Modeling Structural Deviation in 10-K Risk Factors: A Semantic Anomaly Detection and Explainable AI Approach

Fang Sun , Shuangjiang He , Ruiqi Wang , Lingyun Ke , Hongyu Shen , Qiuyue Liao *

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0179.v1

Keywords: risk factors; financial disclosure; structural deviation; anomaly detection; semantic modeling; explainable artificial intelligence; regulatory risk; 10-K analysis

 doi:10.20944/preprints202603.0179.v1

Article

Modeling Structural Deviation in 10-K Risk Factors: A Semantic Anomaly Detection and Explainable AI Approach

Fang Sun [1], Shuangjiang He [2], Ruiqi Wang [3], Lingyun Ke [4], Hongyu Shen [5] and Qiuyue Liao [6,*]

[1] Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA
[2] Information Technology Program, University of the Cumberlands, Williamsburg, KY 40769, USA
[3] College of Graduate and Professional Studies, Trine University, Angola, IN 46703, USA
[4] Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA
[5] School of Computer Science, Cornell Tech, New York, NY 10044, USA
[6] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
* Correspondence: qliao34@gatech.edu

Abstract

This study examines structural changes in regulatory risk disclosure using a semantic modeling framework that integrates sentence embeddings, multivariate anomaly detection, and explainable artificial intelligence. Prior research typically relies on dictionary-based word frequencies, tone indicators, or topic proportions to quantify risk disclosure. While these measures capture disclosure intensity, they do not directly assess whether the internal semantic organization of risk narratives has shifted relative to historical patterns. We propose a structural semantic deviation framework that represents each company-year disclosure using thematic shares and embedding-based dispersion statistics, and evaluates deviations from a historical baseline through unsupervised anomaly detection. Using Item 1A Risk Factors from Wells Fargo and JPMorgan Chase surrounding the 2016 regulatory shock, we demonstrate that traditional lexical metrics fail to isolate structural breaks, whereas embedding-based semantic trajectories reveal substantial narrative reconfiguration. Isolation-based modeling provides stable and discriminative anomaly scores, and SHAP decomposition identifies semantic distance, litigation emphasis, and disclosure contraction as key drivers of deviation in 2025 out-of-sample disclosures. The results suggest that structural semantic modeling captures risk narrative transformation beyond word accumulation, offering an interpretable and scalable framework for regulatory risk assessment.

Keywords: risk factors; financial disclosure; structural deviation; anomaly detection; semantic modeling; explainable artificial intelligence; regulatory risk; 10-K analysis

1. Introduction

Public companies in the United States are required to disclose material risks in the Risk Factors section of Form 10-K filings. These disclosures are intended to inform investors about operational, regulatory, legal, and governance exposures that may affect future performance. Over the past two decades, the length and complexity of these sections have increased substantially, raising concerns about information overload and the diminishing marginal informativeness of additional disclosure text. Prior research has documented both the expansion and the strategic use of risk language in corporate filings, suggesting that risk disclosure serves not only an informational role but also a defensive and reputational function (Campbell et al. 2014; Kravet and Muslu 2013).

The 2016 Wells Fargo unauthorized accounts scandal marked a critical moment in the governance landscape of large U.S. banks. Following regulatory investigations, consent orders, and intensified supervisory oversight, financial institutions faced heightened scrutiny regarding compliance, internal controls, and conduct risk. Empirical studies show that firms often respond to regulatory shocks and litigation exposure by increasing the volume and specificity of risk disclosures (Bao et al. 2018; Beatty

 doi:10.20944/preprints202603.0179.v1

et al. 2010). In the banking sector, post-crisis disclosure patterns frequently reflect shifts in regulatory emphasis and legal vulnerability (Acharya et al. 2012).

Traditional approaches to measuring financial risk disclosure have relied heavily on dictionary-based word counts, sentiment indicators, and topic modeling techniques. The use of domain-specific dictionaries, particularly the financial sentiment lexicon developed by Loughran and McDonald, has become standard practice in accounting and finance research (Loughran and McDonald 2011). Other studies employ topic modeling to identify shifts in thematic emphasis within risk factor sections (Huang et al. 2014) or track changes in uncertainty and negative tone over time (Li 2010). While these methods provide valuable high-level indicators, they treat text primarily as collections of words or topics. As a result, they are limited in their ability to capture structural reorganization within the semantic space of disclosures.

In practice, two firms may exhibit similar levels of negative or uncertainty-related vocabulary while differing substantially in how risk narratives are structured and interconnected. A shift in semantic centrality, dispersion, or narrative emphasis may signal deeper institutional change than incremental increases in word frequency. Recent advances in sentence embedding models enable the representation of disclosure text in continuous semantic space, allowing for the measurement of distributional properties beyond surface-level word counts (Reimers and Gurevych 2019). At the same time, anomaly detection techniques have been increasingly applied in financial contexts to identify unusual patterns relative to historical baselines (Chandola et al. 2009). However, the integration of semantic distribution metrics with explainable anomaly detection remains underexplored in the analysis of regulatory risk disclosure.

This study proposes a framework that combines topic share features, semantic dispersion metrics derived from sentence embeddings, and anomaly detection methods to identify structural deviations in risk narratives. Rather than focusing solely on disclosure intensity, we model baseline stability over rolling historical windows and assess whether subsequent filings represent incremental evolution or structural reframing. To enhance interpretability, we apply feature-level explanation methods to decompose anomaly scores into economically meaningful components.

Using annual 10-K filings from Wells Fargo and JPMorgan Chase surrounding the 2016 regulatory shock, we conduct a comparative analysis of disclosure trajectories. Although both institutions exhibit increases in risk-related language over time, we find substantial differences in semantic dispersion and structural deviation. The results suggest that elevated anomaly scores are more strongly associated with shifts in semantic centrality and litigation emphasis than with raw increases in negative wording.

This paper makes three contributions. First, it extends the literature on financial risk disclosure by introducing semantic dispersion indicators that capture internal distributional change within annual filings. Second, it demonstrates the importance of baseline stability testing in text-based anomaly detection to mitigate false structural signals. Third, it provides interpretable decomposition of anomaly scores in a regulatory-sensitive setting, offering a transparent bridge between machine learning outputs and economic meaning.

By reframing risk disclosure analysis as a problem of structural semantic deviation rather than simple word accumulation, this study provides a more nuanced approach to understanding how financial institutions adapt their public risk narratives in response to regulatory shocks.

2. Related Work

2.1. Dictionary-Based Risk and Tone Measures

A substantial body of literature measures financial disclosure risk using dictionary-based word frequency methods. The most influential contribution in this domain is the financial sentiment dictionary developed by Loughran and McDonald (Loughran and McDonald 2011), which demonstrated that general-purpose sentiment lexicons are poorly suited for financial texts and introduced domain-specific negative, uncertainty, and litigation word categories. Their work established a foundation for quantifying tone in 10-K filings using word intensity metrics.

 doi:10.20944/preprints202603.0179.v1

Subsequent research has applied dictionary-based measures to examine the information content of risk disclosures. Campbell et al. (Campbell et al. 2014) document that mandatory risk factor disclosures contain incremental information for investors. Kravet and Muslu (Kravet and Muslu 2013) show that textual risk disclosures are associated with investor perceptions of firm-level risk. Li (Li 2010) examines forward-looking statements and demonstrates that textual tone conveys economically meaningful information. These studies typically operationalize risk disclosure intensity through counts or proportions of predefined word categories.

Although dictionary-based approaches are transparent and easy to implement, they rely on the assumption that individual word frequencies sufficiently capture disclosure dynamics. They do not account for contextual relationships between sentences, shifts in semantic emphasis, or changes in narrative structure. As a result, firms that maintain similar word intensity levels but reorganize risk narratives in structurally meaningful ways may not be distinguished.

2.2. Topic Modeling and Thematic Analysis

Topic modeling techniques have been widely adopted to explore thematic patterns in corporate disclosures. Latent Dirichlet Allocation introduced by Blei et al. (Blei et al. 2003) provides a probabilistic framework for identifying latent topics within large text corpora. In financial applications, topic modeling has been used to detect changes in disclosure emphasis across time and industries (Hoberg and Phillips 2016).

Huang et al. (Huang et al. 2014) examine tone management strategies and show that firms adjust disclosure emphasis in response to performance outcomes. Other studies use topic distributions to evaluate regulatory focus and litigation exposure in risk factor sections (Bao et al. 2018). Topic proportions are often interpreted as indicators of strategic disclosure adjustments.

While topic modeling allows for a higher-level view of thematic composition compared to dictionary methods, it remains fundamentally based on word co-occurrence patterns. Topic distributions capture relative prevalence of themes but do not directly measure the internal semantic dispersion or structural cohesion of annual filings. Two filings with similar topic proportions may nonetheless differ in how sentences cluster or deviate within embedding space.

2.3. Sentiment Dynamics and Linguistic Uncertainty

Another stream of literature focuses on sentiment dynamics and linguistic uncertainty as proxies for risk perception. Studies have linked uncertainty-related language to macroeconomic conditions and firm-level volatility (Tetlock 2007). Baker et al. (Baker et al. 2016) develop an economic policy uncertainty index based on news text frequency measures, illustrating the broader relevance of uncertainty language.

In accounting research, negative tone and uncertainty measures are commonly associated with future returns, volatility, and litigation risk (Loughran and McDonald 2016). These metrics typically rely on aggregated counts per thousand words, allowing for cross-firm comparison. The appeal of such measures lies in their interpretability and established validation in capital markets research.

However, sentiment and uncertainty indicators remain surface-level statistics. They do not capture whether risk narratives have become more fragmented, more centralized, or structurally rebalanced across themes. Consequently, they are well suited for measuring intensity but less suited for detecting structural reframing.

2.4. Semantic Representation and Anomaly Detection in Financial Text

Recent advances in natural language processing enable richer representations of financial disclosures. Contextual embedding models such as BERT (Devlin et al. 2019) and its sentence-level variants (Reimers and Gurevych 2019) represent text in continuous semantic space, allowing for similarity measurement beyond bag-of-words assumptions. These representations have been applied to earnings call transcripts, financial news, and regulatory filings to capture nuanced language patterns (Yang et al. 2020).

 doi:10.20944/preprints202603.0179.v1

Parallel to advances in representation learning, anomaly detection methods have been increasingly adopted in financial applications. Chandola et al. (Chandola et al. 2009) provide a comprehensive survey of anomaly detection techniques, including isolation-based and density-based approaches. In financial risk monitoring, unsupervised anomaly detection has been used to identify unusual trading patterns, fraud indicators, and abnormal disclosure behavior (Bolton and Hand 2002).

Despite these developments, limited research integrates semantic dispersion metrics with anomaly detection in the context of regulatory risk disclosure. Most existing studies either focus on frequency-based textual indicators or apply anomaly detection to structured financial variables. The question of whether risk factor sections exhibit structural semantic deviations relative to historical baselines remains insufficiently examined.

2.5. Explainable AI in Financial Text Modeling

As machine learning models become increasingly complex, explainability has emerged as a central concern in financial and regulatory applications. Black-box models may achieve strong predictive performance, yet their opacity raises concerns regarding trust, governance, and accountability. This issue is particularly salient in risk disclosure analysis, where interpretability is essential for regulatory relevance and academic rigor.

Recent research has systematically examined the foundations and practical implementations of explainable artificial intelligence. Liu et al. (Liu et al. 2025) provide a comprehensive review of XAI methodologies, distinguishing between intrinsic interpretability and post hoc explanation techniques. The authors emphasize that explainability is not merely a technical add-on but a structural requirement for trustworthy AI systems, especially in high-stakes domains.

In the context of financial reporting, Chen et al. (Chen et al. 2026) propose a privacy-centric and auditable framework for evaluating large language models in automated financial analysis. Their work highlights the importance of transparent model behavior, feature attribution, and reproducible auditing mechanisms. Similarly, Liao et al. (Liao et al. 2025) discuss governance-aware AI deployment in security and compliance settings, stressing that explainability mechanisms are necessary to align model outputs with institutional accountability standards.

Despite these advances, the integration of explainable AI with structural anomaly detection in regulatory disclosure analysis remains limited. Most existing XAI applications focus on classification or prediction tasks, such as sentiment labeling or risk forecasting. Relatively less attention has been paid to explaining unsupervised anomaly scores derived from semantic representations.

In disclosure risk assessment, anomaly detection alone is insufficient without interpretability. A high anomaly score must be decomposed into feature-level contributions in order to distinguish between changes in thematic emphasis, semantic dispersion, or disclosure length. Feature attribution methods such as SHAP provide a principled way to decompose model outputs into additive contributions, enabling transparent comparison across firms and years.

By embedding explainability directly into the anomaly detection pipeline, the present study aligns semantic modeling with governance-aware AI principles. This integration ensures that structural deviation is not only detected but also interpretable in terms of concrete linguistic and semantic drivers.

2.6. Positioning of This Study

The present study builds upon prior work in three ways. First, it retains interpretable topic share and word intensity features to maintain continuity with established disclosure metrics. Second, it introduces semantic dispersion measures derived from sentence embeddings to capture internal distributional properties of annual filings. Third, it applies baseline-calibrated anomaly detection combined with feature-level explanation to distinguish incremental disclosure growth from structural semantic deviation.

By bridging dictionary methods, topic modeling, and embedding-based representation within an explainable anomaly detection framework, this study addresses a gap between surface-level textual indicators and deeper structural analysis of regulatory risk narratives.

 doi:10.20944/preprints202603.0179.v1

3. Methodology

3.1. Conceptual Overview

Prior research typically measures risk disclosure using word frequencies, dictionary-based tone indicators, or topic proportions. These approaches quantify how much risk language appears in a filing and which themes dominate, but they do not directly evaluate whether the internal structure of a disclosure has shifted relative to its historical pattern.

The central premise of this study is that meaningful changes in risk communication often manifest as structural semantic deviations rather than simple increases in word intensity. If a firm substantially reframes its risk narrative, this shift should be observable in the distributional properties of its disclosure text when compared to its own historical baseline.

To operationalize this idea, we construct a structured representation for each company-year and evaluate deviations using an unsupervised anomaly detection framework. We then apply explainable artificial intelligence techniques to decompose anomaly scores into economically interpretable components.

Figure 1 presents the overall analytical pipeline. The framework consists of five sequential modules. First, Item 1A Risk Factors text is collected and segmented into sentence-level observations to form a clean corpus. Second, structured features are constructed by combining interpretable thematic shares, transformer-based sentence embeddings, and embedding-derived dispersion statistics. These features form a company-year representation that captures both topical emphasis and internal semantic geometry. Third, a historical baseline window is selected and features are standardized to construct a stable reference feature space. Fourth, structural deviation is quantified using unsupervised anomaly modeling, and raw anomaly scores are transformed into percentile-calibrated risk scores for cross-year comparability. Finally, an explainable AI layer trains a surrogate regressor and applies SHAP decomposition to attribute structural deviation to specific thematic and semantic drivers.

In addition to the core scoring pipeline, the framework incorporates two validation branches. A baseline internal stability test evaluates model sensitivity using leave-one-year-out scoring and rolling window re-estimation. An out-of-sample evaluation applies the baseline-calibrated model to 2025 disclosures and interprets deviations using SHAP-based feature attribution. Together, these components form a unified and interpretable structural semantic risk assessment framework.

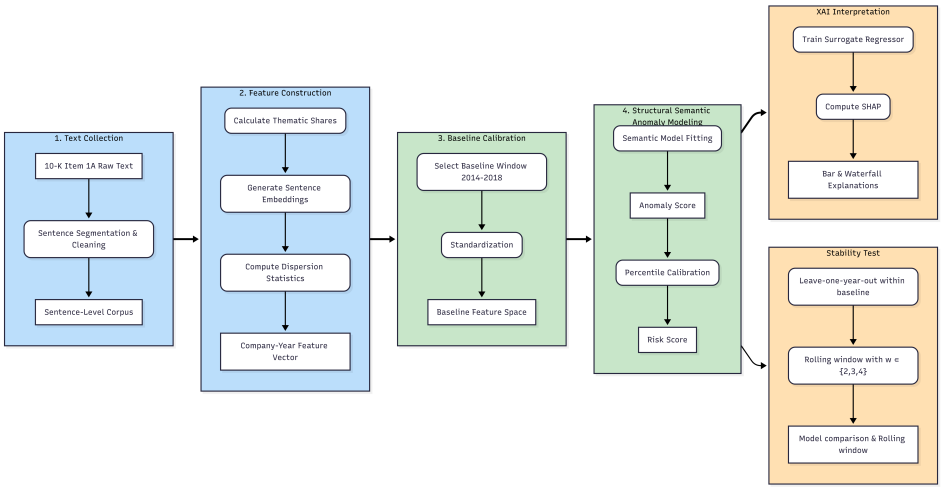

Figure 1. Workflow of the structural semantic risk assessment framework.

3.2. Topic Share Features

To preserve continuity with established disclosure research, we first construct interpretable topic-based features. Using transparent keyword rules, we identify sentences associated with regulatory exposure, litigation risk, conduct-related issues, and internal control matters. For each company-year, we compute the proportion of sentences belonging to each category.

 doi:10.20944/preprints202603.0179.v1

These topic shares capture shifts in thematic emphasis while remaining directly interpretable. Unlike probabilistic topic modeling, this rule-based approach ensures that each dimension corresponds to a clearly defined economic concept. Topic shares therefore provide a structured but transparent layer that connects our framework to traditional disclosure metrics.

3.3. Semantic Dispersion Metrics

While topic shares measure thematic prevalence, they do not capture how sentences are distributed within the broader semantic space of the document. To assess internal structural organization, we represent each sentence using a pretrained sentence embedding model. Each annual filing is thus treated as a collection of vectors in a continuous semantic space.

For each company-year, we compute the centroid of sentence embeddings and evaluate the distribution of distances between individual sentences and this centroid. From this distribution, we extract three summary statistics: the mean distance, the 95th percentile distance, and the standard deviation of distances.

These metrics reflect different aspects of structural organization. The mean distance captures overall semantic dispersion. The upper percentile captures the presence of extreme deviations within the document. The standard deviation reflects heterogeneity in narrative structure. Changes in these measures indicate redistribution or re-centering of risk narratives, even when topic proportions remain stable.

This embedding-based dispersion analysis distinguishes our approach from prior word count and topic modeling methods, which do not incorporate geometric properties of semantic space.

3.4. Baseline Calibration and Anomaly Detection

To determine whether a given company-year represents a structural deviation, we define a historical baseline using prior observations. Rather than comparing individual years directly, we train an unsupervised anomaly detection model on baseline feature vectors. This model learns the joint distribution of topic shares and semantic dispersion metrics under normal historical conditions.

Each company-year is assigned an anomaly score based on its deviation from this learned distribution. To improve interpretability, anomaly scores are transformed into percentile-based risk scores relative to the baseline sample. A score of 90 indicates that the observation is more anomalous than 90% of baseline years.

This baseline-calibrated framework differs from supervised classification models, which require labeled event data. Instead of prespecifying scandal years, deviations are identified relative to internal historical stability. This allows the method to detect structural changes without relying on external event labels.

3.5. Stability Testing

Unsupervised anomaly detection may be sensitive to sample composition. To ensure robustness, we conduct rolling-window stability analyses. For each target year, the model is re-estimated using alternative historical windows. We then evaluate whether anomaly scores remain consistent across these specifications.

This procedure reduces the likelihood that extreme scores arise from sampling noise or model instability. It also allows comparison of firms in terms of structural resilience rather than single-point estimates. Traditional textual disclosure studies rarely incorporate such baseline robustness checks.

3.6. Explainable Artificial Intelligence Layer

Anomaly scores alone do not indicate which features drive deviations. To interpret results in economically meaningful terms, we introduce an explainable artificial intelligence layer based on SHAP values.

Because the isolation-based anomaly model does not directly provide additive feature contributions, we train a tree-based surrogate regression model to approximate the mapping between

 doi:10.20944/preprints202603.0179.v1

standardized feature vectors and anomaly scores. Once this approximation is established, SHAP values are computed to decompose each anomaly score into contributions from individual features.

This decomposition allows us to identify whether deviations are primarily driven by changes in thematic emphasis, shifts in semantic dispersion, or variations in disclosure scale. For example, a high anomaly score may reflect increased litigation-related disclosure, expanded regulatory emphasis, or a structural redistribution of sentence embeddings within the document.

The XAI component is essential for distinguishing incremental disclosure growth from structural reframing. It also enables transparent comparison across firms and years by showing how different feature combinations contribute to similar or divergent anomaly outcomes.

4. Results

4.1. Data and Sample Construction

The empirical analysis is based on annual Form 10-K filings submitted to the U.S. Securities and Exchange Commission. Specifically, we extract the Risk Factors section, which corresponds to Item 1A under Regulation S-K of the Securities Act reporting framework (Campbell et al. 2014). Regulation S-K establishes disclosure requirements for publicly traded firms, including mandatory discussion of material risk factors that may affect financial condition and operations.

The sample focuses on two large U.S. banking institutions, Wells Fargo and JPMorgan Chase. These firms operate under comparable regulatory regimes and are designated as systemically important financial institutions. The 2016 Wells Fargo unauthorized accounts scandal provides a salient regulatory shock within the observation window, allowing comparative analysis of disclosure responses under similar macroeconomic conditions.

We collect annual filings for the years 2014 through 2018 and 2025. The period 2014 to 2018 serves as the historical baseline window for modeling structural stability prior to and immediately following the 2016 regulatory shock. The year 2025 is evaluated relative to this baseline to assess longer-term structural deviation in risk narratives.

For each filing, the Risk Factors section is extracted as plain text and segmented into sentences. Formatting artifacts and excessive whitespace are removed to ensure consistency across years. Sentence-level observations are then aggregated into company-year feature representations for subsequent analysis.

Table 1 reports disclosure length and thematic composition for the Risk Factors sections of Wells Fargo and JPMorgan Chase from 2014 to 2018 and 2025. Table 2 presents embedding-based semantic dispersion statistics for the same company-year observations.

Table 1. Disclosure length and thematic composition in Item 1A risk factor disclosures.

Company	Year	Sentences	AvgLen	Reg	Litig	Conduct	Controls
JPM	2014	225	39.71	0.271	0.138	0.053	0.053
JPM	2015	240	42.11	0.275	0.133	0.058	0.058
JPM	2016	269	44.76	0.305	0.145	0.074	0.067
JPM	2017	395	35.37	0.225	0.122	0.066	0.046
JPM	2018	443	38.21	0.212	0.117	0.077	0.045
JPM	2025	121	41.77	0.273	0.140	0.083	0.058
WFC	2014	460	34.22	0.202	0.046	0.078	0.065
WFC	2015	461	34.41	0.208	0.030	0.078	0.067
WFC	2016	505	35.27	0.240	0.040	0.083	0.077
WFC	2017	504	35.24	0.234	0.044	0.091	0.077
WFC	2018	529	35.43	0.234	0.049	0.095	0.079
WFC	2025	421	36.01	0.264	0.069	0.114	0.121

 doi:10.20944/preprints202603.0179.v1

Table 2. Semantic dispersion statistics derived from sentence embeddings.

Company	Year	MeanDispersion	P95Dispersion	StdDispersion
JPM	2014	0.3806	0.5774	0.1034
JPM	2015	0.3764	0.5771	0.1017
JPM	2016	0.3759	0.5485	0.0965
JPM	2017	0.3135	0.6564	0.1522
JPM	2018	0.3177	0.6494	0.1545
JPM	2025	0.3230	0.6222	0.1500
WFC	2014	0.4665	0.6612	0.1127
WFC	2015	0.4622	0.6653	0.1139
WFC	2016	0.4626	0.6665	0.1130
WFC	2017	0.4607	0.6695	0.1126
WFC	2018	0.4625	0.6738	0.1128
WFC	2025	0.4295	0.6279	0.1126

4.1.1. Disclosure Length

Clear cross-firm differences are visible in disclosure scale. During 2014 to 2018, WFC averages 492 sentences per year, compared to 314 for JPM. In 2018, WFC reports 529 sentences, while JPM reports 443. In 2025, JPM contracts sharply to 121 sentences, representing a reduction of more than 60% relative to its baseline average. WFC, by contrast, remains at 421 sentences in 2025, indicating no comparable structural contraction.

Average sentence length further highlights stylistic differences. JPM reaches 44.76 words per sentence in 2016, whereas WFC remains within a narrower band between 34 and 36 words across most years.

4.1.2. Thematic Composition

Thematic shares reveal structural shifts in narrative emphasis. For JPM, regulatory content peaks at 30.5% in 2016 before declining to 21.2% in 2018 and returning to 27.3% in 2025. Litigation shares remain between 11% and 15% across years.

For WFC, litigation content remains below 5% during the baseline period but rises to 6.9% in 2025. Conduct content increases from 7.8% in 2014 to 11.4% in 2025, while control-related language nearly doubles from 6.5% to 12.1%.

These patterns indicate that WFC's 2025 disclosure reflects a governance-oriented rebalancing, whereas JPM's thematic proportions in 2025 resemble earlier baseline patterns despite large changes in overall length.

4.1.3. Semantic Dispersion

Embedding-based dispersion measures in Table 2 reveal internal structural differences not captured by raw counts. JPM's mean dispersion declines from approximately 0.38 during 2014 to 2016 to 0.31 in 2017 and 2018, while variability increases sharply. Standard deviation rises from 0.0965 in 2016 to above 0.15 in 2017 and 2018. This combination suggests concentration of core narrative themes alongside expansion of semantically extreme sentences.

WFC exhibits consistently higher mean dispersion around 0.46 during the baseline period. In 2025, mean dispersion declines modestly to 0.4295 while variability remains stable. This pattern suggests semantic consolidation rather than fragmentation.

4.2. Traditional Risk Word Metrics

To establish a benchmark against conventional disclosure analysis, we begin with frequency-based risk word measures. These metrics follow the standard dictionary counting approach widely used in accounting and finance research. For each company-year, we compute two statistics. The first is raw risk word count, defined as the total number of occurrences of predefined risk-related terms within the Risk Factors section. The second is risk word intensity, defined as the number of risk words per 1,000

 doi:10.20944/preprints202603.0179.v1

words of disclosure text. The dictionary includes negative, uncertainty, and litigious terms commonly used in prior disclosure studies.

Figure 2 presents risk word intensity from 2014 to 2018. Figure 3 presents raw frequency counts over the same period.

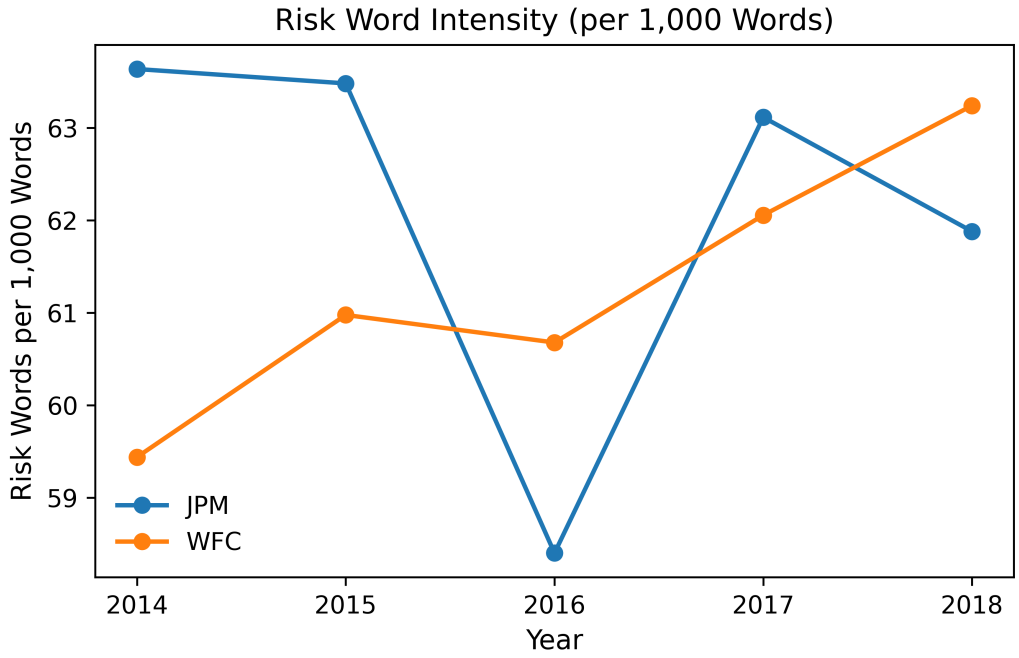

Figure 2. Risk word intensity (per 1,000 words), 2014–2018.

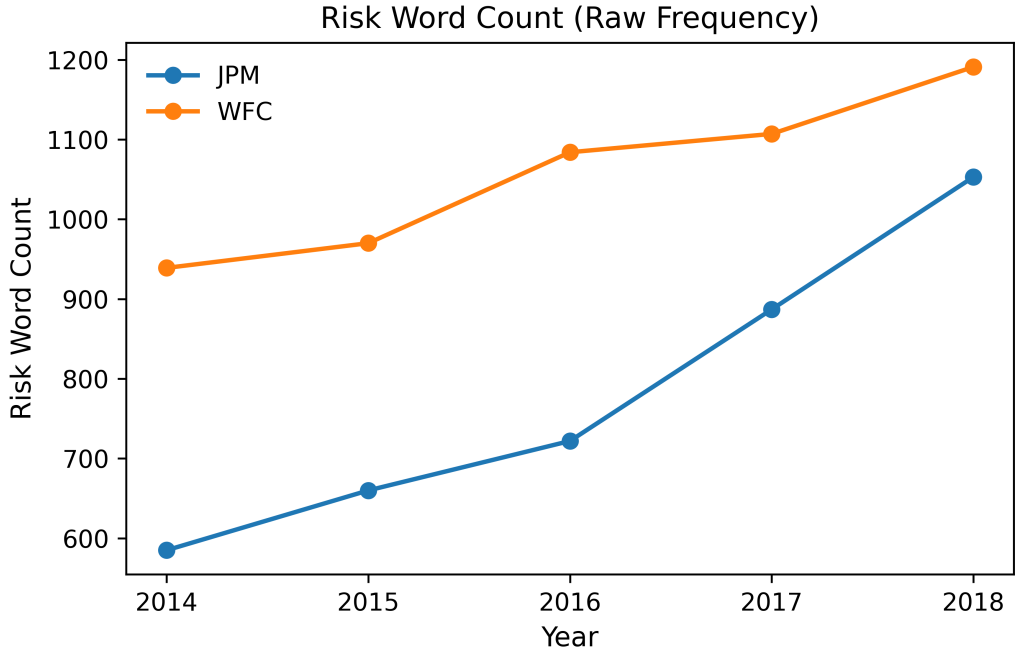

Figure 3. Raw risk word counts, 2014–2018.

4.2.1. Risk Word Intensity

Figure 2 shows that JPMorgan exhibits relatively stable risk word intensity between 2014 and 2018. The measure fluctuates between approximately 58 and 64 words per 1,000 words. Notably, the

 doi:10.20944/preprints202603.0179.v1

year 2016 does not display an abnormal spike. Instead, risk intensity for JPMorgan declines from 63.4 in 2015 to 58.4 in 2016, before increasing again in 2017.

Wells Fargo exhibits a gradual upward trend in intensity, increasing from approximately 59.4 in 2014 to 63.3 in 2018. However, this increase appears smooth and incremental rather than discontinuous.

Importantly, neither firm shows a distinct structural break in 2016 under this metric. The year widely recognized as a major regulatory shock does not emerge as an outlier when measured solely by normalized word frequency.

4.2.2. Raw Risk Word Counts

Figure 3 shows raw counts of risk-related words. For both firms, counts increase over time. JPMorgan's total risk words rise from 587 in 2014 to 1,053 in 2018. Wells Fargo increases from 940 to 1,190 over the same period.

However, these increases closely track expansion in overall disclosure length, as documented in Table 1. When disclosure sections grow longer, raw counts mechanically rise as well. The increase therefore reflects scale rather than structural change in narrative composition.

In 2016, raw risk word counts do not exhibit a disproportionate increase relative to adjacent years. The pattern is gradual and continuous.

4.2.3. Why Traditional Metrics Are Insufficient

These results highlight an important limitation of frequency-based methods. First, raw counts are sensitive to document length. As firms expand their disclosures, counts increase even if thematic structure remains unchanged. Second, normalized intensity measures reduce length bias but still treat all words independently. They do not capture shifts in semantic organization, contextual framing, or internal narrative coherence.

Most notably, the 2016 Wells Fargo scandal, which is widely recognized as a material event in the banking sector, does not produce a clear structural discontinuity in either raw counts or normalized intensity. The metrics suggest incremental variation rather than narrative reconfiguration.

This motivates the need for higher-dimensional semantic modeling. Rather than counting isolated words, embedding-based approaches allow us to examine how sentence meanings reorganize within each year. The following sections therefore move beyond frequency counts and examine structural dispersion and anomaly signals derived from sentence-level embeddings.

4.3. Category-Specific Risk Word Intensity

To further examine whether more granular lexical categories capture structural changes in disclosure, we decompose risk words into three commonly used dimensions: negative tone, uncertainty, and litigious language. These categories are derived from established financial disclosure dictionaries. For each company-year, intensity is measured as the number of category-specific words per 1,000 words of disclosure text.

4.3.1. Negative Word Intensity

Negative word intensity remains relatively stable for both firms across the baseline period. For JPMorgan, the measure fluctuates between 18.0 and 19.1 words per 1,000 words. The value declines from 19.1 in 2015 to 18.0 in 2016, followed by a moderate increase to 18.9 in 2017 and a return to 18.0 in 2018. The variation remains within a narrow band of approximately one word per 1,000 words.

Wells Fargo shows a gradual upward trend, increasing from 16.6 in 2014 to 17.3 in 2018. However, this change appears incremental rather than discontinuous. The year 2016 does not display an abnormal spike or break relative to adjacent years.

 doi:10.20944/preprints202603.0179.v1

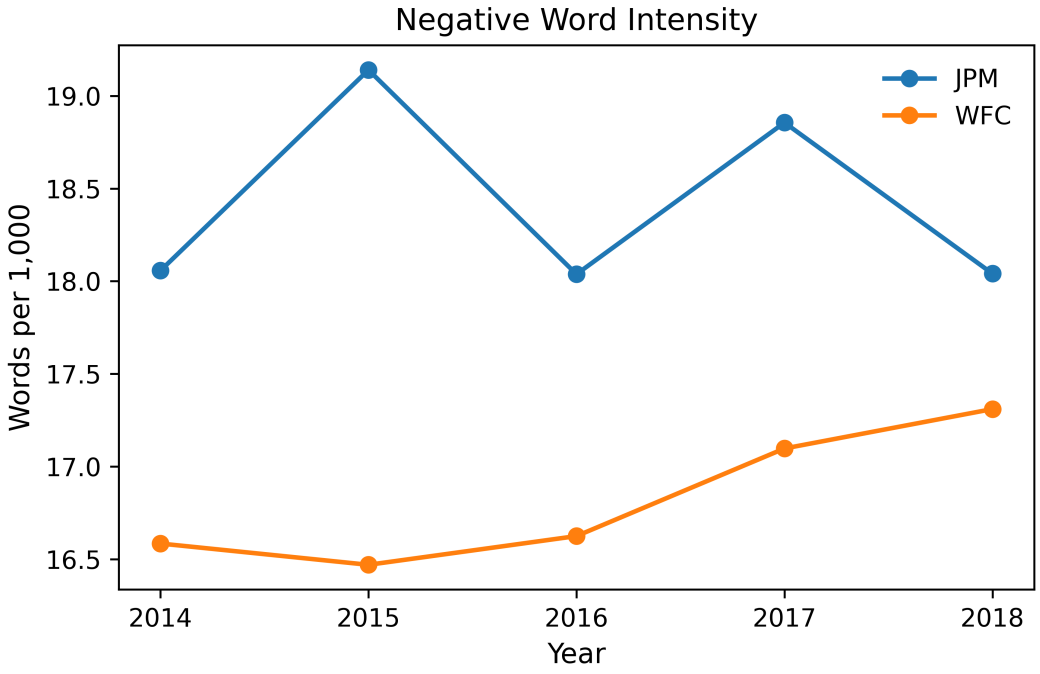

Figure 4. Negative word intensity (per 1,000 words), 2014–2018.

4.3.2. Uncertainty Word Intensity

Uncertainty-related language exhibits a similar pattern. For JPMorgan, intensity peaks in 2015 at 19.3 words per 1,000 words, declines in 2016 to 18.1, and then increases again in 2017 and 2018. Wells Fargo experiences a modest dip in 2016, from 18.3 in 2015 to 17.5 in 2016, followed by gradual increases thereafter.

Again, 2016 does not emerge as an outlier for either institution. Observed variation remains within historical fluctuation ranges.

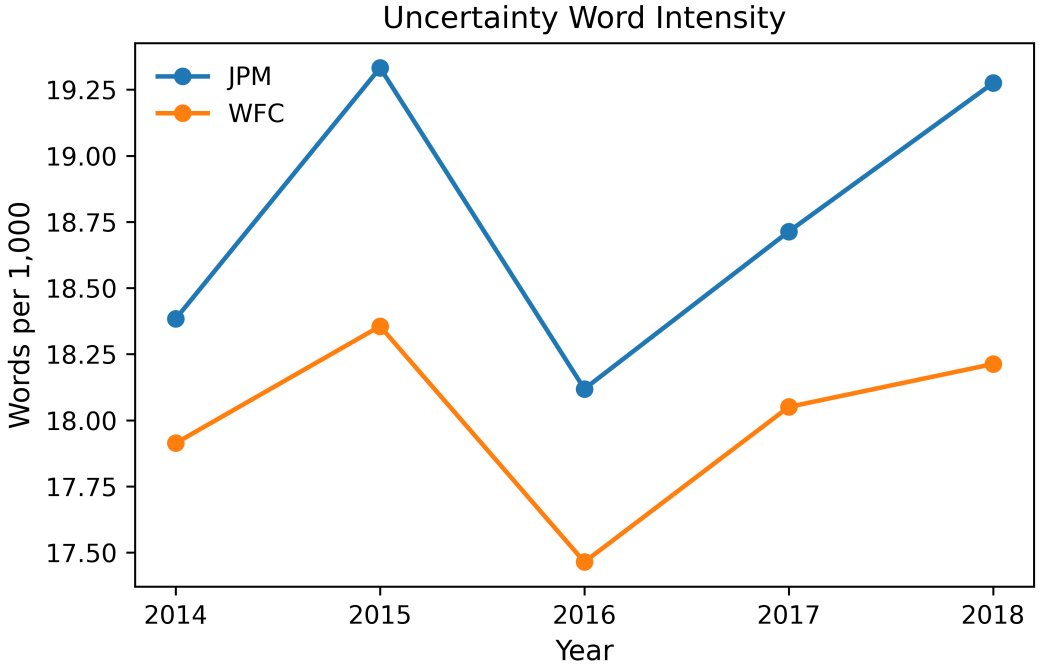

Figure 5. Uncertainty word intensity (per 1,000 words), 2014–2018.

 doi:10.20944/preprints202603.0179.v1

4.3.3. Litigious Word Intensity

Litigious language also fails to produce a distinct structural break. JPMorgan's intensity ranges between 3.5 and 4.4 words per 1,000 words from 2014 to 2018. Although 2016 records 4.4 words per 1,000 words, slightly above 2015, the magnitude of change is modest and does not exceed prior variation.

Wells Fargo maintains substantially lower levels of litigious intensity, remaining below 1.5 words per 1,000 words throughout the period and declining after 2016.

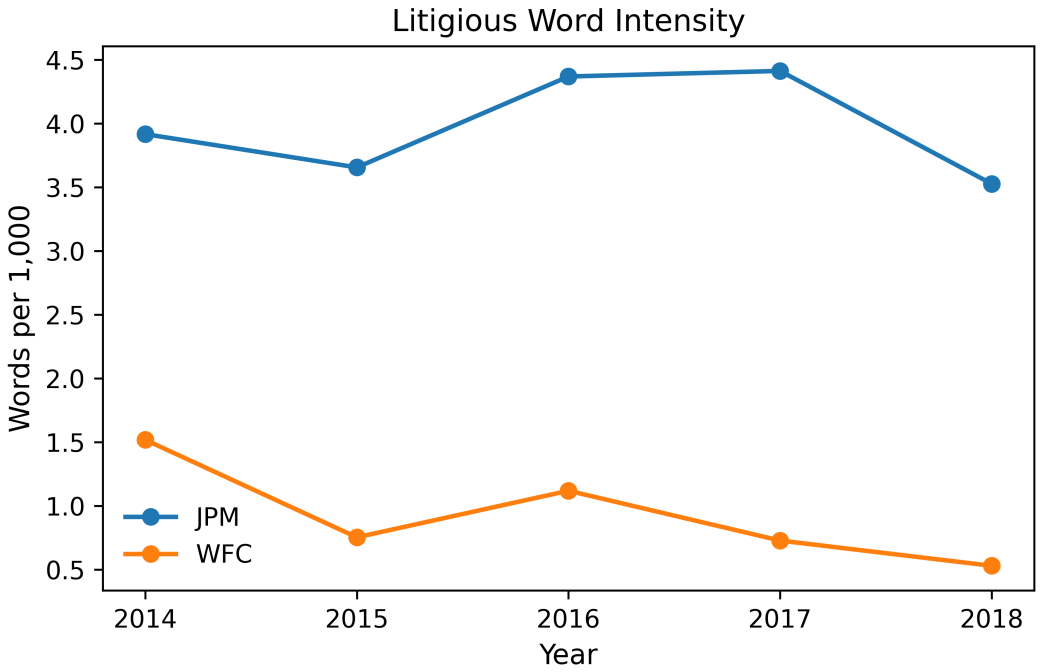

Figure 6. Litigious word intensity (per 1,000 words), 2014–2018.

4.3.4. Interpretation

Across all three lexical dimensions, changes appear smooth and incremental rather than abrupt. The year 2016, widely recognized as a major regulatory event following the Wells Fargo unauthorized accounts scandal and subsequent enforcement actions by federal regulators, does not stand out as a structural outlier under these dictionary-based measures (Consumer Financial Protection Bureau et al. 2016).

This pattern suggests that frequency-based lexical analysis, even when disaggregated into tone-specific categories, may fail to detect deeper narrative restructuring. Structural shifts in disclosure may occur at the level of semantic organization rather than isolated word usage.

5. Structural Semantic Anomaly Modeling

The preceding section demonstrates that traditional lexical metrics, including raw risk word counts and dictionary based intensity measures, do not reliably capture structural discontinuities associated with major regulatory events. Although such methods provide useful descriptive insights, they operate at the level of isolated word frequencies and fail to account for multivariate interactions and semantic configuration.

To address this limitation, we propose a structural semantic anomaly modeling framework that evaluates disclosure shifts in a multidimensional feature space. Rather than focusing on individual word categories, our approach models the joint distribution of thematic shares and semantic dispersion patterns, and identifies deviations from a learned historical baseline.

This section introduces the theoretical foundation, implementation details, and empirical performance of the proposed framework.

 doi:10.20944/preprints202603.0179.v1

5.1. Conceptual Framework

We conceptualize risk disclosure not as a collection of independent word frequencies but as a structured semantic system. Each company year is represented as a feature vector combining thematic composition and embedding based dispersion statistics.

Formally, let

$$\mathbf{x}_t \in \mathbb{R}^d$$

denote the structural representation of a disclosure in year t. The baseline period defines a reference distribution in this feature space. Structural risk emerges when an observation deviates from this historical manifold.

This representation integrates three components: thematic proportions reflecting regulatory, litigation, conduct, and control emphasis; embedding-based semantic dispersion capturing internal heterogeneity of sentences; and disclosure scale variables controlling for text length effects.

By modeling the joint configuration of these dimensions, the framework captures structural reorganization that lexical metrics overlook.

5.2. Model Specification

Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the structural feature vector of a company-year disclosure, where features include thematic shares, semantic dispersion statistics, and disclosure scale controls. The baseline set $\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ corresponds to historical observations from 2014 to 2018.

All features are standardized prior to model fitting. Anomaly detection models are trained exclusively on the baseline distribution and subsequently used to evaluate deviation of target years.

5.2.1. Autoencoder Reconstruction Deviation

The autoencoder captures nonlinear structural dependencies among features. It consists of an encoder function $f_\theta(\cdot)$ and a decoder function $g_\phi(\cdot)$ such that

$$\mathbf{z}_t = f_\theta(\mathbf{x}_t), \qquad \hat{\mathbf{x}}_t = g_\phi(\mathbf{z}_t). \tag{1}$$

The model is trained by minimizing reconstruction loss over the baseline sample:

$$\mathcal{L}(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - g_\phi(f_\theta(\mathbf{x}_i))\|_2^2. \tag{2}$$

The anomaly score for a target observation is defined as

$$s_{\mathrm{AE}}(\mathbf{x}_t) = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2. \tag{3}$$

Large reconstruction error indicates that the observation lies outside the nonlinear structural manifold learned from the baseline distribution (Sakurada and Yairi 2014).

5.2.2. Isolation-Based Structural Deviation

Isolation Forest identifies anomalies through recursive random partitioning (Liu et al. 2008). Let $h(\mathbf{x})$ denote the path length required to isolate observation $\mathbf{x}$ within a tree. The expected path length across trees is $\mathbb{E}[h(\mathbf{x})]$.

The normalization term is

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \tag{4}$$

where $H(\cdot)$ is the harmonic number and n is the subsample size.

The anomaly score is defined as

$$s_{\mathrm{IF}}(\mathbf{x}) = 2^{-\frac{\mathbb{E}[h(\mathbf{x})]}{c(n)}}. \tag{5}$$

 doi:10.20944/preprints202603.0179.v1

Observations that require fewer partitions to isolate receive higher anomaly scores. Unlike distance-based methods, Isolation Forest does not impose distributional assumptions (Liu et al. 2008).

5.2.3. Covariance-Adjusted Distance Deviation

Mahalanobis distance evaluates deviation from the baseline centroid while accounting for feature covariance. Let

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{6}$$

denote the baseline mean vector.

To stabilize covariance estimation under limited sample size, we apply the Ledoit–Wolf shrinkage estimator (Ledoit and Wolf 2004):

$$\boldsymbol{\Sigma}_{\mathrm{LW}} = (1-\lambda)\mathbf{S} + \lambda\mathbf{T}, \tag{7}$$

where $\mathbf{S}$ is the sample covariance, $\mathbf{T}$ is a structured target matrix, and λ is the shrinkage intensity. The structural distance is

$$s_{\mathrm{MD}}(\mathbf{x}_t) = (\mathbf{x}_t - \boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}_{\mathrm{LW}}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}). \tag{8}$$

This formulation provides a parametric benchmark under approximate elliptical structure (Ledoit and Wolf 2004).

5.2.4. Model Explainability via SHAP

To attribute structural deviation to individual features, we apply SHAP values (Lundberg and Lee 2017). For a fitted scoring function $F(\mathbf{x})$, SHAP represents the prediction as

$$F(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d}\phi_j, \tag{9}$$

where ϕ_j measures the marginal contribution of feature j to the anomaly score.

This enables identification of which thematic or semantic dimensions drive structural deviation in target years (Lundberg and Lee 2017).

5.3. Structural Feature Patterns Prior to Anomaly Modeling

Table 3 summarizes semantic dispersion statistics derived from sentence embeddings, and Table 4 reports thematic composition for Item 1A disclosures.

Table 3. Semantic dispersion statistics derived from sentence embeddings.

Company	Year	MeanDispersion	P95Dispersion	StdDispersion
JPM	2014	0.3806	0.5774	0.1034
JPM	2015	0.3764	0.5771	0.1017
JPM	2016	0.3759	0.5485	0.0965
JPM	2017	0.3135	0.6564	0.1522
JPM	2018	0.3177	0.6494	0.1545
WFC	2014	0.4665	0.6612	0.1127
WFC	2015	0.4622	0.6653	0.1139
WFC	2016	0.4626	0.6665	0.1130
WFC	2017	0.4607	0.6695	0.1126
WFC	2018	0.4625	0.6738	0.1128

 doi:10.20944/preprints202603.0179.v1

Table 4. Thematic composition statistics of Item 1A disclosures.

Company	Year	ShareRegulatory	ShareConduct	ShareLitigation	ShareControls
JPM	2014	0.2933	0.0533	0.1378	0.0533
JPM	2015	0.2958	0.0583	0.1333	0.0583
JPM	2016	0.3271	0.0743	0.1450	0.0669
JPM	2017	0.2481	0.0658	0.1215	0.0456
JPM	2018	0.2393	0.0767	0.1174	0.0451
WFC	2014	0.2239	0.0783	0.0457	0.0652
WFC	2015	0.2321	0.0781	0.0304	0.0672
WFC	2016	0.2634	0.0832	0.0396	0.0772
WFC	2017	0.2619	0.0913	0.0437	0.0774
WFC	2018	0.2628	0.0945	0.0491	0.0794

5.3.1. Semantic Dispersion

JPMorgan exhibits a clear structural transition between 2016 and 2017. Mean dispersion declines from approximately 0.376 in 2016 to 0.314 in 2017, indicating substantial semantic concentration. At the same time, both the 95th percentile dispersion and standard deviation increase sharply: P95 rises from 0.5485 to 0.6564 and the standard deviation increases from 0.0965 to 0.1522. This combination indicates that although average sentence similarity increases, extreme deviations become more pronounced; disclosures become more internally concentrated yet exhibit heavier semantic tails.

Wells Fargo displays a markedly different pattern. Mean dispersion remains stable around 0.46 throughout the baseline period, with standard deviation consistently near 0.113. No abrupt structural break appears in dispersion statistics. This contrast suggests that JPMorgan underwent a structural reorganization in disclosure composition after 2016, whereas Wells Fargo maintained greater semantic continuity.

5.3.2. Thematic Composition

Thematic shares further illustrate differentiated structural adjustments. For JPMorgan, ShareRegulatory peaks in 2016 at 0.327 and then declines in subsequent years. ShareLitigation remains elevated in 2014–2016 and declines slightly afterward, while ShareControls decreases after 2016. This pattern indicates heightened regulatory emphasis in 2016 followed by thematic normalization.

For Wells Fargo, ShareRegulatory increases steadily from 0.224 in 2014 to approximately 0.263 in 2016–2018. ShareConduct and ShareControls both trend upward through 2018, while ShareLitigation remains relatively low but increases gradually. Unlike JPMorgan's temporary spike, Wells Fargo exhibits a sustained reweighting of disclosure themes following 2016.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

5.4. Semantic Structural Modeling and Baseline Stability

The limitations of traditional dictionary based metrics become particularly evident when examining structural changes around major regulatory events. As shown in Section 4, risk word intensity and category level lexical shares exhibit only modest variation around 2016. Although 2016 is widely recognized as a critical regulatory year, dictionary based statistics do not clearly isolate it as a structural break. This suggests that purely lexical counting fails to capture deeper semantic reorganization within disclosures.

To overcome this limitation, we construct annual semantic representations using sentence level embeddings derived from a pretrained transformer model. For each company year, sentence embeddings are aggregated to form a year level semantic vector. These high dimensional representations are then projected onto the first two principal components in order to visualize the temporal evolution of disclosure structure.

 doi:10.20944/preprints202603.0179.v1

Figure 7 and Figure 8 present the resulting semantic trajectories for JPMorgan and Wells Fargo. Each point corresponds to a company year, and arrows indicate chronological progression. For JPMorgan in Figure 7 , baseline years 2014 to 2016 cluster tightly in one region of the projection space. In contrast, 2017 and 2018 move sharply toward a distinct region along the first principal component. The displacement between 2016 and 2017 is substantially larger than variation among earlier baseline years, indicating a structural semantic reorganization following the regulatory shock period.

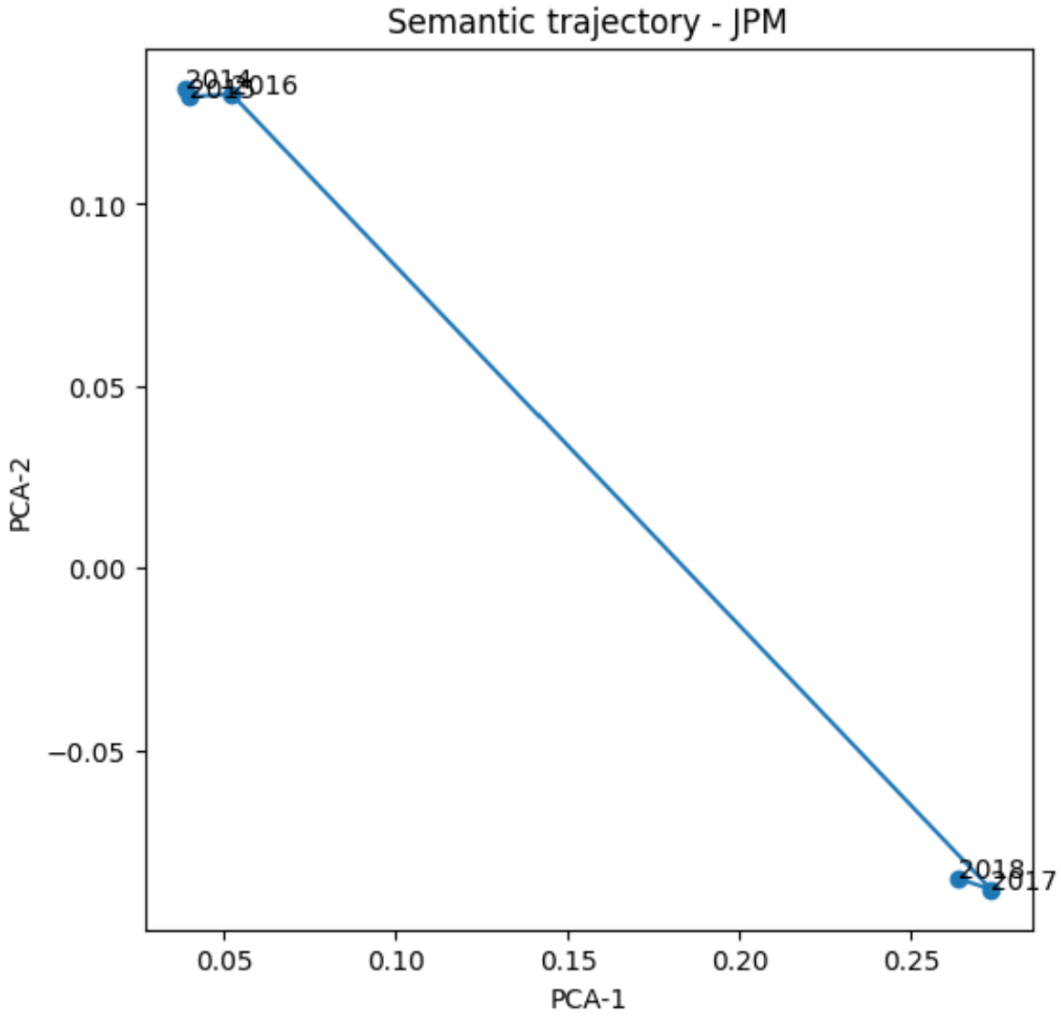

Figure 7. Semantic trajectory projection for JPMorgan (2014–2018).

For Wells Fargo, the trajectory displays a smoother but persistent directional shift from 2014 through 2018 as shown in Figure 8 . The transition from 2015 to 2016 is clearly visible, and subsequent years continue along a coherent semantic path rather than reverting to the pre event region. Unlike dictionary based intensity measures, the embedding space reveals systematic narrative realignment that is consistent with governance and compliance restructuring during this period.

While semantic trajectories provide geometric evidence of structural change, anomaly detection requires a stable and discriminative scoring mechanism. Before interpreting post event deviations, we therefore evaluate the internal stability of candidate detectors within the baseline years. This internal stability assessment serves two purposes. First, it verifies that the scoring function is not dominated by random fluctuations induced by limited sample size. Second, it identifies the detector whose baseline behavior provides sufficient resolution for downstream event analysis.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2026 doi:10.20944/preprints202603.0179.v1

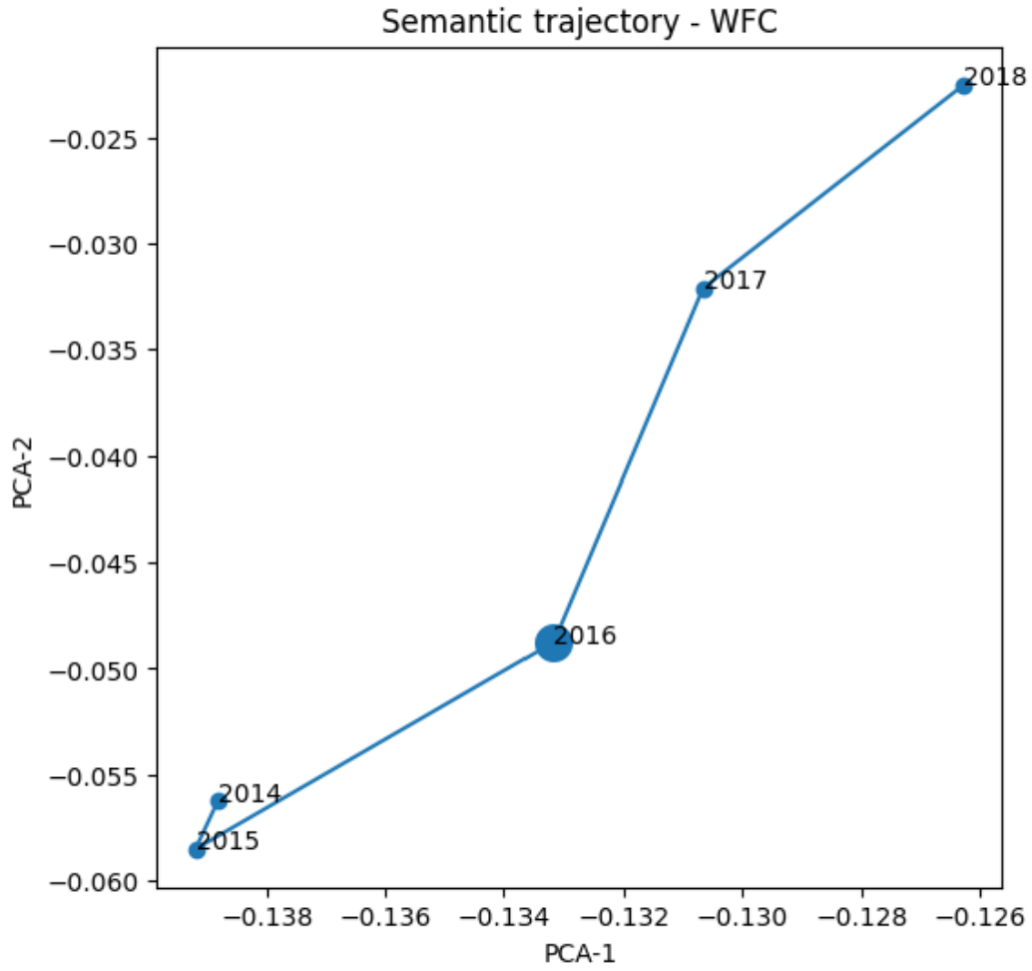

Figure 8. Semantic trajectory projection for Wells Fargo (2014–2018).

Figure 9 reports baseline risk scores under three anomaly detectors: an autoencoder-based reconstruction model, Isolation Forest, and Mahalanobis distance with shrinkage covariance. For each baseline year, the detector is trained on the remaining baseline observations and the held-out year is scored. Scores are transformed into percentile-based risk scores relative to the fitted baseline distribution.

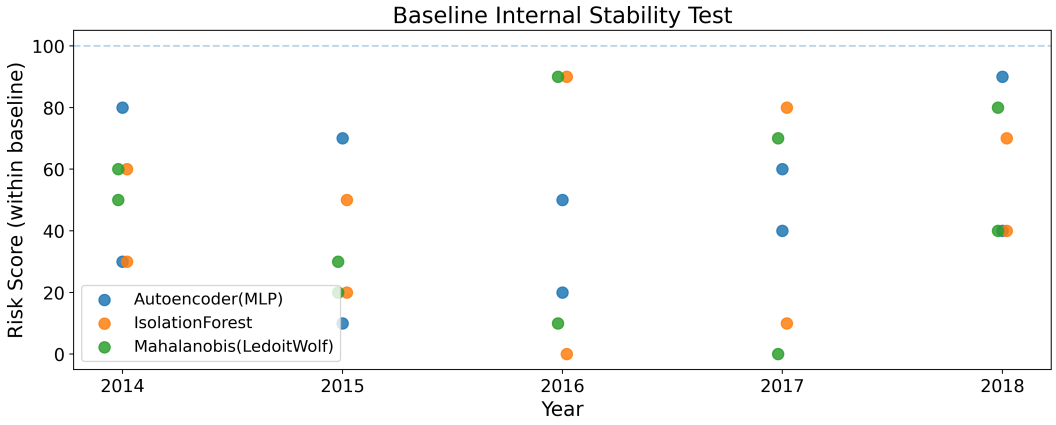

Figure 9. Baseline risk scores across detectors (leave-one-year-out within baseline).

Across baseline years, the autoencoder and shrinkage Mahalanobis model frequently produce saturated or near saturated scores for multiple years. This indicates score compression, where several baseline observations are treated as similarly extreme. Such compression reduces the ability of the

 doi:10.20944/preprints202603.0179.v1

model to distinguish subtle structural variation among baseline disclosures. In contrast, Isolation Forest produces graded and well distributed scores, preserving useful ranking information across years.

To further assess robustness to baseline selection, we conduct a rolling-window stability analysis. Let the baseline window size be $w \in \{2, 3, 4\}$. For each target year $t \in \{2016, 2017, 2018\}$, we construct a baseline set using the w most recent years prior to t, fit the detector on that window, and then score the target year. Scores are again converted into percentile-based risk measures within each window. Figure 10 summarizes these results as heatmaps, with rows corresponding to window sizes and columns corresponding to target years.

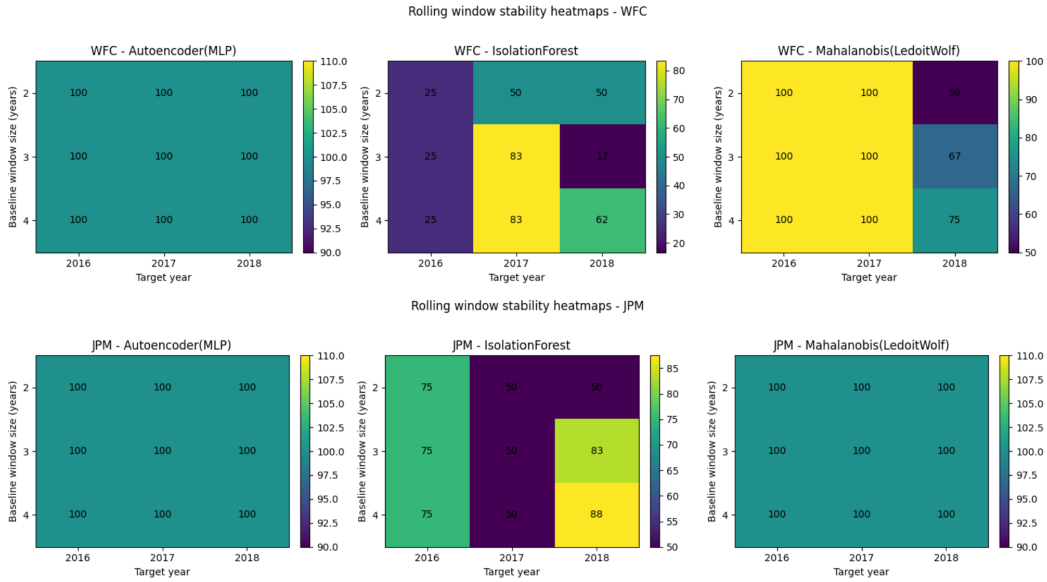

Figure 10. Rolling-window stability heatmaps of baseline risk scores.

The rolling window analysis reinforces the earlier conclusion. The autoencoder yields nearly constant scores across window sizes and target years, indicating limited sensitivity under the current configuration. The shrinkage Mahalanobis model shows moderate stability but still exhibits score compression in several settings. Isolation Forest provides nontrivial variation across both window size and target year, without collapsing observations into uniform scores. This behavior reflects a desirable balance between stability and responsiveness to structural change.

Taken together, the semantic trajectory analysis and the baseline stability evaluation demonstrate two key points. First, embedding based representations capture structural reorientation that lexical metrics overlook. Second, among the tested detectors, Isolation Forest offers the most discriminative and stable anomaly scoring framework for identifying disclosure shifts associated with regulatory stress.

5.5. Out-of-Sample Structural Risk Assessment: 2025

After validating the semantic representation and selecting Isolation Forest as the primary detector, we evaluate the 2025 disclosures as out-of-sample observations relative to the historical baseline.

Table 5 reports the anomaly score, percentile-based risk score, and selected structural features.

Table 5. Out-of-sample structural risk scores for 2025.

Firm	RiskScore	AnomalyScore	SemMeanDist	nSent
WFC 2025	60	0.0158	0.4295	421
JPM 2025	90	0.0522	0.3230	121

 doi:10.20944/preprints202603.0179.v1

5.5.1. Overall Structural Deviation

The percentile-based risk score places JPMorgan in the upper decile of deviation relative to its historical disclosure distribution, while Wells Fargo remains in the upper-middle range. Importantly, this ranking emerges from multivariate structural comparison rather than single-feature differences.

The raw anomaly score further illustrates the magnitude of deviation. JPM's anomaly score exceeds that of WFC by more than a factor of three. This difference cannot be attributed to a single variable. Instead, it reflects joint movement across semantic dispersion, thematic composition, and disclosure length.

One striking feature is the sharp contraction in JPM's sentence count. With only 121 sentences, its 2025 disclosure is structurally compressed relative to its historical baseline. In contrast, WFC maintains 421 sentences, much closer to its prior disclosure volume. This structural compression changes the proportional distribution of semantic and thematic signals in the embedding space.

5.5.2. Feature-Level Explanation via SHAP

To interpret these anomaly scores, we compute SHAP values for the Isolation Forest model. Figure 11 presents feature-level contributions for both firms.

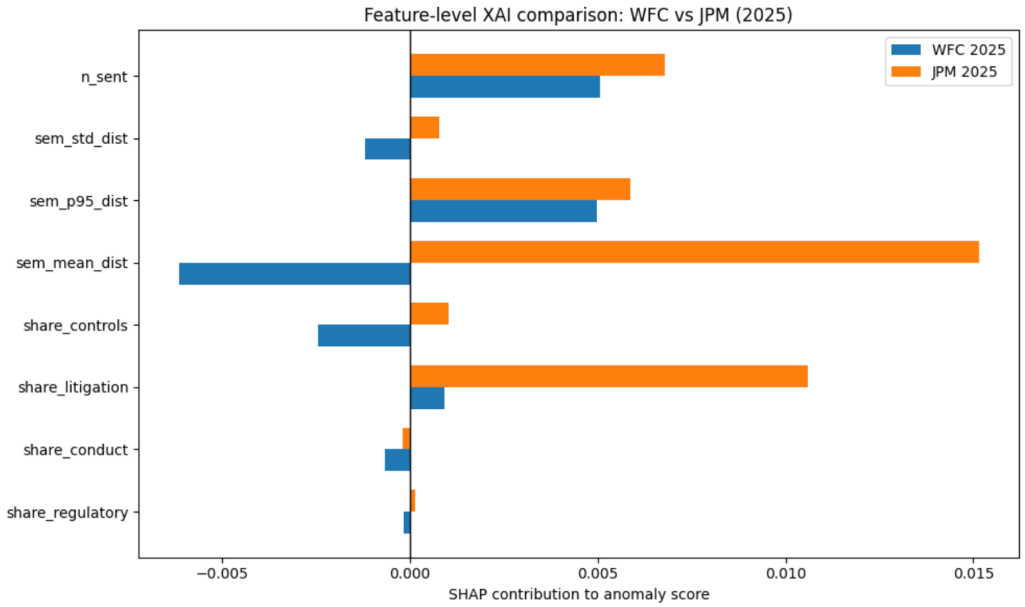

Figure 11. SHAP-based feature contributions for JPMorgan and Wells Fargo (2025).

For JPMorgan, the dominant positive contributors to anomaly are semantic mean distance, litigation share, upper-tail semantic dispersion, and sentence count contraction. The SHAP value for semantic mean distance is the largest contributor, indicating that JPM's 2025 disclosure lies geometrically farther from its historical embedding centroid. This implies narrative reconfiguration rather than incremental lexical adjustment. The elevated litigation share further increases deviation, suggesting stronger legal framing relative to historical norms. The positive SHAP contribution of upper-tail semantic dispersion indicates greater heterogeneity among sentences, meaning that certain portions of the disclosure deviate strongly from the historical semantic center. This pattern is consistent with selective emphasis or reframing in specific risk domains.

In contrast, Wells Fargo exhibits a more balanced SHAP profile. While regulatory share and control-related language contribute positively to anomaly, semantic mean distance contributes less dramatically. The overall structure remains closer to its baseline manifold. Importantly, the absence of a dominant single feature suggests incremental structural adjustment rather than abrupt narrative transformation.

 doi:10.20944/preprints202603.0179.v1

5.5.3. Comparative Interpretation

The joint reading of Table 5 and Figure 11 highlights three important insights.

First, anomaly magnitude differs substantially between firms even when thematic shares appear superficially similar. Both firms exhibit comparable regulatory share, yet JPM receives a much higher structural risk score. This demonstrates that frequency-based metrics alone cannot explain the divergence.

Second, semantic distance measures play a central role in differentiating firms. Embedding-based dispersion captures shifts in framing, emphasis, and contextualization that are not observable through word counting.

Third, the interaction between disclosure length and thematic concentration amplifies structural deviation. JPM's shortened disclosure alters proportional weighting across semantic features, leading to a compounded anomaly effect.

Taken together, these findings confirm that the proposed semantic anomaly framework provides interpretable and economically meaningful structural risk assessment. The model not only assigns differential risk scores but also identifies the precise structural drivers underlying each firm's deviation from historical norms.

6. Conclusions

This study examines structural risk disclosure dynamics using a semantic modeling framework that integrates sentence-level embeddings with anomaly detection. We begin by evaluating traditional dictionary-based metrics and show that risk word intensity and category-level frequency statistics provide limited sensitivity to structural shifts around major regulatory events. Although 2016 is widely recognized as a pivotal year, lexical metrics alone do not clearly isolate it as a structural break.

To address this limitation, we construct year-level semantic representations using transformer-based sentence embeddings. Semantic trajectory analysis reveals clear geometric displacement in embedding space following regulatory stress periods. These structural shifts are not merely changes in word frequency but reflect broader narrative realignment in disclosure framing and emphasis.

We then introduce a multivariate anomaly detection framework and conduct baseline internal stability tests to ensure methodological robustness. Among the tested detectors, Isolation Forest demonstrates the most favorable balance between stability and discriminative power. Rolling-window analysis confirms that the selected model is robust to baseline definition and avoids score compression observed in alternative approaches.

Applying the validated framework to 2025 disclosures, we identify substantial heterogeneity in structural deviation across firms. JPMorgan exhibits a high percentile-based risk score and pronounced semantic displacement relative to its historical baseline, while Wells Fargo shows moderate but controlled structural adjustment. Feature-level SHAP analysis provides interpretable explanations, highlighting the roles of semantic distance, litigation emphasis, and disclosure length contraction in driving anomaly scores.

Overall, the findings demonstrate that embedding-based structural modeling captures disclosure transformation that traditional lexical counting overlooks. The proposed framework provides a scalable and interpretable approach for structural risk assessment, enabling both detection and explanation of narrative shifts in corporate reporting.

Future research may extend this framework to broader firm samples, cross-industry comparisons, and forward-looking prediction of market or regulatory outcomes based on structural disclosure dynamics.

Funding: research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202603.0179.v1

References

Acharya, V.V.; Engle, R.; Richardson, M. Capital shortfall: A new approach to ranking and regulating systemic risks. *Am. Econ. Rev.* 2012, 102, 59–64.

Baker, S.R.; Bloom, N.; Davis, S.J. Measuring economic policy uncertainty. *Q. J. Econ.* 2016, 131, 1593–1636.

Bao, Y.; Datta, A.; Liu, Z. Litigation risk and voluntary disclosure: Evidence from risk factor disclosures. *J. Account. Public Policy* 2018, 37, 1–23.

Beatty, A.; Liao, S.; Weber, J. The effect of private information and monitoring on the role of accounting quality in investment decisions. *Contemp. Account. Res.* 2010, 27, 17–47.

Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 2003, 3, 993–1022.

Bolton, R.J.; Hand, D.J. Statistical fraud detection: A review. *Stat. Sci.* 2002, 17, 235–255.

Campbell, J.L.; Chen, H.; Dhaliwal, D.S.; Lu, H.M.; Steele, L.B. The information content of mandatory risk factor disclosures in corporate filings. *Rev. Account. Stud.* 2014, 19, 396–455.

Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* 2009, 41, 1–58.

Chen, Y.; Song, L.; Liu, Z.; Yao, J.; Liu, K.; Liao, Q. TrustLLM-Fin: A Privacy-Centric and Auditable Impact Assessment Framework for Large Language Models in Automated Financial Reporting. 2026.

Consumer Financial Protection Bureau; Office of the Comptroller of the Currency; Office of the Los Angeles City Attorney. Wells Fargo Bank, N.A. Consent Order and Stipulation and Consent to the Issuance of a Consent Order; CFPB Administrative Proceeding No. 2016-CFPB-0015; Washington, DC, USA, 2016.

Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL HLT 2019*, Minneapolis, USA, 2–7 June 2019; pp. 4171–4186.

Hoberg, G.; Phillips, G. Text-based network industries and endogenous product differentiation. *J. Polit. Econ.* 2016, 124, 1423–1465.

Huang, X.; Teoh, S.H.; Zhang, Y. Tone management. *Account. Rev.* 2014, 89, 1083–1113.

Kravet, T.; Muslu, V. Textual risk disclosures and investors' risk perceptions. *Rev. Account. Stud.* 2013, 18, 1088–1122.

Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* 2004, 88, 365–411.

Li, F. The information content of forward-looking statements in corporate filings. *J. Account. Res.* 2010, 48, 1049–1102.

Liao, Q.; Chen, Y.; He, S.; Wang, R.; Xu, W.; Chu, W. Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience. 2025.

Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In *Proceedings of the 2008 IEEE International Conference on Data Mining*; IEEE: 2008; pp. 413–422.

Liu, X.; Huang, D.; Yao, J.; Dong, J.; Song, L.; Wang, H.; Yao, C.; Chu, W. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* 2025, 6, 285.

Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-K filings. *J. Finance* 2011, 66, 35–65.

Loughran, T.; McDonald, B. Textual analysis in accounting and finance: A survey. *J. Account. Res.* 2016, 54, 1187–1230.

Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; 2017; pp. 4765–4774.

Reimers, N.; Gurevych, I. Sentence BERT: Sentence embeddings using Siamese BERT networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.

Sakurada, M.; Yairi, T. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis*; ACM: 2014; pp. 4–11.

Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Finance* 2007, 62, 1139–1168.

Yang, X.; You, Y.; Zhang, Y. Financial text mining using deep learning: A review. *Expert Syst. Appl.* 2020, 157, 113544.