Explainable Machine Learning for Predicting Stock Price Volatility from Corporate Risk Disclosures

Kean Birry
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
seanwork@oregonstate.edu

FanLong Pan
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
fanp@ucf.edu

Pradeep M. Mishra
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hellopradeep@uc.edu

Abstract

Predicting stock price volatility from corporate risk disclosures presents a complex socio-technical problem that sits at the intersection of financial economics, natural language processing, systems engineering, and regulatory governance. This paper develops a systems-level examination of explainable machine learning architectures designed to forecast volatility using qualitative risk factor narratives in corporate filings. Rather than proposing a single algorithmic solution, the analysis focuses on structural trade-offs across data ingestion, textual representation, volatility target selection, predictive modeling, explanation generation, and audit infrastructure. The discussion emphasizes that explainability in financial machine learning must operate across multiple institutional audiences, including model developers, risk managers, regulators, and investors. It further examines how semantic anomaly detection in risk disclosures can function as a governance safeguard to flag structural deviations that may undermine model confidence. The paper addresses robustness under concept drift, sustainability of large language model pipelines, fairness across firms and sectors, and policy implications for algorithmic accountability. By integrating perspectives from accounting, finance, machine learning, and regulatory science, the paper offers a forward-looking framework for designing, deploying, and governing explainable volatility prediction systems that balance predictive performance with institutional legitimacy, auditability, and operational resilience.

Keywords

explainable machine learning; volatility forecasting; corporate risk disclosures; financial text analysis; model governance; algorithmic accountability.

1. Introduction

The modern financial information environment is increasingly shaped by unstructured corporate narratives that influence investor expectations, risk assessments, and trading behavior. Among these narratives, mandatory risk factor disclosures in periodic filings have attracted sustained scholarly attention because they are formally regulated, periodically updated, and directly relevant to forward-looking uncertainty. Early text-based research in

finance demonstrated that general-purpose dictionaries often misclassify financial language, while domain-specific lexicons can better capture the nuanced meaning of terms such as liability, risk, and uncertainty [1]. Subsequent survey work has shown that textual analysis has become a mature but still evolving methodology in accounting and finance, with important implications for measuring disclosure tone, complexity, and specificity [2]. In parallel, empirical studies have established that the information content of mandatory risk factor disclosures is associated with changes in firm-level risk perceptions and market outcomes [3]. Investor reactions to textual risk disclosures further suggest that qualitative narratives convey incremental information beyond quantitative accounting variables [4].

Within this context, predicting stock price volatility from corporate risk disclosures is not simply a forecasting exercise. Volatility is central to option pricing, risk management, portfolio construction, and regulatory capital calculation. A system that translates long-form qualitative risk narratives into forward-looking volatility signals must therefore satisfy demanding institutional requirements. These requirements include predictive accuracy, interpretability, stability over time, robustness to changing disclosure practices, and compatibility with audit expectations. The explainability imperative arises because financial institutions must justify model outputs to internal risk committees, external auditors, and regulators. As text-as-data research has formalized the inferential and measurement challenges associated with unstructured data [5], there is a corresponding need to examine how machine learning systems can be designed to remain transparent under conditions of high dimensionality and rapid market adaptation. This paper addresses that gap by providing a system-level examination of explainable machine learning for volatility prediction from corporate risk disclosures, with particular attention to architecture, governance, deployment, fairness, and sustainability.

2. Corporate Risk Disclosures as Prediction Inputs

Corporate risk disclosures, particularly the risk factor section of annual reports, differ from news articles, analyst reports, and social media content in several important respects. They are produced under regulatory pressure, reviewed by legal counsel, and shaped by litigation concerns. The language in these disclosures tends to be dense, repetitive, and organized around broad categories such as competition, regulation, cybersecurity, supply chain disruption, credit conditions, and operational hazards. Although some disclosures rely heavily on boilerplate language, deviations from standard templates can be informative. Prior studies have shown that textual risk disclosures influence investor risk perceptions, even after controlling for quantitative risk measures [4]. Mandatory risk factor disclosures also provide incremental information about firm-level uncertainty and can affect trading activity and market valuations [3].

From a representation perspective, the high dimensionality of corporate risk language creates immediate challenges. Early bag-of-words and dictionary approaches reduce text into counts or sentiment categories, but they often discard syntactic relationships, negation, and contextual meaning. Transformer architectures introduced self-attention mechanisms that allow models to relate different parts of a document without relying on sequence-aligned recurrence [6]. Bidirectional pretraining further enabled contextual representations that capture word meaning in relation to surrounding text [7]. Domain-adapted models such as FinBERT have extended these representations to financial language by pretraining on financial corpora, improving performance on sentiment and classification tasks relevant to investment analysis [8]. In a volatility prediction system, contextual encoders can convert risk

disclosure sections into dense vector representations. However, these representations are not inherently interpretable, and their usefulness depends on how the predictive layer consumes them. The system architect must decide whether the encoder remains frozen, is fine-tuned, or is replaced by simpler count-based representations for specific governance-sensitive tasks.

3. Volatility as a Prediction Target

Stock price volatility is a latent construct that must be approximated through observable proxies. Common proxies include realized volatility computed from high-frequency returns, implied volatility extracted from option prices, and idiosyncratic volatility estimated relative to factor models. Empirical work has documented stylized facts such as volatility clustering, heavy tails, and asymmetric responses to negative return shocks [10]. The behavior of conditional variance is also affected by leverage and interest rate effects, making volatility a dynamic and persistent object of study [11]. These statistical properties have direct consequences for machine learning systems. A model trained to predict a specific volatility proxy may not transfer cleanly to another proxy, and performance may differ across calm and turbulent market regimes.

Machine learning has become an influential tool in empirical asset pricing and financial prediction. Large-scale studies have found that machine learning methods can improve return and volatility forecasts by capturing nonlinear interactions and high-dimensional predictors [9]. Nevertheless, forecasting gains in financial markets are frequently unstable because the informational content of predictors shifts as market participants adapt. A risk disclosure phrase that once signaled elevated uncertainty may become normalized over time, diminishing its predictive value. This adaptive character of financial systems must be embedded into the design of volatility prediction infrastructure. The target definition, evaluation window, and recalibration policy should be treated as first-order design decisions rather than technical afterthoughts.

4. System Architecture and Structural Trade-offs

A responsible volatility prediction system can be organized as a layered pipeline consisting of document ingestion, section identification, text normalization, semantic encoding, volatility modeling, explanation generation, and governance reporting. Each layer introduces distinct failure modes. Ingestion failures can produce missing or duplicated filings. Section identification errors can place irrelevant text into the risk factor corpus. Text normalization errors can alter the meaning of negative statements. Encoder drift can render a once-valid representation stale. Model miscalibration can generate misleading confidence intervals. Governance failures can allow model updates to occur without corresponding updates to documentation and explanation artifacts.

The system must therefore be designed for adaptive market behavior. The predictive content of disclosures changes over time as regulations evolve, as industries develop new risk vocabularies, and as investors adjust their expectations [12]. Structural trade-offs arise when deciding between end-to-end deep learning pipelines and modular architectures that separate representation from forecasting. End-to-end models may optimize predictive performance jointly, but they often produce opaque internal representations that are difficult to audit. Modular systems can use a frozen text encoder followed by a more transparent statistical learner, thereby enabling clearer attribution of volatility predictions to specific sections or topics. Hybrid systems may fine-tune an encoder but distill its behavior into simpler surrogate models for explanation purposes. These choices affect not only accuracy but also

computational cost, retraining frequency, and the stability of explanations. A system that optimizes only for predictive accuracy may fail to maintain institutional trust when its explanations shift unpredictably after each retraining.

5. Explainability Layers and Interpretation Methods

Explainability in financial machine learning must serve multiple audiences with different analytical needs. Model developers require diagnostic attributions to identify overfitting or spurious correlations. Risk managers require stable feature importance measures that can be reviewed against economic intuition. Regulators require documentation that connects model outputs to controlled inputs and defined limitations. Investors and analysts require plain-language explanations that do not overstate causal claims. No single explanation technique satisfies all these audiences simultaneously.

Shapley value explanations offer a theoretically grounded approach to feature attribution, but their exact computation becomes intractable for high-dimensional textual inputs, and approximations may be sensitive to the choice of background data [13]. Local surrogate methods provide contrastive explanations by perturbing input text and observing model response, yet perturbed text can produce unnatural sequences that lie outside the training distribution [14]. Concept-level explanation methods attempt to move beyond token-level attribution by testing whether internal model states align with human-defined concepts such as liquidity, regulatory uncertainty, or cybersecurity exposure [15]. In volatility prediction, concept-level explanations are often more meaningful for governance purposes than token-level salience. A particular phrase may matter less than the broader risk topic it activates.

A mature explainability architecture should therefore support hierarchical aggregation. Token-level attributions can be rolled up to clause-level signals, then to risk topic categories, and finally to a volatility impact statement. It should also distinguish between lexical explanations and temporal explanations. A prediction may change because the disclosure language changed, because peer behavior changed, or because the model mapping was updated. If these sources of change are not separated, the system may produce technically correct but operationally misleading explanations.

6. Operational Governance, Documentation, and Anomaly Detection

Explainability is inseparable from documentation and lifecycle governance. Structured model documentation practices, such as model cards, provide a mechanism for recording intended use, training data, evaluation procedures, and known limitations [16]. Financial regulators have similarly highlighted the importance of data quality, model risk management, and accountability when artificial intelligence and machine learning are used in financial services [17]. In a volatility prediction system, documentation should include the volatility target definition, the disclosure corpus version, the labeling window, the text encoder version, the predictive model class, and the explanation technique. Without this documentation, an explanation produced by one model version can easily become detached from the model currently deployed in production.

An important governance capability involves detecting structural anomalies in risk disclosures themselves. Recent research has proposed modeling structural deviation in 10-K risk factors through semantic anomaly detection and explainable artificial intelligence [18]. Such methods compare a firm's current disclosure against its own historical patterns and against peer templates. Structural deviation may indicate emerging risk, a change in litigation strategy, or managerial obfuscation. From a systems perspective, anomaly detection should not

necessarily replace standard volatility forecasts. Instead, it can operate as a gating mechanism that reduces confidence when a filing is highly anomalous or routes it for human review. This safeguard aligns predictive automation with operational resilience and provides a concrete link between explainability and institutional oversight.

7. Robustness, Drift, and Sustainability

Financial text models are particularly vulnerable to concept drift because disclosure practices, accounting standards, and market vocabulary evolve over time. A model trained on one period may perform well in backtests but degrade when filing styles change, when new risk categories emerge, or when market regimes shift. Robustness engineering requires continuous monitoring of input distributions, periodic model retraining, and stress testing against crises that may not appear in the training window. Algorithmic audit frameworks emphasize that accountability should be continuous rather than episodic [19]. This means that explanation stability should be monitored alongside predictive performance. If a model update improves accuracy but changes attribution patterns in ways that cannot be explained to risk managers, its operational value may be reduced.

Regulatory sustainability also requires attention to legal expectations for automated decision-making. In jurisdictions influenced by European data protection and algorithmic transparency debates, there is growing interest in whether individuals and institutions have a right to explanation for significant automated decisions, although the legal scope remains uncertain [20]. For volatility prediction from corporate disclosures, the primary affected parties are firms, investors, and market intermediaries rather than individual consumers. Nevertheless, policy-oriented research has cautioned that predictive modeling for policy and regulation should be supplemented by causal analysis and institutional context [21]. A sustainable system should therefore communicate predictive associations without implying causal claims about managerial intent or legal liability. This distinction is essential for regulatory acceptance and ethical legitimacy.

8. Fairness, Policy, and Cross-Sectoral Implications

Fairness in volatility prediction from corporate risk disclosures extends beyond individual-level anti-discrimination. It involves equitable treatment of firms, sectors, and disclosure styles. Smaller firms may use less standardized language, making their filings more likely to be classified as anomalous or high risk. Firms with limited disclosure history may be underrepresented in training data. International filers may exhibit linguistic patterns that language models pretrained primarily on United States filings do not represent well. Large pretrained language models can also encode social and economic biases from their training corpora, which raises concerns about the fairness and environmental sustainability of increasingly large models [22]. A responsible system should therefore report disaggregated performance across firm size, sector, filing length, and complexity. It should also evaluate whether prediction errors are concentrated among particular groups of firms and whether those errors could influence capital allocation or regulatory scrutiny.

Cross-sectoral comparisons reveal that the same explainability requirements appear in credit risk, insurance underwriting, and fraud detection, though the specific fairness definitions differ. Deep learning architectures provide flexible representations that support multiple output tasks, but they require disciplined engineering to maintain stability and interpretability [23]. In volatility prediction, peer comparison can be enriched by product market similarity networks, which capture firm relationships beyond industry classification [24]. Such networks

improve contextualization but also introduce correlated errors when product market boundaries change. Systematic training and evaluation practices remain essential for maintaining generalization and avoiding silent failures [25]. Policy frameworks should therefore encourage model documentation, anomaly detection, disaggregated evaluation, and human review channels rather than prescribing a single algorithmic technique. This systems-level policy orientation is more likely to support innovation while protecting market integrity.

9. Conclusion

Explainable machine learning for predicting stock price volatility from corporate risk disclosures requires a deliberate synthesis of textual representation, financial econometrics, system architecture, and governance. The central challenge is not merely to improve predictive accuracy but to design a system that remains intelligible across multiple institutional contexts. Risk disclosure narratives are high-dimensional, adaptive, and shaped by legal and regulatory forces. Volatility itself is latent and sensitive to market conditions. A layered architecture that separates ingestion, encoding, forecasting, explanation, anomaly detection, and governance can help manage this complexity. At the same time, such modularity introduces trade-offs in predictive performance and computational cost. Explainability must be treated as a cross-cutting property supported by structured documentation, hierarchical interpretation, and semantic anomaly detection. Future work should examine how disclosure practices evolve across jurisdictions, how language model updates affect volatility forecasts, and how audit frameworks can be made sufficiently flexible to accommodate rapid technological change. By adopting a systems perspective, researchers and practitioners can build volatility prediction infrastructure that is not only accurate but also robust, fair, and institutionally legitimate.

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. Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187–1230.

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

4. Kravet, T., & Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. Review of Accounting Studies, 18(4), 1088–1122.

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

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

7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). Association for Computational Linguistics.

8. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

9. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5), 2223–2273.

10. Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance, 1(2), 223–236.

11. Christie, A. A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. Journal of Financial Economics, 10(4), 407–432.

12. Lo, A. W. (2017). Adaptive Markets: Financial Evolution at the Speed of Thought. Princeton University Press.

13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).

14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In International Conference on Machine Learning (pp. 2668–2677). PMLR.

16. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of FAT 2019 (pp. 220–229). ACM.

17. Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services. Financial Stability Board.

18. 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.

19. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of FAccT 2020 (pp. 33–44). ACM.

20. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. AI Magazine, 38(3), 50–57.

21. Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324), 483–485.

22. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of FAccT 2021 (pp. 610–623). ACM.

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

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

25. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.