# Computational Finance Approaches to Detecting Information Asymmetry in Corporate Disclosures

Tongshan Zhu
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
zhu2003@ucf.edu

Blake A. Horton
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
blakework@unr.edu

**Abstract**

Corporate disclosure is the central mechanism through which public firms communicate value-relevant information to external stakeholders, but disclosure narratives are also strategic artifacts shaped by managerial discretion. Language choice, structural ordering, emphasis, and omission can preserve information asymmetry even within mandatory reporting regimes. This paper develops a systems-level analysis of computational finance approaches for detecting information asymmetry in corporate disclosures. It examines how textual analysis, natural language processing, and machine learning transform unstructured filings, risk-factor sections, earnings calls, and regulatory submissions into measurable signals of semantic deviation, tone, specificity, and anomaly. The discussion moves beyond isolated model evaluation to consider representation design, detection architecture, interpretability, deployment cadence, and human oversight as interconnected components of a disclosure monitoring infrastructure. In doing so, the paper analyzes structural trade-offs among prediction accuracy, fairness, robustness, regulatory legitimacy, and operational sustainability. It draws on evidence from accounting, finance, machine learning, and algorithmic governance to show that effective detection systems must be understood as socio-technical infrastructures rather than purely statistical instruments. The analysis concludes with implications for regulators, audit institutions, and platform designers seeking to reduce information asymmetry without creating new forms of opacity, bias, or systemic fragility.

**Keywords**

corporate disclosure, information asymmetry, computational finance, natural language processing, anomaly detection, explainable artificial intelligence, regulatory technology, algorithmic fairness.

## 1. Introduction

Information asymmetry in corporate disclosure arises when managers possess material information that outside investors cannot directly observe or verify. Mandatory reporting rules reduce some of this asymmetry by standardizing financial statements and requiring periodic filings, yet substantial discretion remains in the narrative portions of corporate communication. Early computational finance research began to address this problem by applying dictionary-based textual analysis to regulatory filings and financial news, showing that the tone and word choice of corporate narratives carry information about future performance and valuation [1]. A parallel line of work demonstrated that media sentiment has measurable effects on stock market outcomes, reinforcing the view that language is not a neutral transmission channel for financial information [2]. These foundational contributions established the intellectual basis for using computational text analysis to detect managerial signaling and selective disclosure.

Over the past two decades, the field has shifted from simple word-count methods toward more sophisticated representations of corporate language, including machine learning classifiers, topic models, contextual embeddings, and anomaly detection architectures. This evolution reflects both the increasing volume of unstructured financial text and the recognition that information asymmetry is often encoded in subtle structural properties of disclosure rather than in isolated keywords. The present paper adopts a systems perspective, treating disclosure monitoring as an infrastructure that integrates data ingestion, representation learning, detection algorithms, explanation mechanisms, governance controls, and institutional oversight. Such a perspective is necessary because the practical value of any computational approach depends not only on statistical discrimination but also on its reliability, auditability, equity, and capacity to function under shifting regulatory and economic conditions.

## 2. Information Asymmetry in Corporate Disclosure: A Systems Perspective

Corporate disclosure can be understood as an information infrastructure through which firms signal their current condition and future prospects to capital markets. The quality of this infrastructure is shaped by accounting standards, legal liability, exchange listing requirements, and the strategic incentives of managers. Computational work on forward-looking statements in corporate filings demonstrated that the information content of narrative disclosure is substantial and that machine learning methods can recover managerial signals even when those signals are embedded in complex language [3]. Mandatory risk factor disclosures subsequently became a central object of study because they require firms to discuss material risks, yet they also permit considerable latitude in how risks are framed and aggregated. Research on mandatory risk factor disclosures found that the length and content of these sections are associated with information asymmetry, firm risk, and market behavior [4]. However, disclosure volume alone does not guarantee transparency. Specificity is critical, and more granular risk-factor language has been linked to reduced information asymmetry and improved capital market outcomes [5].

The persistence of asymmetry in mandatory narratives reflects the fact that managers can use generic language, boilerplate repetition, and strategic omission to comply with regulatory requirements while withholding decision-relevant detail. Later research advanced beyond static keyword lists to discover and quantify latent risk types directly from textual risk disclosures, illustrating the value of unsupervised and semi-supervised methods for mapping the semantic structure of corporate risk reporting [6]. Early predictive systems translated such textual variation into risk scores using regression-based architectures, and these systems showed that text-derived features can complement traditional accounting variables in predicting financial distress and risk outcomes [7]. From a systems perspective, however, these models are only one component of a broader monitoring environment that includes reporting incentives, enforcement intensity, auditor scrutiny, and investor attention.

## 3. Computational Representations of Corporate Disclosure

The construction of a computational representation is the first architectural decision in any disclosure monitoring system. Research on initial public offering prospectuses treated the narrative as a high-dimensional content space in which meaningful variation across firms can be extracted and compared [8]. In adjacent work, product descriptions in annual filings were used to construct text-based industry networks, demonstrating that corporate language can reveal competitive positioning and sectoral relationships that are not directly captured by standard industrial classification [9]. These studies illustrate that representation choices determine which dimensions of information asymmetry become visible to downstream models.

A separate line of inquiry used 10-K text to gauge financial constraints, showing that narrative complexity and tone can serve as proxy signals for financing frictions when quantitative variables are ambiguous or incomplete [10]. Methodological reviews of text-as-data approaches have emphasized that representation validity depends on the match between the linguistic model and the institutional context of financial reporting, including genre conventions, temporal drift, and litigation sensitivity [11]. Direct managerial sentiment measures extracted from earnings calls added predictive information about stock returns, further confirming that richer representations can reveal private information beyond standard accounting numbers [12]. More recently, structural deviation in 10-K risk factors has been modeled through semantic anomaly detection and explainable artificial intelligence, focusing on how firms depart from their own historical disclosure patterns and from peer norms [13]. This line of work shifts attention from absolute word frequency to relative structural deviation, which is often more diagnostic of strategic disclosure behavior.

## 4. Detection Architectures for Asymmetric Information Signals

Modern detection architectures increasingly rely on deep learning to model the sequential and contextual nature of corporate narratives. Attention-based sequence models have transformed natural language processing by allowing systems to learn long-range dependencies and context-sensitive representations without manual feature engineering [14]. Deep learning architectures more generally have shown that hierarchical feature learning can uncover latent regularities in unstructured data, although their full potential in financial disclosure monitoring depends on careful task specification and evaluation design [15]. These architectures are particularly suited to detecting subtle deviations in risk-factor sections, earnings call transcripts, and management discussion and analysis.

However, architectural sophistication carries governance risks. Very large language models can generate fluent but unreliable summaries of corporate disclosures, and their deployment in high-stakes financial settings raises concerns about stochastic outputs, auditability, and the reproduction of unintended biases [16]. Detection systems must therefore be designed with constraints that preserve human interpretability, enforce domain-specific validation, and subject model outputs to institutional review. This is not a purely technical matter. The fairness properties of machine learning systems are shaped by training data, label definitions, and evaluation metrics, and disclosure monitoring systems can inherit biases from historical reporting practices or from imbalanced regulatory enforcement [17]. In the context of sentiment analysis and textual measurement, even widely used tools have been shown to encode demographic bias, which could distort assessments of managerial tone across different executive groups [18]. These issues indicate that detection architecture must be evaluated not only by aggregate accuracy but also by its distributional consequences across firms, industries, and stakeholders.

## 5. Structural Trade-offs in Model Design and Deployment

System designers face persistent trade-offs between predictive power and interpretability. Deep contextual models may improve detection of subtle narrative manipulation, but they often function as black boxes that are difficult for auditors, regulators, and courts to evaluate. Conversely, simpler lexical or regression-based systems offer greater transparency but may miss structural anomalies that do not reduce to individual keyword shifts. Fairness-aware design compounds this trade-off because constraints intended to reduce group-level bias can alter model performance in unexpected ways [17]. Bias in textual measurement can emerge from training corpora, annotation protocols, or the choice of sentiment lexicons, and such bias can produce misleading evaluations of disclosure quality across different executive and industry contexts [18].

Deployment conditions introduce further structural constraints. Corporate disclosure monitoring must operate under strict latency requirements during earnings seasons, handle heterogeneous filing formats and languages, and adapt to regulatory changes across jurisdictions. The integration of machine learning into corporate governance decisions, such as director selection, illustrates both the efficiency potential and the legitimacy challenges of algorithmic decision support in contexts where fiduciary duties and legal standards remain qualitative [19]. A detection system that flags anomalous disclosure must produce outputs that can be explained, contested, and audited by human reviewers. Without this institutional scaffolding, algorithmic flags may be ignored or over-trusted, both of which can weaken the overall monitoring regime.

## 6. Governance, Fairness, and Regulatory Implications

Corporate disclosure monitoring is embedded within a larger regulatory environment that increasingly includes environmental, social, and governance disclosures. Investors use ESG information to assess long-term risk and management quality, but the heterogeneity of ESG reporting creates new opportunities for selective disclosure and greenwashing [20]. Mandatory CSR and sustainability reporting can reduce information asymmetry, but it also imposes compliance burdens and may generate information overload if the disclosed metrics are not standardized or comparable [21]. Computational monitoring systems must therefore be capable of comparing textual and quantitative ESG claims across firms and over time, while respecting the evolving nature of sustainability standards.

Ensemble-based monitoring architectures offer one way to integrate heterogeneous signals from different models and data sources while limiting the fragility associated with any single detection method [22]. In a disclosure monitoring context, ensemble designs can combine lexical, semantic, structural, and temporal anomaly detectors to produce more stable indications of asymmetric information. The governance challenge is to preserve the diversity of these model components so that their errors do not become correlated. Regulatory authorities may also require that disclosure monitoring systems used for enforcement or market surveillance undergo independent validation, bias audits, and periodic recalibration. These requirements reflect a broader shift toward algorithmic accountability in financial markets, where computational tools are increasingly viewed as part of the regulatory infrastructure rather than neutral analytical aids.

## 7. Sustainability and Robustness of Disclosure Monitoring Infrastructure

A sustainable disclosure monitoring infrastructure must be robust to changes in reporting practices, economic regimes, and modeling assumptions. Corporate language evolves as accounting standards change, as new risk categories emerge, and as firms adapt their disclosure strategies in response to detection systems themselves. This adversarial dynamic creates a feedback loop in which monitored firms may learn to avoid flagged patterns while preserving underlying information asymmetry through more subtle rhetorical strategies. Detection systems that rely on static lexicons or fixed training windows are likely to degrade over time. Continuous recalibration, adversarial evaluation, and periodic human review are therefore essential components of a resilient monitoring architecture.

Sustainability also has an institutional dimension. Monitoring systems require sustained access to high-quality filing data, adequate computational resources, and personnel capable of interpreting model outputs in legal and financial contexts. Smaller regulatory agencies or audit teams may lack the capacity to maintain sophisticated machine learning pipelines, which can create uneven enforcement across jurisdictions. Cloud-based shared infrastructure and open benchmarking datasets can mitigate some of these disparities, but they also raise concerns about data governance, confidentiality, and vendor lock-in. A well-designed disclosure monitoring system must balance technical sophistication with operational simplicity so that it can function reliably over long time horizons without requiring continuous expert intervention.

## 8. Future Research Directions

Future research should examine multimodal disclosure monitoring that integrates textual narratives with tabular financial data, earnings call audio, visual presentations, and market microstructures. The combination of these modalities may reveal information asymmetries that are not detectable in any single source. Research is also needed on causal identification in textual anomaly detection, because many apparent anomalies may reflect legitimate changes in business conditions rather than strategic concealment. Natural experiments, regulatory changes, and external shocks can provide identification opportunities for separating informative deviation from benign variation.

Further attention should be directed to the explainability of disclosure anomaly systems. Explanations that highlight specific passages, numeric inconsistencies, or temporal deviations can support auditor judgment, but they can also create confirmation bias if presented without appropriate uncertainty information. The development of standardized evaluation protocols for disclosure monitoring systems, including measures of fairness, robustness, and drift, would help regulators compare tools from different vendors and research groups. Finally, interdisciplinary collaboration among accounting researchers, computer scientists, legal scholars, and market regulators is necessary to align technical progress with the institutional goals of transparency, fairness, and investor protection.

## 9. Conclusion

Computational finance approaches to detecting information asymmetry in corporate disclosures have moved from simple lexicon-based sentiment analysis to contextual anomaly detection and explainable machine learning. These methods can reveal managerial signals embedded in narrative structure, risk-factor specificity, tone, and cross-sectional deviation. However, the effectiveness of such systems depends on more than model accuracy. It depends on the representation of corporate language, the design of detection architectures, the integration of human oversight, and the governance mechanisms that ensure fairness, robustness, and institutional legitimacy.

A systems-level perspective clarifies that disclosure monitoring is a socio-technical infrastructure rather than an isolated analytical task. Structural trade-offs among interpretability, fairness, timeliness, and operational sustainability must be managed explicitly. Regulatory and audit institutions have a growing role in shaping the norms and standards under which these systems operate. As corporate disclosure continues to expand in volume and complexity, computational monitoring will become an increasingly important component of capital market transparency. Its development should therefore be guided by careful attention to both empirical performance and the broader institutional consequences of algorithmic detection in financial governance.

**References**

1. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35-65.

2. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139-1168.

3. Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049-1102.

4. Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H. M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19(1), 396-455.

5. Hope, O.-K., Hu, D., & Lu, H. (2016). The benefits of specific risk-factor disclosures. Review of Accounting Studies, 21(4), 1005-1045.

6. Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60(6), 1371-1391.

7. Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference, 272-280.

8. Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. Review of Financial Studies, 23(7), 2821-2864.

9. Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423-1465.

10. Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. Journal of Financial and Quantitative Analysis, 50(4), 623-646.

11. Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535-574.

12. Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. Journal of Financial Economics, 132(1), 126-149.

13. Sun, F., He, S., Wang, R., Ke, L., Shen, H., & Liao, Q. (2026). Modeling Structural Deviation in 10-K Risk Factors: A Semantic Anomaly Detection and Explainable AI Approach. Risks, 14(4), 87.

14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

18. Thelwall, M. (2018). Gender bias in sentiment analysis. Online Information Review, 42(1), 45-59.

19. Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting directors using machine learning. Review of Financial Studies, 34(7), 3226-3264.

20. Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. Financial Analysts Journal, 74(3), 87-103.

21. Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. Review of Accounting Studies, 26(3), 1176-1248.

22. Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple Classifier Systems (pp. 1-15). Springer.