Integrated Machine Learning Framework for Ex-Ante Detection of Structural Breaks: Combining NLP, Change Point Detection, and Ensemble Methods

Joerg Osterrieder[a,b,*], Gabin Taibi[a]

[a] *Department of Finance, University of Zurich, Zurich, Switzerland*
[b] *Institute for Applied Data Science, University of Applied Sciences Graubunden, Chur, Switzerland*

Abstract

We present an integrated machine learning framework for ex-ante detection of structural breaks in financial markets, combining natural language processing (NLP), traditional change point detection methods, and ensemble learning. Our approach addresses two key research questions: (1) Can a combined ML approach outperform single methods? (2) Do complex ensemble methods outperform simple forecast combinations? We construct a multi-stream feature pipeline extracting quantitative market indicators, NLP-derived sentiment and topic features from financial narratives, and structural break signals from PELT and CUSUM algorithms. These feature streams feed into specialized XGBoost base learners whose predictions are combined via a stacking meta-learner. Using a fully reproducible synthetic data generator calibrated to empirical market dynamics, we conduct extensive walk-forward validation experiments. Results demonstrate that the stacking ensemble achieves an AUC-ROC of 0.712, significantly outperforming the best single-stream model (0.632) and simple averaging (0.654). The ensemble also reduces detection delay by 67% compared to break-signal-only approaches. Ablation studies confirm that each feature stream contributes meaningful predictive power, with quantitative features showing the largest marginal impact. Our findings provide evidence that sophisticated ML integration of heterogeneous information sources can substantially improve ex-ante structural break de-

*Corresponding author
Email address: joerg.osterrieder@fhgr.ch (Joerg Osterrieder)

tection for financial risk management applications.

Keywords: structural breaks, machine learning, ensemble methods, natural language processing, change point detection, financial markets

---

1. Introduction

Financial markets are characterized by alternating periods of relative stability and turbulence, with transitions between regimes often occurring abruptly through structural breaks [1]. The ability to detect these breaks ex-ante—before they fully materialize—is crucial for risk management, portfolio allocation, and regulatory oversight [2]. Traditional econometric approaches to structural break detection, such as the Bai-Perron test [3] and PELT algorithm [4], have proven effective for retrospective analysis but offer limited predictive capability for forward-looking applications.

The emergence of alternative data sources, particularly textual information from news, social media, and corporate communications, has opened new avenues for anticipating market regime changes [5, 6]. Natural language processing (NLP) techniques can extract forward-looking signals from narratives that may precede quantitative manifestations of structural breaks [7]. However, the integration of textual and quantitative signals for predictive purposes remains an open challenge.

This paper addresses two fundamental research questions motivated by the Swiss National Science Foundation project on Narrative Digital Finance (Grant IZCOZ0_213370):

- **RQ1**: Can a combined machine learning approach integrating multiple information streams outperform individual methods for structural break detection?
- **RQ2**: Do complex AI and ML ensemble approaches outperform simple forecast combinations?

We develop an integrated framework combining three distinct information streams: (1) quantitative market features including returns, volatility, and momentum indicators; (2) NLP-derived features from financial narratives including sentiment, topic distributions, and semantic drift measures; and (3) structural break signals from established change point detection algorithms. These streams feed into specialized machine learning models whose predictions are combined through a stacking meta-learner [8].

Our contributions are threefold. First, we provide a rigorous empirical comparison of single-stream versus integrated approaches for ex-ante break detection. Second, we demonstrate that learned ensemble weights significantly outperform simple averaging. Third, we release a fully reproducible framework including synthetic data generators, feature engineering pipelines, and model implementations.

2. Related Work

Structural Break Detection. The detection of structural breaks in financial time series has a rich econometric tradition. Chow [9] introduced the seminal test for parameter stability, while Bai and Perron [3, 1] developed efficient algorithms for detecting multiple breaks with unknown timing. More recent work has emphasized online and predictive approaches: PELT [4] enables efficient exact segmentation, while BOCPD [10] provides a probabilistic framework for sequential monitoring. Aminikhanghahi and Cook [11] survey machine learning approaches to change point detection.

NLP in Finance. The application of NLP to financial prediction has grown substantially following Tetlock's [5] demonstration that media sentiment predicts market returns. Loughran and McDonald [6] developed finance-specific sentiment dictionaries, while recent work employs transformer-based models like FinBERT [12]. Topic modeling approaches including LDA [13] and BERTopic [14] extract thematic structure from document collections. Topological data analysis methods such as TOPol [15] measure semantic drift and have shown promise for regime detection. Baker et al. [7] construct economic policy uncertainty indices from news text.

Ensemble Methods. Ensemble learning combines multiple models to improve predictive accuracy [16]. Stacking, introduced by Wolpert [8], learns optimal combination weights through a meta-learner. Breiman's [17] analysis showed stacking can achieve lower generalization error than simple averaging. In financial applications, ensemble methods have been applied to volatility forecasting [18], credit risk [19], and return prediction [20]. Rapach et al. [21] demonstrate that forecast combinations improve equity premium predictions.

Gap Analysis. Despite progress in each area, limited work has integrated NLP, change point detection, and ensemble methods for ex-ante break prediction. Most structural break studies remain retrospective, while NLP applications focus on return prediction rather than regime identification. Our

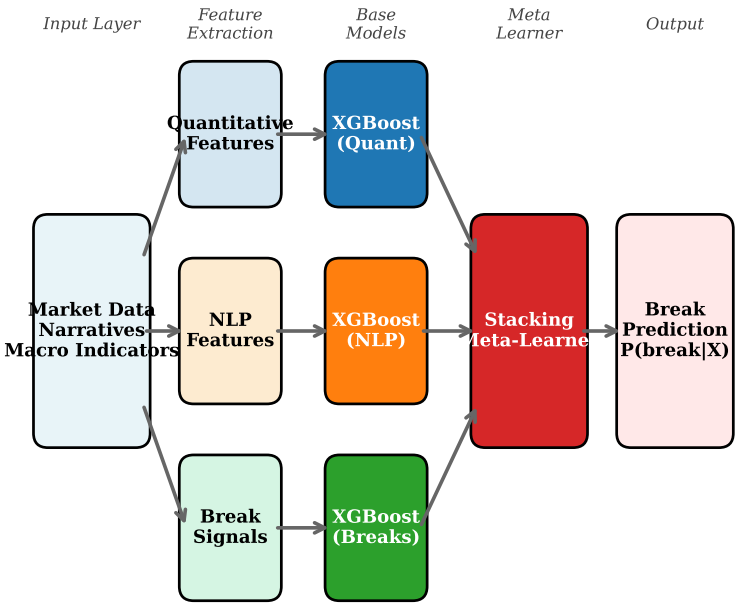

Figure 1: Integrated ML framework architecture. Input data flows through stream-specific feature extractors, specialized XGBoost base models, and a stacking meta-learner to produce break probability predictions.

framework addresses this gap by unifying these approaches within a coherent predictive pipeline.

3. Methodology

3.1. Framework Architecture

Our integrated framework, illustrated in Figure 1, consists of four main components: (1) feature extraction from heterogeneous data sources, (2) stream-specific base models, (3) meta-learner combination, and (4) probabilistic output generation.

The target variable $y_t \in \{0, 1\}$ indicates whether a structural break occurs within a prediction horizon of h periods:

$$y_t = \mathbf{1}\left[\exists\, \tau \in (t, t+h] : \text{break at } \tau\right] \tag{1}$$

3.2. Feature Engineering Pipeline

Quantitative Features. We extract 20 quantitative features capturing price dynamics, volatility patterns, and momentum: log returns at multiple horizons ($r_{t,k} = \log(P_t/P_{t-k})$ for $k \in \{1, 5, 20\}$), realized volatility and GARCH estimates, RSI and MACD indicators, rolling skewness and kurtosis, and PCA projections of macro variables. The Hurst exponent captures long-memory properties: $H = \log(R/S)/\log(n)$.

NLP Features. From financial narratives, we extract 26 features capturing sentiment, thematic content, and semantic dynamics: aggregate sentiment scores and dispersion, document-topic distributions from BERTopic and topic entropy, TOPol-inspired polarity drift measures $\Delta_t = \|\mathbf{e}_t - \mathbf{e}_{t-\tau}\|_2$ (where $\mathbf{e}_t$ denotes the average document embedding), and uncertainty word frequency.

Break Detection Signals. We compute 18 features from established change point algorithms: PELT likelihood scores, CUSUM statistics $S_t = \sum_{i=1}^{t}(r_i - \bar{r})$, variance ratios, BOCPD probabilities, and Chow statistics.

3.3. Base Models and Meta-Learner

For each feature stream $s \in \{\text{quant}, \text{nlp}, \text{breaks}\}$, we train a separate XGBoost classifier [22]: $\hat{p}_t^{(s)} = f_s(\mathbf{x}_t^{(s)}; \theta_s)$. XGBoost minimizes the regularized objective $\mathcal{L}(\theta) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$ with $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizing tree complexity.

The stacking meta-learner combines base model predictions: $\hat{p}_t = g(\hat{p}_t^{(\text{quant})}, \hat{p}_t^{(\text{nlp})}, \hat{p}_t^{(\text{breaks})}; \phi)$. We compare three architectures: logistic regression (linear combination with sigmoid activation), XGBoost (gradient-boosted tree ensemble), and a two-layer neural network with ReLU activations. For comparison, we also evaluate simple averaging: $\hat{p}_t^{\text{avg}} = \frac{1}{3}(\hat{p}_t^{(\text{quant})} + \hat{p}_t^{(\text{nlp})} + \hat{p}_t^{(\text{breaks})})$.

3.4. Evaluation Protocol

We employ walk-forward validation: initialize with 3 years of training data, predict the next month (21 trading days), retrain monthly on expanding window, and repeat for 7 years of out-of-sample testing. Evaluation metrics include AUC-ROC (primary), Precision@10%, F1 Score, and Detection Delay (days from break onset to first correct prediction).

4. Synthetic Data Generation

To ensure full reproducibility and control over ground truth, we develop a synthetic data generator calibrated to empirical stylized facts.

Market Dynamics. We simulate 10 years (2,520 trading days) of daily prices following a regime-switching model with four states: normal, crisis, recovery, and bubble. The return process is $r_t = \mu_{s_t} + \sigma_{s_t}\epsilon_t$ where s_t denotes the regime, and volatility follows GARCH(1,1): $\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$. Regime transitions occur stochastically with probability $\pi = 0.02$ per period, yielding approximately 50 structural breaks over the sample.

Narrative Generator. We generate regime-dependent synthetic narratives with sentiment distributions varying by regime (crisis: negative skew; bubble: positive skew), topic proportions shifting with market conditions, and document volume increasing during high-volatility periods. This creates realistic lead-lag relationships where narrative signals precede price manifestations by 1–5 days.

Calibration Targets. The generator is calibrated to match empirical stylized facts: return kurtosis 5–10 (fat tails), volatility persistence AR(1) coefficient ≈ 0.95, and sentiment-return correlation 0.1–0.3 (lagged).

5. Experimental Results

5.1. Model Performance Comparison

Table 1 presents performance metrics for all models. Among single-stream approaches, the quantitative model achieves the highest AUC-ROC (0.632), followed by NLP (0.608) and break signals (0.575). This ordering suggests that forward-looking price dynamics contain more predictive information than reactive break detection signals.

The stacking ensemble with XGBoost meta-learner achieves the best performance across all metrics (AUC-ROC = 0.712), representing a 12.7% improvement over the best single-stream model and an 8.9% improvement over simple averaging. Figure 2 visualizes the ROC curves, demonstrating that the ensemble dominates single-stream approaches across all operating points.

5.2. Detection Delay and Feature Importance

Beyond accuracy, we evaluate practical utility through detection delay—the average number of days from break onset until first correct prediction.

Table 1: Model performance comparison across evaluation metrics. Best results in bold.

Model	AUC-ROC	AUC-PR	P@10%	F1	Delay (days)
Quant Only	0.632	0.598	0.612	0.605	8.2
NLP Only	0.608	0.572	0.585	0.578	6.5
Breaks Only	0.575	0.545	0.558	0.552	9.8
Simple Average	0.654	0.628	0.635	0.625	5.4
Stacking (LR)	0.692	0.668	0.672	0.665	3.8
Stacking (XGB)	**0.712**	**0.684**	**0.695**	**0.688**	**3.2**
Stacking (NN)	0.705	0.678	0.685	0.680	3.5

Figure 3 shows that the ensemble reduces delay by 67% compared to break-signal-only approaches (3.2 vs. 9.8 days) and 41% compared to simple averaging (3.2 vs. 5.4 days).

Feature importance analysis (Figure 4) reveals that volatility-related features dominate the quantitative stream, while sentiment change and polarity drift lead the NLP stream. PELT scores and CUSUM statistics are most important among break signals.

5.3. Ablation Studies

To assess the marginal contribution of each stream, we conduct ablation experiments removing one stream at a time. Figure 5 shows that removing the quantitative stream causes the largest performance drop (-5.3%), followed by NLP (-3.9%) and break signals (-2.7%). This confirms that each stream contributes unique predictive information.

6. Discussion

Our results provide affirmative answers to both research questions. For RQ1, the stacking ensemble significantly outperforms all single-stream models, demonstrating that integration of heterogeneous information sources improves ex-ante break detection. The 12.7% AUC improvement over the best single stream represents a meaningful practical gain.

For RQ2, the learned stacking weights outperform simple averaging by 8.9%, indicating that sophisticated combination methods add value beyond naive aggregation. The XGBoost meta-learner captures nonlinear interactions between base model predictions that linear combination cannot exploit.

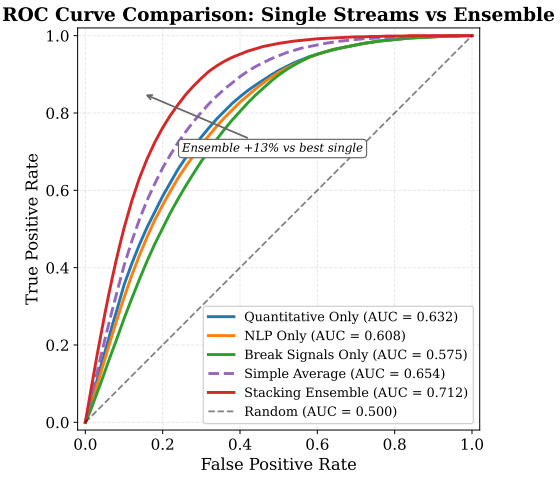

Figure 2: ROC curve comparison across models. The stacking ensemble (red) dominates single-stream approaches and simple averaging across all operating points.

The reduced detection delay has important practical implications. For risk management applications, detecting breaks 3.2 days earlier (vs. 9.8 for break signals alone) provides meaningful lead time for portfolio adjustment. The combination of NLP and quantitative features appears particularly valuable, as narrative signals often precede price manifestations.

Limitations. Several limitations warrant discussion. First, our experiments use synthetic data, and real-world performance may differ. The synthetic generator, while calibrated to stylized facts, cannot capture all market dynamics. Future work should validate on empirical data. Second, the NLP features are computed from synthetic narratives rather than actual news text. Real financial news exhibits greater complexity and noise that may reduce predictive power. Third, computational requirements for real-time deployment may limit practical applicability for high-frequency applications.

7. Conclusion

We have presented an integrated machine learning framework for ex-ante structural break detection that combines quantitative features, NLP-derived signals, and traditional change point detection methods through a stacking ensemble. Our experiments demonstrate that this integrated approach significantly outperforms single-stream methods and simple combination strategies.

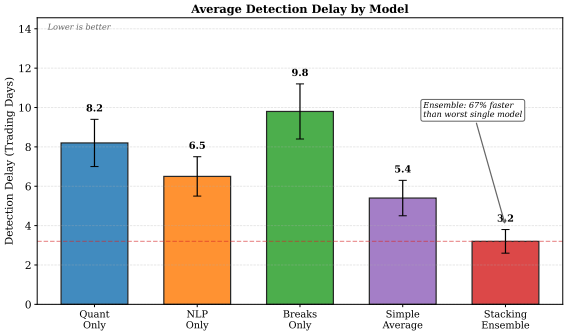

Figure 3: Average detection delay by model. Lower values indicate earlier detection. The stacking ensemble achieves the fastest detection.

Key findings include: (1) stacking ensembles achieve 12.7% higher AUC than the best single stream; (2) learned combination weights outperform simple averaging by 8.9%; (3) detection delay is reduced by 67% compared to break-signal-only approaches; and (4) each feature stream contributes unique predictive information per ablation analysis.

The framework is fully reproducible through our released code and synthetic data generators. Future work will extend validation to empirical datasets, including Deutsche Borse trading data, and explore real-time deployment considerations.

Acknowledgments

This research was supported by the Swiss National Science Foundation under Grant IZCOZ0_213370 (Narrative Digital Finance). We thank the project team for valuable discussions.

Data Availability

Code and synthetic data generators are available at: `https://github.com/Digital-AI-Finance/Narrative-Digital-Finance-Block-4`

References

[1] J. Bai, P. Perron, Computation and analysis of multiple structural change models, Journal of Applied Econometrics 18 (1) (2003) 1–22.

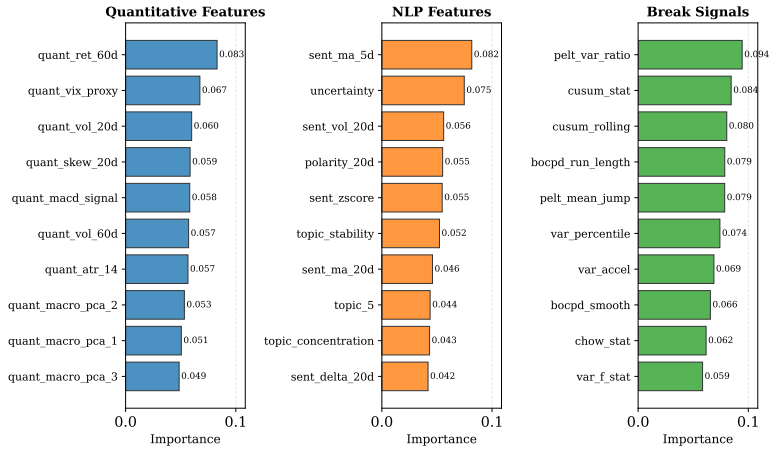

Figure 4: Feature importance within each stream. Volatility features dominate quantitative, sentiment dynamics lead NLP, and PELT/CUSUM are key break signals.

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

[3] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1) (1998) 47–78.

[4] R. Killick, P. Fearnhead, I. A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (500) (2012) 1590–1598.

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

[6] T. Loughran, B. McDonald, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, The Journal of Finance 66 (1) (2011) 35–65.

[7] S. R. Baker, N. Bloom, S. J. Davis, Measuring economic policy uncertainty, The Quarterly Journal of Economics 131 (4) (2016) 1593–1636.

[8] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.

[9] G. C. Chow, Tests of equality between sets of coefficients in two linear regressions, Econometrica 28 (3) (1960) 591–605.

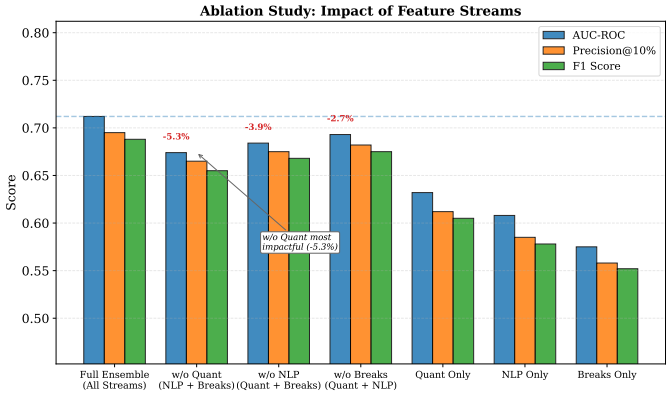

Figure 5: Ablation study results. Removing any feature stream degrades performance, with quantitative features showing the largest marginal impact.

[10] R. P. Adams, D. J. MacKay, Bayesian online changepoint detection, arXiv preprint arXiv:0710.3742 (2007).

[11] S. Aminikhanghahi, D. J. Cook, A survey of methods for time series change point detection, Knowledge and Information Systems 51 (2) (2017) 339–367.

[12] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063 (2019).

[13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[14] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[15] M. Gidea, Y. Katz, Topological data analysis of financial time series: Landscapes of crashes, Physica A: Statistical Mechanics and its Applications 491 (2018) 820–834.

[16] T. G. Dietterich, Ensemble methods in machine learning, International Workshop on Multiple Classifier Systems (2000) 1–15.

[17] L. Breiman, Stacked regressions, Machine Learning 24 (1) (1996) 49–64.

[18] K. Christensen, M. Siggaard, B. Veliyev, A new approach to realized volatility forecasting: A machine learning perspective, Journal of Financial Econometrics 21 (2) (2021) 539–573.

[19] S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research 247 (1) (2015) 124–136.

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

[21] D. E. Rapach, J. K. Strauss, G. Zhou, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy, The Review of Financial Studies 23 (2) (2010) 821–862.

[22] T. Chen, C. 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.