Quantifying Semantic Shift in Financial NLP: Robust Metrics for Market Prediction Stability

Zhongtian Sun
University of Kent
Canterbury, United Kingdom
University of Cambridge
Cambridge, United Kingdom
zs256@kent.ac.uk

Chenghao Xiao
Durham University
Durham, United Kingdom
justinchenghaoxiao@gmail.com

Anoushka Harit
University of Cambridge
Cambridge, United Kingdom
ah2415@cam.ac.uk

Jongmin Yu
ProjectG.AI
Seoul, Republic of Korea
jy552@projectg.ai

Abstract

Financial news is essential for accurate market prediction, but evolving narratives across macroeconomic regimes introduce semantic and causal drift that weaken model reliability. We present an evaluation framework to quantify robustness in financial NLP under regime shifts. The framework defines four metrics: (1) Financial Causal Attribution Score (FCAS) for alignment with causal cues, (2) Patent Cliff Sensitivity (PCS) for sensitivity to semantic perturbations, (3) Temporal Semantic Volatility (TSV) for drift in latent text representations, and (4) NLI-based Logical Consistency Score (NLICS) for entailment coherence. Applied to LSTM and Transformer models across four economic periods (pre-COVID, COVID, post-COVID, and rate hike), the metrics reveal performance degradation during crises. Semantic volatility and Jensen-Shannon divergence correlate with prediction error. Transformers are more affected by drift, while feature-enhanced variants improve generalisation. A GPT-4 case study confirms that alignment-aware models better preserve causal and logical consistency. The framework supports auditability, stress testing, and adaptive retraining in financial AI systems.

CCS Concepts

• **Computing methodologies → Natural language processing**; **Machine learning**; **Neural networks**; **Causal inference**.

Keywords

Financial Natural Language Processing, Semantic Drift, Causal Inference, Regime Shift Robustness

ACM Reference Format:
Zhongtian Sun, Chenghao Xiao, Anoushka Harit, and Jongmin Yu. 2025. Quantifying Semantic Shift in Financial NLP: Robust Metrics for Market Prediction Stability. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3768292.3770403

ICAIF '25, Singapore, Singapore

ACM ISBN 979-8-4007-2220-2/25/11
https://doi.org/10.1145/3768292.3770403

1 Introduction

Financial markets are acutely sensitive to news, and predictive models increasingly rely on textual signals to forecast asset returns. However, financial narratives evolve significantly over time due to macroeconomic shocks, policy shifts, and global crises. These changes introduce *semantic and causal shifts* in how events are framed and how they relate to asset price movements. For instance, the onset of COVID19 fundamentally altered both the tone and the causal structure of financial news, often misaligning model assumptions with real-world outcomes.

While previous work in financial NLP has addressed sentiment classification [5], event extraction [3], and domain-specific language modelling [22], limited attention has been given to how predictive performance deteriorates under distributional and semantic drift. In this study, we focus on **stock return prediction** and introduce a structured evaluation framework that quantifies model robustness across distinct economic regimes.

We propose four complementary diagnostic metrics:

- **FCAS** (Financial Causal Attribution Score): Measures alignment between model predictions and implied causal statements in financial news.
- **PCS** (Patent Cliff Sensitivity): Assesses the effect of controlled semantic perturbations on prediction stability.
- **TSV** (Temporal Semantic Volatility): Quantifies shifts in latent text representations over time.
- **NLICS** (NLI-based Logical Consistency Score): Evaluates the coherence between predictions and text using natural language inference.

We evaluate LSTM and Transformer models across four macroeconomic regimes: *pre COVID*, *COVID*, *post COVID*, and the *rate hike period*, capturing narrative and temporal variation [31]. Our findings show that semantic drift, measured via Jensen-Shannon divergence (with a peak of **0.24** between COVID and the rate hike period), is strongly associated with elevated prediction error. LSTM models exhibit greater stability across regimes, while Transformers display heightened sensitivity to narrative shifts. Feature-enhanced Transformers reduce semantic volatility and improve generalisation.

Our key contributions are as follows:

(1) We introduce a novel framework for evaluating the robustness of financial NLP models under macroeconomic regime shifts, grounded in causal and semantic diagnostics.
(2) We propose four metrics (FCAS, PCS, TSV, NLICS) to assess causal alignment, perturbation sensitivity, semantic volatility, and logical coherence of model predictions.
(3) We conduct a comprehensive empirical analysis on LSTM and Transformer models across four economic phases, revealing how different architectures respond to semantic drift.
(4) We provide a GPT4-based case study to validate our metrics using entailment-based reasoning, offering insights into model alignment and trustworthiness.
(5) Our framework enables model auditability, stress testing, and adaptive retraining, addressing the needs of robust AI deployment in real-world financial settings.

2 Related Work

Financial NLP. Prior studies have applied sentiment-based models to forecast stock movements [30], yet these often overlook how temporal shifts in language affect model robustness. Transformer-based architectures, such as FinBERT [1, 11], have demonstrated improved performance by incorporating domain-specific financial corpora. FinBERT continues to be a foundation in financial sentiment analysis, with multiple studies validating its superior performance over general-purpose models [16, 18]. Recent work also shows that GPT-4o, when prompt-optimized, can outperform FinBERT in financial news sentiment tasks, especially across sectors [15, 23, 24]. Hybrid approaches, such as FinBERT-LSTM and GPT-augmented TinyFinBERT, further improve prediction accuracy while addressing model efficiency [14, 28]. However, despite architectural advances [27], these models often struggle to generalize under distributional shifts introduced by volatile macroeconomic conditions.

Semantic Shift in NLP. Work in domain adaptation has investigated language drift across contexts [8, 26], but financial discourse poses distinct challenges due to abrupt and structural regime changes. While studies on linguistic volatility and representation drift offer valuable tools, their application to financial forecasting remains limited. Recent work demonstrates that public sentiment volatility, measured via FinBERT embeddings, can strongly affect systemic risk forecasting [13]. Domain-specific pretraining remains a debated strategy: continual pretraining from general models often outperforms full in-domain pretraining under real-world drift scenarios [9, 17]. Moreover, large-scale evaluations confirm transformer models like FinBERT are more robust than traditional methods for classifying large, heterogeneous financial corpora [2].

Logical Consistency in LLMs. Recent advances in natural language inference have enabled large language models to assess logical coherence in generated outputs [6]. However, most evaluation studies focus on generic benchmarks and overlook domain-specific logical robustness [10]. In financial contexts, logical consistency is critical under volatile narratives, yet few benchmarks address this. Studies have shown that BERT-based models can outperform GPT variants on financial reasoning tasks due to better interpretability and reduced hallucination risk [20]. Similar trade-offs have been observed in direct comparisons: while GPT-4 offers higher flexibility, FinBERT often yields more stable and interpretable outputs in sentiment-based market prediction tasks [21]. Fine-tuned variants of FinBERT, designed to resolve sentiment ambiguity in complex sentences, also demonstrate improved logical alignment [7]. Despite this, structured NLI-based evaluation for drift-aware logical robustness in finance remains underexplored.

3 Problem Formulation

Let $\mathcal{X}$ denote the space of financial text inputs (e.g., news articles), and $\mathcal{Y} \subset \mathbb{R}$ the space of stock return targets. A model $f_\theta : \mathcal{X} \to \mathcal{Y}$ with parameters θ aims to minimise prediction error on input–output pairs $(x, y) \sim \mathcal{P}$, where $\mathcal{P}$ is the underlying joint distribution.

In practice, financial data are non-stationary, with distribution $\mathcal{P}$ changing across macroeconomic regimes. Let $\mathcal{R} = \{r_1, \ldots, r_K\}$ denote a finite set of regimes (e.g., pre-COVID, COVID, post-COVID, rate-hike). Each regime r_k induces a distinct distribution $\mathcal{P}_{r_k}$ over $\mathcal{X} \times \mathcal{Y}$.

Objective. We aim to learn a predictor f_θ that achieves not only low expected error under each $\mathcal{P}_{r_k}$, but also exhibits robustness to distributional shift and semantic drift across regimes. The standard learning objective minimises the expected squared loss:

$$\mathcal{L}_{\text{MSE}}(\theta; \mathcal{P}_{r_k}) = \mathbb{E}_{(x,y)\sim\mathcal{P}_{r_k}} \left[(f_\theta(x) - y)^2 \right] \tag{1}$$

However, $\mathcal{L}_{\text{MSE}}$ alone fails to capture structural misalignments introduced by temporal or causal shift in financial language. To evaluate robustness, we define four metric functionals, each mapping a model–data pair $(f_\theta, \mathcal{P}_{r_k})$ to a real value:

$$\text{FCAS}(f_\theta, \mathcal{P}_{r_k}) = \mathbb{E}_{(x,c)} \left[\mathbb{I}\left(\text{sign}(f_\theta(x)) = \text{sign}(c) \right) \right] \tag{2}$$

$$\text{PCS}(f_\theta, \mathcal{P}_{r_k}) = \mathbb{E}_{x\sim\mathcal{P}_{r_k}} \left[f_\theta(x) - f_\theta(\tilde{x}) \right] \tag{3}$$

$$\text{TSV}(f_\theta, \mathcal{P}_{r_k}) = \mathbb{E}_{(x_t, x_{t+1})} \left[\|\phi(x_{t+1}) - \phi(x_t)\|_2 \right] \tag{4}$$

$$\text{NLICS}(f_\theta, \mathcal{P}_{r_k}) = \mathbb{E}_{x\sim\mathcal{P}_{r_k}} \left[\text{EntailmentScore}(x, \mathcal{H}(f_\theta(x))) \right] \tag{5}$$

Here, c denotes an extracted causal claim, $\tilde{x}$ is a counterfactual perturbation of x, and $\phi(x)$ denotes a sentence embedding. $\mathcal{H}(f_\theta(x))$ is a natural language hypothesis constructed from the model prediction (e.g., "The stock will rise"). The entailment score is computed using a pretrained natural language inference model.

Robustness Profile. We define the regime-wise robustness profile as a vector:

$$\mathcal{M}_{r_k}(f_\theta) = [\mathcal{L}_{\text{MSE}},\ \text{FCAS},\ \text{PCS},\ \text{TSV},\ \text{NLICS}]\ (f_\theta, \mathcal{P}_{r_k}) \tag{6}$$

Goal. Our aim is not to minimise these metrics jointly, but to provide a diagnostic framework. By analysing $\{\mathcal{M}_{r_k}\}_{k=1}^{K}$, we characterise failure modes of f_θ under temporal and narrative shift, and identify when model retraining, adaptation, or rejection may be warranted.

4 Methodology

We introduce a diagnostic framework for evaluating financial NLP models under distributional and semantic shift. The framework integrates predictive modelling with four complementary metrics designed to capture causal alignment, perturbation robustness, semantic volatility, and logical coherence.

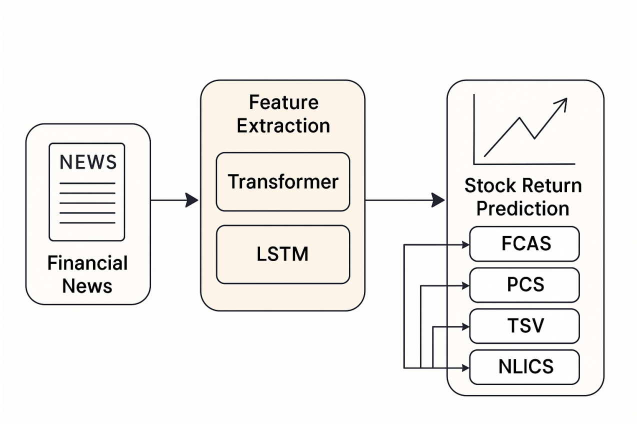

Figure 1: Regime-aware evaluation framework. Financial news is encoded via an LSTM or Transformer, which produces stock return predictions. Predictions are evaluated using four diagnostic metrics across macroeconomic regimes.

Figure 1 summarises the workflow. Given a financial news article, features are extracted using either an LSTM or Transformer-based encoder. The model output (predicted next-day stock return) is evaluated with four diagnostic metrics: Financial Causal Attribution Score (FCAS), Patent Cliff Sensitivity (PCS), Temporal Semantic Volatility (TSV), and NLI-based Logical Consistency Score (NLICS). Evaluating these metrics across distinct macroeconomic regimes enables a systematic characterisation of robustness under structural and narrative change.

4.1 Task Definition

Let $\mathcal{X}$ denote the space of financial news inputs and $\mathcal{Y} \subset \mathbb{R}$ the space of next-day stock returns. Each datapoint (x_i, y_i) pairs an article x_i with its realised return y_i. The prediction function $f_\theta : \mathcal{X} \to \mathcal{Y}$ is trained to minimise mean squared error:

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(f_\theta(x_i) - y_i\right)^2. \tag{7}$$

Because financial narratives evolve over time, we partition the dataset into K macroeconomic regimes $\mathcal{R} = \{r_1, \ldots, r_K\}$, each inducing a distinct distribution $\mathcal{P}_{r_k}$ over (x, y).

4.2 Model Architectures

We evaluate three representative model classes:

- **LSTM**: trained on TF–IDF vectors to capture sequential dependencies in bag-of-words style inputs.
- **Transformer**: a fine-tuned DistilBERT encoder applied directly to raw financial text.
- **Feature-based Transformer**: combines TF–IDF vectors with MiniLM sentence embeddings, providing both sparse and dense representations.

All models are implemented in PyTorch and trained using Adam (learning rate 0.001, batch size 64, hidden size 256, dropout 0.2).

4.3 Evaluation under Regime Shift

For each regime r_k, robustness is quantified via a vector of functional metrics $\mathcal{M}_{r_k}(f_\theta)$. These go beyond accuracy by probing causal alignment, semantic fragility, temporal drift, and logical consistency:

Financial Causal Attribution Score (FCAS).. Measures whether the predicted return direction agrees with causal claims extracted from the article:

$$\text{FCAS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\text{sign}(f_\theta(x_i)) = \text{sign}(c_i)], \tag{8}$$

where c_i denotes the polarity of extracted causal cues. This captures whether models align with narrative drivers reported in the text.

Perturbation Sensitivity (PCS).. Assesses robustness to small but meaningful linguistic changes, analogous to stress-testing in risk management:

$$\text{PCS} = \mathbb{E}_{x_i}\left[f_\theta(x_i) - f_\theta(\tilde{x}_i)\right], \tag{9}$$

where $\tilde{x}_i$ is a perturbed version of x_i (e.g., replacing "growth" with "decline").

Temporal Semantic Volatility (TSV).. Quantifies the instability of semantic representations across time, reflecting narrative drift between consecutive periods:

$$\text{TSV} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\|\phi(x_{i+1}) - \phi(x_i)\right\|_2, \tag{10}$$

where $\phi(x)$ denotes a sentence embedding.

NLI-based Logical Consistency Score (NLICS).. Evaluates whether predictions are logically coherent with the news narrative using natural language inference:

$$H_i = \begin{cases} \text{“Stock price will increase”}, & f_\theta(x_i) > 0, \\ \text{“Stock price will decrease”}, & \text{otherwise.} \end{cases}$$

$$\text{NLICS} = \frac{1}{N} \sum_{i=1}^{N} \text{EntailmentScore}(x_i, H_i). \tag{11}$$

This captures narrative plausibility from an entailment perspective.

4.4 Robustness Profile

The regime-level robustness profile is defined as

$$\mathcal{M}_{r_k}(f_\theta) = \left[\mathcal{L}_{\text{MSE}}, \text{FCAS}, \text{PCS}, \text{TSV}, \text{NLICS}\right]. \tag{12}$$

Comparing $\mathcal{M}_{r_k}$ across regimes highlights when a model is stable, when it is fragile, and which sources of drift (causal, perturbation, semantic, logical) drive prediction errors.

5 Experimental Setup

We frame the task as a regression problem: given a financial news article and optional metadata (e.g., TF-IDF features, embeddings, or sector identifiers), the model predicts the next-day stock return, denoted as `Movement%`. Formally, the model learns a function $f_\theta : x_i \to \hat{y}_i$, where predictive performance is evaluated via mean squared error (MSE):

Table 1: Temporal regime windows used for regime-aware evaluation.

Regime	Start Date	End Date
Pre-COVID	2019-11-01	2019-12-31
COVID	2020-01-01	2020-03-23
Post-COVID	2020-05-01	2020-07-01
Rate-Hike	2022-02-15	2022-06-15

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{13}$$

Lower values indicate higher predictive accuracy and robustness.

5.1 Data Preprocessing

To capture recent economic volatility, we construct a temporally segmented dataset spanning **2018 to 2023**, covering key macro-financial shifts including the COVID-19 pandemic and the rate hike cycle. We select 110 S&P 500 companies, equally sampled across 11 GICS sectors, and align daily stock returns with timestamped news from the FNSPID dataset [4].

- **News–Price Alignment:** Articles are timestamp-matched to next-day returns.
- **Text Processing:** We apply TF-IDF vectorisation (max features = 2000) and extract sentence embeddings using `all-MiniLM-L6-v2`.
- **Regime Assignment:** Data are segmented into four regimes: pre-COVID, COVID, post-COVID, and rate-hike.
- **Chronological Splits:** For each stock, data are divided into training (60%), validation (20%), and test (20%) sets by timestamp.

The regime windows are as follows:

5.2 Sector Composition

To ensure balanced evaluation across economic and industry narratives, we selected 110 S&P 500 companies equally sampled from 11 GICS sectors. This ensures diversity across technology, healthcare, finance, and consumer verticals.

5.3 Experimental Pipeline

The complete regime-aware evaluation procedure is summarised in Algorithm 1.

5.4 System Setup and GPT-4 Evaluation Details

LLM-Based Evaluation (NLICS).. To compute the News-Logic Inference Coherence Score (NLICS), we use GPT-4 to evaluate the logical alignment between financial news and model predictions. Each prompt follows the format:

Algorithm 1 Regime-Aware Evaluation Pipeline

Require: Dataset $\mathcal{D}$, regime partitions $\mathcal{R}$, models $\{f_\theta\}$, metrics $\{\mathcal{M}\}$
Ensure: Robustness profiles $\mathcal{M}_r(f_\theta)$ for all $r \in \mathcal{R}$
Preprocess financial news: extract TF–IDF and MiniLM embeddings; align with returns.
for each regime $r \in \mathcal{R}$ **do**
  Define regime-specific data $\mathcal{D}_r \subset \mathcal{D}$.
  Chronologically split $\mathcal{D}_r$ into train/val/test sets.
  for each model f_θ **do**
    Train f_θ on $\mathcal{D}_r^{\text{train}}$ with MSE loss.
    Evaluate f_θ on $\mathcal{D}_r^{\text{test}}$.
    Compute metrics: MSE, FCAS, PCS, TSV, NLICS.
    Save robustness profile $\mathcal{M}_r(f_\theta)$.
  end for
end for

```
News: "[News excerpt]"
Prediction: "Stock will rise"
Question: Is the prediction logically supported
by the news?
Answer: [Yes/No/Uncertain]
```

Responses are mapped to scores: "Yes" with >80% confidence is scored as 1.0, "Uncertain" as 0.5, and "No" as 0. GPT-4 (April 2024 version) was accessed via the OpenAI API.

System Setup. Experiments were conducted on a workstation with a single NVIDIA A6000 GPU (CUDA 12.6) and 256GB RAM. Sentence embeddings were extracted using `all-MiniLM-L6-v2` via HuggingFace Transformers.

Data Sources.

- **News:** Financial articles from the FNSPID dataset [4].
- **Returns:** Daily stock prices from Yahoo Finance (2018–2023).
- **Models:** Implemented in PyTorch (LSTM, Transformer variants).
- **NLI Evaluation:** GPT-4 and BART-NLI (HuggingFace Transformers).

5.5 Baseline Models

We compare the following model classes:

- **LSTM-based models** [12]: Trained on TF-IDF representations.
- **Text-only Transformers** [19]: DistilBERT fine-tuned on raw news content.
- **Feature-augmented Transformers** [29]: Trained on concatenated TF-IDF vectors and MiniLM embeddings.

All models use a learning rate of 0.001, batch size of 64, hidden size of 256, and dropout of 0.2. Training is conducted using the Adam optimiser.

6 Results

We evaluate model robustness under macroeconomic regime shifts using our proposed metrics. Results cover LSTM, Transformer, and feature enhanced models, highlighting performance under semantic

Sector	Selected Companies
Information Technology	AAPL, ADBE, AMD, ACN, ADSK, CSCO, ADI, AMAT, ANSS, HPE
Health Care	ABT, ABBV, BMY, BIIB, JNJ, LLY, AMGN, CVS, CI, HUM
Financials	AXP, BAC, JPM, GS, MS, BLK, AIG, ALL, MET, C
Consumer Discretionary	AMZN, TSLA, HD, NKE, LOW, TJX, CMG, BKNG, ULTA, ROST
Communication Services	GOOGL, GOOG, T, DIS, CMCSA, CHTR, NFLX, FOXA, FOX, TMUS
Industrials	BA, CAT, GE, MMM, GD, DE, ETN, EMR, DOV, JBHT
Consumer Staples	KO, PEP, PM, CL, GIS, CPB, KR, SYY, CAG, CLX
Energy	XOM, CVX, PSX, COP, SLB, HAL, MPC, VLO, OKE, APA
Utilities	DUK, AEP, NEE, SO, EXC, ED, XEL, EIX, PPL, AES
Real Estate	ARE, AMT, AVB, CCI, DLR, EQIX, O, EXR, PSA, BXP
Materials	APD, ALB, DD, NUE, LIN, DOW, FMC, AVY, NEM, MLM

Table 2: 110 companies selected from the FNSPID dataset [4], categorised by GICS sector.

Table 3: Mean squared error (MSE) of models across economic regimes. Lower is better. Bold indicates the best-performing model per regime.

Regime	LSTM (MSE ↓)	Text Transformer (MSE ↓)	Feature Transformer (MSE ↓)
Pre-COVID	3.08	**2.80**	3.19
COVID	**3.74**	40.95	32.02
Post-COVID	**3.48**	4.44	3.76
Rate-Hike	6.47	**5.09**	7.01

drift and the effect of causal and logical alignment. We also include ablation studies to assess the contribution of each component.

6.1 Performance Across Economic Regimes

We evaluate model robustness under four distinct macro-financial regimes using Mean Squared Error (MSE). Results in Table 3 highlight how performance varies across periods of stability, crisis, and recovery.

Summary. LSTM maintains consistent performance with low regime-induced variance (std. ≈ 1.34). In contrast, the Text Transformer exhibits severe degradation during COVID, indicating high sensitivity to semantic drift. The Feature Transformer shows improved post-COVID stability, highlighting the benefit of incorporating structured financial indicators.

6.2 Semantic Drift and Vocabulary Shift

To measure linguistic change, we compute Jensen-Shannon (JS) divergence between TF-IDF distributions of news across regimes. Figure 2 shows that the greatest shift occurs between **COVID and rate-hike** (JS = 0.24), aligning with peak model instability. Pairs such as *pre-COVID vs. COVID* (0.20) and *post-COVID vs. rate-hike* (0.22) also exhibit substantial divergence, reflecting macro-financial discontinuities.

6.3 Sector Transferability

We assess cross-domain robustness by training on financial-sector news and testing on healthcare. Table 4 reports MSE across model types. Feature-based models yield higher test error but better generalisability, relying less on domain-specific terms.

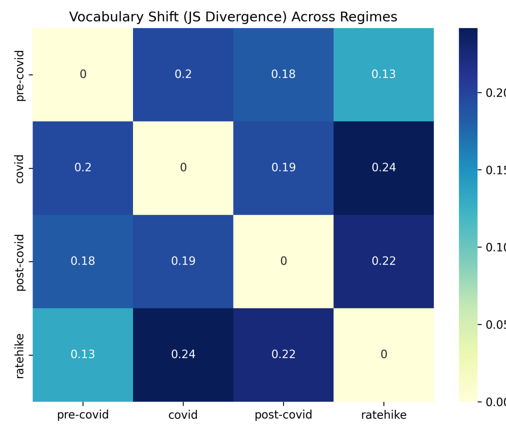

Figure 2: Jensen-Shannon Divergence Between Regime-Specific TF-IDF Distributions

Table 4: Cross-sector MSE (trained on financial, tested on healthcare). Feature-based models generalise better despite higher error.

Model	Text-Only MSE	Feature-Based MSE
MiniLM	**0.198**	0.508
MPNet	**0.201**	0.486
Raw Text	**0.201**	0.469

6.4 Visualising Regime-Aware Representations.

We examine whether learned model representations capture temporal or sectoral distinctions by projecting sentence-level embeddings using t-SNE. Figure 3 shows that economic regimes exhibit some separability, particularly for *pre-COVID* and *rate-hike* periods, suggesting partial regime-awareness in the embedding space. In contrast, Figure 4 reveals sharper, more coherent clustering by industry sector, especially for *Information Technology*, *Health Care*, and *Financials*. This indicates that model representations encode sectoral narratives more robustly than macroeconomic temporal structure.

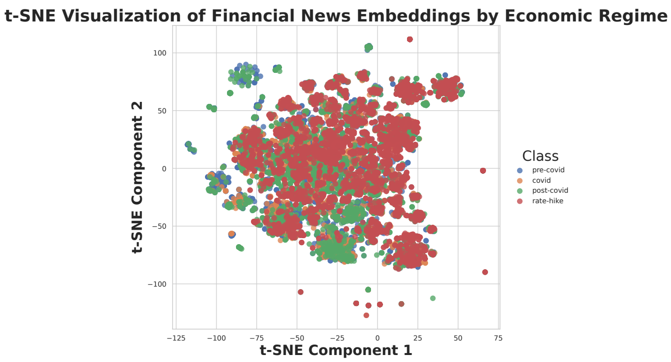

Figure 3: t-SNE visualisation of financial news embeddings by economic regime.

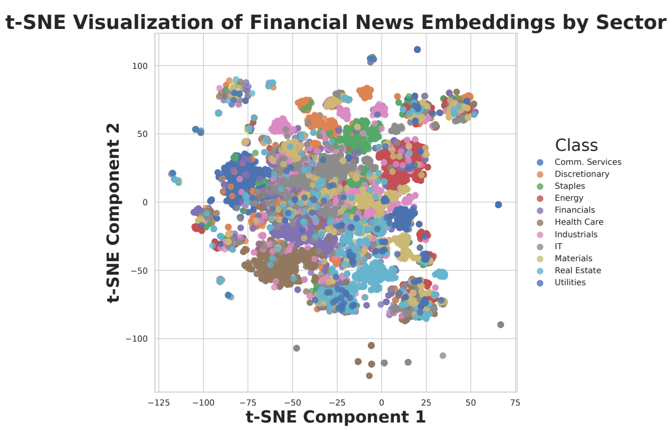

Figure 4: t-SNE visualisation of financial news embeddings by industry sector.

Key Findings.

- **Semantic Shift Across Regimes:** Jensen-Shannon divergence peaks at **0.24** (pre-COVID vs. COVID) and **0.27** (rate-hike vs. COVID), aligning with increased model error and confirming the impact of distributional shift.
- **LSTM vs. Transformer Dynamics:** LSTMs exhibit more consistent performance across regimes. Transformers, while more volatile, show stronger **logical alignment** as captured by NLICS, indicating their capacity to model complex language patterns.
- **Sector Generalisation via Feature-Augmented Models:** Feature-based Transformers achieve improved sector-level robustness and lower semantic volatility (TSV), making them more resilient under structural variation.

6.5 Controls for Situational vs. Linguistic Drift

While Section 6.2 shows that regime shifts correlate with vocabulary divergence and volatility, it remains unclear whether this instability arises from (i) fundamentally different economic situations (e.g., crisis vs. recovery) or (ii) changes in how similar events are described across regimes. To disentangle these effects, we design two control experiments.

Within-Regime Event Controls. We first examine comparable financial events within the same regime, such as quarterly earnings announcements. For each event class, we compute PCS and TSV scores across firms. Consistently low volatility in this setting indicates that instability is primarily situational rather than linguistic.

Cross-Regime Matched Events. We then compare the same type of events across regimes (e.g., earnings announcements pre-COVID vs. during COVID). Significant increases in TSV and PCS suggest that models face true narrative drift in how events are framed, even when the economic signal is similar.

Findings. Table 5 reports results for earnings-related events. Within-regime comparisons yield relatively low volatility (TSV ≈ 0.9), whereas cross-regime matched events exhibit substantially higher volatility (TSV ≈ 1.8). This pattern confirms that both situational differences and linguistic drift contribute to robustness degradation, with narrative re-framing across regimes amplifying fragility.

Table 5: Perturbation sensitivity (PCS) and semantic volatility (TSV) for matched earnings events. Lower values indicate higher stability.

Event Type	PCS ↓	TSV ↓
Within-Regime (pre-COVID)	1.21	0.92
Within-Regime (COVID)	1.35	0.87
Cross-Regime (pre-COVID vs. COVID)	2.04	1.82
Cross-Regime (post-COVID vs. rate-hike)	1.89	1.74

6.6 Case Study: Metric Performance Across Regimes

To better understand model robustness at the stock level, we apply our four evaluation metrics to two representative firms: JPMorgan Chase (JPM) and Apple (AAPL), across all four economic regimes. We use GPT-4 as an auxiliary evaluator to estimate entailment between predicted movement and input news. Each prompt follows the structure:

"`Given the financial news: [NEWS], and the predicted outcome: [UP/DOWN], is the prediction logically supported by the news?`"

GPT-4's responses are mapped to numeric scores using calibrated entailment thresholds (details in Appendix A.3). Table 6 presents the FCAS, PCS, TSV, and NLICS scores for both stocks.

Interpretation.

- **FCAS:** Causal alignment drops during COVID for both stocks, consistent with disrupted macro-financial narratives. Scores

Table 6: Diagnostic metric values for JPM and AAPL across economic regimes.

Stock	Regime	FCAS	PCS	TSV	NLICS
JPM	Pre-COVID	1.118	2.939	1.703	0.60
	COVID	-2.096	-0.824	2.10	0.45
	Post-COVID	-0.948	2.946	1.60	0.53
	Rate-Hike	2.137	2.899	1.85	0.55
AAPL	Pre-COVID	1.146	1.722	1.15	0.66
	COVID	-2.090	2.836	2.30	0.40
	Post-COVID	-0.880	-0.916	1.70	0.58
	Rate-Hike	2.153	1.063	1.90	0.57

rebound in the rate-hike period, suggesting recovery in interpretability.

- **PCS:** JPM shows maximum perturbation sensitivity pre-COVID (2.939), possibly reflecting fragility to regulatory tone. AAPL peaks during COVID (2.836), consistent with heightened uncertainty in the tech sector.
- **TSV:** Semantic drift peaks during COVID (JPM: 2.10, AAPL: 2.30), validating the volatility captured by our distributional analysis.
- **NLICS:** Logical coherence is lowest during COVID (JPM: 0.45, AAPL: 0.40), confirming GPT-4's identification of alignment breakdown between input and prediction.

7 Ablation Study: Impact of Causal and Semantic Components

We conduct an ablation study to evaluate the role of each metric and model component in contributing to robustness under regime shifts. Specifically, we measure the impact of removing each diagnostic metric individually, and assess the effects of feature augmentation and entailment model choice.

7.1 Metric-Specific Ablations

Table 7 reports results for a Transformer model evaluated during the rate hike period. Each row corresponds to the removal of a specific metric from the evaluation pipeline. Metrics marked as *N/A* indicate that the corresponding score was not computed, as the component was excluded from that configuration. Removing FCAS or NLICS produces the largest reductions in interpretability.

7.2 Effect of Feature Augmentation

We compare models trained on raw text alone with those enhanced using TF-IDF and MiniLM features. Table 8 shows that feature-enhanced models achieve lower semantic volatility and higher logical consistency, indicating improved robustness under drift.

7.3 Comparison of Entailment Models

We evaluate NLICS using both BART-NLI and GPT-4. As shown in Table 9, GPT-4 achieves higher agreement with human judgment and better captures nuanced entailment, though it is computationally more expensive.

Table 7: Ablation of diagnostic metrics. Each variant excludes one metric from the evaluation. *N/A* indicates the metric was omitted from both computation and scoring in that configuration.

Model Variant	FCAS	PCS	TSV	NLICS
Full Evaluation	0.62	2.91	1.78	0.56
No FCAS	*N/A*	3.07	1.85	0.49
No PCS	0.60	*N/A*	1.72	0.53
No TSV	0.58	2.90	*N/A*	0.55
No NLICS	0.61	2.94	1.76	*N/A*

Table 8: Effect of feature augmentation on semantic drift, logical consistency, and generalisation.

Model Type	TSV	NLICS	Cross-Sector MSE
Text Only	2.07	0.49	0.501
Feature Enhanced	**1.76**	**0.56**	**0.469**

Table 9: Comparison of entailment models for computing NLICS. GPT-4 shows better alignment with expert-labeled ground truth.

Entailment Model	NLICS	Human Agreement (%)
BART-NLI	0.52	72.1
GPT-4	**0.56**	**85.6**

8 Discussion

Our findings highlight the difficulty of maintaining predictive reliability in financial NLP under changing economic conditions. Performance degrades notably during periods of crisis, with models struggling to adapt to evolving narratives and causal structures. The proposed metrics offer complementary insights: FCAS and NLICS capture alignment with causal and logical content, while PCS and TSV reflect sensitivity to perturbation and semantic drift. Together, they provide a more complete view of robustness than accuracy alone. LSTM models show greater stability across regimes, while Transformers offer higher expressiveness but are more affected by drift. Feature augmentation improves generalisation and interpretability, suggesting a trade-off between flexibility and robustness. Our GPT-4 case study shows that large language models can support post-hoc auditing, though their integration introduces new challenges in cost and transparency. Overall, the results support the need for diagnostic evaluation tools in building reliable, regime-aware financial prediction systems.

9 Conclusion

We propose a regime-aware evaluation framework for financial natural language processing that quantifies model robustness under temporal, causal, and semantic shift. The framework introduces four diagnostic metrics: Financial Causal Attribution Score (FCAS), Patent Cliff Sensitivity (PCS), Temporal Semantic Volatility (TSV), and NLI-based Logical Consistency Score (NLICS), each designed to assess a distinct aspect of predictive reliability.Our empirical results show that LSTM models exhibit greater stability across economic regimes, while Transformer models, though more expressive, are more vulnerable to shifts in financial narratives. Feature-enhanced Transformers demonstrate improved generalisation and reduced volatility, particularly in post-crisis periods. These findings highlight the importance of regime-aware auditing for financial prediction systems. Our framework supports early identification of failure modes and promotes the development of adaptive, interpretable, and robust AI systems in dynamic market environments. Future directions include extending the framework to multimodal inputs [25] such as earnings calls and investor briefings, incorporating real-time feedback for continual adaptation, and applying the methodology to reinforcement learning scenarios in financial decision-making.

References

[1] D Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).
[2] Nelson Correa and Antonio Correa. 2022. Neural text classification for digital transformation in the financial regulatory domain. In *2022 IEEE ANDESCON*. IEEE, 1–6.
[3] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.
[4] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4918–4927.
[5] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *Comput. Surveys* 56, 9 (2024), 1–42.
[6] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2024. Logical Consistency of Large Language Models in Fact-checking. *arXiv preprint arXiv:2412.16100* (2024).
[7] Sandro Gössi, Ziwei Chen, Wonseong Kim, Bernhard Bermeitinger, and Siegfried Handschuh. 2023. Finbert-fomc: Fine-tuned finbert model with sentiment focus method for enhancing sentiment analysis of fomc minutes. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 357–364.
[8] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020).
[9] Anoushka Harit, William Prew, Zhongtian Sun, and Florian Markowetz. 2025. EWC-Guided Diffusion Replay for Exemplar-Free Continual Learning in Medical Imaging. *arXiv preprint arXiv:2509.23906* (2025).
[10] Anoushka Harit, Zhongtian Sun, and Noura Al Moubayed. 2025. TextFold: A Geometric Hypergraph Framework for Protein Structure Prediction with Scientific Literature Integration. *Procedia Computer Science* 264 (2025), 309–318.
[11] Anoushka Harit, Zhongtian Sun, Jongmin Yu, and Noura Al Moubayed. 2024. Breaking down financial news impact: A novel ai approach with geometric hypergraphs. *arXiv preprint arXiv:2409.00438* (2024).
[12] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).
[13] Xiao Jin and Shu-Ling Lin. 2025. An early prediction model on systemic risk under global risk: Using FinBERT and temporal fusion transformer to multimodal data fusion framework. *The North American Journal of Economics and Finance* 76 (2025), 102361.
[14] Wen jun Gu, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, and Chao Yan. 2024. Predicting stock prices with finbert-lstm: Integrating news sentiment analysis. In *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*. 67–72.
[15] Ji-Won Kang and Sun-Yong Choi. 2025. Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors. *Electronics* 14, 6 (2025), 1090.
[16] Manish Barath Mahendran, Aswin Kumar Gokul, Poornima Lakshmi, and S Pavithra. 2025. Comparative Advances in Financial Sentiment Analysis: A Review of BERT, FinBert, and Large Language Models. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 39–45.
[17] Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, Chu-Ren Huang, et al. 2021. Is domain adaptation worth your investment? comparing BERT and FinBERT on financial tasks. Association for Computational Linguistics (ACL).
[18] S Baghavathi Priya, Madhav Kumar, Nitheesh Prakash JD, et al. 2025. Advanced financial sentiment analysis using FinBERT to explore sentiment dynamics. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 889–897.
[19] V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
[20] Edward Sharkey and Philip Treleaven. 2024. BERT vs GPT for financial engineering. *arXiv preprint arXiv:2405.12990* (2024).
[21] Olamilekan Shobayo, Sidikat Adeyemi-Longe, Olusogo Popoola, and Bayode Ogunleye. 2024. Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A data-driven approach. *Big Data and Cognitive Computing* 8, 11 (2024), 143.
[22] Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology* 73, 9 (2022), 1314–1335.
[23] Zhongtian Sun and Anoushka Harit. 2025. RicciFlowRec: A Geometric Root Cause Recommender Using Ricci Curvature on Financial Graphs. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 1284–1289.
[24] Zhongtian Sun, Anoushka Harit, Alexandra I Cristea, Jingyun Wang, and Pietro Lio. 2023. Money: Ensemble learning for stock price movement prediction via a convolutional network with adversarial hypergraph model. *AI Open* 4 (2023), 165–174.
[25] Zhongtian Sun, Anoushka Harit, Alexandra I Cristea, Jialin Yu, Noura Al Moubayed, and Lei Shi. 2022. Is unimodal bias always bad for visual question answering? a medical domain study with dynamic attention. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 5352–5360.
[26] Zhongtian Sun, Anoushka Harit, Alexandra I Cristea, Jialin Yu, Lei Shi, and Noura Al Moubayed. 2022. Contrastive learning with heterogeneous graph attention networks on short text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.
[27] Zhongtian Sun, Anoushka Harit, Jongmin Yu, Jingyun Wang, and Pietro Liò. 2025. Advanced hypergraph mining for web applications using sphere neural networks. In *Companion Proceedings of the ACM on Web Conference 2025*. 1316–1320.
[28] Graison Jos Thomas. 2024. Enhancing tinybert for financial sentiment analysis using gpt-augmented finbert distillation. *arXiv preprint arXiv:2409.18999* (2024).
[29] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
[30] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[31] Yi-Chi Wu, Ching-Sung Chen, and Yu-Jiun Chan. 2020. The outbreak of COVID-19: An overview. *Journal of the Chinese medical association* 83, 3 (2020), 217–220.