Reviewer 2 Comments

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

---

Summary

This paper introduces NarrativeBreak, which combines NLP sentiment analysis with regime detection for portfolio optimization. The framework is well-designed and the reproducibility infrastructure is commendable. However, the empirical validation is weak due to reliance on synthetic data, and the statistical significance is marginal.

Major Issues

E1. No Real Data Validation

My primary concern is the complete absence of real-data validation. While I understand the reproducibility motivation, several issues arise:

1. The lead time advantage may be "programmed in" to the synthetic data generator
2. Real NLP pipelines have entity recognition errors, sarcasm, and other noise not captured
3. No comparison to published benchmarks on real data

Required: Either (a) validate on public data (GDELT, EDGAR), or (b) provide a formal comparison showing synthetic statistics match published empirical results.

E2. Test Period Concerns

The test period (2020-2023) includes extreme events (COVID crash, 2022 rate hiking cycle). Questions:

1. How does performance vary in calmer periods (e.g., 2017-2019)?
2. Is the outperformance driven by a few extreme events or consistent across time?
3. Would rolling window analysis show stable improvement?

E3. Statistical Significance Marginal

The Ledoit-Wolf p-value of 0.034 is concerning:

1. With multiple strategy comparisons (6 baselines), Bonferroni correction gives p=0.20
2. The 95% CI for Sharpe difference likely includes zero
3. Bootstrap with 1,000 iterations may underestimate variance

Required: Report Bonferroni-corrected p-values and bootstrap CIs for the Sharpe ratio itself (not just the difference).

E4. NLP Method Comparison

The comparison between FinBERT, VADER, and LM Dictionary is valuable, but:

1. What about more recent LLMs (GPT-4, Claude)?
2. The accuracy labels (next-day return sign) are noisy; how does this affect conclusions?
3. How were ensemble weights learned? Grid search? Optimization objective?

Minor Issues

e1. Missing Benchmarks

Consider adding momentum and risk parity baselines.

e2. Portfolio Constraints

Sensitivity to the 50% max position constraint would be helpful.

e3. Transaction Costs

10bp seems low for a weekly strategy. Test with 20-30bp.

e4. Replication Concerns

Code should be made available. Pre-registration would strengthen credibility.

Assessment

The framework is innovative and the reproducibility infrastructure is excellent. However, the lack of real-data validation and marginal statistical significance are significant weaknesses. A strong revision addressing these concerns would make this a publishable paper.

Note: This is a simulated review for demonstration purposes.