# SentState: Incremental Event-Driven Tracking of Financial Market Sentiment with Small Language Models

Arnab Dev, Imtiaz Karim
The University of Texas at Dallas

## Introduction and Methodology

**The Problem**:
Financial markets react to news within seconds. The machines that read that news fall into two camps, each with a structural flaw:

- Memoryless classifiers (FinBERT, FinLlama, FinDPO) score each headline in isolation and forget the unfolding story — fast and cheap, but blind to context.
- Agent-memory LLMs (FinMem, StockMem) remember event history but pay for it in tokens and latency, running only at daily cadence.

No system offers contextual memory at streaming cost.

### Our Idea: What SentState Does

Gives each stock a compact, continuously-updated **sentiment state**: the **memory of an agent** at the **per-event cost of a classifier**

→ **One small-model inference per headline** → **Deterministic math updates the state**

Temporal decay and event reinforcement follow a marked **Hawkes process**

Does an **incremental sentiment state** carry information that **per-headline scoring** does not — beyond what a conventional volatility model and raw news volume already explain?

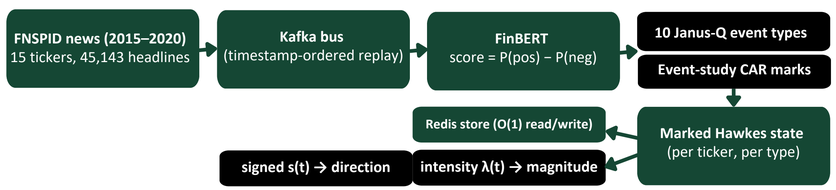



**Fig 1. Deterministic · O(1) per event · p50 28.4 ms end-to-end (CPU).**

**The Methodology:**

- Data: FNSPID headlines replayed through Kafka → Redis (2015–2020; 15 coverage-audited large-caps; 45,143 headlines).
- Scoring: pre-trained FinBERT; continuous score from the classification head, not generated numbers.
- Event marks: Janus-Q's 10 event types; market-grounded weights from an event-study CAR layer, estimated only on strictly-past windows.
- State: marked Hawkes; decay + self-excitation fitted per event type by maximum likelihood.
- Circularity control: 11.1% of headlines are "non-events" (price echoes, market recaps) — they exist because the price already moved, so we drop them.

## Analysis and Findings

### Why the naive metric fails

Correlating sentiment with forward volatility is confounded: realized volatility is highly autocorrelated, so any series tracking recent volatility scores well while knowing nothing about the news. A pure trailing-volatility series (no news) scores 4× the best sentiment arm; even raw headline count beats every sentiment arm. We therefore measure incremental $R^2$ over HAR-RV (Corsi 2009) with Newey–West t-statistics — the variance a conventional volatility model does not already explain.

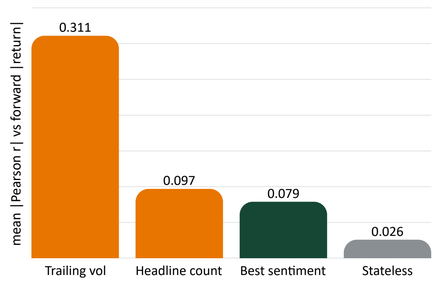
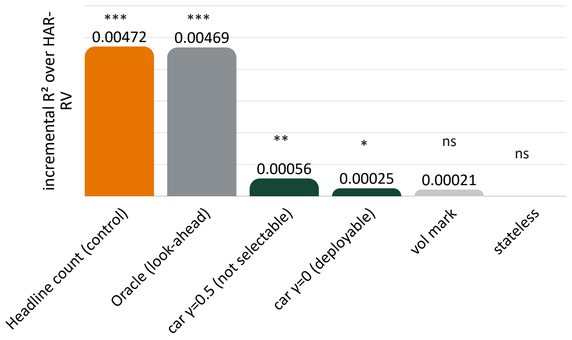



Fig 2. On the raw metric (non-events dropped), no-news baselines (trailing volatility, headline count) beat sentiment — so raw correlation is the wrong yardstick.

### News volume vs. news content




*** p<0.001 · ** p<0.01 · * p<0.05 · n.s. = not significant

Fig 3. Incremental variance over HAR-RV(non-events dropped, n ≈ 20,265) — **news volume beats news content ≈ 8×.**

Once volatility persistence is controlled for, the incremental state is what makes sentiment content statistically detectable: stateless sentiment is insignificant (t = 0.14), while the deployable state arm reaches t = 2.08 (p = 0.038). But the effect is small — news volume beats news content ≈ 8×, and even a look-ahead oracle only ties raw headline count, bounding the entire content channel.

### Event typing is not the bottleneck

We first believed a better event classifier would close the gap to the oracle — and falsified our own claim. Event type explains only ~1% of how large an event turns out to be; the oracle's advantage lives within types (19.5× within-type spread vs 1.90× between). Improving typing 57.5% → 48.3% moved the signal by +0.001. The real lever is per-headline magnitude, not the type label.

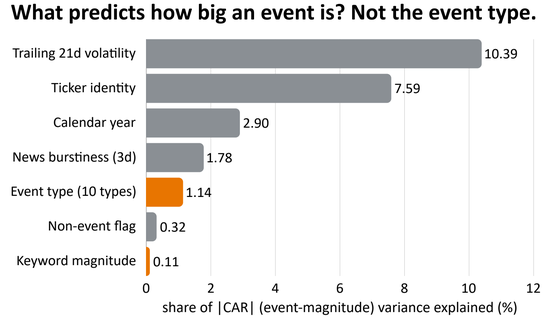



Fig 4. Share of |CAR| (event-magnitude) variance by candidate mark — event type carries ~1%, so typing is not the bottleneck.

## Conclusion & Reference

**What We Found:**

- **Feasible:** deterministic, O(1)-per-event streaming sentiment; **p50 28.4 ms** end-to-end on CPU.
- The incremental state makes sentiment **content statistically detectable** where per-headline scoring does not — a small but real effect.
- **Strongest architecture result:** the intensity channel λ(t) beats the signed level in **5 of 6 disjoint years**.
- **Honest ceiling:** news *volume* beats news *content* ≈ 8×; even a look-ahead oracle only ties headline count.
- **Two earlier claims were falsified by our own controls** — reported here as retractions. That discipline is the result we most stand behind.

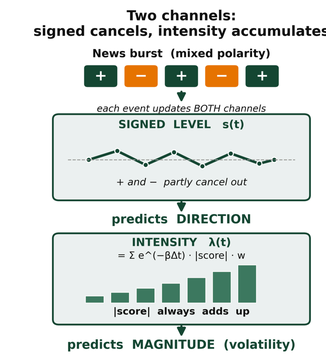



Fig 5. **Signed cancels, intensity accumulates** — one channel per target.

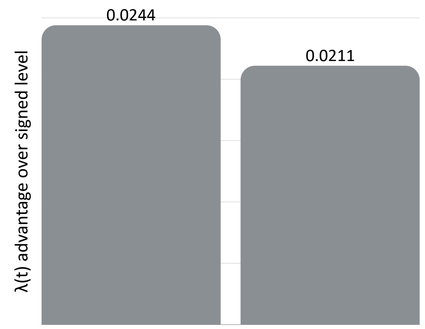



Fig 6. Intensity channel λ(t) − signed level, per split half

**Limitations**

- **News volume dominates content** at daily granularity — the central limitation.
- The look-ahead oracle only *ties* headline count, bounding the entire content channel.
- Pre-trained FinBERT (no fine-tuning); keyword typing leaves 48% untyped (no longer the bottleneck).
- Daily correlation despite intraday state; no transaction-cost backtest.

**Future Work**

- **Per-headline magnitude prediction** (Janus-Q Stage II CAR regression), distilled small to keep the streaming latency claim.
- Learned event classifier; live Alpaca WebSocket feed; context-stuffing baseline; transaction-cost-aware backtest.

### References

Corsi (2009) *J. Financial Econometrics*; Tetlock (2007) *J. Finance*; Newey & West (1987) *Econometrica*; Janus-Q (2026) arXiv:2602.19919; FinDPO (2025); FinLlama (2024); StockMem (2025); FNSPID (2024).

## Contact

**Arnab Dev**
arnab.dev@utdallas.edu