Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 16, No. 3, 2025

Predicting Internal Control Deficiencies from Investor Online Interaction Texts

Dongjie Lin*

School of Public Finance and Taxation, Central University of Finance and Economics, Beijing 102206, China
*Corresponding author email: rayldj@foxmail.com

Abstract: This study adopts a text-based perspective to transform Q&A on exchange-hosted investor interaction platforms into forward-looking signals and fuses them with financial ratios within a unified deep learning framework to predict internal control deficiencies. Empirical results show that, relative to a no-text baseline, the early-fusion model that incorporates textual information improves Average Precision (AP) and Area Under the ROC Curve (ROC-AUC) by approximately 14.6% and 4.1%, respectively; further adopting hierarchical text fusion yields cumulative improvements of 18.3% and 5.1% over the baseline. The findings document the incremental predictive value of text data for internal control deficiencies and provide effective tools for regulatory early-warning, audit sampling, and corporate self-assessment, with important theoretical and practical implications.

Keywords: Internal Control Deficiencies, Investor Online Interaction Platforms, Text Analysis.

1. Introduction

Internal control underpins the reliability of financial reporting and the credibility of capital markets. Identifying deficiencies solely from annual reports and audit outcomes is constrained by disclosure lags and inconsistent definitions, which impede timely and forward-looking detection. In China's capital market, exchange-hosted investor online interaction platforms have long been open to the public and provide channels for information exchange between listed firms and investors. To address the scarcity of identified deficiency cases and the timeliness problem, this study converts the Q&A texts on these investor interaction platforms into forward-looking, firm–year risk signals and, on that basis, develops a text-driven model to predict internal control deficiencies.

This study's primary contribution is to systematically leverage public Q&A content from investor online interactions to construct forward-looking firm–year risk signals, using them to predict subsequent internal control deficiencies. Empirically, relative to a baseline that relies only on structured features such as financial ratios, incorporating investor-interaction texts yields substantial gains in ranking performance. By building an internal control deficiency prediction model grounded in text analysis of investor interaction content, the paper enriches the literature on internal control effectiveness and offers practical tools for regulatory early warning, audit sampling, and corporate self-assessment.

2. Institutional Background

Investor online interaction platforms are open to the public and primarily serve as centralized, convenient, and standardized Q&A channels between listed firms and investors. The Shenzhen Stock Exchange defines its platform as an online venue for real-time interaction between investors and issuers, while the Shanghai Stock Exchange provides functions such as posing questions to listed firms, responding to investors, and conducting live interviews. Both platforms maintain firm-specific pages and consolidated Q&A feeds to improve communication efficiency among investors, listed companies, and regulators.

Exchanges impose explicit rules on information posted to these platforms and emphasize that interactive Q&A does not substitute for statutory disclosure. The Shanghai Stock Exchange specifies that its interaction platform aims to facilitate communication but prohibits the release of undisclosed material information to prevent information asymmetry and selective disclosure. At the same time, platform content supplies leads for frontline supervision, enhancing ex post oversight and risk identification. These provisions stress that the platforms are for communication and inquiries, whereas material information or matters requiring disclosure must be reported through designated media in accordance with applicable rules.

Q&A is the core of these platforms, complemented by modules such as push notifications, search, thematic interviews, and roadshows. Official user guides enumerate key functions and common workflows, including submitting questions to issuers, company responses, personalized subscriptions, and interaction rankings; Q&A detail pages and issuer homepages are publicly accessible. The Shenzhen platform likewise centers on firm Q&A and event records, supports online results briefings, and offers searchable history. These features create structured anchors in the text—firm identifier, time, question–answer pairs, and topic tags—which facilitate the construction of firm–year datasets for subsequent text modeling and rolling evaluation.

3. Theoretical Analysis

The predictive content of investor-interaction Q&A texts for internal control deficiencies hinges on their continuous, fine-grained exposure of "process information" and "control frictions." First, internal control deficiencies are systematically associated with accounting quality, risk premia, and fraud propensity. Firms that disclose material internal control weaknesses tend to exhibit poorer accrual quality and larger estimation errors [1], and face higher systematic and idiosyncratic risk as well as a higher cost of equity capital [2]. A weaker control environment is also significantly linked to

subsequent fraud revelations [3]. These findings imply that follow-up questioning on high-risk areas—such as revenue recognition, receivables collection, related-party transactions, expense capitalization, and remediation progress—together with evasive wording and response behavior, can leave measurable traces in interaction texts prior to statutory disclosure.

Second, the Q&A setting itself provides incremental information and is relatively "unscripted." Compared with prepared managerial statements, the language tone, specificity, and interaction structure of Q&A more strongly explain market reactions and subsequent risk. The Q&A portion of earnings calls carries independent information in tone and specificity [4], and deception-related linguistic cues in managerial answers can help detect potential reporting problems [5].

Accordingly, when weaknesses exist in the design or operation of internal control, they manifest as stable, learnable signals in the topic structure, semantic style, and response behavior observed in investor-interaction texts, thereby endowing the texts with testable predictive power for internal control deficiencies.

4. Construction of the Internal Control Deficiency Prediction Model

4.1. Prediction Task and Data Sources

The analysis is conducted at the firm–year level. Using public information available during the prior fiscal year, the model predicts internal control deficiencies disclosed in the subsequent year's annual report. Textual data are drawn from the Q&A records of online investor interaction platforms. Before ingestion, the Q&A texts undergo preprocessing that standardizes templated expressions, removes noise and duplicates, and normalizes placeholders for links and emojis. Firms with zero interaction are not excluded; instead, they are flagged with an explicit indicator variable to reflect genuine heterogeneity in interaction intensity and disclosure preferences across firms. Sample splitting follows a firm-grouping principle: training, validation, and test sets are isolated by firm to reduce cross-split information leakage for the same issuer and to enhance the external validity of the results.

4.2. Model Specification and Text Integration

Building on deep learning, three comparable models are specified within a unified framework to identify text signals.

First, the baseline model uses only structured features. The network adopts a deep architecture suitable for tabular data and outputs a firm–year probability of deficiency, serving as the performance benchmark without text.

Second, the early-fusion model augments the baseline with an annual text representation. A Chinese pre-trained language model encodes each individual question and answer into sentence- or segment-level vectors; multiple Q&A items within the same year are then aggregated by mean pooling, max pooling, or attention pooling to obtain a firm–year text vector. This vector is concatenated with structured features and fed into the same classification head, thereby assessing the direct incremental value of incorporating text.

Third, the hierarchical text model treats the multiple Q&A items within a year as a time-ordered sequence. Each Q&A retains its independent representation, which is then aggregated at the annual level by a hierarchical Transformer to capture structural information such as topic evolution, questioning order, and salient fragments.

4.3. Feature Engineering

The structured inputs primarily comprise financial-statement ratios. Organized by economic meaning, the features cover liquidity and short-term solvency (current ratio, quick ratio, current liabilities/total assets), profitability (return on assets and operating margin), operating efficiency (total asset turnover, inventory turnover, and accounts receivable turnover), capital structure and long-term leverage (debt-to-equity and noncurrent liabilities-to-equity), cash quality (operating cash flow/total assets, operating cash flow/revenue, and net cash flow/total assets), asset composition (the shares of intangible assets and of property, plant and equipment), and size and scale (the logarithms of total assets and revenue).

When financial data are subject to restatement or correction, the model uniformly adopts the originally disclosed values (pre-restatement/pre-correction) for training and evaluation so as to mirror the information set actually available at the time.

All inputs are strictly aligned in a forward-looking manner: to predict internal control deficiencies in period t+1, the model uses only information disclosed in period t or earlier. To limit the influence of outliers on estimation and training, continuous variables are winsorized at the 1st and 99th percentiles at the annual level. Text enters the model via either the annual-level representation or the hierarchical aggregation representation described earlier. Disclosures of material, significant, and general internal control deficiencies in internal control evaluation reports are uniformly coded as "deficiency present."

4.4. Training Procedure and Evaluation Protocol

The training procedure follows standard practice for imbalanced classification: a class-weighted loss is employed, with dropout and early stopping to control overfitting. Given potential domain differences between general-purpose corpora and the investor-interaction context, the pre-trained language model is used in a frozen-parameter mode to obtain text representations. When cross-year distributional shift is evident, small-step domain adaptation is conducted without compromising comparability. The evaluation strategy centers on firm-level out-of-sample assessment: the validation set is used for threshold selection and hyperparameter tuning, and the test set reports the out-of-sample results. Reporting focuses on AP as the primary metric, with ROC-AUC provided in parallel.

5. Performance Evaluation and Analysis of the Internal Control Deficiency Prediction Model

5.1. Data and Sample

Financial data for listed companies are obtained from CSMAR for the 2010–2022 period. The sample is filtered as follows: first, firms in the financial and insurance industries are excluded; second, observations with missing key financial variables are removed. Within the retained observations, approximately 5.6% disclose the presence of internal control

deficiencies. Investor-interaction texts are sourced from the Shenzhen Stock Exchange's and the Shanghai Stock Exchange's online investor Q&A platforms. Text data are processed on an annual window, including de-templating, de-duplication, and cleaning, while retaining firms with zero interaction.

5.2. Out-of-Sample Results

Relative to a no-text baseline that uses only structured features such as financial and governance variables (M0), the early-fusion model that incorporates investor-interaction texts (M1) delivers consistent improvements on the same test-set basis, as reported in Table 1. Average Precision (AP) increases from 0.082 to 0.094 (+14.63%), and ROC-AUC rises from 0.626 to 0.652 (+4.15%). The AP gain reflects enhanced ranking performance under class imbalance, indicating that text information improves the model's ability to identify potential deficiency cases. The concurrent improvement in ROC-AUC suggests a clearer decision boundary and better discrimination across thresholds.

Table 1. Out-of-Sample Performance Comparison (M0 vs. M1)

Metric	M0: No-text baseline	M1: Early-fusion model	Change (M1 vs. M0)
AP	0.082	0.094	+14.63%
ROC-AUC	0.626	0.652	+4.15%

Under a common training–validation–test split and the same evaluation protocol, the hierarchical text model (M2) delivers additional gains; Table 2 reports the results. Average Precision (AP) is 0.097, representing an 18.29% improvement over M0 and a further 3.19% over M1. ROC-AUC reaches 0.658, a 5.11% increase relative to M0 and an additional 0.92% over M1. These findings indicate that, relative to simple annual mean pooling, hierarchical aggregation of within-year Q&A sequences better exploits topic and order information embedded in the texts, thereby further improving the ranking of high-risk cases while maintaining overall discrimination.

Table 2. Out-of-Sample Performance of the Hierarchical Text Model (M2)

Metric	M2: Hierarchical text model	Change vs. M0	Change vs. M1
AP	0.097	+18.29%	+3.19%
ROC-AUC	0.658	+5.11%	+0.92%

6. Conclusion

Building on the exchange-based system of investor online interactions, this paper proposes a text-driven framework for predicting internal control deficiencies. Q&A content from investor-interaction platforms is transformed into forward-looking firm–year risk signals and fused with financial ratios within a unified deep learning framework, and effectiveness is tested through out-of-sample evaluation grouped by firm. Empirical evidence shows that, relative to a baseline without text, both the early-fusion model and the hierarchical text model deliver consistent improvements in Average Precision (AP) and ROC-AUC, indicating that investor-interaction texts reliably enhance the predictive performance for internal control deficiencies.

Acknowledgements

This work was supported by the Humanities and Social Sciences Research Youth Fund of the Ministry of Education of China (Project title: "Machine Learning for Internal Control Deficiency Prediction: Model Development and Applications"; Grant No. 19YJC790072).

References

[1] Doyle J T, Ge W, McVay S E. Accruals Quality and Internal Control over Financial Reporting [J]. The Accounting Review, 2007, 82(5): 1141–1170.

[2] Ashbaugh-Skaife H, Collins D W, Kinney W R Jr, LaFond R. The Effect of SOX Internal Control Deficiencies on Firm Risk and Cost of Equity [J]. Journal of Accounting Research, 2009, 47(1): 1–43.

[3] Donelson D C, Ege M, McInnis J M. Internal Control Weaknesses and Financial Reporting Fraud [J]. Auditing: A Journal of Practice & Theory, 2017, 36(3): 45–69.

[4] Price S M, Doran J S, Peterson D R, et al. Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone [J]. Journal of Banking & Finance, 2012, 36(4): 992–1011.

[5] Larcker D F, Zakolyukina A A. Detecting Deceptive Discussions in Conference Calls [J]. Journal of Accounting Research, 2012, 50(2): 495–540.