (Hughes and Thakor 1992; Drake and Vetsuypens 1993; Lowry and Shu 2002), while still being subject to high levels of monitoring.

Second, the initial S-1 filing is the first public disclosure event for the firm, therefore the filing has not yet gone through an amendment process, allowing for more nuanced linguistic cues to be potentially embedded within the filing (Hanley and Hoberg 2012). I focus on a linguistic-based approach as opposed to using quantitative financial information as my research question

posits the relevant information content will consist of subtle nuances or cues contained within the textual firm disclosures (Loughran and McDonald 2016). Additionally, prior literature finds that IPO-firms lack a history of tangible information (i.e., past positive earnings streams, revenues, or even dividends), and therefore many studies analyze the qualitative content of the S-1 and find it to be incrementally informative (Hanley and Hoberg 2010; Arnold et al. 2010; Ferris et al. 2013; Loughran and McDonald 2013; Brau et al. 2016).

To examine the nuanced ways in which written communication is indicative of future fraud, I create a novel measure, *fraud_cues*, using the Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) section of the initial S-1 filing. Hanley and Hoberg (2010) demonstrate the informativeness of the MD&A section, finding this section has the highest absolute value of residuals from a regression estimating standard content (Hanley and Hoberg 2010). Using the content of S-1 filings taken from EDGAR, I employ a machine learning approach using a support vector machine (SVM) to create *fraud_cues*. Specifically, I utilize a "text classifier" machine learning model that employs a deep learning layer within the vectorization stage. I use this SVM approach for several reasons. First, the advantage of using deep learning vectorization in this setting is the ability to collectively examine sentence construction, word co-location, and individual word choice to classify the text vectors based on relevancy (Guo et al. 2016). Thus, *fraud_cues* does not limit classification to a pre-identified bag-of-words and allows for a nuanced analysis of the information content within the MD&A of the S-1 filing.[2]

Second, popular press has posited that sophisticated investors are basing investment decisions on charismatic language intended to mislead investors as opposed to firm fundamentals

[2] A key limitation to using a "bag-of-words" approach is that even when using n-grams the method does not capture the context in which words appear and misses syntactic and semantic properties (Ling et al. 2015; Ramesh et al. 2015).

(Galloway 2019). SVM performs the data transformations necessary to discern whether nuanced information (e.g., charismatic language) can be used to identify firms as future-fraud firms. Third, this approach utilizes publicly available financial statement information and can therefore be replicated, which has been cited as a challenge by prior textual analysis literature (Loughran and McDonald 2016).

I first evaluate the validity of my measure by showing that *fraud_cues* has high within-sample detection of financial manipulation. In additional analyses, I show that my measure is not simply identifying underpricing in the IPO period or directly proxying for a "bag-of-words" approach. Further analyses assess the usefulness of *fraud_cues* in identifying future fraud firms out-of-sample. I compare the performance of my measure against four financial statement fraud prediction measures from the literature and one new measure. As there are currently no studies that examine the independent information content within pre-manipulation period disclosure, I compare my measure with the most similar available measures from prior literature in order to provide context for the predictive accuracy of my model. Namely, I compare my approach to the Dechow et al. (2011) *f-score*, the Bao et al. (2020) raw ensemble method *RUSBoost*, the Hoberg and Lewis (2017) measure of abnormal disclosure, and the Brown et al. (2020) *topic* analysis measure. I also create a new measure, *IPO_f-score*, using a set of quantitative variables likely associated with pre-fraud firms. The primary difference between my measure and the benchmark measures is they all utilize manipulation-period disclosures in their original construction and therefore predict the likelihood that a manipulation is presently occurring.[3] I compare my measure to the benchmark measures both at the time of the IPO and again at the time of the first

[3] While I acknowledge using these measures in a setting for which they were not intended is subject to limitations, they provide evidence that my measure is not subsumed by pre-existing fraud prediction measures.

misstatement to determine how my measure of the likelihood of future financial manipulation compares to extant measures of current financial manipulation.

I find the out-of-sample performance results for my measure, *fraud_cues,* are significantly better than the performance of a random guess with an AUC of 0.694 for predicting future AAER's, providing evidence that the linguistic cues contained within the S-1 filing are predictive of whether a firm will manipulate its financial statements in the future. When comparing the six measures using only the information available within the S-1, *fraud_cues* has the highest AUC, showing prior measures of fraud prediction cannot be used in place of my measure in the S-1 setting, and that developing a distinct measure in this setting is incrementally informative. I also show that qualitative information outperforms quantitative information in the pre-fraud period. These results provide evidence of the predictive benefit of using linguistic cues instead of contextual, raw financial, or financial ratio-based manipulation predictors. When comparing the measures using the S-1 information to construct *fraud_cues* and then updating the other measures to include the first misstated 10-K filing, I find *fraud_cues* continues to outperform the benchmark measures at predicting AAERs even without being updated relative to the other measures. This suggests the information in pre-fraud disclosures can yield better fraud prediction than the use of financial ratios or raw data items.

To exploit the differences in the time periods of the measures, I consider the incremental usefulness of *fraud_cues* in conjunction with the benchmark prediction measures identified by prior literature. The combined results produce an AUC of 0.752. I also combine the measures most informative in the pre-fraud period, and produce an AUC of 0.793. The joint analysis results suggest combining linguistic cues, content, abnormal disclosure, raw financial data, and financial based ratios increases overall predictive ability.

Next, as AAER's are relatively rare, not all the firms identified by *fraud_cues* as having a high likelihood of future fraud will receive one. As such, I consider whether my fraud prediction measure has implications for future earnings management and future information asymmetry as well. Prior literature finds earnings management contributes to analyst forecast errors, reduces shareholder value, is associated with lower returns following economic events, and generally misleads stakeholders about the fundamental value of the firm (Abarbanell and Lehavy 2003; DuCharme et al. 2004; Chou et al. 2006; Healy and Wahlen 1999). Therefore, the ability to predict firms that will engage in earnings management enables monitors to preserve not only shareholder value but the usefulness and value relevance of financial reporting. Specifically, I examine whether *fraud_cues* identifies firms more likely to engage in earnings management to meet or beat earnings targets. I find the firms identified by my measure as exhibiting a greater likelihood of receiving an AAER are also more likely to *just* meet or beat earnings targets, consistent with prior research examining earnings management.

Next, I examine whether an increased likelihood of financial statement fraud is associated with higher levels of information asymmetry. As private information is obtained by informed traders, the firm will experience lower volume around earnings announcements as liquidity investors exit the market, resulting in higher bid-ask spreads (Copeland and Galai 1983). I find my measure identifies firms with significantly higher bid-ask spreads and lower trading volume around earnings announcements during the 5-year period following the S-1 filing, suggesting these firms have lower liquidity (Affleck-Graves et al. 2002; Chae 2005). Collectively these findings show that the same textual cues that are associated with future AAERs are *also* associated with both earnings management and information asymmetry measures that result in undesirable

outcomes for regulators and capital market participants by reducing value relevance and resulting in lower liquidity.

Additionally, I consider whether my measure has predictive ability over future returns. I do not find results, potentially due to the relative rarity of AAER's subsuming any abnormal returns. I consider an alternative measure of fraud, 10(b)-5 lawsuits, that is slightly less rare both to examine abnormal returns and assess the robustness of my original measure. When re-calculating *fraud_cues* based on the S-1 filings of firms that are named in section 10(b)-5 lawsuits, within the 5-year period following the IPO, I find that the measure based on cues again outperforms the benchmark measures from prior literature, with an AUC of 0.561. While the predictive power of this measure is lower than the AAER counterpart, this measure does have some predictive power over abnormal returns. I find that, after constructing yearly portfolios based on linguistic cues, the cumulative abnormal return of the lowest fraud-likelihood portfolio is positive and significant. This result holds after considering risk factors. Additionally, I consider whether including a larger population of financial statement line items from the S-1 filing improves the predictive accuracy of a quantitative fraud prediction model and find that the measure based on *fraud_cues* continues to have higher predictive accuracy.

This study provides the first evidence that the rich qualitative information in the IPO prospectus can be useful for more than just examining underpricing, but also in the prediction of (1) future financial statement manipulation, (2) earnings management, and (3) information asymmetry. My study contributes to the extant literature at the intersection of both textual analysis and financial manipulation prediction. I first extend this literature by documenting that textual analysis of the first publicly available, pre-manipulation period, qualitative data can be used to predict financial statement manipulation without requiring the use of manipulated financial

information. While prior literature suggests firms will commit fraud when there is opportunity, motivation, and the pressure to do so, my study suggests a persistent element overlooked by prior literature that results in a predictable likelihood measure.

Second, my study connects the spectrum of financial misconduct activities by training a model on AAERs and using it to predict earnings management and other financial misconduct. While other studies (Dechow et al. 2011; Bao et al. 2020) have examined the relationship between accruals-based earnings management and the likelihood of future AAERs, my study is the first to do so using textual analysis, showing that verbal cues are an important avenue for researchers to consider.

Third, I expand on the burgeoning research using machine learning approaches in the fraud prediction literature (Cecchini et al. 2010; Goel et al. 2010; Purda and Skillicorn 2015; Bao et al. 2020; Bertomeu et al. 2020; Brown et al. 2020). My study extends prior research by using machine learning to predict the future likelihood of manipulation and does so using nuanced textual disclosures, answering the call from Loughran and McDonald (2016, p.1188) to examine whether there are "subtle cues in managements' [disclosures] that computers can discern better than [market participants]."

Lastly, the assessment of the future likelihood of fraud is of significant interest to capital market participants, creditors, governmental regulators, and researchers. The ability to predict the likelihood of future fraud, up to five years prior to the actual manipulation, should be of significant interest as it is in alignment with regulatory goals and investors' incentives. The SEC has already begun to use text-based tools to detect anomalies in firm disclosures, therefore this research is in tandem with their regulatory objectives (Eaglesham 2013; Bauguess 2018).

2. Literature Review and Hypothesis Development

A large literature examines whether it is possible to detect financial statement manipulation using quantitative information. Given the lack of any formal theoretical predictors of fraud, Green and Choi (1997) were of the first to employ an algorithmic approach, ad-hoc selecting five financial ratios and using an NN model to predict fraud using a hold-out sample. While this model achieved a 74 percent accuracy rate for detecting manipulation, the analysis performed was within-sample, and the NN approach did not provide insight into which ratios were the most useful and where the cutoffs occurred. After the passage of Statement of Auditing Standards (SAS) 99 in 2002, external auditors were tasked with providing reasonable assurance that financial statements are free from material fraud. As such, auditors developed fraud checklists and began to consider the theoretical underpinnings of fraudulent reporting. Early literature finds the financial items on the audit fraud checklist are useful in detecting fraud but rely on proprietary data sources and are therefore not generalizable (Bell and Carcello 2000; Asare and Wright 2004).

With both generalizability and increased theorization as primary concerns, Dechow et al. (2011) formally theorize why certain financial statement information is associated with manipulation, such as the misstatement of receivables to improve sales growth. In their study, Dechow et al. (2011) utilize not only financial statement ratios but other market-related, off-balance sheet, and nonfinancial variables. Their final measure, *f-score*, is widely regarded as one of the most powerful fraud prediction tools and achieved an accuracy rating of 71.53 percent when tested on a sample of 29,159 firm-years. However, Dechow et al. (2011) achieve a significant amount of their overall statistical power from including firm-years post manipulation in their sample to provide insight into how financial statements revert post-manipulation. Overall, the Dechow et al. (2011) study provided a richer understanding of the financial statement accounts

and other metrics most likely to be affected by fraud, but *f-score* is limited in its ability to predict this fraud out-of-sample.

In order to predict fraud out of sample, algorithmic approaches became more common. Concurrent with the Dechow et al. (2011) study, Cecchini et al. (2010) utilize a financial kernel (FK) by examining restated financial accounts to extract the line items most commonly restated. This approach to create FK results in the use of many of the same line items found within the *f-score* ratios. Cecchini et al. (2010) utilize a support vector machine learning algorithm that separates fraud firms from non-fraud firms based on changes to the FK measures over time. Bertomeu et al. (2020) analyze whether accounting variables or complimentary audit variables better capture whether a firm misstates earnings. They find the array of accounting variables studied has the greatest importance on the detection of serial misstatements and consider whether their measure has look-ahead predictive power over a 1-year and 2-year horizon. Bertomeu et al. (2020) find lower predictive power in their look-ahead tests, and it is important to note that they do not exclude serial misstatement firms in their sample, thus their measure is not capturing the pre-manipulation period exclusively. Bao et al. (2020) employ an ensemble learning method but depart from both the benchmark *f-score* and FK measures by choosing to not make ex-ante predictions involving financial ratios and instead use raw accounting data. Bao et al. (2020) construct a list of raw financial variables based on the raw data items used to construct the ratios used by Cecchini et al. (2010) and Dechow et al. (2011), resulting in a twenty-eight-variable measure, *RUSBoost*. When *RUSBoost* predicts fraud out-of-sample, it outperforms both benchmark measures. Despite the consideration of raw data items, Bao et al. (2020) find significantly less predictive power when considering all the raw data items available for the firms examined, showing that theoretical guidance is important in fraud prediction.

One potential concern with relying on quantitative information is the difficulty in identifying a comprehensive list of attributes of financial manipulation. Another concern is that while relying on certain line items found in misstatements is useful in predicting financial manipulation, there are a wide variety of financial attributes available to perpetrators, enabling them to potentially conceal fraud in new ways (Cecchini et al. 2010; Allee et al. 2021). As such, a separate but related stream of literature examines the ways in which qualitative information, such as textual disclosures, can be used to detect financial manipulation. Textual disclosures have been found to contain incrementally informative content to quantitative disclosures in a variety of mediums. More complex annual reports have been shown to indicate poorer performing firms and lead to greater analyst dispersion and lower forecast accuracy (Li 2008, Lehavy et al. 2011). Press releases with a more optimistic tone overall are associated with a higher return on assets (Davis et al. 2012), while press releases with a more optimistic tone relative to their accompanying MD&A are indicative of managers attempting to strategically increase market perceptions when reporting bad news (Davis and Tama-Sweet 2012). In conference calls, tone dispersion and managerial affective states have shown to contain incremental information about future firm performance (Allee and DeAngelis 2015; Mayew and Venkatachalam 2012). Collectively, this research suggests there is nuanced information content within qualitative disclosure, and that this information concerns not just what is said, but how it is said.

With respect to fraud, recent literature has taken advantage of the incremental informativeness of qualitative disclosure content to compare manipulation-period and non-manipulation period disclosure to identify a new list of fraud attributes. Cecchini et al. (2010) find that classifying financial text into dictionaries enables a VSM measure to detect manipulated statements from non-manipulated statements 75 percent of the time. Firms that manipulate their

financial statements are also shown to use fewer general references, lower non-extreme positive tone, and more optimistic language (Rogers et al. 2011; Larcker and Zakolyukina 2012). While these papers hypothesized ex-ante which linguistic attributes perpetrators might employ, Purda and Skillicorn (2015) utilize an SVM model to construct word lists and find these lists outperform pre-defined measures, a key difference from the collective findings of the quantitative literature stream.

The thematic content of qualitative disclosure is also informative, as fraud firms under-report detail and overstate performance and growth potential (Hoberg and Lewis 2017; Brown et al. 2020).[4] The Brown et al. (2020) study provides evidence that thematic content is informative relative to the other commonly studied features used within textual analysis. Their study examines proxies for length, complexity, variation, readability, tense, word choice, and emphasis to ultimately conclude that "what" managers say is more important than how they say it. Lastly, in computer science, syntactic stylometry, or the study of linguistic patterns independent of context, has been useful in detecting deceptive practices (Zhou et al. 2003; Feng et al. 2012). In their 2010 study, Goel et al. (2010) use stylometry to detect manipulation in misstated financial statements, using not just the features examined by Brown et al. (2020) but also a deeper linguistic analysis using markers such as pronouns, prepositions, and conjunctions to achieve an 89.51 percent accuracy rating when detecting misconduct.

While existing research has made progress using qualitative disclosure to identify a series of theoretical, thematic, and stylistic attributes associated with and predictive of fraud, I consider a fundamentally different research question by examining pre-fraud disclosure as opposed to fraud-period disclosure. Literature that examines financial statement fraud documents an

[4] Thematic content is defined following Brown et al. (2020) as the topics identified within qualitative disclosure.

association with weaknesses in firms' internal governance and monitoring that creates an opportunity to commit fraud (Beasley1996; Beasley et al. 2000). Following that intuition, certain firms who may otherwise decide to manipulate their financial statements are unable to as they are constrained by monitors. As such, there exist three firm subgroups: a) fraud firms, b) non-fraud firms, and c) would be fraud-firms subject to prohibitive constraints. Groups b) and c) are systematically different, with group c) having a greater likelihood of committing future fraud than group b) upon experiencing a change in the level of monitoring. In order to examine whether these systematic differences exist, I focus on linguistic patterns in the pre-fraud period for two primary reasons. First, stylometric differences have been identified as one of the strongest fraud predictors and second, quantitative information is more likely to be heavily scrutinized by external monitors and is primarily backward looking (Bonsall et al. 2014). Therefore, I focus on qualitative linguistic cues and state my hypothesis as follows:

Hypothesis 1: There are linguistic cues within qualitative disclosures issued prior to the occurrence of fraud that improve the prediction of fraud, relative to quantitative features and content-based information.

I next examine whether linguistic cues from disclosures issued prior to the occurrence of fraud improve the prediction of future earnings management. Earnings management is the process of altering financial reporting to either mislead investors about actual underlying performance or to favorably influence contractual outcomes that rely on financial reporting (Healy and Wahlen 1999; Schipper 1989; Lo 2008). Earnings management therefore has the potential to erode the value relevance of financial information and result in disclosure that is not relevant nor faithfully represented, a violation of FASB Concept Statement No. 8. Earnings management also impacts shareholder value as firms that manipulate earnings are more likely to be sued (DuCharme et al.

2006). Additionally, earnings management can result in higher analyst forecast errors as analysts are not able to consistently identify firms that manage earnings (Abarbanell and Lehavy 2003; Burghstahler and Eames 2006). Lastly, firms that manage earnings around security offerings have lower one-year aftermarket returns than their counterparts and firms that manage earnings around mergers have higher post-merger announcement lawsuits (Chou et al. 2006; Teoh et al. 1998).

Prior literature examines and finds proxies for earnings management are significant predictors of financial statement fraud (Beneish 1999; Dechow et al. 2011). Building on this logic, I consider the relationship between linguistic cues within qualitative disclosure and future earnings management. If earnings management is also the result of systematic differences between firms, I expect to find that firms more likely to commit fraud are also more likely to engage in higher levels of earnings management. I state my hypothesis as follows:

Hypothesis 2: There are linguistic cues within qualitative disclosures issued prior to the occurrence of fraud that are associated with higher levels of future earnings management.

Next, I consider the relationship between future fraud and future information asymmetry. Information asymmetry exists when insiders have a deeper insight into the firm's information environment than outsiders, creating an information disparity (Aboody and Lev 2000). Information asymmetry has negative implications for capital markets as it is negatively correlated with trading volume as liquidity traders exit the market (Foster and Viswanathan 1990). As information asymmetry spreads to the secondary market, larger traders will hold fewer shares, increasing the risk for non-institutional investors and lowering the equilibrium price of the firm (Lambert et al. 2011). Lastly, this decrease in liquidity trading volume can amplify adverse selection, leading to higher bid-ask spreads (Copeland and Galai 1983). Therefore, information asymmetry can be seen

as a type of corporate misconduct, potentially resulting in lower liquidity. Specifically, I consider whether linguistic cues issued prior to egregious misconduct (i.e. fraud) can be used to predict firms with more asymmetric information. I state my hypothesis as follows:

Hypothesis 3: There are linguistic cues within qualitative disclosures issued prior to the occurrence of fraud that are associated with higher levels of future information asymmetry.

Lastly, a key component of the theory underlying Hypothesis 1 is that timelier prediction of misconduct is useful, and thus incrementally informative to existing measures used to detect it. While the identification of firms in the pre-fraud period based on linguistic cues allows for more targeted monitoring, this identification also has potential implications for asset pricing. If investors fail to price the nuanced linguistic cues contained within the firm filings, or the lack thereof, they will potentially misprice the stock of firms that commit fraud and not realize the benefits of firms that are not likely to commit future fraud. That is, if and when a fraud is revealed, this mispricing will correct, and the investors will experience predictable losses. As such, a portfolio trading strategy exploiting this mispricing should result in positive abnormal stock returns. Conversely, Fama and French (1993) find that it is possible to explain the variation in the cross-sectional stock returns using a set of risk factors. As such, it is possible that any pricing anomaly identified through the aforementioned portfolio strategy might be better explained as a lack of an adjustment for risk. If investors are in fact pricing the likelihood of future fraud, the required rate of return would be higher as the likelihood of future fraud increases. As such, a portfolio trading strategy would have to account for risk to show the abnormal returns generated are the result of mispricing. Therefore, I state my hypothesis as follows:

Hypothesis 4: A risk-adjusted trading strategy based on the likelihood of future financial manipulation generates positive abnormal stock returns.

3. Setting, Data, Sample Selection, and Methodology

3.1 IPO Setting

In order to test my hypotheses, it is necessary to utilize a setting where firms are subject to a high level of external monitoring and scrutiny. For this reason, I focus on disclosure during the IPO process. The IPO setting is uniquely appropriate for several reasons. First, the IPO process is a time when firms must balance disclosing less negative information to not leave money on the table (Skinner 1994; Healy and Palepu 2001; Ritter and Welch 2002) with increasing disclosure to avoid litigation risk (Hanley and Hoberg 2012). This creates a dichotomy as the IPO event and its related disclosures are a time when information asymmetry between firms and investors is especially high (Loughran and McDonald 2013) but the IPO firm is also heavily monitored by management and underwriters concerned with potential litigation.

Second, the S-1 filing issued during the IPO process is the first publicly available firm disclosure, serving as the earliest possible fraud predictor. The IPO prospectus must only contain three years of audited income statements and two years of audited balance sheets (with only two years required for emerging growth companies). This lack of a long history of tangible information has resulted in the consideration of qualitative textual disclosures when examining underpricing (Hanley and Hoberg 2010; Hanley and Hoberg 2012; Loughran and McDonald 2013).

Third, it has been suggested by the popular press that IPO firms have been using charismatic words and phrases to avoid focusing on financial information (Galloway 2019). If firms have incentives to manipulate disclosures to increase investment, yet are limited by monitors and litigation risk, how will they react when they are past this heavily scrutinized firm event? As

such, the IPO setting provides a unique opportunity to evaluate whether firms respond in discernable ways that provide predictive evidence of their future propensity to manipulate financial statements when they are not monitored as heavily as they are during the IPO process.

3.2 Data and Sample Selection

I begin my sample selection with the original S-1 filing for the 5,048 completed U.S. IPO's from 1996-2015 taken from the intersection of the Audit Analytics and Compustat databases. I start with 1996 as it is the first full year after Congress enacted the Private Securities Litigation Reform Act of 1995, which codified "loss causation", requiring the plaintiff to carry the burden of proving loss causation for all actions arising from section 10(b) of the 1934 Act, and fundamentally changed the definition of fraud. The sample period ends in 2015 to allow for a five-year post-IPO financial manipulation window.[5] To remain consistent with the prior literature examining the textual components of S-1 filings, I eliminate American Depositary Receipts, unit issues, Real Estate Investment Trusts, Special Purpose Acquisition Companies, and financial industry firms with SIC codes between 6000-6999. I use a web algorithm to extract the MD&A section and eliminate any observations without this section, which results in 2,454 completed U.S. firms remaining the final sample.

I next match the sample based on CIK to classify an S-1 filing as belonging to a firm that will commit fraud. In order to identify frauds, I utilize the Accounting and Auditing Enforcement Release (AAER) database, used by Dechow et al. (2011). The AAER database is compiled by the University of Southern California and provides details on firms subject to SEC enforcement actions. To remain in my sample, the AAER must not name the IPO year and the period named

[5] During my sample period, the JOBS Act was signed into law (April 5, 2012). The Act allows Emerging Growth Companies, firms with less than $1 billion in annual revenues, to file draft IPO registration statements. As these statements are required to be filed publicly within 21 days before a road show, the JOBS Act does not interfere with the sample selection of this study.

during the AAER must not exceed the 5-year window post IPO-filing. This is done to concentrate analyses on firms in the period just after they complete their IPO, as this period is potentially less heavily monitored. Additionally, firms are most likely to have a similar organizational team in place as during the S-1 process. Research finds 35% of IPO firm CEOs leave within the first five years indicating that the majority of my sample retains its CEO throughout the entire sample period (Mitsuhashi and Welbourne 1999). As all AAER observations list the dates of the impacted financial statements, I remove any observations where the S-1 filing was listed as an impacted statement, to focus exclusively on pre-fraud qualitative content.

As shown in Table 1, there were 68 IPO-firms that received an AAER within the 5-year post IPO period between 1995-2015. The overall percentage of fraud is 3 percent. Also shown in Table 1, the percentage of firms that received AAERs has been decreasing over time. This downward trend in AAERs during the sample period is consistent with statements from SEC officials that post-financial crisis resources were diverted away from accounting fraud investigations in order to focus on financial crisis cases such as those involving Ponzi schemes (Ceresney 2013), and not necessarily that fraud has been decreasing over time. A limitation of this study is that there are potential fraud firms during the sample period that were not identified by the SEC, which would result in an overstated Type 1 error rate.

3.3 Linguistic Cues Measure

My measure to classify the qualitative properties of financial statement manipulation within the S-1 filing, *fraud_cues*, is based on a text classification approach. Text classification is a classic component of Natural Language Processing (NLP), that involves classifying text into tags (e.g. Fraud/ Non-Fraud). I choose to utilize a support vector machine learning algorithm as this measure has been widely used within accounting literature and has been shown to outperform both

logistic regressions and naïve-bayes approaches (Cecchini et al. 2010; Goel et al. 2010; Purda and Skillicorn 2015). Consistent with prior literature, the first step in machine learning based textual analysis is vectorization, where the S-1 text is transformed into its numerical representation based on feature extraction (i.e. word choice, frequency, and co-location). My study departs from this prior literature through the use of an advanced deep learning architectural layer during the feature extraction step, to determine which linguistic attributes are most relevant to fraudulent firms based on the data from within the MD&A section of the S-1 filing. Collectively, extant literature suggests tone, word choice, grammatical voice, and distribution of parts of speech to assist in the detection of fraud (Goel et al. 2010, Rogers et al. 2011, Larcker and Zakolyukina 2012, Purda and Skillicorn 2015. The proprietary deep learning architectural layer I use allows me to simultaneously examine all of these linguistic properties. The proprietary nature is a result of using a commercially available machine learning product, MonkeyLearn, shown by extant literature to out-perform other web-based service models (Basmmi et al. 2020). MonkeyLearn is a web-based service model that enables a user to upload a set of training text for analysis and performs vectorization and machine learning analysis. The model can then process a set of test data to output classification as well as accuracy statistics.

I utilize a standard support vector machine (SVM) model that uses a resampling procedure to fit the model, consistent with prior literature (Witten and Frank 2005, Hastie, Tibshirani, and Friedman 2003, Larcker and Zakolyuikna 2012). The resampling procedure splits the training data into four subsets of equal size to perform cross-validation and then trains the final model using the entire dataset. The cross-validation measure allows the model to produce accuracy metrics on its within-sample predictive ability that are an average across subsets, reducing the potential bias from using just one training session. I train the model with a training sample of 1996-2000, with 2001-

2015 serving as a hold-out sample. Table 1 reports summary statistics related to training and testing samples. The training sample represents approximately 80% of the AAER firms while the testing sample represents about 20%. This was done in order to ensure the machine learning algorithm had a sufficient amount of data to train the AAER model.[6] Relying on a hold-out sample enables *fraud_cues* to serve as an out-of-sample prediction score as opposed to a detection score.

To address the class imbalance problem caused by the relative rarity of fraud, I perform an under-sampling technique following prior literature (Perols 2011; Perols et al. 2017). When uncorrected for, relative rarity results in an algorithm biased toward the majority class (Maloof 2003). To avoid this bias, I use Multi-Subset Observation Undersampling (OU) which keeps the total number of fraud observations in my training sample constant and randomly selects an equal number of non-fraud observations without replacement, following Perols et al. (2017).

Table 2 presents baseline statistics related to within-sample fraud detection. As discussed previously, the SVM classifier was trained using four-fold cross-validation. These statistics indicate that SVM performed better without adjusting for stop words in my test period, which appears reasonable considering the measure is constructed to take all words into account when creating the cues-based measure. The table presents four detailed accuracy results: (1) accuracy, (2) F1 score, (3) precision, and (4) recall. Accuracy represents the percentage of firms that were predicted with the correct tag while F1 score is the combination of precision and recall. Precision and recall are specific to the *Fraud* and *Non-Fraud* samples. For the Fraud sample, the percentages are calculated as follows: precision = (the number of true Fraud firms/ number of firms guessed as Fraud firms) and recall = (the number of true Fraud firms/ number of overall Fraud firms). The

[6] I replicate this process using a 10-year training period and a 10-year test period and note that accuracy metrics on within-sample predictability yield consistent inferences with the 5-year training period and 15-year test periods chosen.

results are in alignment with those found in Goel et al. (2010) with the exception of the recall rate of 74% significantly outperforming their recall rate of 41.5%. Alternatively stated, this means that of the total number of Fraud firms in the training population, my measure predicted 74% correctly. The comparability of these statistics suggests my SVM model is achieving high within-sample accuracy and performing consistently with models from prior literature. Additional analyses related to measure validation are reported in Section 5.

3.4 *Benchmark Measures*

I consider three benchmark measures from prior literature to provide context for my measure, as well as two measures constructed specifically for the IPO period. First, I consider three measures designed to predict financial statement fraud using fraud period disclosure. Despite the lack of direct comparability, it is important to examine whether *fraud_cues* is identifying pre-fraud attributes that are different from pre-existing fraud period attributes.

3.4.1 f-score

The Dechow et al. (2011) *f-score* measure is considered the benchmark fraud prediction measure and serves as the starting point for comparing the out-of-sample performance of my measure. I calculate the *f-score* for the firms in my sample following measure 1 of table 9 of Dechow et al. (2011):

$$AAER = B_0 + B_1 RSST\ Accruals_{i,t} + B_2 Change\ in\ Receivables_{i,t} + B_3 Change\ in\ Inventory_{i,t} + B_4 \%\ Soft\ Assets_{i,t} + B_5 Change\ in\ Cash\ Sales_{i,t} + B_6 Change\ in\ ROA_{i,t} + B_7 Actual\ issuance_{i,t} + B_8 BTM_{i,t} + e. \quad (1)$$

Where t is the IPO year. All variables are defined as in Dechow et al. (2011), and where *AAER* is a variable equal to one if the observation is an IPO firm that receives an AAER within 5-years of the IPO date for the IPO date sample and zero otherwise. I first estimate equation (2) for the testing window of 1996 to 2000 and use the estimated coefficients to construct *f-score* for my out-of-

sample period of 2001 to 2015. I note that I omit two variables found in Dechow et al. (2011), lagged market-adjusted stock returns and the existence of operating leases. There are no lagged market-adjusted stock returns available at the S-1 period as the firm was not publicly traded, and the inclusion of the operating leases variable drastically reduces the sample size.[7] If firms begin to manipulate their financial statement information in advance of committing fraud, it is possible the *f-score* will have predictive power in the pre-fraud period. In order to compare my pre-fraud disclosure measure with the fraud-period detection ability of the *f-score*, I re-estimate equation (1) with AAER equal to one if the observation year is the first year covered by the AAER and is equal to zero otherwise for a sample of firm-years beginning with the IPO year and ending with the first year covered by the AAER.

3.4.2 RUSBoost

For my second benchmark measure I construct *RUSBoost*, following Bao et al. (2020), by applying an ensemble learning algorithm to raw financial data items. I begin with the list of 28 raw financial data items selected by Bao et al. (2020). A full list of these variables can be found in Panel A of Table 3A, where *RUSBoost* is indicated in the Benchmark column. I use the same training and test year conventions as Bao et al. (2020), using all years prior to the IPO-year as training data with a 2-year gap (1996 to 1999 as training for 2001, 1996 to 2000 for 2002, etc.). I follow the same methodology as Bao et al. (2020), using an ensemble learning method (as opposed to SVM) combined with a random under-sampling (RUS) technique to address the problem of class imbalance.[8] I calculate *RUSBoost* using the financial information within the S-1 to remain consistent with *fraud_cues*. Similar to the *f-score*, if firms begin to manipulate financial statements

[7] Inferences remain quantitatively and qualitatively similar when operating leases are included.
[8] In un-tabulated analyses, I recompute *RUSBoost* using an SVM model and the same training and testing period as *fraud_cues*, inferences reported in Section 4 remain unchanged.

line items in advance of committing fraud, or if these line items are systematically different between the two pre-fraud groups (b) and (c), I expect *RUSBoost* to have out-of-sample predictive power. I also re-calculate *RUSBoost* using the first financial statement containing the fraud, to better conform to the original setting the measure was developed for.

3.4.3 Abnormal Disclosure

For the third benchmark measure I calculate abnormal disclosure, *AbDisc*, following Hoberg and Lewis 2017. This study was the first to consider whether the thematic content of manipulated disclosures differed from non-manipulated disclosures. They find firms engaging in fraud produce abnormal disclosure relative to their peers, after controlling for size, age, fixed effects, and controls based on quantitative fraud models. I follow the cosine similarity approach seen in models (1) through (4) of Hoberg and Lewis (2017) to replicate a measure of cosine similarity between fraudulent and non-fraudulent S-1 filing MD&A sections. I calculate the measure using the base vocabulary from 1996 to 2000 to fit the model, reserving 2001 to 2015 as the out of sample test period. If future fraud firms produce abnormal disclosure during the IPO period, I expect *AbDisc* to have predictive power in the pre-fraud period. I re-estimate the measure using all annual filings starting with the S-1 through the first fraudulent filing to assess the out-of-sample predictive power of *AbDisc* in the fraud-period.

3.4.4 Topic

For the fourth benchmark measure I use *IPO_topic*, following Brown et al. (2020). *IPO_topic* is a measure based on the narrative content within the S-1 filings that builds off the measure used in Hoberg and Lewis (2017). In order to use *IPO_topic* as a comparison to linguistic cues, I perform LDA topic analysis on the MD&A section of the S-1 filings over rolling 5-year windows following the original study. This methodology accounts for changes in the relevant

topics over time. Consistent with the methodology employed in Brown et al. (2020), I select 31 topics from each window. After selecting the topics, I follow models (1) and (2) from page 260 of Brown et al. (2020) to apply the estimated coefficients from a regression of *AAER* on vectors of *topic* for each window of my test sample. The construction of *IPO_topic* is distinct from the measures discussed thus far as it is the first to be re-thought for the IPO period. While the S-1 has been used to construct the other benchmark measures, the construction of *IPO_topic* allows me to specifically test whether pre-fraud thematic content differs between future-fraud and future non-fraud firms. The comparison of *IPO_topic* and *fraud_cues* allows for the direct comparison of whether pre-fraud topic content or pre-fraud linguistic attributes are better able to predict future fraud.

3.4.5 IPO f-score

While *f-score* and *RUSBoost* are quantitative measures useful in establishing whether pre-fraud disclosure content is different from fraud period disclosure content, these measures are not directly comparable to *fraud_cues*. In order to examine whether quantitative information is useful in predicting future-fraud, I select a set of quantitative variables that I hypothesize to be associated with future-fraud.

$$AAER = B_0 + B_1 RSST\ Accruals_{i,t} + B_2 Change\ in\ Receivables_{i,t} + B_3 Change\ in\ Inventory_{i,t} + B_4 \%\ Soft\ Assets_{i,t} + B_5 Startup_{,t} + B_6 VC_{i,t} + B_7 Age_{i,t} + B_8 Hot\_IPO_{i,t} + e. \quad (2)$$

I include *RSST accruals*, *Change in Receivables*, and *Change in Inventory* as higher accruals have been documented in the pre-manipulation period (Dechow et al. 2013). I include *% Soft Assets* as firms with greater net operating assets have more accounting flexibility, making it easier for these firms to commit future fraud (Barton and Simko 2002). The next group of variables proxy for the monitoring environment surrounding the IPO. *Startup* is an indicator variable equal to one if the

annualized pre-IPO revenues are less than $1 million while *VC* is an indicator variable equal to 1 if the firm has venture capital backing per the SDC database (Bochkay et al. 2018). Compared to other investor types, venture capitalists have more expertise in funding start-ups and impose high monitoring. As such, I would anticipate a negative coefficient on B_5 and B_6. *Age* is the number of years between the founding date and the IPO date (Loughran and Ritter 2004). The longer a firm exists, the less need for external monitoring, thus I anticipate a positive coefficient on B_7. Lastly, *Hot_IPO* is a dummy variable equal to one if a firm went public during a period when the inverse of the industry median IPO book-building period falls in the top two quintiles (Wang et al. 2010). A shorter book-building period indicates a hot IPO market, and more optimism in the firm's prospects. As such, I would expect firms in a hot IPO market have less monitoring as the market is more confident in their ability to generate positive returns, and therefore expect a positive coefficient on B_8. In order to compute *IPO_f-score*, I first estimate equation (2) for the testing window of 1996 to 2000 and use the estimated coefficients to construct *f-score* for my out-of-sample period of 2001 to 2015.

3.5 Receiver Operating Characteristics

Following prior research, I use the area under the receiver operating characteristics (ROC) curve to evaluate the out-of-sample prediction ability of *fraud_cues* (Larcker and Zakolyukina 2012; Bao et al. 2020; Brown et al. 2020). The ROC curve plots the two-class true positive rate against the false positive rate at various threshold settings. The area under the curve (AUC) reports the area under the ROC curve, which is the probability that the measure will rank a manipulation observation correctly. With no known information, the probability of correctly assigning an observation as a misconduct is 0.5. The AUC report values of 0 – 1.0 depending on the accuracy of the measure. Following prior literature, I calculate the AUCs using pooled data for all test years.

3.6 Earnings Management

For my market-based tests of Hypothesis 2 and Hypothesis 3, I utilize the test sample from Table 1 of 1,406 firms as I require every firm to have a *fraud_cues* score. For every firm in my sample, I collect annual filings starting with the first post-IPO filing and ending 5 years following the IPO date to remain consistent with my AAER convention. For each firm-year, I merge my sample with the Center for Research in Security Prices (CRSP) database to collect return variables, the Compustat database to collect firm specific variables, and analyst variables from I/B/E/S. I first empirically test Hypothesis 2 using the following linear regression where the unit of observation is firm-year:

$$MBE(inv)_{i,t} = \beta_0 + \beta_1 Fraud\_cues_i + \beta_2 Size_{i,t-1} + \beta_3 BM_{i,t-1} + \beta_4 Lev_{i,t-1} + \beta_5 AnalystCount_{i,t} + \beta_6 AnalyistDisp_{i,t} + \varepsilon_{i,t} \quad (3)$$

Where *MBE(inv)* is an indicator equal to 1 if the firm failed to meet or beat (missed) analyst targets, and zero otherwise. As Hypothesis 2 posits firms identified as having a higher likelihood of future fraud will engage in higher levels of earnings management, I expect a negative coefficient on β_1. I include *Size*, the log of market equity value, to control for size. I include the book-to-market ratio, *BM* as growth affects investors' response to earnings performance. Additionally, I include leverage, *Lev,* which is measured as long-term debt relative to assets, to control for a firms' proximity to debt default. Lastly, I include *AnalystCount*, calculated as the number of analysts issuing forecasts, and *AnalystDisp*, a measure of analyst dispersion. I run pooled ordinary least squares (OLS) regressions with one dimensional clustering by firm, including year fixed effects. To examine whether the firms are managing earnings to beat earnings forecasts, I perform a second analysis where I limit my sample to only those firms that meet or beat analyst forecasts. I estimate the following model:

$$MBE_{i,t} = \beta_0 + \beta_1 Fraud\_cues_i + \beta_2 Size_{i,t-1} + \beta_3 BM_{i,t-1} + \beta_4 Lev_{i,t-1} + \beta_5 AnalystCount_{i,t} + \beta_6 AnalyistDisp_{i,t} + \varepsilon_{i,t} \quad (4)$$

Where all variables are defined above and *MBE* is a continuous variable measured as the difference between the actual EPS and the median EPS forecast for values of zero or greater. Higher values of *MBE* indicate that the firm beat earnings targets by a larger magnitude, therefore if firms are managing earnings to narrowly meet or beat analyst targets' I expect to find a negative coefficient on β_1. Both my definition of *MBE(inv)* and the refinement to create *MBE* for values greater and equal to zero are in alignment with the prior literature that examines earnings management (Doyle et al. 2013).

3.7 Information Asymmetry

For my market-based tests of Hypothesis 3, I consider whether firms with a higher likelihood of future fraud have higher future information asymmetry by examining the bid-ask spreads around earnings announcements (Glosten and Harris 1988; Coller and Yohn 1997). In order to examine this relationship, I estimate the following model:

$$AbSpread_{i,t} = \beta_0 + \beta_1 Fraud\_cues_i + \beta_2 Size_{i,t-1} + \beta_3 BM_{i,t-1} + \beta_4 Lev_{i,t-1} + \beta_5 AnalystCount_{i,t} + \beta_6 AnalyistDisp_{i,t} + \varepsilon_{i,t} \quad (5)$$

Where *AbSpread* is calculated as the average daily bid-ask spread over trading days 0 through 1 relative to the earnings announcement date minus the average daily bid-ask spread over the trading days [-41, -11], following Chi and Shanthikumar (2017), and all remaining variables are as previously defined. I calculate the average daily bid-ask spreads as the difference between the quoted offer price and the quoted bid price, divided by the midpoint and multiplied by 100 (Bushee 2010). If firms more likely to manipulate in the future produce future disclosures with a higher degree of information asymmetry, I expect a positive coefficient on β_1. Additionally, I examine

whether there is abnormal trading volume around the earnings announcement by estimating the following model:

$$AbVol_{i,t} = \beta_0 + \beta_1 Fraud\_cues_i + \beta_2 Size_{i,t-1} + \beta_3 BM_{i,t-1} + \beta_4 Lev_{i,t-1} + \beta_5 AnalystCount_{i,t} + \beta_6 AnalyistDisp_{i,t} + \varepsilon_{i,t} \quad (6)$$

Where all variables except *AbVol* are defined above and *AbVol* is calculated as the stock's average daily trading volume over days 0,1 relative to the earnings announcement minus the average over the trading days [-41, -11]. Daily trading volume is calculated as the log of the product of the closing price and the number of shares traded. If firms that are likely to engage in future fraud also reveal more pre-announcement private information, there will be less new information revealed during the earnings announcement window and lower relative volume, therefore I predict a negative coefficient on β_1.

3.8 Market Returns

To test Hypothesis 4, I examine whether *fraud_cues* is associated with future excess returns. In this specification, I estimate cross-sectional regressions of excess returns on the likelihood that a firm will commit future fraud and other firm-level characteristics:

$$Return_{i,t} = \beta_0 + \beta_1 Fraud\_cues_i + \beta_2 Size_{i,t-1} + \beta_3 BM_{i,t-1} + \varepsilon_{i,t} \quad (7)$$

Where all variables are defined above and *Return* is calculated both yearly and cumulatively over the five years following the IPO.

4. Results

4.1 Descriptive Results

Table 3, Panels A and C present summary statistics of the financial variables for the firm-years in my sample at both the IPO date and fraud date respectively. Panel B details the differences between the Fraud and Non-Fraud observations at the IPO date. With respect to the financial ratios used to construct *f*-score, firms that will eventually receive AAERs have greater changes in

receivables and a larger percentage of soft assets at the time of their S-1, but the remaining variables are generally consistent between the two samples. The percentage of soft assets is noteworthy as the variable represents firms with more accounting flexibility, suggesting these firms have the ability to report positive earnings surprises using this line item (Barton and Simko 2002). Additionally, firms that commit future fraud are less likely to have venture capital backing, consistent with the discussion of VC's acting as strong monitors. Lastly, firms in hot IPO markets are more likely to commit future fraud, consistent with my ex-ante expectations. As most of the financial variables examined were originally conceived with the fraud-period in mind, it appears reasonable that several of the variables are not statistically different between the two samples. When examining the 28 raw variables used in the *RUSBoost* measure, there are no statistically significant differences between the fraud and non-fraud sample. When examining the firms at their fraud date in Panel D, more of the variables exhibit statistically significant differences between the two groups. This appears reasonable as this is the time-period in which the variables were originally hypothesized to detect manipulated reporting.

I examine Pearson correlations between the variables, un-tabulated for brevity. Correlations significant at the 10 percent level are in bold. The overall lack of significant correlations between the financial variables and the occurrence of future fraud suggests historical qualitative measures of manipulated financial reporting are less informative in the pre-fraud period. The significant relationship between *fraud_cues* and *RSST_accruals*, *change in inventory*, *percentage of soft assets*, and *change in return on assets* suggests the cues measure is identifying similar firms as the Dechow et al. (2011) *f-score* measure. However, as the *f-score* was not originally intended to apply to the pre-manipulation period I do not offer a discussion on the sign of the correlation. *Fraud_cues* is also significantly correlated with several of the raw financial

variables used to calculate *RUSBoost*, but as many of the raw variables are not correlated with the misconduct cues measures it suggests the cues measure is also picking up different attributes of manipulation.

4.2 Tests of H1

4.2.1 Benchmark Comparison – IPO Period

In table 4, I present out-of-sample tests of the predictive role of *fraud_cues*. The AUC statistic of 0.694 is greater than a random classification measure. This result indicates that linguistic cues within the S-1 filing are predictive of future fraud. Next, I present results comparing *fraud_cues* to the benchmark measures (*f-score*, *RUSBoost*, and *AbDisc)* to examine whether financial variables identified by prior literature are also useful in predicting future-fraud. *P-values* reported for the benchmark measures are based on a two-tailed t-test between each measure and *fraud_cues*. Panel A represents AUC statistics for measures fitted with information from the S-1 filing. The AUC indicates that the *f-score*, *RUSBoost*, and *AbDisc* measures are not statistically greater than a random guess (AUC = 0.50). This result shows provides evidence that conventional measures used to predict fraud using fraud-period disclosure are less informative in the pre-fraud period, necessitating the use of new measures such as *fraud_cues*.

The AUC for *IPO_topic* reflects a predictive gain of 8 percent over a random measure. Despite this predictive gain, *fraud_cues* is statistically better at predicting future fraud ($p < 0.05$). Lastly, the AUC for *IPO_f-score* is 0.652, which makes it the second-best predictor of future fraud following *fraud_cues* which remains weakly statistically superior ($p < 0.09$). The finding that linguistic cues exhibit incremental predictive power over quantitative variables and topical content supports Hypothesis 1. In their best *topic* measure designed specifically for their setting, Brown et al. 2020 achieve a predictive accuracy AUC of 0.680, while Bao et al. 2020 achieve a predictive

accuracy AUC of 0.725 for their best *RUSBoost* measure. Therefore, with an AUC of 0.694, *fraud_cues* is in alignment with prior literature in terms of predictive ability, with the key distinction of identifying fraud up to 5 years in advance. Additionally, I find that my measure of *fraud_cues* predicts 11 out of the 13 (84 percent) future frauds correctly, which suggests the results of my measure are economically significant.

4.2.2 Benchmark Comparison – Misconduct Period

Panel B of Table 4 presents out-of-sample tests of the predictive role of the benchmark measures at the fraud date. The AUC statistic for *fraud_cues* remains unchanged as this variable is not updated. The AUC for *f-score* is lower than at the IPO date, potentially attributable to both high Type 1 and Type 2 errors. *AbDisc* remains statistically indistinguishable from a random guess in both panels. *RUSBoost* performs significantly better when calculated at the fraud date, with an AUC of 0.601. As such, my next analyses examine whether the predictive power of *fraud_cues* combined with the predictive power of the benchmark measures are incrementally predictive over the stand-alone measures.

4.2.3 Joint Predictability

Table 5, Panel A, presents the logistic regression of the determinants of future fraud using the fraud measures suggested by prior literature. In column (1), I present the results with just the *f-score* measure, column (2) includes the addition of my *fraud_cues* measure, column (3) further includes *RUSBoost*, column (4) adds *IPO_topic*, column (5) adds *AbDisc*, and column (6) is a combination of all four measures. In column (2), *fraud_cues* remains weakly significant, ($p < 0.10$), when included in a model with *f-score* and significant, ($p < 0.05$), when included with all other measures in column (5). The table also presents AUC statistics for all columns. Adding *fraud_cues* to the *f-score* measure results in a predictive gain of 28.9 percent. The inclusion of

RUSBoost, column (3), only represents a predictive gain of seven percent while *topic*, column (4), produces a gain of 23 percent, and *AbDisc*, column (5), does not represent a predictive gain. Combining all four measures results in an AUC of 75.2%, which suggests *fraud_cues* combined with pre-existing measures achieves a higher prediction rate.

Table 5, Panel B, presents the logistic regression using the three measures designed for the pre-fraud period. Topical content was originally hypothesized to change with fraudulent disclosure, but was refit using topics from the S-1 filing, therefore this measure is hybrid and included in both Panel A and Panel B. Column (1) presents the results with the *IPO_f-score* measure, Column (2) includes *fraud_cues*, Column (3) adds *IPO_topic*, and Column (4) combines the three measures. Of the four specifications, a combined *fraud_cues* and *IPO_f-score* measure has the greatest predictive power, with an AUC of 0.793. These results suggest that a combination of linguistic cues and quantitative information increases overall predictive ability, with no gains from the inclusion of topical content.

4.3 Tests of H2

The results from tests of Hypothesis 1 imply that linguistic cues taken from the S-1 filing are useful in the prediction of future financial fraud. As the instances of fraud remain relatively rare, Hypothesis 2 extends the usefulness of the measure further to investigate whether *fraud_cues* is associated with future earnings management. Table 6 presents the results from estimating Equations (3) and (4). Consistent with my predictions, the coefficient on *fraud_cues* is negative and highly statistically significant ($p < .01$) for both equations. Table 6, Column (1), shows that firms with a higher likelihood of receiving an AAER are less likely to miss analyst earnings targets. Table 6, Column (2) examines only firms that meet or beat analyst earnings targets and shows that firms with a higher likelihood of receiving an AAER are more likely to *just* meet or beat earnings

targets.[9] These results show that a model trained to predict out-of-sample future financial fraud also identifies firms more likely to engage in future earnings management. This is important for two reasons. First, this finding strengthens the link between earnings management and fraud and second, this finding shows how *fraud_cues* has broader usefulness that extends beyond just the prediction of fraud.

4.4 Tests of H3

Hypothesis 3 posits that firms more likely to receive a future AAER are also more likely to engage in other forms of disclosure-based misconduct, which will result in increased information asymmetry. Table 7 presents the results from estimating Equations (5) and (6). In Table 7, Column (1), the coefficient on *fraud_cues* is positive and highly statistically significant ($p < .01$). As *fraud_cues* increases, the bid-ask spread increases, suggesting higher levels of information asymmetry between investors. This finding is consistent with that of Keung et al. (2010) who document that earnings management is associated with higher mean average spreads and greater information asymmetry. In Table 7, Column (2), the coefficient on *fraud_cues* is negative and highly statistically significant ($p < .01$). As *fraud_cues* increases, abnormal trading volume decreases, which suggests a weaker response to the earnings announcement driven by pre-announcement private information leakage. Collectively, the findings in Table (7) and Table (8) show that *fraud_cues* can be used to identify a wide array of financial reporting misconduct and that there is an association between the different types of misconduct. These findings also show that the identification of future fraud is especially important for regulators and all capital market participants as *fraud_cues* identifies an array of future misconduct.

[9] My results are robust to an alternative dependent variable, *JMBE*, calculated as a truncated sample of firms with earnings surprises between -4 cents per share and +3 cents per share, following Doyle et al. 2013.

4.5 Tests of H4

Panel A of Table 8 presents the results of Equation (7) with yearly returns as the dependent variable. With the exception of year three, the coefficients on fraud cues are not statistically different from zero. Panel B of Table 8 presents the results using cumulative returns as the dependent variable, consisting only of firms that have returns every year during 5-year period. Collectively, these results suggest that *fraud_cues* is not useful in predicting future returns. This finding is consistent with Larcker and Zakolyukina (2012), who do not find significant excess returns for their AAER sample. I believe that the lack of results is due do the rarity of AAER's, making it hard to detect any potential mispricing correction by the market. As linguistic cues in the pre-fraud period are unable to predict future returns, I do not have cause to further examine whether the relationship can be explained by underlying risk factors.

5. Supplemental Analysis

In this section, I report several additional analyses, including an alternative measure of future fraud, validation of the original measure, and an additional benchmark measure based on an exhaustive set of financial statement line items.

5.1 Lawsuit Sample

To check the robustness of my results, I re-estimate my results based on whether a lawsuit under Section 10(b)-5 was brought against the firm. Many studies have used lawsuits as a proxy for the presence of corporate financial fraud (Srinivasan 2005, Peng and Roell 2008, Wang et al. 2010). I focus on Section 10(b)-5 lawsuits as opposed to Section 11 lawsuits as Section 10(b)-5 lawsuits require proof of intent and therefore have a higher threshold. I obtain the population of 10(b)-5 lawsuits from the Stanford Class Action Clearinghouse (SCAC) and require the same time-period constraints as the AAER dataset. I next hand collect information from the SCAC to classify

each lawsuit as “settled”, “dismissed”, or “ongoing”. To remain in my sample, the lawsuit must be tagged as either “settled” or “ongoing”. All lawsuits labeled as “dismissed” are excluded. Table 10 reports the distribution of 10(b)-5 lawsuits. There are 313 IPO-firms that were sued under Section 10(b)-5 within the 5-year post IPO period between 1995-2015.

Table 10 reports the out-of-sample performance for the 10(b)-5 lawsuit sample. At the IPO date, the area under the ROC curve for the *fraud_cues* measure based on lawsuits is 0.561. For the lawsuit sample, the AUC’s for *f-score*, *AbDisc*, *IPO_topic*, and *IPO_f-score* are 0.452, 0.313, 0.348, and 0.485 respectively. None of these prediction measures outperform a random guess of 0.50. With an AUC of 0.552, *RUSBoost* is the only other model that performs better than a random guess. At the fraud date, *RUSBoost* outperforms *fraud_cues* with an AUC 0f 0.691. This appears reasonable as *RUSBoost* was originally conceived to predict fraud using the manipulated financial statements. Despite lower predictive accuracy relative to the AAER measure, this alternative specification based on 10(b)-5 lawsuits shows that it is possible to adapt the cues-based measure to another sample, providing additional evidence that there is incrementally informative content found using lexical cues to predict future manipulation.

Table 9 documents the larger percentage of lawsuit frauds compared with AAER frauds. I re-estimate Equation (7) to take advantage of this increased fraud sample. Panel A of Table 11 presents the results with yearly returns as the dependent variable. Similar to the results reported for the measure based on AAERs, the coefficients on *fraud_cues* in columns (1) through (5) are not statistically different from zero. Panel B of Table 11 presents the results with cumulative returns as the dependent variable. The coefficients on *fraud_cues* are negative and significant ($p < 0.01$) for the cumulative five-year period following the IPO. These results show that linguistic fraud cues are associated with returns.

Next, I investigate whether the predictive power of *fraud_cues* can be explained by underlying risk factors. I estimate time-series regressions of excess returns on Fama-French (1993) factors and the Carhart (1997) momentum factor:

$$ExRet_t=\beta_0+\beta_1(Mkt\text{-}RF)_t + \beta_2 SMB_t+ \beta_3 HML_t+\beta_4 WML_t+e_t \quad (8)$$

I first sort firms into decile portfolios each year based on their *fraud_cues* score, rebalancing yearly. I then calculate value-weighted returns of each portfolio and subtract T-bill rates to get the excess returns.[10] I run three-factor plus momentum model time-series regressions by fraud cues deciles and present the results in Table 12. Confirming my earlier findings, the main conclusion from Table 12 is that the predictive ability of lawsuit *fraud_cues* remains after adjusting for the factors. The result is driven by the low portfolio, firms that I predict are less likely to receive 10(b)-5 lawsuits. It is possible that firms receive 10(b)-5 lawsuits from shareholders based on declining performance as suggested by Hanley and Hoberg (2012), and *fraud_cues* has predictive ability in identifying these firms. I also note that a trading strategy based on *fraud_cues* would only generate economically meaningful positive abnormal returns from a long strategy and not a hedge strategy, as the abnormal returns for the high portfolio are not statistically significant.

5.2 Measure Validation

While prior literature has used advanced computational models to perform textual analysis extraction, there is a wide array of literature that relies on bag-of-words lists.[11] I investigate whether my original *fraud_cues* measure based on AAER's is associated with word lists used by prior literature. Specifically, I use the Loughran-McDonald word lists from their 2013 study of Initial Public Offerings as it is one of the most widely used. The results of this analysis are un-

[10] For robustness, I calculate the equally-weighted returns of the portfolios and find results consistent with the value-weighted portfolio.
[11] See Loughran and McDonald (2016) for a review of the literature.

tabulated for brevity. There are no significant associations except that weak modal words are negative and significantly associated with *fraud_cues* ($p < 0.01$). Weak modal words are hypothesized to make it more difficult for investors to assess the value of a firm. This study makes no predictions as to whether a firm would be more or less likely to use ambiguous language in the pre-fraud period. This finding and lack of theoretical interpretation does however shed light upon the difficulty of forming ex-ante predictions based on word lists in a pre-fraud setting. Overall, this result lends support for the decision to employ an SVM approach based on deep vectorization instead of a logistic approach based on pre-determined word lists.

I next examine whether *fraud_cues* is associated with underpricing. Prior literature suggests that riskier IPOs will be underpriced more than their less-risky counterparts, referred to as the changing risk composition hypothesis (Ritter 1984). If IPO period risk is associated with future-fraud risk, I expect to find a significant relationship between *fraud_cues* and underpricing. I calculate underpricing, *fdret*, following Ritter (1984) and use the offer price from the Securities Data Company (SDC) and the first closing price from CRSP. The results of this analysis are un-tabulated for brevity. I do not find a significant relationship between my linguistic cues measure and underpricing, further providing evidence that future-fraud is an independent construct and that my study is not related to the literature that examines IPO underpricing.

5.3 *Additional Raw Data*

Next, I consider whether adding more of the raw data items available within the S-1 filing will help improve the predictive accuracy of a quantitative model. There are two primary reasons for performing this analysis. First, there are a lack of formal theoretical predictors of fraud and second, none of the existing fraud prediction literature considers which theoretical predictors in the pre-fraud period are the most informative. In order to perform the most complete analysis of

which quantitative variables are associated with future fraud, I include all available raw financial statement items for the S-1 filing available within Compustat. Table 13 reports the results. The value of the AUC is lower for both the AAER sample and the lawsuit sample. These results suggest that adding more quantitative variables does not improve prediction accuracy relative to lexical cues and echoes the concerns of prior literature that the lack of a long history of tangible information makes it difficult to use the quantitative information within the S-1 filing (Loughran and McDonald 2013). As such, the results support the decision to focus on qualitative information to construct the fraud prediction measure I use in my study.

6. Conclusion

Financial statement fraud can result in significant losses to shareholder value for the fraud firms as well as threaten the efficiency of capital markets by impairing overall trust. The current literature examining fraud prediction has focused on using the actual manipulated financial statements without considering the content of pre-fraud financial statements as a standalone information source. In this study I employ a machine learning algorithm to examine whether there is information content within pre-manipulation disclosures that provides predictive evidence on whether a firm will manipulate their financial statements and receive an AAER in a future period.

I develop a unique measure, *fraud_cues*, by using a machine learning algorithm to detect subtle lexical cues within the earliest pre-fraud disclosure, the S-1 filing. I perform out-of-sample analyses using historical data as a training sample to predict future manipulation and achieve an accuracy rating of 0.694 in my primary analysis, out-performing benchmark models for predicting manipulation in the IPO setting. My standalone linguistic cues measure with an AUC of 0.694 when predicting AAERs is not only the highest performing measure, but I achieve this predictive

ability up to five years in advance of the fraud, enabling regulatory monitors such as the SEC the ability to better allocate their limited resources.

Further, I find that when combining my measure with quantitative measures, I am able to achieve an AUC of 0.793 for predicting AAERs out-of-sample. Additionally, I provide evidence that lexical cues used to predict the future receipt of an AAER are also associated with both higher future earnings management and higher future information asymmetry. Firms with a higher *fraud_cues* score are more likely to manage earnings to *just* meet or beat analyst forecasts and to experience increased information asymmetry around earnings announcements within the first 5-years following the IPO. Earnings management and information asymmetry can erode the value relevance of financial reporting and lower liquidity by forcing those investors without access to private information out of the market.

I find that my inferences are robust to using another definition of fraud, 10(b)-5 filings, and that the measure based on these lawsuits has some predictive power over future abnormal returns. When conducting a portfolio analysis, I find firms identified by *fraud_cues* as least likely to commit future fraud generate positive abnormal returns, even after adjusting for common risk proxies.

This is the first study to consider whether it is possible to utilize information prior to the occurrence of fraud as an independent predictor, and successfully predicts 84 percent of the future AAERs received by firms that went through an initial public offering after 2000. As such, the results of this study are not only economically valuable in detecting future cases of financial manipulation, but in providing a timelier measure of detection. These findings are relevant to a growing accounting literature that uses the data from public filings in fraud prediction.

Additionally, these findings should be of interest to regulators and capital market participants as they are in alignment with SEC enforcement objectives.

7. References

Abarbanell, Jeffery, and Reuven Lehavy. 2003. "Biased Forecasts or Biased Earnings? The Role of Reported Earnings in Explaining Apparent Bias and over/Underreaction in Analysts' Earnings Forecasts." *Journal of Accounting and Economics* 36 (1–3): 105–46.

Aboody, David, and Baruch Lev. 2000. "Information Asymmetry, R&D, and Insider Gains." *The Journal of Finance* 55 (6): 2747–66.

Affleck-Graves, John, Carolyn M. Callahan, and Niranjan Chipalkatti. 2002. "Earnings Predictability, Information Asymmetry, and Market Liquidity." *Journal of Accounting Research* 40 (3): 561–83.

Allee, Kristian D., Bok Baik, and Yongoh Roh. 2021. "Detecting Financial Misreporting with Real Production Activity." *Working Paper*.

Allee, Kristian D., and Matthew D. Deangelis. 2015. "The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion: The Structure of Voluntary Disclosure Narratives." *Journal of Accounting Research* 53 (2): 241–74.

American. 2016. "Five Financial Heroes and Villains: How a Master of Science in Accounting at American University Can Turn You into a Fiscal Superstar." *American University* (blog). October 4, 2016. https://au.blogs.american.edu/accounting/five-financial-heroes-and-villains-how-a-master-of-science-in-accounting-at-american-university-can-turn-you-into-a-fiscal-superstar/.

Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff, and Richard Sloan. 2018. "Financial Reporting Fraud and Other Forms of Misconduct: A Multidisciplinary Review of the Literature." *Review of Accounting Studies* 23 (2): 732–83.

Arnold, Tom, Raymond P.H. Fishe, and David North. 2010. "The Effects of Ambiguous Information on Initial and Subsequent IPO Returns." *Financial Management* 39 (4): 1497–1519.

Asare, Stephen K., and Arnold M. Wright. 2004. "The Effectiveness of Alternative Risk Assessment and Program Planning Tools in a Fraud Setting*." *Contemporary Accounting Research* 21 (2): 325–52.

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2020. "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach." *Journal of Accounting Research* 58 (1): 199–235.

Barton, Jan, and Paul J. Simko. 2002. "The Balance Sheet as an Earnings Management Constraint." *The Accounting Review* 77 (s-1): 1–27.

Basmmi, Ain Balqis Md Nor, Shahliza Abd Halim, and Nor Azizah Saadon. 2020. "Comparison of Web Services for Sentiment Analysis in Social Networking Sites." *IOP Conference Series: Materials Science and Engineering* 884 (July): 012063.

Bauguess, Scott W. 2018. "The Role of Machine Readability in an AI World." Presented at the SEC Keynote Address, Financial Information Management Conference 2018, May 3. https://www.sec.gov/news/speech/speech-bauguess-050318.

Beasley, Mark S. 1996. "An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud." *The Accounting Review* 71(4): 443-465.

Beasley, Mark S., Joseph V. Carcello, Dana R. Hermanson, and Paul D. Lapides. 2000. "Fraudulent Financial Reporting: Consideration of Industry Traits and Corporate Governance Mechanisms." *Accounting Horizons.* 14(4): 441-454.

Bell, Timothy B., and Joseph V. Carcello. 2000. "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting." *AUDITING: A Journal of Practice & Theory* 19 (1): 169–84.

Beneish, Messod D. 1997. "Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance." *Journal of Accounting and Public Policy* 16 (3): 271–309.

Bertomeu, Jeremy, Edwige Cheynel, Eric Floyd, and Wenqiang Pan. 2020. "Using Machine Learning to Detect Misstatements." *Review of Accounting Studies*, October.

Bochkay, Khrystyna, Roman Chychyla, Srini Sankaraguruswamy, and Michael Willenborg. 2018. "Management Disclosures of Going Concern Uncertainties: The Case of Initial Public Offerings." *The Accounting Review* 96(3): 29-59.

Bonsall, S. B., Z. Bozanic, and K. J. Merkley. 2014. "What Do Forward and Backward-Looking Narratives Add to the Informativeness of Earnings Press Releases?" *Working Paper*.

Brau, James C., James Cicon, and Grant McQueen. 2016. "Soft Strategic Information and IPO Underpricing." *Journal of Behavioral Finance* 17 (1): 1–17.

Brown, Nerissa C., Richard M. Crowley, and W. Brooke Elliott. 2020. "What Are You Saying? Using *Topic* to Detect Financial Misreporting." *Journal of Accounting Research* 58 (1): 237–91.

Burgstahler, David, and Michael Eames. 2006. “Management of Earnings and Analysts’ Forecasts to Achieve Zero and Small Positive Earnings Surprises.” *Journal of Business Finance Accounting* 33 (5–6): 633–52.

Bushee, Brian J., Ian D. Gow, and Daniel J. Taylor. 2018. “Linguistic Complexity in Firm Disclosures: Obfuscation or Information?: LINGUISTIC COMPLEXITY IN FIRM DISCLOSURES.” *Journal of Accounting Research* 56 (1): 85–121.

Carhart, Mark M. 1997. “On Persistence in Mutual Fund Performance.” *The Journal of Finance* 52 (1): 57–82.

Cecchini, Mark, Haldun Aytug, Gary J. Koehler, and Praveen Pathak. 2010. “Detecting Management Fraud in Public Companies.” *Management Science* 56 (7): 1146–60.

Ceresney, Andrew. 2013. “Financial Reporting and Accounting Fraud.” Speech at American Law Institute Continuing Legal Education. Washington, DC.

Chae, Joon. 2005. “Trading Volume, Information Asymmetry, and Timing Information.” *The Journal of Finance* 60 (1): 413–42.

Chi, Sabrina, and Devin Shanthikumar. 2017. “Local Bias in Google Search and the Market Response around Earnings Announcements.” *The Accounting Review* 92(4): 115-143.

Chou, De-Wai, Michael Gombola, and Feng-Ying Liu. 2006. “Earnings Management and Stock Performance of Reverse Leveraged Buyouts.” *The Journal of Financial and Quantitative Analysis* 41 (2): 407–38.

Coller, Maribeth and Teri Yohn. 1997. “Management Forecasts and Information Asymmetry: An Examination of Bid-Ask Spreads.” *Journal of Accounting Research* 35(2): 181-191.

Copeland, Thomas E., and Dan Galai. 1983. “Information Effects on the Bid-Ask Spread.” *The Journal of Finance* 38 (5): 1457–69.

Davis, Angela K., Jeremy M. Piger, and Lisa M. Sedor. 2012. “Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*: Content of Earnings Press Release Language.” *Contemporary Accounting Research* 29 (3): 845–68

Davis, Angela K., and Isho Tama-Sweet. 2012. “Managers’ Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A*: Language in Earnings Press Releases vs. MD&A.” *Contemporary Accounting Research* 29 (3): 804–37.

Dechow, Patricia M., Weili Ge, Chad R. Larson, and Richard G. Sloan. 2011. “Predicting Material Accounting Misstatements*: Predicting Material Accounting Misstatements.” *Contemporary Accounting Research* 28 (1): 17–82.

Doyle, Jeffrey, Jared Jennings, and Mark Soliman. 2013. "Do managers define non-GAAP earnings to meet or beat analyst forecasts." *Journal of Accounting and Economics* 56: 40-56.

Drake, Philip D., and Michael R. Vetsuypens. 1993. "IPO Underpricing and Insurance against Legal Liability." *Financial Management* 22 (1): 64.

DuCharme, Larry L, Paul H Malatesta, and Stephan E Sefcik. 2004. "Earnings Management, Stock Issues, and Shareholder Lawsuits." *Journal of Financial Economics* 71 (1): 27–49.

Eaglesham, Jean. 2013. "Accounting Fraud Targeted." *The Wall Street Journal*, May 27, 2013. https://www.wsj.com/articles/SB10001424127887324125504578509241215284044.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Feng, Song, Ritwik Banerjee, and Yejin Choi. 2012. "Syntactic Stylometry for Deception Detection." Presented at the 50th Annual Meeting of the Association for Computational Linguistics, Republic of Korea. https://www.aclweb.org/anthology/P12-2034.pdf.

Ferris, Stephen P., (Grace) Qing Hao, and (Stella) Min-Yu Liao. 2013. "The Effect of Issuer Conservatism on IPO Pricing and Performance*." *Review of Finance* 17 (3): 993–1027.

Foster, F. Douglas, and S. Viswanathan. 1990. "A Theory of the Interday Variations in Volume, Variance, and Trading Costs in Securities Markets." *Review of Financial Studies* 3 (4): 593–624.

Galloway, Scott. 2019. "Yogababble." *No Mercy/ No Malice* (blog). September 27, 2019. profgalloway.com/yogababble.

Gande, Amar, and Craig M. Lewis. 2009. "Shareholder-Initiated Class Action Lawsuits: Shareholder Wealth Effects and Industry Spillovers." *Journal of Financial and Quantitative Analysis* 44 (4): 823–50.

Glosten, L. and L. Harris. 1988. "Estimating the components of the Bid-ask spread." *Journal of Financial Economics* 21: 123-142.

Goel, Sunita, Jagdish Gangolly, Sue R. Faerman, and Ozlem Uzuner. 2010. "Can Linguistic Predictors Detect Fraudulent Financial Filings?" *Journal of Emerging Technologies in Accounting* 7 (1): 25–46.

Green, Brian P., and Jae H. Choi. 1997. "Assessing the Risk of Management Fraud Through Neural Network Technology." *Auditing: A Journal of Practice and Theory* 16 (1): 14–28.

Guo, Yanming, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. 2016. "Deep Learning for Visual Understanding: A Review." *Neurocomputing* 187 (April): 27–48.

Hanley, Kathleen Weiss, and Gerard Hoberg. 2010. "The Information Content of IPO Prospectuses." *Review of Financial Studies* 23 (7): 2821–64.
———. 2012. "Litigation Risk, Strategic Disclosure and the Underpricing of Initial Public Offerings." *Journal of Financial Economics* 103 (2): 235–54.

Hastie, T., R. Tibshirani, and J. Friedman. 2003. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Corrected. New York: Springer-Verlang.

Healy, Paul M, and Krishna G Palepu. 2001. "Information Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature." *Journal of Accounting and Economics* 31 (1–3): 405–40.

Healy, Paul M., and James M. Wahlen. 1999. "A Review of the Earnings Management Literature and Its Implications for Standard Setting." *Accounting Horizons* 13 (4): 365–83.

Hoberg, Gerard, and Craig Lewis. 2017. "Do Fraudulent Firms Produce Abnormal Disclosure?" *Journal of Corporate Finance* 43 (April): 58–85.

Hughes, Patricia J., and Anjan V. Thakor. 1992. "Litigation Risk, Intermediation, and the Underpricing of Initial Public Offerings." *Review of Financial Studies* 5 (4): 709–42.

Karpoff, Jonathan M., D. Scott Lee, and Gerald S. Martin. 2008. "The Cost to Firms of Cooking the Books." *Journal of Financial and Quantitative Analysis* 43 (3): 581–611.

Keung, Edmund, Zhi-Xing Lin, and Michael Shih. 2010. "Does the Stock Market See a Zero or Small Positive Earnings Surprise as a Red Flag?" *Journal of Accounting Research* 48 (1): 105–36.

Lambert, Richard A., Christian Leuz, and Robert E. Verrecchia. 2012. "Information Asymmetry, Information Precision, and the Cost of Capital*." *Review of Finance* 16 (1): 1–29.

Larcker, David F., and Anastasia A. Zakolyukina. 2012. "Detecting Deceptive Discussions in Conference Calls: Detecting Deceptive Discussions in Conference Calls." *Journal of Accounting Research* 50 (2): 495–540.

Lehavy, Reuven, Feng Li, and Kenneth Merkley. 2011. "The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts." *The Accounting Review* 86 (3): 1087–1115.

Li, Feng. 2008. “Annual Report Readability, Current Earnings, and Earnings Persistence.” *Journal of Accounting and Economics* 45 (2–3): 221–47.

Ling, Wang, Lin Chu-Cheng, Yulia Tsvetkov, Silvio Amir, Ramon Astudillo, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. “Not All Contexts Are Created Equal: Better Word Representations with Variable Attention.” *The 2015 Conference on Empirical Methods in Natural Language Processing*, 1367–72.

Lo, Kin. 2008. “Earnings Management and Earnings Quality.” *Journal of Accounting and Economics* 45(2-3): 350-357.

Loughran, Tim, and Bill McDonald. 2013. “IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language.” *Journal of Financial Economics* 109 (2): 307–26.

Loughran, Tim, and Bill Mcdonald. 2016. “Textual Analysis in Accounting and Finance: A Survey: TEXTUAL ANALYSIS IN ACCOUNTING AND FINANCE.” *Journal of Accounting Research* 54 (4): 1187–1230.

Loughran, Tim, and Jay Ritter. 2004. “Why Has IPO Underpricing Changed over Time?” *Financial Management* 33 (3): 5–37.

Lowry, Michelle, and Susan Shu. 2002. “Litigation Risk and IPO Underpricing.” *Journal of Financial Economics* 65 (3): 309–35.

Maloof, Marcus. 2003. “Learning When Data Sets Are Imbalanced and When Costs Are Unequal and Unknown.” Presented at the Learning from Imbalanced Data Sets II, Department of Computer Science, Georgetown University.

Mayew, William J., and Mohan Venkatachalam. 2012. “The Power of Voice: Managerial Affective States and Future Firm Performance.” *The Journal of Finance* 67 (1): 1–43.

Mitsuhashi, Hitoshi, and Theresa Welbourne. 1999. “Chief Executive Officer (CEO) Tenure in Initial Public Offering (IPO) Firms: An Event History Analysis of the Determinants of Turnover.” *CAHRS Working Paper #99-01*, Cornell University, .

Palmrose, Zoe-Vonna, Vernon J. Richardson, and Susan Scholz. 2004. “Determinants of Market Reactions to Restatement Announcements.” *Journal of Accounting and Economics* 37 (1): 59–89.

Peng, Lin, and Aisla Roell. “Executive Pay and Shareholder Litigation.” *Review of Finance* 12(1): 141-184.

Perols, Johan. 2011. “Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms.” *AUDITING: A Journal of Practice & Theory* 30 (2): 19–50.

Perols, Johan L., Robert M. Bowen, Carsten Zimmermann, and Basamba Samba. 2017. “Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction.” *The Accounting Review* 92 (2): 221–45.

Purda, Lynnette, and David Skillicorn. 2015. “Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection.” *Contemporary Accounting Research* 32 (3): 1193–1223.

Ramesh, Bharath, Cheng Xiang, and Tong Heng Lee. 2015. “Shape Classification Using Invariant Features and Contextual Information in the Bag-of-Words Model.” *Pattern Recognition* 48 (3): 894–906.

Ritter, Jay R., and Ivo Welch. 2002. “A Review of IPO Activity, Pricing, and Allocations.” *The Journal of Finance* 57 (4): 1795–1828.

Rogers, Jonathan L., Andrew Van Buskirk, and Sarah L. C. Zechman. 2011. “Disclosure Tone and Shareholder Litigation.” *The Accounting Review* 86 (6): 2155–83.

Schipper, Katherine. 1989. “Commentary on Earnings Management.” *Accounting Horizons* 3 (4): 91–102.

Skinner, Douglas J. 1994. “Why Firms Voluntarily Disclose Bad News.” *Journal of Accounting Research* 32 (1): 38.

Srinivasan, Suraj. 2005. “Consequences of Financial Reporting Failure for Outside Directors: Evidence from Accounting Restatements and Audit Committee Members.” *Journal of Accounting Research* 43(2): 291-334.

Stoughton, Neal M., and Josef Zechner, 1998.“IPO-mechanisms, monitoring and ownership structure, Journal of Financial Economics.” 49: 45-77.

Teoh, Siew Hong, Ivo Welch, and T.J. Wong. 1998. “Earnings Management and the Long-Run Market Performance of Initial Public Offerings.” *The Journal of Finance* 53 (6): 1935–74.

Wang, Tracy Yue, Andrew Winton, and Xioyun Yu. 2010. “Corporate Fraud and Business Conditions: Evidence from IPOs.” *Journal of Finance* 65(6): 2255-2292.

Witten, I., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd ed. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Zhou, L., D.P. Twitchell, Tiantian Qin, J.K. Burgoon, and J.F. Nunamaker. 2003. "An Exploratory Study into Deception Detection in Text-Based Computer-Mediated Communication." In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of The*, 10 pp. Big Island, HI, USA: IEEE.

8. Appendix

Table 1: Distribution of IPOs and IPO AAERs.

Year	IPO	AAER	%
1996	105	20	19%
1997	122	5	4%
1998	176	6	3%
1999	358	16	4%
2000	287	8	3%
2001	56	2	0%
2002	50	0	2%
2003	44	1	1%
2004	139	2	2%
2005	123	2	2%
2006	115	2	1%
2007	116	1	0%
2008	18	0	0%
2009	38	0	1%
2010	81	1	2%
2011	86	2	0%
2012	98	0	0%
2013	151	0	0%
2014	180	0	0%
2015	111	0	0%
Total	2,454	68	3%
Percentage	100%	3%	
Training Sample	1,048	55	5%
Test Sample	1,406	13	1%

Table 2: The within-sample performance evaluation metrics for the test period 1996-2000. This tables presents within-sample performance statistics. (1) Accuracy is the percentage of texts that were predicted with the correct tag. (2) F1 Score is the combination of precision and recall for all tags. (3) Precision refers to the percentage of texts the classifier correctly predicted. (4) Recall refers to the percentage of texts the classifier predicted.

Features/Dataset	Accuracy	F1 Score	Precision	Recall	Precision	Recall
AAER (filter stop words)	72%	72%	69%	76%	75%	68%
AAER (w/o stop words)	73%	73%	71%	74%	74%	72%

Table 3: Descriptive Statistics.

Variable	N	Benchmark	Mean	S.D.	Min	Q1	Q2	Q3	Max
Fraud_AAER	1,406		0.01	0.09	0.00	0.00	0.00	0.00	1.00
Fraud_cues	1,406		-0.20	0.69	-0.1	-0.7	-0.6	0.65	0.98
RSST accruals	1,406	f-score	0.61	0.67	-0.7	0.09	0.42	1.02	1.92
Change in Receivables	1,406	f-score	0.03	0.06	-0.1	0.00	0.01	0.05	0.32
Change in Inventory	1,406	f-score	0.02	0.04	-0.0	0.00	0.00	0.02	0.21
% Soft Assets	1,406	f-score	0.39	0.29	0.01	0.13	0.33	0.64	0.97
Change in Cash Sales	1,406	f-score	0.48	1.43	-2.1	0.00	0.17	0.52	9.80
Change in return on assets	1,406	f-score	0.10	0.34	-0.6	-0.0	0.00	0.12	1.29
Actual Issuance	1,406	f-score	0.98	0.16	0.00	1.00	1.00	1.00	1.00
Book-to-market	1,406	f-score	0.30	0.27	-0.5	0.13	0.26	0.42	1.26
Startup	1,406	IPO f-score	0.57	0.50	0.00	0.00	1.00	1.00	1.00
VC	1,406	IPO f-score	0.50	0.59	0.00	0.00	0.00	1.00	1.00
Age	1,406	IPO f-score	14.47	18.35	-1.0	4.00	8.00	17.00	93.00
Hot_IPO	1,406	IPO f-score	0.56	0.49	0.00	0.00	1.00	1.00	1.00
Current Assets, Total	1,406	RUSBoost	252.37	652.7	0.00	49.67	101	202.6	1126
Accounts payable, trade	1,406	RUSBoost	53.23	231.5	0.00	2.11	6.57	26.03	5001
Assets, total	1,406	RUSBoost	895.06	2951	0.00	89.46	189	599.2	5579
Common/ordinary equity, total	1,406	RUSBoost	255.87	917.5	-82	45.91	102	234.3	2347
Cash and short-term equivalents	1,406	RUSBoost	111.30	330.2	0.00	20.37	6.57	107.2	9626
Cost of goods sold	1,406	RUSBoost	473.68	1809	0.00	18.69	189	234.3	3979
Common shares outstanding	1,406	RUSBoost	52.88	114.6	000	17.99	103	107.2	2372
Debt in current liabilities, total	1,406	RUSBoost	23.36	163.5	0.00	0.00	54.9	234.5	3157
Long-term debt issuance	1,406	RUSBoost	217.45	1340	-0.1	0.00	54.3	52.01	4257
Long-term debt, total	1,406	RUSBoost	344.54	1522	0.00	0.00	28.2	5.29	2564
Depreciation and Amortization	1,406	RUSBoost	34.79	109.2	0.00	1.21	0.59	68.50	1642

Table 4: The out-of-sample performance evaluation metrics for the test period 2001-2015. This table shows fraud prediction models' performance comparison using the area under the receiver operating characteristics (ROC) curve (AUC). The ROC is the standard technique used to select classifiers. The p-values in parentheses are based on a two-tailed t-test of fraud_cues vs. the other model and are therefore not available for fraud_cues.

Panel A: IPO Date	**(1)**
Method	AAER
1) fraud_cues	0.694
2) f-score	0.403 (0.01)
3) RUSBoost	0.5204 (0.05)
4) AbDisc	0.2951 (0.00)
5) IPO_topic	0.587 (0.05)
6) IPO_f-score	0.652 (0.09)
N	1,406
Panel B: Misconduct Date	
7) f-score	0.201 (0.00)
8) RUSBoost	0.601 (0.06)
9) AbDisc	0.341 (0.00)
N	9,493

Table 5: Logistic Regression of the determinants of misconduct at IPO date. This table presents the regression coefficients from estimating the likelihood of receiving a future AAER. The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively. The AUC represents the area under the receiver operating characteristics (ROC) curve (AUC).

Panel A						
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	-5.875*** (-3.97)	-5.555*** (-3.87)	-5.846*** (-4.04)	-5.602*** (-3.40)	-5.201*** (-3.39)	-4.622*** (-3.14)
Fraud_cues		0.860* (1.71)				1.005** (1.99)
RUSBoost			-0.793 (-0.93)			-0.761 (-0.91)
IPO_topic				-0.001 (-1.27)		-0.00103 (-1.34)
AbDisc					-0.000 (-1.51)	-0.001 (-1.52)
RSST Accruals	-0.145 (-0.22)	-0.147 (-0.21)	-0.199 (-0.29)	-0.418 (-0.63)	-0.149 (-0.22)	-0.624 (-0.84)
Change in Receivables	3.246 (0.94)	3.889 (1.17)	3.674 (1.05)	1.794 (0.52)	3.011 (0.88)	2.487 (0.76)
Change in Inventory	3.086 (0.62)	3.542 (0.69)	4.575 (0.88)	2.053 (0.41)	1.899 (0.36)	2.431 (0.44)
% Soft Assets	3.170** (2.46)	2.762** (2.06)	3.328* (2.54)	3.557** (2.71)	3.288** (2.54)	3.623*** (2.58)
Change in Cash Sales	0.0577 (0.25)	0.0684 (0.27)	0.047 (0.19)	0.0615 (0.26)	0.0957 (0.41)	0.115 (0.42)
Change in return on assets	-0.677 (-0.43)	-0.653 (-0.39)	-0.654 (-0.42)	-0.414 (0.29)	-0.639 (-0.42)	-0.101 (-0.07)
Actual Issuance	-1.243 (-0.99)	-1.509 (-1.26)	-0.1254 (-1.03)	-1.096 (-0.87)	-1.134 (-0.92)	-1.268 (-1.06)
Book-to-Market	0.962 (1.06)	0.964 (1.04)	0.959 (1.07)	1.102 (1.19)	1.191 (1.24)	1.313 (1.32)
Pseudo R2	0.118	0.144	0.126	0.160	0.145	0.214
AUC	0.403	0.692	0.4758	0.6347	0.2951	0.752
Misconduct Observations	13	13	13	13	13	13
Non-Misconduct Observations	1,393	1,393	1,393	1,393	1,393	1,393

Table 5 (Cont.)

Panel B				
Variable	(1)	(2)	(3)	(4)
Intercept	-7.837*** (-4.67)	-7.616*** (-4.38)	-7.549*** (-4.42)	-7.404*** (-4.17)
Fraud_cues		0.007** (2.29)		0.008** (2.38)
IPO_topic			-0.001 (-0.70)	-0.001 (-0.90)
RSST Accruals	-0.722 (-1.22)	-0.875 (-1.39)	-0.652 (-1.11)	-0.811 (-1.29)
Change in Receivables	2.942 (0.74)	3.344 (0.83)	1.838 (0.46)	2.147 (0.53)
Change in Inventory	-3.683 (-0.44)	-3.421 (-0.34)	-4.738 (-0.58)	-4.695 (-0.56)
% Soft Assets	3.058* (1.83)	2.644 (1.51)	3.127* (1.88)	2.831 (1.62)
Startup	-0.247 (-0.31)	-0.337 (-0.43)	-0.254 (-0.32)	-0.318 (-0.39)
VC	-1.302 (-1.14)	-1.315 (-1.13)	-1.423 (-1.23)	-1.438 (-1.23)
Age	0.00177 (0.10)	-0.00134 (-0.08)	0.00837 (0.47)	0.00464 (0.25)
Hot_IPO	1.484 (1.35)	1.470 (1.32)	1.523 (1.38)	1.535 (1.36)
Pseudo R2	0.132	0.156	0.156	0.138
AUC	0.652	0.793	0.617	0.763
Misconduct Observations	13	13	13	13
Non-Misconduct Observations	1,393	1,393	1,393	1,393

Table 6: Earnings management around earnings announcements. This table presents the regression coefficients from estimating the likelihood of missing earnings or meeting/beating earnings. The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively.

Variables	**MBE(inv)**	**MBE**
Fraud_cues	-0.045***	-0.069***
	(-4.06)	(-3.05)
Size	0.001	-0.001
	(-0.69)	(-0.03)
BM	0.156***	0.158**
	(4.14)	(1.99)
Lev	-0.001	-0.001
	(-0.18)	(-0.28)
AnalystCount	-0.011***	-0.012***
	(-6.62)	(-3.56)
AnalystDisp	0.001	0.312***
	(0.14)	(70.11)
Year FE's	Yes	Yes
Adj. R-Squared	0.028	0.694
Obvs	4,070	2,929

Table 7: Information asymmetry around earnings announcements. This table presents the regression coefficients from estimating the likelihood of abnormal spread or abnormal volume. The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively.

Variables	**AbSpread**	**AbVol**
Fraud_cues	0.015***	-0.064***
	(5.88)	(-5.59)
Size	0.000***	0.000***
	(25.42)	(31.69)
BM	0.406***	-1.658***
	(43.13)	(-37.02)
Lev	-0.000***	0.001***
	(-19.76)	(18.61)
AnalystCount	-0.003	0.155***
	(-0.23)	(23.93)
AnalystDisp	0.001	-0.307
	(0.69)	(-0.77)
Year FE's	Yes	Yes
Adj. R-Squared	0.000	0.156
Obvs	4,070	4,070

Table 8: Logistic regression of the ability of fraud_cues to predict future returns. This table presents the regression coefficients from estimated abnormal returns measured both yearly (Panel A) and cumulatively (Panel B). The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively.

Panel A	**RetOne**	**RetTwo**	**RetThree**	**RetFour**	**RetFive**
Fraud_Cues	-0.018	0.004	-0.067	0.004	0.036
	(-0.62)	(0.13)	(-2.12)	(0.16)	(1.32)
Size	-0.000	0.000	0.000*	-0.000	0.000*
	(-0.88)	(0.62)	(1.89)	(-0.45)	(2.08)
BM	-0.032	-0.056	0.037	0.034	-0.106*
	(-0.50)	(-0.71)	(0.39)	(0.40)	(-2.14)
Year FEs	Yes	Yes	Yes	Yes	Yes
Obs	1,254	1,107	997	832	646
Adj. R2	0.075	0.096	0.084	0.075	0.078
Panel B	**RetOne**	**RetTwo**	**RetThree**	**RetFour**	**RetFive**
Fraud_Cues		0.005	-0.021	-0.044	-0.036
		(0.10)	(-0.33)	(-0.62)	(-0.49)
Size		-0.000	0.000	0.000	0.000
		(-0.63)	(1.36)	(1.08)	(1.36)
BM		-0.023	-0.063	-0.121	-0.237
		(-0.23)	(-0.46)	(-0.67)	(-1.22)
Year FEs		Yes	Yes	Yes	Yes
Obs		646	646	646	646
Adj. R2		0.083	0.044	0.015	0.020

Table 9: Distribution of IPOs and IPO Lawsuits.

Year	IPO	Lawsuit	%
1996	105	23	22%
1997	122	14	11%
1998	176	29	16%
1999	358	41	11%
2000	287	22	8%
2001	56	4	7%
2002	50	4	8%
2003	44	3	7%
2004	139	13	9%
2005	123	15	12%
2006	115	11	10%
2007	116	14	12%
2008	18	2	11%
2009	38	8	21%
2010	81	14	17%
2011	86	14	16%
2012	98	11	11%
2013	151	31	21%
2014	180	29	16%
2015	111	11	10%
Total	2,454	313	13%
Percentage	100%	13%	
Training Sample	1,048	129	12%
Test Sample	1,406	184	13%

Table 10: The out-of-sample performance evaluation metrics for the test period 2001-2015. This table shows fraud prediction models' performance comparison using the area under the receiver operating characteristics (ROC) curve (AUC). The ROC is the standard technique used to select classifiers. The p-values in parentheses are based on a two-tailed t-test of fraud_cues vs. the other model and are therefore not available for fraud_cues.

Panel A: IPO Date	**(1)**
Method	Lawsuit
1) fraud_cues	0.561
2) f-score	0.452 (0.01)
3) RUSBoost	0.552 (0.70)
4) AbDisc	0.313 (0.00)
5) IPO_topic	0.348 (0.01)
6) IPO_f-score	0.485 (0.00)
N	1,406
Panel B: Misconduct Date	**(1)**
7) f-score	0.269 (0.00)
8) RUSBoost	0.691 (0.03)
9) AbDisc	0.512 (0.08)
N	9,493

Table 11: Logistic regression of the ability of fraud_cues to predict future returns. This table presents the regression coefficients from estimated abnormal returns measured both yearly (Panel A) and cumulatively (Panel B). The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively.

Panel A	**RetOne**	**RetTwo**	**RetThree**	**RetFour**	**RetFive**
Fraud_Cues	-0.016	-0.069	-0.013	0.043	-0.028
	(-0.56)	(-1.69)	(-0.27)	(1.39)	(-0.67)
Size	-0.000	0.000	0.000*	-0.000	0.000*
	(-0.75)	(0.13)	(1.92)	(-0.32)	(1.97)
BM	-0.033	-0.068	0.036	0.026	-0.125**
	(-0.52)	(-0.86)	(0.37)	(0.28)	(-2.46)
Year FEs	Yes	Yes	Yes	Yes	Yes
Obs	1,254	1,107	997	832	646
Adj. R2	0.075	0.098	0.081	0.074	0.079
Panel B	**RetOne**	**RetTwo**	**RetThree**	**RetFour**	**RetFive**
Fraud_Cues		-0.109***	-0.119***	-0.160***	-0.181**
		(-3.25)	(-3.03)	(-3.38)	(-2.93)
Size		-0.000	0.000	0.000	0.000
		(-0.76)	(1.52)	(1.03)	(1.32)
BM		-0.040	-0.081	-0.147	-0.278
		(-0.41)	(-0.56)	(-0.77)	(-1.36)
Year FEs		Yes	Yes	Yes	Yes
Obs		646	646	646	646
Adj. R2		0.008	0.050	0.020	0.025

Table 12: Excess returns associated with a fraud_cues based portfolio trading strategy.
This table presents the regression coefficients from a portfolio strategy based on fraud_cues. The t-statistics are in parentheses. All tests are two-tailed *, **, *** denoted significance at the 0.1, 0.05, and 0.01 levels, respectively.

Decile	RetOne	RetTwo	RetThree	RetFour	RetFive
Low	0.128**	0.212**	0.381***	0.577***	0.605***
	(2.12)	(2.39)	(3.43)	(3.70)	(3.88)
2	0.002	0.094	0.256**	0.345***	0.423***
	(0.04)	(1.15)	(2.51)	(2.86)	(3.00)
3	0.093	0.204*	0.377***	0.637***	0.827***
	(0.24)	(1.96)	(2.98)	(3.93)	(3.97)
4	-0.070	0.107	0.327**	0.616***	0.568***
	(-1.06)	(1.08)	(2.42)	(3.58)	(3.01)
5	-0.017	-0.109	0.037	0.116	0.154
	(-0.25)	(1.22)	(0.33)	(0.75)	(1.01)
6	0.023	0.000	0.264*	0.333*	0.343
	(0.35)	(0.00)	(1.76)	(1.79)	(1.63)
7	0.006	0.071	0.117	0.139	0.239
	(0.09)	(0.57)	(0.82)	(0.83)	(1.18)
8	-0.097	0.092	0.262**	0.314**	0.262
	(-1.40)	(0.97)	(2.21)	(2.43)	(1.61)
9	0.062	0.081	0.274**	0.392**	0.357*
	(0.72)	(0.70)	(2.02)	(2.19)	(1.84)
High	-0.038	-0.040	0.011	0.071	0.107
	(-0.52)	(-0.42)	(0.11)	(0.57)	(0.48)

Table 13: The out-of-sample performance evaluation metrics for the test period 2001-2015. This table shows fraud prediction models' performance comparison using the area under the receiver operating characteristics (ROC) curve (AUC). The ROC is the standard technique used to select classifiers. The p-values in parentheses are based on a two-tailed t-test of fraud_cues vs. the other model and are therefore not available for fraud_cues.

Panel A: IPO Date	**(1)**	**(2)**
Method	AAER	Lawsuit
1) fraud_cues	0.694	0.561
2) all_variables	0.208 (0.00)	0.323 (0.00)
N	1,406	1,406