Reviewer 1 Comments

Manuscript: NarrativeBreak: Integrating Structural Break Detection with Multi-Source NLP Signals for Dynamic Portfolio Optimization **Recommendation:** Major Revision **Confidence:** High

Summary

This paper proposes NarrativeBreak, a framework that integrates NLP-based sentiment analysis with Hidden Markov Model regime detection for portfolio optimization. The core contribution—using narrative signals to detect regime changes before they manifest in prices—is novel and potentially valuable.

However, I have significant concerns about the empirical validation and the theoretical grounding of the lead time advantage.

Major Comments

M1. Synthetic Data Calibration

The paper relies entirely on synthetic data, which raises concerns about external validity. While the reproducibility benefits are acknowledged, I need to see:

1. A table comparing the calibration targets to achieved synthetic statistics
2. Evidence that the lead-lag relationship between sentiment and returns is calibrated to published empirical findings, not arbitrary
3. Sensitivity analysis showing results are robust to calibration parameter choices

M2. Lead Time Mechanism

The claimed 5.7-day lead time is central to the paper's contribution. However:

1. Is this lead time consistent across different market regimes (bull/bear)?
2. Could this be an artifact of the synthetic data generation process?
3. What is the standard error of the 5.7-day estimate?

Please provide a deeper theoretical discussion of why sentiment should lead prices, grounded in behavioral finance or information diffusion literature.

M3. HMM Specification

1. Why three regimes? Was this validated via BIC/AIC?
2. The Gaussian emission assumption may not hold for sentiment data, which can be fat-tailed
3. How sensitive are results to HMM hyperparameters, particularly the transition matrix prior?

M4. Ablation Study Interpretation

The finding that removing multi-source integration improves Sharpe by +0.17 is counterintuitive. This needs explanation:

1. Does FinBERT alone outperform the ensemble in all regimes?
2. Is this a result of the confidence calibration down-weighting too aggressively?
3. Why keep multi-source in the framework if single-source is better?

Minor Comments

m1. Notation Consistency

The notation switches between z_t and $z^{(t)}$ for regime. Please unify.

m2. Statistical Tests

The Diebold-Mariano test should use HAC-robust standard errors. Also, 1,000 bootstrap iterations may be insufficient; consider 10,000.

m3. Missing Details

- When is rebalancing triggered exactly? End of day Friday?
- Transaction costs: per trade or per dollar traded?
- MVO lookback period for covariance estimation?

m4. Figure Quality

Some figures are difficult to read when printed in grayscale. Consider using distinct line patterns in addition to colors.

Assessment

The paper addresses an interesting problem and proposes a creative solution. However, the reliance on synthetic data and the unexplained ablation results are significant concerns. With major revisions addressing these issues, the paper could make a valuable contribution.

Note: This is a simulated review for demonstration purposes.