Summary of Unearthing Financial Statement Fraud: Insights from News Coverage Analysis
(Zhou Zheng; October 12, 2025)

ILO1: How to critically evaluate the quality of a research idea

1) The research questions:

1. How to construct measurement indicators based on news coverage that more effectively capture the characteristics of financial statement fraud?

2. How to utilize a comprehensive feature set and advanced machine learning algorithms?

2) Why are the research questions interesting?

In order to improve the accuracy and economic value of detecting financial statement fraud among Chinese listed companies

3) What is the paper's contribution? (1. Find the literature; 2. Summarize the literature; 3. Summarize the marginal contributions to the literature)

Summary of the literature:

Prior studies have primarily relied on disclosed and confirmed financial statement fraud cases for measurement, resulting in sample selection bias and limiting their ability to capture fraudulent behaviors that remain undisclosed or unconfirmed.

Marginal contributions to the literature:

1. This study enhances the accuracy and practical value of financial fraud detection by adopting a news-based Financial Statement Fraud Propensity (FSFP) index as the label to capture fraudulent behaviors that remain undisclosed or unconfirmed.

2. To capture peer contagion effects, this study incorporates peer firms' characteristics into the feature set, thereby enriching the determinants of fraud detection.

3. Leveraging complex network techniques, this study constructs an association network from business description texts in annual reports to improve the peer-grouping method.

4. By further expanding the feature set and applying more advanced machine learning algorithms, the proposed approach significantly improves the recall and overall performance of financial fraud detection.

ILO2: How to develop a testable research hypothesis

4) What hypotheses are tested in the paper? list them explicitly.

1. News reports can reveal the propensity for financial statement fraud.

2. The firm-grouping approach based on complex network techniques performs better than traditional industry or regional classifications.

3. Peer or related firms exhibit contagion effects that influence fraudulent behaviors.

4. Multi-dimensional risk factors — including financial, governance, market, and sentiment variables — help explain the risk of financial fraud. Machine learning models prove effective in detecting financial statement fraud

a) Do these hypotheses follow from and answer the research questions?

Hypothese 1 answers Question 1, Hypothese 2、3、4 answer Question 2

b) Do these hypotheses follow from theory or are they otherwise adequately developed? Please explain the logic of the hypotheses (use visualization if possible).

The rationale for Hypothesis 1 is that media reports capture the market's and regulators' attention and scrutiny of a firm's financial condition, thereby indirectly signaling the risk of financial statement fraud.

The rationale for Hypothesis 2、3 is that firms are subject to information transmission and imitation effects, so closely related firms may display similar behavioral patterns, including financial statement fraud.

The rationale for Hypothesis 4 is that individual financial indicators alone cannot comprehensively reflect complex fraudulent behaviors, and that combining multi-source information with advanced algorithms allows for more effective detection of abnormal patterns.

ILO3: How to create a sound research design to test a research hypothesis

5) Sample: comment on the appropriateness of the sample selection procedures.

1. The study sample comprises listed companies in the China A-share market that traded between 2001 and 2022 and had at least three consecutive years of trading, ensuring temporal coverage and continuity.

2. Financial institutions, such as banks and insurance companies, were excluded due to the significant differences in their financial statements compared to other firms, thereby avoiding excessive sample heterogeneity.

3. The dataset was split in a 70:30 ratio, and a year-wise block cross-validation method was applied to effectively reduce bias and variance associated with sample partitioning.

6) Dependent and independent variables: comment on the appropriateness of variable definition and measurement (focus on the key dependent variables and independent variables).

The dependent variable is measured using the Financial Statement Fraud Propensity (FSFP) index constructed from news coverage, which allows the classification of firms into non-fraud, low-fraud and high-fraud categories. This approach avoids reliance solely on potentially manipulated financial statement data. A Difference-in-Differences model leveraging the exogenous shock of the U.S.–China trade war is employed. Since FSFP relies on media coverage and sentiment analysis, it may be subject to reporting biases and textual noise.

The independent variables encompass four categories of fundamental factors (financial, governance, market and sentiment) and are further augmented with peer contagion, peer comparison, time trends and quarterly factors, resulting in a 605-dimensional feature set. Peer relationships are constructed based on annual report text similarity, replacing heterogeneous industry classifications, allowing for a continuous characterization of business similarity and dynamic changes, which aligns theory with data more closely.

7) Regression/prediction model specification: comment on the appropriateness of the regression/prediction model specification.

1. The FSFP label constructed based on news coverage can capture fraud signals revealed by the media and facilitate early detection .

2. Predictions are compared with confirmed fraud cases from the China Securities Regulatory Commission, and net detection value is calculated to provide evidence of real-world effectiveness .

3. The feature set covers financial, governance, market, sentiment, peer contagion, and peer comparison factors, with the inclusion of quarterly granularity, fully representing multiple risk dimensions.

ILO4: How to interpret and communicate the results of the research design

8) What difficulties arise in drawing inferences from the empirical work?

1. News-based fraud measurements and labels rely on keyword extraction and text preprocessing, which are prone to omissions or noise, affecting the magnitude and significance of estimated effects.

2. Fraud measures driven by ex-post events exhibit selection bias, limiting sample representativeness and external validity.

3. The sample includes only A-share firms, excludes financial institutions, and requires a minimum trading history, constraining the generalizability of conclusions to other markets and industries.

9) Describe at least one publishable and feasible extension of this research.

1. Construct a correction mechanism to address biases in news reporting.

2. Expand the news sources for the FSFP label.

Summary of Tone at the Bottom: Measuring Corporate Misconduct Risk from the Text of Employee Reviews
(Zhou Zheng; October 12, 2025)

ILO1: How to critically evaluate the quality of a research idea

1) The research questions:

Whether statistical text analysis of Glassdoor employee review texts can be used to extract indicators measuring corporate misconduct risk?

Whether these indicators provide incremental predictive value for out-of-sample forecasts of future violations, fines and whistleblower reports relative to commonly observed features?

2) Why are the research questions interesting?

Demonstrating the leading indicators of corporate misconduct risk can be extracted from employee reviews on Glassdoor. These indicators significantly improve the prediction of future violations and whistleblower complaints beyond traditional observable features. This provides investors, boards, and regulators with more timely and effective tools for risk monitoring.

3) What is the paper's contribution? (1. Find the literature; 2. Summarize the literature; 3. Summarize the marginal contributions to the literature)

Summary of the literature:

Prior research has largely focused on detecting accounting fraud and financial misstatements, or explaining corporate misconduct using market and firm characteristics. While existing studies highlight the role of media and whistleblowers in uncovering corporate misconduct, such information is often disclosed ex post and is associated with high individual costs and selection bias. Early studies using social platform data have mostly relied on Glassdoor star ratings to predict financial and market outcomes.

The marginal contributions to the literature:

1. Construct a "misconduct risk tone" indicator from Glassdoor employee review texts, using inverse regression to learn word weights and aggregating them at the firm-year level, thereby capturing internal information on control environment and culture to measure corporate misconduct risk.

2. The text-based indicator significantly improves out-of-sample prediction of future violations, the number of violations, and fines.

3. The study shows that the text indicator is particularly effective for long-term risk, predicting the transition from "few historical violations" to "serial violations" and "criminal violations".

4. The approach does not rely on pre-specified dictionaries, distinguishes among three textual contexts ("pros," "cons," and "advice to management"), and controls for public information such as media coverage, thereby isolating "internal risk information" independent of ratings in employee reviews.

ILO2: How to develop a testable research hypothesis

4) What hypotheses are tested in the paper? list them explicitly.

Hypothesis 1: Employee review texts on Glassdoor contain internal information related to future corporate misconduct risk, which can be aggregated into a "miscon duct risk word index" for measurement.

Hypothesis 2: The word index constructed from review texts can significantly predict out-of-sample whether violations will occur in the following year, improving predictive performance metrics such as AUC and pseudo-R^2.

Hypothesis 3: The text-based index provides incremental information for predicting misconduct relative to easily observable variables, such as Glassdoor ratings, firm size, leverage, ROA, industry, past violations, media coverage, and sentiment, rather than merely serving as a substitute for these features.

a) Do these hypotheses follow from and answer the research questions?

Typothesis 1 answer Question 1, Typothesis 2、3 answer Question 2

b) Do these hypotheses follow from theory or are they otherwise adequately developed? Please explain the logic of the hypotheses (use visualization if possible).

Logic of Hypothesis 1: Employee insights are captured in text, Employees are closest to daily operations, controls, and culture, and face low costs when anonymously posting on Glassdoor. Thus, their review texts should contain observable leading signals of future misconduct risk, ahead of external channels.

Logic of Hypothesis 2: Text provides incremental value beyond ratings. Five-point ratings are coarse and noisy, whereas textual comments are detailed and can be data-driven weighted. therefore, text-based indicators should provide significant incremental predictive information for misconduct even after controlling for ratings and common financial features.

Logic of Hypothesis 3: Contextual sections matter. The same word may carry different semantic meanings across "Pros," "Cons," and "Advice to Management"

sections. Modeling and weighting these sections separately preserves contextual differences and enhances the effectiveness of risk signals.

ILO3: How to create a sound research design to test a research hypothesis

5) Sample: comment on the appropriateness of the sample selection procedures.

1. Only U.S.-listed firms were selected and matched with Compustat, RavenPack, and Violation Tracker, ensuring the availability and consistency of governance, news, and violation data.

2. Firms not listed in the U.S. but with U.S. operations were excluded to ensure consistency in regulatory and violation coverage, though this reduces the external validity of conclusions for multinational firms.

3. Using 2008–2011 as the training period and 2012–2017 as the prediction period helps avoid information leakage and enables true out-of-sample testing.

4. Non-English, extremely common, or extremely rare words were filtered, and features were partitioned by Pros/Cons/Advice sections, enhancing interpretability and controlling noise.

6) Dependent and independent variables: comment on the appropriateness of variable definition and measurement (focus on the key dependent variables and independent variables).

The dependent variables (violation indicator, number of violations, and fines) are based on enforcement records from Violation Tracker, which are observable and have significant economic consequences, effectively proxying realized corporate misconduct . High-visibility violations are selected as a subsample to reflect risks most observable in employees' daily experience; however, keyword-based classification may introduce some measurement error.

The core independent variable is the Glassdoor text-based "misconduct word index," constructed via polynomial inverse regression to reduce high-dimensional word frequencies into three section-specific indicators (Pros/Cons/Advice). This method is data-driven and does not rely on a pre-specified dictionary. Word weights are learned from 2008–2011 and applied to predict outcomes in 2012–2017, preventing contemporaneous overfitting and ensuring forward-looking validity.Modeling Pros, Cons, and Advice sections separately allows semantic distinctions and enhances the ability of text-based variables to capture risk signals.

7) Regression/prediction model specification: comment on the appropriateness of the regression/prediction model specification.

1. The "misconduct word index" is constructed via polynomial inverse regression to reduce dimensionality and extract incremental information. In the absence of a standardized dictionary, this methodological choice is reasonable.

2. Gradient-boosted trees are trained using an 80/20 train-test split, incorporating shrinkage parameters and limited-depth interactions. Out-of-sample evaluation with pseudo-R^2 and AUC demonstrates robust predictive performance, supporting the model's validity.

3. Compared with firm size, prior violations, media coverage, and Glassdoor ratings, the text-based index provides significant out-of-sample fit improvements and variable influence, indicating that the model captures unique risk signals from employee reviews.

ILO4: How to interpret and communicate the results of the research design

8) What difficulties arise in drawing inferences from the empirical work?

1. Enforcement records are based on the timing of enforcement actions rather than the actual occurrence of misconduct, making it difficult to establish temporal order and capture lag structures, which limits causal inference.

2. The data include only violations detected and recorded by enforcement agencies; undetected or unenforced misconduct is unobserved, resulting in measurement error and selection bias.

3. Textual dimensionality reduction relies on the strong assumption of word independence; violations of this assumption may compromise the adequacy of reduction and the validity of subsequent inference.

9) Describe at least one publishable and feasible extension of this research.

Extend the text-based indicators over multiple years.