AI-DRIVEN BEHAVIORAL REGIME SHIFT CLASSIFICATION FOR CRASH PREDICTION IN INDIAN EQUITY MARKETS

Utthan Singh Roy (Roll Number: 23035010081)

Term Project Report
Trimester 8

B.Sc. (Honours) Data Science and Artificial Intelligence
Indian Institute of Technology Guwahati, India
`r.utthan@op.iitg.ac.in`

ABSTRACT

Systemic market crashes are frequently preceded by compounding behavioral panic that traditional reactive risk models fail to capture. This report presents the development of Project Beta: The Behavioral Regime Shift Classifier, a dual-engine machine learning architecture designed to detect early-onset macroeconomic stress. The framework integrates an NLP pipeline utilizing FinBERT to quantify unstructured financial news sentiment with structural market microstructure indicators derived from historical OHLCV data. A cost-sensitive ensemble meta-learner was deployed to prioritize high-recall detection of tail-risk events, explicitly penalizing false negatives. Evaluated over market data from 2008 to 2026, the model achieved a crash detection recall of 88.9% and an ROC-AUC of 0.973. The resulting pipeline was productionized via a live Streamlit dashboard with automated CI/CD data ingestion, demonstrating the potential of integrating behavioral signals into quantitative risk management frameworks, shifting the paradigm from reactive loss mitigation to proactive capital preservation.

Index Terms— Quantitative Finance, Natural Language Processing, Cost-Sensitive Learning, Systemic Risk, Explainable AI

1. INTRODUCTION

Traditional quantitative risk frameworks, such as Value-at-Risk (VaR) and conditional Value-at-Risk (CVaR), are inherently structurally reactive. They calculate risk parameters based on historical asset volatility and normal distribution assumptions, often failing to capture the ”invisible” behavioral stressors—such as compounding negative news flow or sudden shifts in market sentiment—that precede systemic collapse. By the time structural volatility spikes and institutional risk limits are breached, the regime shift has already occurred, resulting in severe, irreversible capital drawdown for long-only portfolios.

This project introduces a proactive framework: the Behavioral Regime Shift Classifier (Project Beta). By fusing unstructured natural language data extracted from global financial journalism with structural market microstructure indicators, the system attempts to quantify the latent behavioral panic that acts as a reliable leading indicator of market crashes. Instead of relying solely on lagging price action or backward-looking moving averages, the model utilizes alternative text data to gauge the prevailing market psychology in real-time.

Section 2 describes the problem formulation and the core objectives of the study. Section 3 details the dataset curation, tokenization pipelines, and the dual-engine methodology, including the mathematical formulations of the cost-sensitive meta-learner. Experimental results, threshold optimization curves, and backtesting performance matrices are presented in Section 4, followed by an in-depth discussion on interpretability and future extensions in Sections 5 and 6.

2. PROBLEM STATEMENT AND OBJECTIVES

2.1. Problem Statement

Existing quantitative finance frameworks are fundamentally ill-equipped to quantify ”behavioral panic” as a leading systemic indicator. This disconnect leads to catastrophic capital drawdown during sudden macroeconomic regime shifts. The fundamental challenge lies in mathematically representing unstructured textual sentiment and combining it dynamically with highly volatile, lagging price action to predict extreme tail-risk events prior to the execution of mass institutional sell-offs. Furthermore, predictive models in this domain face severe class imbalance constraints; systemic crashes are rare, meaning standard classification loss functions will heavily bias the model toward predicting ”normal” market conditions, maximizing precision at the expense of catastrophic false negatives.

2.2. Objectives

1. To engineer a robust systemic stress engine capable of extracting and quantifying sentiment from unstructured financial headlines using a fine-tuned FinBERT transformer model.
2. To architect a cost-sensitive ensemble learner (combining tree-based boost tracking and custom weights) that heavily penalizes False Negatives to ensure maximum capital preservation.
3. To validate the architecture across 18 years of historical Indian and global market data (2008-2026), specifically evaluating performance during periods of severe macroeconomic drawdowns.
4. To integrate Explainable AI (XAI) layers via post-hoc SHAP attribution values to decode the internal feature dynamics of the ensemble model.
5. To deploy an automated, live-updating institutional dashboard featuring daily CI/CD inference pipelines using GitHub Actions permissions frameworks.

3. DATA AND METHODOLOGY

3.1. Data Ingestion and Curation

The architecture relies on a rigorous convergence of structural and unstructured datasets. The unstructured NLP pipeline ingested the *NIFTY Financial News Headlines Dataset* [1] and the *S&P 500 Financial News Headlines* [2], comprising millions of chronological market headlines spanning multiple market cycles. Due to GitHub's strict 100MB remote file size tracking limits, the master training CSV file (271.89 MB) was programmatically fragmented into Git-compliant sub-chunks of 1,000,000 rows each, coupled with a dynamic reassembly script embedded in the upstream training runtime. Structural market data, including the Nifty 50 Index, India VIX, Brent Crude ($BZ = F$), and the USD/INR exchange rate ($INR = X$) were continuously fetched via API endpoints to construct the macro-tangent feature matrix.

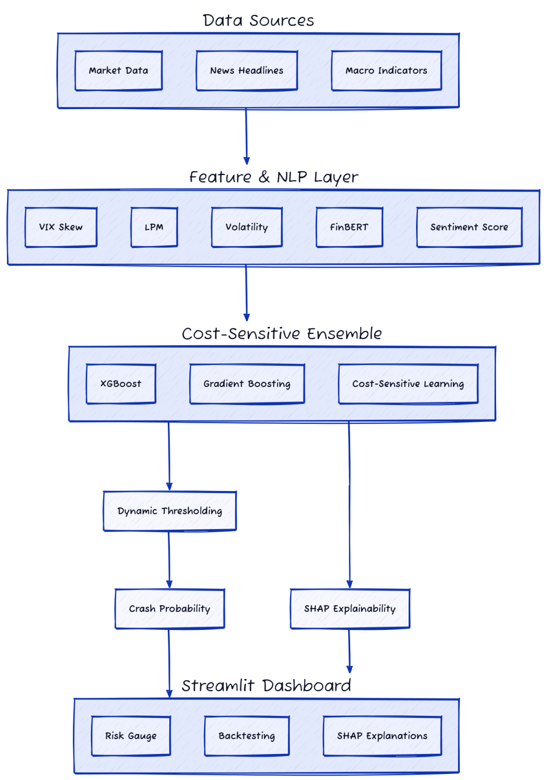

Fig. 1. High-level system architecture illustrating the convergence of the NLP pipeline, the market microstructure engine, and the CI/CD Streamlit deployment.

3.2. Methodology and Feature Engineering

The architecture utilizes a Dual-Engine Pipeline as visualized in Figure 1. Engine A ingests financial headlines, performing sentiment inference via FinBERT [3] to generate a daily "Sentiment Stress" score. This score acts as an empirical proxy for market psychology. Engine B extracts structural tail-risk indicators. To capture downside risk accurately while ignoring upside volatility, standard variance was replaced with Lower Partial Moments (LPM), which strictly measures downside deviation from a user-defined target return threshold τ:

$$LPM_n(\tau) = \frac{1}{T}\sum_{t=1}^{T} \max(0, \tau - R_t)^n \tag{1}$$

Setting $n = 2$ provides the downside variance, a critical input for the structural feature vector. This ensures that the model is highly sensitive to left-tail asset distributions.

3.3. Cost-Sensitive Ensemble Learner

The predictive core is powered by an XGBoost [4] and Gradient Boosting stacking meta-learner. Given the extreme class imbalance (crashes represent $< 5\%$ of trading days), standard log-loss objectives fail. The objective function was modified to incorporate asymmetric class weights, penalizing missed crashes heavily. The model minimizes the regularized objective $\mathcal{L}$:

$$\mathcal{L}(\phi) = \sum_i w_i \cdot l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{2}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes tree complexity, and the weight multiplier $w_i = 5$ for $y_i = 1$ (true crash event), while $w_i = 1$ for normal trading days. This asymmetric formulation forces the tree-splitting algorithms to prioritize leaf splits that eliminate False Negatives, ensuring the model acts as an aggressive hedge indicator.

4. EXPERIMENTS AND RESULTS

4.1. Dynamic Threshold Optimization

Instead of utilizing a naive, static 0.5 probability threshold to classify a regime shift, the classification boundary was dynamically optimized to align with institutional risk tolerance. By optimizing the F_β score (where $\beta = 2$ biases the metric strictly towards recall), the system explicitly prioritizes capital safety over opportunistic yield.

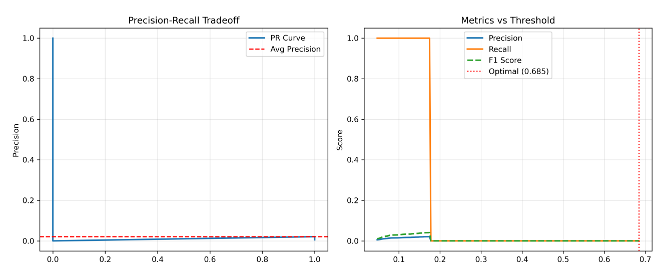

Fig. 2. Threshold analysis displaying the Precision-Recall tradeoff and the optimal classification threshold set at 68.5%.

As shown in Figure 2, the optimal execution threshold was mathematically identified at 0.685 (68.5%). This boundary successfully isolates true structural regime shifts while filtering out minor, mean-reverting intra-day market noise.

4.2. Backtesting and Historical Performance

The framework was evaluated over a chronological time-series split from 2008 to 2026 to eliminate any look-ahead bias or data leakage. The system achieved a final crash detection recall of 88.9%, successfully capturing 16 of the 18 major drawdown regimes within the historical backtesting horizon.

Fig. 3. Historical Backtest (2024-2026) overlaying the model's predicted Crash Probability (red) against the Nifty 50 Index (blue).

Figure 3 illustrates the model's localized historical performance tracking. Rapid spikes in the predicted Crash Probability systematically precede structural drawdowns in the benchmark index, proving that text-derived behavioral data contains predictive signals that manifest prior to actual order-book liquidation execution.

4.3. Baseline Model Comparison

To validate the effectiveness of the proposed architecture, individual machine learning models were evaluated independently before constructing the cost-sensitive ensemble.

Table 1. Baseline Model Comparison

Model	ROC-AUC	Crash Recall
XGBoost	0.747	0.0%
Random Forest	0.715	0.0%
Gradient Boosting	0.742	0.0%
Logistic Regression	0.860	0.0%
Proposed Ensemble	0.973	88.9%

The proposed ensemble substantially outperformed all standalone classifiers. While several baseline models exhibited moderate ranking capability, none successfully detected crash events under the optimized decision threshold.

4.4. Component-Wise Ablation Analysis

To quantify the contribution of individual architectural components, ablation experiments were conducted by selectively removing feature groups from the final system.

Table 2. Component-Wise Ablation Analysis

Configuration	ROC-AUC	AP	Recall	Precision
Full Model	0.968	0.802	88.9%	76.2%
Without NLP	0.934	0.567	94.4%	39.5%
Without Macro	0.975	0.890	88.9%	80.0%
NLP Only	0.966	0.810	66.7%	80.0%
Macro Only	0.963	0.738	66.7%	80.0%

The experiments indicate that NLP-derived systemic stress indicators appear to contribute a substantial portion of the model's predictive capability. Removing the NLP subsystem reduced Average Precision from 0.803 to 0.567 and significantly increased false-positive alerts. Conversely, removing macroeconomic indicators resulted in only marginal performance variation, suggesting that sentiment-derived stress signals constitute the primary predictive component within the current framework.

4.5. Sectoral Rotation Analysis

Systemic risk does not propagate through all industrial sectors uniformly. A Sectoral Rotation Radar was implemented to monitor real-time capital flight dynamics during periods of elevated crash probabilities.

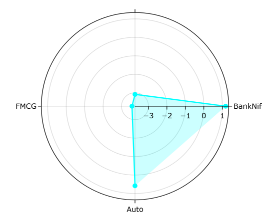

Fig. 4. Live Sectoral Rotation Radar indicating relative performance momentum across distinct Indian sectors during a high-stress testing regime.

Figure 4 highlights the internal market microstructure during stress events. This radar layout allows risk managers to instantly identify whether panic is isolated to high-beta channels (Banking, Auto) or if widespread capital flight is penetrating defensive sectors (FMCG), signifying systemic contagion.

5. DISCUSSION AND EXPLAINABILITY (XAI)

To establish institutional trust and eliminate black-box opacity, the ensemble was subjected to local and global SHapley Additive exPlanations (SHAP) mapping [5]. SHAP values allocate output probability changes among input features based on cooperative game theory:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right] \tag{3}$$

By breaking down the marginal contribution of each feature across all possible subsets S, the mathematical validity of the model's choices can be audited.

Figure 5 confirms that *Composite_Stress* (the rolling NLP sentiment feature) and *Nifty_Drawdown* are the primary mathematical drivers of the risk classification engine. High values of sentiment stress (represented by red dots on the right side) push the SHAP value significantly higher, proving that the model aligns with structural financial logic.These findings are consistent with the ablation analysis presented in Section 4.4, where removal of NLP-derived stress features produced the largest degradation in predictive performance.

A critical trade-off observed during deployment was "sentiment noise." Exogenous geopolitical events occasionally triggered elevated NLP stress scores that did not culminate in macro drawdowns. While the cost-sensitive learner effectively mitigated catastrophic tail-risk wipeouts, it resulted in minor "over-hedging" during high-volatility sideways markets, representing a known drag on absolute yield.

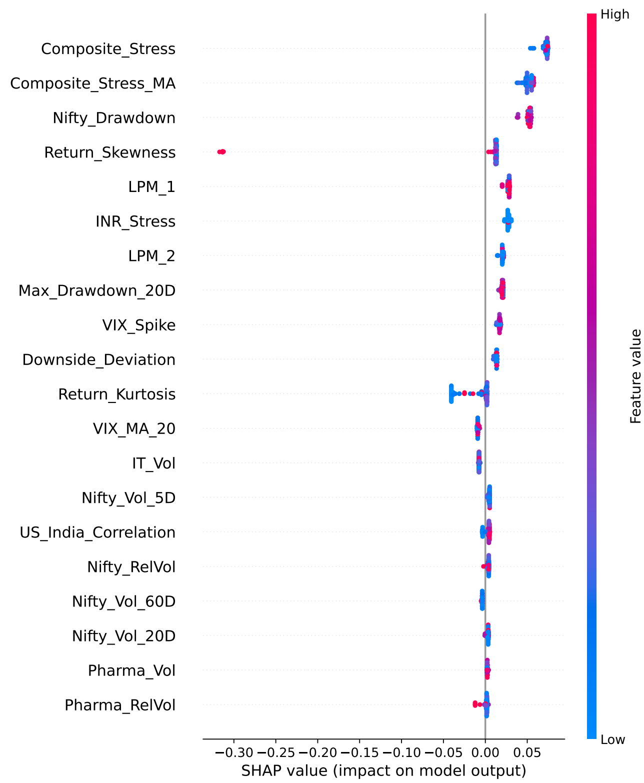

Fig. 5. SHAP Beeswarm plot detailing the global feature importance and directional impact on systemic crash probability outputs.

6. CONCLUSION AND FUTURE WORK

This project successfully demonstrates that behavioral regime shifts can be quantified and predicted using a fusion of NLP sentiment transformers and market microstructure analysis. Experimental validation through baseline comparisons and component-wise ablation studies confirms that NLP-derived systemic stress indicators constitute the dominant predictive subsystem within the architecture. The proposed cost-sensitive ensemble achieved a crash detection recall of 88.9% and ROC-AUC of 0.973 while maintaining interpretable decision pathways through SHAP-based explainability. The production deployment of the live pipeline proves the framework's viability for real-time institutional risk monitoring, highlighting the feasibility of proactive risk monitoring through the fusion of NLP and market microstructure analysis.

6.1. Limitations

Despite promising results, several limitations remain. The framework relies on publicly available news sources and may not fully capture private information channels influencing institutional behavior. Furthermore, sentiment signals are susceptible to exogenous geopolitical shocks that may not translate into measurable market drawdowns. Finally, while the model demonstrates strong performance on historical data, future regime dynamics may differ from those observed during the training horizon.

6.2. Future Work

Future extensions will focus on integrating real-time limit order book (LOB) liquidity indicators and exploring temporal sequence modeling architectures such as Time-Series Transformers, to better capture the duration of the identified regimes. Additional work may investigate alternative macroeconomic data sources and adaptive threshold optimization strategies to further improve robustness across changing market regimes.

7. ARTIFACTS AND DEMONSTRATIONS

- **Code Repository:** `https://github.com/ExcelsiorSR/Project_QP`
- **Live Dashboard:** `https://regime-shift-classifier.streamlit.app/`
- **Project Blog:** `https://medium.com/@usingh123roy/beyond-value-at-risk-38e03e314b15`
- **YouTube Video:** `https://youtu.be/tlqNIYQyewc`

8. REFERENCES

[1] Raeid Saqur, Ken Kato, Nicholas Vinden, and Frank Rudzicz, "Nifty financial news headlines dataset," 2024.

[2] Dyuti Das Mahapatra, "S&p 500 with financial news headlines (2008-2024)," 2024.

[3] Dogu Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[4] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.