

When Words Move Markets: Sentiment and Semantic Complexity in Central Bank Communication and Equity Volatility

Arush Sachdeva

IIIT Hyderabad

arush.sachdeva@research.iiit.ac.in

Anirban Dasgupta

IIIT Hyderabad

anirban.dasgupta@iiit.ac.in

Abstract

Central bank communications influence market expectations, but their linguistic attributes are inadequately examined in emerging markets. We examine 337 Reserve Bank of India Monetary Policy Committee documents (2010–2025) and relate their textual features to minute-level and daily volatility in NIFTY indices. **Our load-bearing result concerns sentiment.** Within a pre-specified nine-test confirmatory family, negative-sentiment probability is associated with event-window equity volatility and CLS-based sentiment with event-window turbulence in India VIX; both survive Benjamini–Hochberg correction, rank correlation, bootstrap intervals and regime-conditional controls, both strengthen under a time-matched volatility baseline, and both outperform a Loughran–McDonald lexicon on an identical sample. The sentiment model is a fine-tuned FinBERT that we validate directly on human-annotated central-bank text (81.3% accuracy, macro- $F_1 = 0.784$, Cohen’s $\kappa = 0.714$ against a human majority vote). **Semantic complexity, measured with the DAST framework, plays a secondary, construct-level role:** it is largely orthogonal to classical readability ($|r|_{\max} = 0.21$) but behaves as an ordinal measure of generated semantic structure, and at the daily horizon a sentiment $\times$ complexity interaction yields only small exploratory in-sample gains that are not significant under heteroskedasticity-robust inference and do not improve out-of-sample prediction, even after controlling for document length. A third candidate, attention concentration, is Bonferroni-significant under Pearson correlation yet reverses sign under Spearman and has a bootstrap interval spanning zero; we report it as a cautionary methodological result rather than a finding. A pre-specified February 20 Chow test and an endogenous break search near March 11, 2020

indicate pronounced COVID-period regime dependence. Overall, textual features add in-sample explanatory power over a GARCH plus policy-rate baseline but no incremental out-of-sample predictive value on any index.

1 Introduction

Financial markets react not only to policy decisions but to how those decisions are communicated. The clarity, tone, and complexity of central bank disclosures can amplify or dampen market reactions (Hansen and McMahon, 2016). While quantitative policy changes receive extensive scrutiny, the textual characteristics of accompanying statements remain comparatively understudied, particularly in emerging markets where information asymmetries are pronounced, and retail investor participation is high.

The Reserve Bank of India (RBI) releases detailed Monetary Policy Committee (MPC) minutes following each meeting, containing not only rate decisions but also extensive economic assessments, forward guidance, and policy rationale that market participants must parse quickly. When communications are linguistically complex or semantically ambiguous, interpretative difficulty may amplify market uncertainty (Anand et al., 2023), potentially elevating volatility even in the absence of substantive policy surprises. Prior work has established that monetary policy announcements affect stock returns (Ioannidis and Kontonikas, 2008), but focuses predominantly on quantitative variables. Hansen and McMahon (Hansen and McMahon, 2016) showed that Federal Reserve statement complexity affects volatility using traditional readability metrics; Anand et al. (Anand et al., 2023) operationalized semantic complexity via valence shifters; and advances in domain-specific NLP have enabled fine-grained sentiment analysis through models like FinBERT (Araci, 2019). Four gaps remain: (i) limited coverage of emerging markets, where higher

Dataset: https://github.com/GeekArush1/Dataset_Research
Code: https://github.com/GeekArush1/Research_Code

retail participation, thinner liquidity and greater information asymmetry may amplify the effects of communication ambiguity; (ii) no multi-scale analysis separating rapid intraday price discovery from slower multi-day digestion of linguistically dense documents; (iii) unexplored sentiment–complexity interactions; and (iv) unknown stability of text–volatility relationships across regime shifts such as crises. Institutional and socioeconomic factors mediating these gaps in the Indian context are discussed in Appendix C.

This paper addresses these gaps using the DAST framework (Besharati and Izadi, 2019) for semantic complexity and multi-method FinBERT sentiment scoring across 337 RBI MPC documents.

Central claim, stated up front. The two constructs in our title play deliberately asymmetric roles, and we report them on different evidentiary terms throughout. What survives our full inference protocol is a *sentiment*–volatility relationship at the announcement window; semantic complexity enters as a distinct, validated *construct* whose downstream market contribution is exploratory and, out of sample, approximately zero. We therefore organise every empirical claim into three tiers—confirmatory sentiment associations, a cautionary attention non-finding, and exploratory complexity results—and state which tier a result belongs to wherever it is reported (§4.1 gives the summary map).

Where semantic complexity fits. DAST is a measurement contribution here rather than a predictive channel, and we are explicit about this. The index is largely orthogonal to fourteen classical readability indices ($|r|_{\max} = 0.21$; Appendix B.14) and ranks first among fifteen complexity features for three announcement-change targets, so it is not redundant with surface metrics. But as implemented it is close to a monotone transform of generated lattice size (Appendix B.19), it contributes approximately zero out of sample (Appendix B.2), and it enters the daily sentiment $\times$ complexity interaction only as a small exploratory in-sample pattern (§4.4). No headline claim in this paper rests on it. Our contributions, in order of evidentiary strength, are:

(1) *The primary finding.* Two textual associations that survive a pre-specified multiplicity correction *and* rank-based, bootstrap and regime-conditional robustness—negative sentiment with equity volatility through a *polarity channel*, and CLS-based sentiment with implied-volatility turbu-

lence.

(2) *An explicit failure case reported alongside it:* an attention-concentration feature that is Bonferroni-significant under Pearson correlation but reverses sign under rank correlation, recorded because the diagnostic separating the two is not standard practice here.

(3) A consistent separation of in-sample association from out-of-sample prediction, including a GARCH plus policy-rate baseline over which text adds in-sample explanatory power but no out-of-sample gain. We additionally document substantial COVID-period regime dependence that helps explain why pooled analyses yield weak predictability.

(4) A released corpus and end-to-end pipeline for RBI monetary-policy communication, with two independent human validation studies—of the LLM-generated semantic logic underlying DAST (Fleiss’ $\kappa = 0.782$) and of the fine-tuned sentiment model on central-bank sentences (Cohen’s $\kappa = 0.714$ against a human majority vote)—a level of domain validation rarely reported when general-purpose financial sentiment models are transferred to central-bank text; and a multi-scale design separating immediate intraday reactions from multi-day information digestion.

2 Related Work

Financial sentiment analysis has moved from lexicons such as Loughran–McDonald (Loughran and McDonald, 2011) and concept-level resources like SenticNet (Cambria et al., 2010) to transformers: BERT (Devlin et al., 2018) enabled contextual representations and Araci (Araci, 2019) fine-tuned it into FinBERT. Following Rizinski et al. (Rizinski et al., 2024), we fine-tune FinBERT on six domain datasets (Malo et al., 2014; Maia et al., 2018; Sinha and Khandait, 2021; Wilson et al., 2005; Loughran and McDonald, 2011) and adopt multiple scoring methods; we retain the lexicon as an explicit baseline rather than only as training data.

Semantic complexity in financial text has been studied through multiple lenses. Besharati and Izadi (Besharati and Izadi, 2019) introduced the DAST framework, quantifying complexity via lattice-based representations of meaning hierarchies, while Anand et al. (Anand et al., 2023) operationalized complexity through valence shifters, adjectives, adverbs, adversative conjunctions, and negators, that heighten interpretative difficulty in

financial disclosures. We build on both approaches by applying DAST to central bank communications in an emerging market where information asymmetry is more pronounced.

Long financial documents exceed the 512-token limit of standard transformers (Devlin et al., 2018), which Longformer (Beltagy et al., 2020) and LayoutLMv3 (Huang et al., 2022) address; our pipeline uses LlamaParse (LlamaIndex Team, 2024) for extraction. Empirically, monetary-policy announcements move stock returns internationally (Ioannidis and Kontonikas, 2008), with heterogeneous sectoral responses in India (Prabu et al., 2020), and FOMC statement complexity affects volatility under readability metrics (Hansen and McMahon, 2016). For volatility measurement, Whaley (Whaley, 1993) introduced the VIX methodology and Rekabsaz et al. (Rekabsaz et al., 2017) showed that embedding-based text features complement traditional indicators. We relate text to realized and implied volatility at multiple time scales in an emerging market, and—departing from common practice—subject every headline correlation to distribution-free inference before reporting it.

3 Methodology

Our methodology integrates semantic complexity analysis using the DAST (Deciding About Semantic Complexity of a Text) framework with fine-tuned sentiment analysis to comprehensively characterize RBI MPC communications and their relationship to market volatility. The pipeline consists of four main stages: document collection and preprocessing, semantic complexity computation, sentiment analysis, and market data integration. A detailed six-step methodology pipeline is presented in Appendix A.1.

3.1 Data Collection

3.1.1 MPC Document Corpus

We collected 337 RBI monetary-policy communications spanning January 28, 2010 to June 6, 2025 from the official RBI website (<https://rbi.org.in>), across seven document types: Bulletins (131), Macroeconomic and Monetary Developments (54), Monetary Policy Statements (48), MPC Meeting Minutes (42), Governor Press Statements (32), Technical Advisory Committee Minutes (23) and Monetary Policy Reports (7) (Appendix A.1). The corpus therefore spans both policy-decision communications (Monetary Pol-

icy Statements, MPC Minutes and Governor Press Statements; 122 documents) and periodic publications released on a regular calendar (Bulletins, Macroeconomic and Monetary Developments and Monetary Policy Reports; 192 documents). We use “event date” throughout to mean a trading date on which at least one such communication was released, not exclusively a rate-decision date, and we make no claim that every event date carries a policy surprise. Documents were downloaded programmatically and parsed with the LlamaParse API (LlamaIndex Team, 2024).

3.1.2 Market Volatility Data

We use one-minute OHLC data for NIFTY 50, NIFTY BANK and India VIX on communication-release dates and matched non-release days, and daily OHLC data for the same indices over the study period via the yfinance API (Aroussi, 2024).

3.2 Document Preprocessing Pipeline

A comprehensive eight-stage preprocessing pipeline (tokenization, case normalization, stop-word removal, noise reduction, spell correction, POS tagging, lemmatization, stemming) ensures high-quality text for downstream analysis; full details are provided in Appendix A.2.

3.3 Sentiment Analysis

3.3.1 FinBERT Fine-tuning

We fine-tune on six financial sentiment datasets (Financial PhraseBank (Malo et al., 2014), FiQA (Maia et al., 2018), Loughran-McDonald (Loughran and McDonald, 2011), MPQA (Wilson et al., 2005), SEntFiN (Sinha and Khandait, 2021) and Senti-DD), ~30,560 annotated instances, with 80/20 stratified splits and AdamW (Appendix A.8). Because these are general financial text rather than central-bank communication, we validate the fine-tuned model on a human-annotated sample of RBI sentences (Appendix A.7).

3.3.2 Multi-Method Sentiment Scoring

We use three complementary scoring approaches—logit-based, CLS-token-based and attention-weighted—combined by confidence-weighted aggregation. We treat the composite as *exploratory* rather than calibrated; Appendix B.7 shows the downstream findings are insensitive to the weighting, and Appendix A.9 gives formal definitions.

3.4 Semantic Complexity Analysis: DAST Framework

We implement the DAST (Deciding About Semantic Complexity of a Text) framework (Besharati and Izadi, 2019) to quantify document semantic complexity through lattice-based representation of meaning hierarchies.

3.4.1 Semantic Logic Generation

DAST requires semantic logic: principal intuitions, an initial text representation, and derivation rules of the form [concept1, concept2] $\rightarrow$ [derived_concept]. We generate this automatically with Google’s Gemini 2.0 Flash from structured prompts. The LLM receives the output of the preprocessing pipeline of §3.2 rather than the raw document, so the generated logic is derived from lemmatised, stopword-filtered text. SCI variance across repeated generations is low (3–5%; Appendix A.4), and a human validation of the generated logic (three annotators; Fleiss’ $\kappa = 0.782$; 83.3% of rule-sets judged valid; $\approx 91.5\%$ pattern coverage) is reported in Appendix A.6.

3.4.2 Lattice Construction

We construct a directed acyclic graph in which nodes are concepts and edges are derivation relationships, creating new nodes whenever all prerequisite symbols of a rule are present, until no further derivations are possible. Algorithmic details appear in Appendix A.3.

3.4.3 Complexity Calculation

Semantic complexity is computed recursively using the recursive aggregation implemented in our DAST-based lattice pipeline:

$$\text{Complexity}(n) = \begin{cases} 1.0 & \text{if } n \text{ is a leaf node} \\ 1.0 + \prod_{c \in \text{children}(n)} \text{Complexity}(c) & \text{otherwise} \end{cases} \quad (1)$$

Document-level complexity aggregates all node complexities:

$$\text{Complexity}_{\text{doc}} = \sum_{n \in V} \text{Complexity}(n) \quad (2)$$

This is intended to capture hierarchical depth and breadth of semantic relationships, but the generated lattices vary considerably in shape: roughly 30% contain no derivation edges, so Equation 1 returns 1.0 at every node and SCI reduces to the node

count, and across the corpus SCI is close to a monotone transform of lattice size (Spearman $\rho = 0.98$ with node count; Appendix B.19). We therefore treat SCI as an *ordinal* measure of generated semantic structure rather than a calibrated cardinal score, and—consistent with the tiering announced in §1—we do not treat any complexity-based result as confirmatory. It nonetheless remains distinct from surface readability: across fourteen classical indices the maximum absolute correlation with DAST is $|r| = 0.21$ (Appendix B.14).

3.5 Market Volatility Metrics

We employ a comprehensive suite of volatility measures to capture different dimensions of market reactions to MPC announcements, operating at both minute-level and daily frequencies.

3.5.1 Minute-Level Volatility Analysis

For the high-frequency analysis, we use one-minute OHLC data for NIFTY 50, NIFTY BANK, and India VIX and compute a Parkinson high–low volatility measure. For the two equity indices this is realized price volatility. India VIX is itself an implied-volatility index, so the same estimator applied to it measures the intraday *variability of implied volatility* rather than its level; we therefore call the India VIX minute-level target implied-volatility *turbulence* throughout, and reserve “implied volatility” for the daily India VIX level-change target of Appendix A.13. Our primary response variable is the percentage difference between average Parkinson volatility in a standardized 10:00–11:00 AM IST event window and the corresponding non-event-day clock-time baseline. Because publication timestamps are not consistently available, this is a standardized window rather than exact release-time alignment. We also examine the 9:15–10:00 AM pre-window and later intraday windows (Appendix A.12).

3.5.2 Daily Volatility Metrics

At the daily horizon the event-window regressions use annualized realized volatility (RV): the sample standard deviation of daily log returns within each window, scaled by $\sqrt{252}$, over three post-MPC windows (Days 0–1, 2–3, 4–5) plus a pre-event window. We also compute rolling annualized standard deviations over 5-, 10- and 21-day windows and use the India VIX level as a forward-looking implied-volatility measure (Appendix A.13).

3.5.3 Volatility Differential Metrics

To isolate MPC-specific effects we additionally compute a daily volatility spike (post-announcement minus pre-announcement volatility), a volatility-decay measure and a time-to-mean-reversion measure; all three are defined in Appendix A.14.

3.6 Statistical Analysis Framework

Robust inference. The minute-level spike distribution is strongly right-skewed (Table 3), so Pearson p -values alone are inadequate: a few high-leverage event dates can drive a correlation to nominal significance. For *every* test in the confirmatory family we therefore also report Spearman’s ρ , a bias-corrected and accelerated (BCa) bootstrap interval on Pearson r ($B = 10,000$), and the partial correlation after residualising both variables on a post-COVID indicator. We retain an association as a finding only if it clears all four (Appendix A.15). Multiplicity within the nine-test family is handled by Bonferroni and Benjamini–Hochberg adjustment, and regression inference uses HC3 standard errors (Appendix A.20). This four-criterion protocol is the standard against which the *confirmatory* tier of §4.1 is judged; results that do not clear it are labelled exploratory or non-findings wherever they appear.

Event-window regressions. We estimate separate OLS regressions of event-window RV on sentiment and SCI for each post-announcement window (Appendix A.16), and test whether the association between sentiment and the market response varies with complexity via

$$V_t = \beta_0 + \beta_1 S_t + \beta_2 \text{SCI}_t + \beta_3 (S_t \times \text{SCI}_t) + \epsilon_t, \quad (3)$$

assessing β_3 under HC3 standard errors and treating the interaction as exploratory. We also test cross-validated polynomial regressions (degrees 2–3; Appendix A.17).

Other analyses. ACF/PACF diagnostics assess serial dependence in the event-level series and support event-level rather than cumulative specifications (Appendix A.18). Random Forest and Gradient Boosting regressions provide descriptive feature-importance rankings (Appendix A.19); we do not interpret them as evidence of out-of-sample contribution.

3.7 Minute-Level Event Analysis

We compute volatility differentials between MPC and non-MPC days within matched clock-time windows and regress minute-level spikes on sentiment, SCI and attention metrics. Because the baseline in Equation 24 is a full-sample non-event average, we also recompute every spike against a *time-matched* baseline built from the 20 non-event trading days nearest in calendar time to each event, and re-run the confirmatory family on it (Appendix B.17). Full specifications are in Appendix A.21.

4 Results and Discussion

MPC textual signals are detectable only at short horizons and add only incremental in-sample signal over strong market baselines. We treat the pre-specified event-window correlations and COVID break test as *confirmatory*, and interactions, feature-importance rankings, stratification and polynomials as *exploratory*. Confirmatory findings must also survive rank-based, bootstrap and regime-conditional inference; institutional mechanisms are discussed in Appendix C.

4.1 What the Analysis Establishes

Because this section reports many tests, we state first which results are load-bearing and which are not, and where each is developed.

(i) Confirmatory: sentiment and event-window volatility. Two associations clear all four criteria of §3.6: negative-sentiment probability with minute-level equity volatility spikes (NIFTY 50 $r = 0.194$; NIFTY BANK $r = 0.189$) and CLS-based sentiment with India VIX event-window turbulence ($r = -0.224$). Both strengthen under a time-matched baseline and both outperform a Loughran–McDonald lexicon on an identical sample (§4.3.2, §4.6). This is the paper’s central empirical result.

(ii) Exploratory: semantic complexity. SCI is interpreted ordinally throughout. It enters the confirmatory story only as a conditional coefficient in the India VIX joint model, which we flag as exploratory because it was not pre-specified (§4.6); at the daily horizon the sentiment $\times$ complexity interaction produces only small in-sample gains that are non-significant under HC3 inference and negative out of sample (§4.4).

(iii) A cautionary non-finding: attention concentration. The strongest equity signal under Pearson correlation reverses sign under rank correlation, has

a bootstrap interval spanning zero, and disappears under a time-matched baseline. We report it as a methodological caution, not a channel (§4.5).

(iv) **Confirmatory: regime non-stationarity.** A pre-specified Chow test and an endogenous break search indicate substantial COVID-period regime dependence, which helps account for the weak fit of pooled models (§4.7).

(v) **Prediction.** Out of sample, only volatility persistence contributes; textual features—sentiment included—add no incremental out-of-sample R^2 over a GARCH plus policy-rate baseline (§4.8). The confirmatory associations in (i) are therefore statements about incremental event responses, not about forecasting.

4.2 Descriptive Statistics and Correlation Structure

Semantic complexity is extremely right-skewed: the median SCI is 21 and the minimum 5, but because Equation 1 aggregates child complexities multiplicatively, a few deep lattices take values orders of magnitude larger, so all regressions use an outlier-trimmed SCI and all correlations are interpreted ordinally. Sentiment centres near neutrality (mean dynamic score ≈ -0.06), consistent with central-bank communication balancing cautious and optimistic tones. Minute-level spikes are strongly right-skewed: standard deviations exceed 80% for both NIFTY indices, and both have a positive mean but a *negative* median, so the event-window response is heterogeneous rather than uniformly elevated (Appendix B.1).

Correlation structure: Bivariate associations are generally modest (most $|r| < 0.25$), although the confirmatory rank analysis identifies a stronger monotone CLS–India VIX association ($\rho = -0.299$). We therefore test conditional, non-linear, regime-dependent and multivariable relationships; the 2010–2025 sentiment–VIX co-evolution (Appendix, Figure 7) shows episodic, stress-concentrated explanatory power.

4.3 Minute-Level Analysis: Event Window and Rapid Mean Reversion

Minute-level analysis reveals distinct intraday dynamics in which elevated volatility around the MPC announcement window dissipates rapidly, a pattern consistent with rapid price discovery.

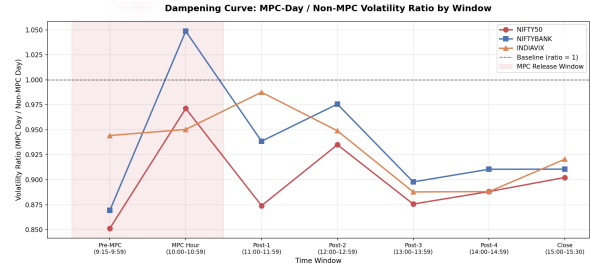


Figure 1: Intraday volatility lifecycle: ratio of mean MPC-day to mean non-MPC-day Parkinson volatility by clock-time window. Volatility rises into the standardized 10:00–11:00 AM window on all three indices and mean-reverts thereafter, but only NIFTY BANK exceeds its matched non-event baseline (ratio > 1). The shaded region is a standardized window, not exact release-time alignment.

4.3.1 Localized Impact and Rapid Mean Reversion

Minute-level volatility rises from the pre-announcement window into the standardized 10:00–11:00 AM event window for all three indices and then mean-reverts within roughly two hours (Figure 1). The elevation is relative rather than absolute: against the matched non-event baseline only NIFTY BANK exceeds unity in the event window (ratio ≈ 1.05), while NIFTY 50 (≈ 0.97) and India VIX (≈ 0.95) remain below it; NIFTY 50 rises sharply from a pre-window level of ≈ 0.85 while India VIX is essentially flat (≈ 0.94 to ≈ 0.95). The event-date distribution is correspondingly skewed: mean spikes are positive for both equity indices (+9.5%, +14.3%) while medians are negative (−8.5%, −10.2%). We therefore do not claim a uniform announcement-day elevation; what the minute-level analysis provides is a temporally localized, mean-reverting response whose *cross-sectional* variation is the object of the analysis below (Appendix B.3).

4.3.2 Primary Finding: Sentiment and Minute-Level Volatility

Table 1 presents the confirmatory family: three textual features (negative-sentiment probability, attention concentration, CLS-based sentiment) against minute-level volatility spikes for two equity indices and India VIX, with Pearson, bootstrap, rank and regime-conditional inference reported side by side. Textual features are averaged at the event-date level when several communications map to the same trading date, giving $N = 184$ event dates.

Five of the nine tests are nominally significant

Table 1: Pre-specified confirmatory family: textual features against standardized 10:00–11:00 AM Parkinson-volatility spikes, event-date aggregated ($N = 184$). BCa intervals use $B = 10,000$ paired resamples; the partial correlation residualises both variables on a post-March-2020 indicator. **Robust** marks tests clearing all four criteria (§3.6). All five raw-significant tests survive Benjamini–Hochberg correction (Table 5); three individual tests, corresponding to two distinct textual channels, survive the full four-criterion protocol.

Textual Feature	Index	Pearson r	p	95% BCa CI	Spearman ρ	ρ p	Partial r (p)	Robust
Prob. Negative	NIFTY 50	0.194	0.0084	[0.063, 0.303]	0.186	0.011	0.151 (0.041)	✓
	NIFTY BANK	0.189	0.0101	[0.065, 0.300]	0.168	0.023	0.151 (0.041)	✓
	India VIX	0.077	0.2972	[−0.016, 0.198]	0.006	0.932	0.087 (0.241)	
Attention Std	NIFTY 50	0.198	0.0071	[−0.082, 0.542]	−0.118	0.111	0.335 (< 0.001)	
	NIFTY BANK	0.218	0.0029	[−0.066, 0.547]	−0.097	0.192	0.340 (< 0.001)	
	India VIX	−0.041	0.5811	[−0.165, 0.200]	−0.030	0.683	−0.060 (0.416)	
CLS Score (Norm)	NIFTY 50	−0.094	0.2063	[−0.188, 0.039]	−0.064	0.389	−0.122 (0.099)	
	NIFTY BANK	−0.071	0.3386	[−0.165, 0.056]	−0.050	0.503	−0.094 (0.204)	
	India VIX	−0.224	0.0022	[−0.357, −0.083]	−0.299	< 0.001	−0.221 (0.003)	✓

and all five survive Benjamini–Hochberg FDR correction ($q = 0.05$; all adjusted $p_{BH} \leq 0.0181$), two of them also under Bonferroni (Appendix B.4). Multiplicity correction controls the family-wise error rate but says nothing about leverage, and the remaining columns of Table 1 split the five sharply. The two that survive everything are both sentiment channels.

Negative sentiment: an equity polarity channel.

Negative-sentiment probability is positively associated with equity volatility spikes but not with India VIX, and it clears every criterion: bootstrap intervals excluding zero, rank correlations agreeing in sign and significance, and survival of the regime control. It is unchanged when the predictable component of sentiment is removed—replacing the level by its AR(1) innovation leaves the correlations at $r = 0.194$ ($p = 0.0082$) and $r = 0.189$ ($p = 0.0102$; Appendix B.10)—and it roughly doubles under the time-matched baseline ($r = 0.292$, $p = 0.0001$ and $r = 0.340$, $p < 0.0001$ on the data-quality-filtered sample of $N = 181$ event dates; Table 15, Appendix B.17). The estimates are flat in the baseline window length ($r = 0.29$ – 0.34 for $k \in \{10, 20, 40, 60\}$; Table 16).

CLS sentiment: implied-volatility turbulence.

CLS-based sentiment is negatively associated with India VIX and with nothing else, and it is the strongest association in the family: its rank correlation exceeds its Pearson correlation in magnitude and the regime-conditional partial correlation is essentially unchanged. The AR(1) innovation behaves identically ($r = -0.208$, $p = 0.0046$). The target is the intraday *variability* of the implied-volatility index, not its level: the same score is uncorrelated with the daily India VIX level change

($r = 0.035$, $p = 0.54$), so the channel means calmer implied-volatility dynamics around the announcement rather than a lower level of implied volatility.

4.4 Semantic Complexity at the Daily Horizon: Digestion Effects (Exploratory)

At the daily horizon we test whether semantic complexity and sentiment combine non-additively over post-event windows (0–1, 2–3, 4–5 days). This is the principal place where complexity carries any market-facing claim, and the result is weak, so we state its status before the numbers: it is an exploratory in-sample pattern only. High-sentiment and high-complexity groups tend to show higher average volatility, though not universally. Adding a sentiment $\times$ complexity interaction to a sentiment+complexity baseline raises in-sample R^2 in all three windows, but the gains are small ($\Delta R^2 = 0.03$, 0.27, 0.81 percentage points for Days 0–1, 2–3, 4–5; Table 8) and shrink further to 0.01, 0.18, 0.78 percentage points once $\log(1 + \text{word count})$ is controlled. After that control, no interaction coefficient is significant under HC3-robust inference (minimum $p = 0.113$), and no target shows a significant out-of-sample gain under permutation testing (minimum $p = 0.289$). Random-forest importance ranks sentiment above complexity and is seed-stable, but is descriptive only, and polynomial expansions improve in-sample fit while degrading out-of-sample performance. We therefore treat the daily interaction as an exploratory in-sample pattern rather than a predictive effect (Appendix B.13). Taken with §4.3.2 the pattern across horizons is consistent: textual signal is detectable at the announcement window

and is not detectable, in or out of sample, at longer horizons.

4.5 A Cautionary Non-Finding: Attention Concentration

Under Pearson correlation, attention standard deviation is the strongest equity signal and the only feature to survive Bonferroni on an equity index, but we do not report it as a finding. Its bootstrap intervals span zero, its rank correlations are negative and non-significant, and the association vanishes under the time-matched baseline (NIFTY 50 $r = 0.086$, $p = 0.25$; NIFTY BANK $r = 0.115$, $p = 0.12$; $N = 181$, Table 15). The large regime-conditional partial correlations (0.335, 0.340) do not overturn this diagnosis: residualising on the post-COVID indicator sharpens the linear fit, while the BCa intervals remain about 0.62 wide ($\approx 2.6\times$ the polarity intervals) and the rank statistic retains the opposite sign. The *measurement* claim does survive: exact-document recomputation reproduces the released statistic ($r = 1.000$), correlates strongly negatively with normalized attention entropy ($r = -0.958$), and all 295 reconstructed inputs contain exactly 512 model-visible tokens. Thus *attention_std* measures concentration rather than dispersion or length; the market association fails, not the construct (Appendix A.11).

4.6 Joint Models and Lexicon Baseline

Joint model. Negative sentiment and attention concentration are moderately correlated at the event-date level ($r = 0.451$, $p < 10^{-9}$), so we estimate the joint specification of Equation 25 ($N = 181$ complete events per index). No equity content feature is individually significant after adjustment: negative sentiment is borderline for NIFTY 50 ($\beta = 0.131$, HC3 $p = 0.049$) and attention concentration is far from significance on both indices ($p = 0.546$ and $p = 0.456$), consistent with the bivariate diagnosis above. For India VIX, CLS sentiment remains strong ($\beta = -0.228$, HC3 $p = 0.0010$) and semantic complexity carries a significant negative conditional coefficient ($\beta = -0.136$, HC3 $p = 0.0009$) that we report as exploratory, since it was not a pre-specified channel (Appendix B.5). This conditional coefficient is the only place where complexity appears alongside a confirmatory result, and we do not promote it beyond the exploratory tier.

Lexicon baseline. A fine-tuned transformer is worth its cost only if it improves on a dictionary.

On an identical sample of 164 event dates, FinBERT negative-sentiment probability attains $r = 0.292$ (NIFTY 50) and $r = 0.326$ (NIFTY BANK), whereas the Loughran–McDonald negative-word share attains $r = -0.011$ ($p = 0.89$) and $r = -0.074$ ($p = 0.34$). The two measures are nearly uncorrelated with each other ($r = 0.090$, $p = 0.25$), so the fine-tuned model is not recovering lexicon counts (Appendix B.18).

4.7 COVID Structural Break: Regime Non-Stationarity

A Quandt–Andrews sup- F search, which locates the break endogenously rather than imposing it, places the maximum at March 11, 2020 for both NIFTY 50 and NIFTY BANK (69 pre-break, 115 post-break event dates). The statistic exceeds the 10% reference critical value for NIFTY 50 (sup- $F = 10.14$) but not the 5% value, and stays below conventional thresholds for NIFTY BANK (7.64); the independent searches nevertheless select the same date on both indices. A Chow test at the pre-specified February 20, 2020 boundary strongly rejects parameter stability for both ($F = 9.91$, $p = 8.3 \times 10^{-5}$; $F = 7.52$, $p = 7.3 \times 10^{-4}$). Together these support substantial COVID-period regime dependence (Appendix B.6).

Level shift versus slope change. A Chow test rejects intercept and slope stability jointly and so cannot say which changed. The distinction matters because the spike in Equation 24 is normalized by a full-sample non-event baseline, so any change in the overall level of market volatility shifts the intercept mechanically. We therefore estimate

$$\text{Spike}_{i,j} = \alpha_0 + \alpha_1 D_i^{\text{post}} + \beta_0 S_i + \beta_1 (S_i \times D_i^{\text{post}}) + \epsilon_{i,j},$$

with D^{post} indicating the post-March-2020 regime, and test $H_0 : \beta_1 = 0$ under HC3-robust inference. Under the full-sample baseline the level shift is large and highly significant ($\hat{\alpha}_1 = 47.6$, $p < 10^{-4}$ on both indices) while the slope change is not separately identified ($p = 0.228$ and $p = 0.256$). Under the time-matched baseline the level shift shrinks by roughly two-thirds ($\hat{\alpha}_1 = 14.7$ and 15.6 , $p \leq 0.0008$) and the slope change becomes clearly significant ($\hat{\beta}_1 = 54.0$, $p = 0.0005$; $\hat{\beta}_1 = 74.1$, $p = 0.0002$); robust Wald tests of joint stability reject in every specification ($F = 18.8$ – 23.7 , $p < 10^{-4}$). We therefore describe the COVID transition as a substantial upward shift in the level of

event-window volatility *together with* a steepening of the sentiment–volatility slope that is identified once the baseline is time-matched (Appendix B.17). Either way, the non-stationarity helps account for the weak fit of pooled models.

4.8 Feature-Group Contribution: Volatility Persistence Dominates

To separate fitted-model importance from genuine predictive contribution we evaluate feature groups with an out-of-sample drop-group test (Appendix B.2). Persistence and lagged-volatility features are the only group with a significant out-of-sample contribution, adding 5.5–33.1 percentage points of ΔR^2_{OOS} ; sentiment, complexity and attention groups contribute approximately zero or negative increments and are never significant. Volatility persistence therefore dominates prediction, and the textual associations of §4.3.2 characterize incremental event responses rather than standalone forecasting power.

Among content-derived features, sentiment probability ranks above semantic complexity for minute-level prediction, consistently across both model types, with no single content feature exceeding 10% importance and cumulative content contribution of 15–25% (Appendix B.11). This ordering is descriptive and in-sample only.

Econometric baseline. Against a GARCH(1,1) one-step-ahead conditional-volatility forecast augmented with the announced repo-rate change (taken from the RBI’s official policy-rate record), text adds significant *in-sample* explanatory power (nested F : India VIX $p = 0.035$, NIFTY BANK $p = 0.005$) but does *not* improve forward-chaining out-of-sample R^2 on any index ($\Delta R^2_{\text{OOS}} < 0$ throughout; paired squared-loss differences non-significant, $p = 0.28$ – 0.73), and the in-sample result is robust to an alternative rate encoding ($p \leq 0.03$ across indices). Text therefore carries associational signal but no incremental out-of-sample value over a simple econometric baseline (Appendix B.16).

5 Conclusion

Across 337 RBI monetary-policy communications (2010–2025) we find statistically detectable short-horizon textual associations and incremental in-sample explanatory value, but no incremental out-of-sample predictive value over market-history baselines.

The primary finding is about sentiment. At minute resolution the event-window response is temporally localized and mean-reverts within about two hours. Two associations survive the full inference protocol—nominal significance, multiplicity correction, a bootstrap interval excluding zero, a sign- and significance-consistent rank correlation, and a post-COVID regime control: negative-sentiment probability with equity volatility spikes, and CLS-based sentiment with India VIX event-window turbulence (Table 1). Both strengthen under a time-matched baseline, and both outperform a Loughran–McDonald lexicon on the same sample.

Semantic complexity and the daily horizon. At the daily horizon the sentiment $\times$ complexity term yields only small in-sample gains that are not significant under HC3 inference and do not improve out-of-sample prediction, even after length control.

An explicit non-finding. A third candidate, attention concentration, is Bonferroni-significant under Pearson correlation but reverses sign under rank correlation; we report it as a non-finding. The construct is well identified—entropy and Gini diagnostics on a fixed 512-token reconstruction confirm it measures concentration rather than dispersion or input length—so the failure is one of association, not measurement. We record it explicitly because Pearson inference on heavy-tailed event-study outcomes is standard practice here and would have licensed the opposite conclusion.

Prediction and regime. The feature-group out-of-sample test attributes significant predictive contribution solely to volatility persistence, and against a GARCH plus policy-rate baseline text adds in-sample explanatory power but negative incremental out-of-sample R^2 on every index. An endogenous break near March 11, 2020 together with pre-specified Chow tests indicates substantial COVID-period regime dependence: a large upward level shift accompanied by a steeper sentiment–volatility slope that is identified once the baseline is time-matched.

Overall. RBI communication carries real but modest, conditional and largely in-sample textual signal—informative for how disclosure is evaluated, yet not a standalone volatility predictor.

Limitations

Data and scope. Our analysis is observational and may be affected by confounding from concurrent macroeconomic events, and it focuses on equity

and implied-volatility indices rather than other asset classes. We discuss in Appendix C how institutional features of Indian markets—high retail participation, algorithmic news parsing, and the RBI’s dual role as regulator and communicator—mediate the text–volatility relationship, and why the effects we document are incremental rather than dominant.

Level of aggregation. Our analysis is conducted predominantly at the index level, so we cannot tell whether the volatility response is broad-based or mechanically driven by a few heavyweight constituents. Three extensions would sharpen the interpretation: (i) a *firm-level* analysis of major individual stocks, especially the heavyweight banks that tend to move the index; (ii) a *sector-level* analysis comparing banking volatility against other sectors rather than relying only on the broad index; and (iii) a *control corpus* drawn from regional or private-bank communications, which would help separate market-structure effects from textual effects and make the interpretation considerably cleaner.

Inference and specification. Our headline correlations are event-level associations on a right-skewed outcome, and we have shown that Pearson inference alone can be misleading in exactly this setting: one member of the pre-specified family is Bonferroni-significant under Pearson correlation and reverses sign under Spearman. We report the full inference protocol for every test rather than only for the tests that survive it, but the same caution applies to any single-statistic result in this literature. The minute-level spike is normalized by a non-event baseline; because a full-sample baseline absorbs the low-frequency level of market volatility, we report both the full-sample and a time-matched ($k = 20$ nearest non-event days) construction throughout, together with sensitivity to k (Appendix B.17). The two constructions agree on which associations survive; they differ in the magnitude of the estimated COVID level shift, and we describe both.

Semantic complexity. SCI as implemented is close to a monotone transform of generated lattice size, and roughly 30% of lattices contain no derivation edges, so the index behaves ordinally rather than cardinally and the “hierarchical depth” interpretation should be applied with care (Appendix B.19). It remains largely orthogonal to classical readability ($|r|_{\max} = 0.21$), so it is not redundant with surface metrics, but it contributes approximately zero out of sample.

Reproducibility. Our full pipeline and dataset are publicly released and the pipeline is end-to-end reproducible, so updated documents and market-data feeds can be substituted to obtain an ongoing *descriptive* assessment of the textual characteristics historically associated with volatility. Given the out-of-sample nulls in Appendices B.2 and B.16, we do not present it as a forecasting system.

Methodological caveats. The daily sentiment $\times$ complexity interaction (Table 8) yields only small exploratory in-sample fit gains and does not enhance out-of-sample prediction (Appendix B.15). Explicitly controlling for $\log(1 + \text{word count})$ preserves the interaction sign across all nine market-response targets but leaves every HC3 interaction coefficient non-significant and produces no significant permutation-tested out-of-sample gain (Appendix B.14.1); we therefore do not interpret the interaction as a robust predictive effect. The confidence-weighted sentiment composite is exploratory: the confidence weights are near-constant and do not track inter-method agreement, so we do not claim they are well-calibrated, though replacing the composite with a simple equal-weight average leaves all downstream associations essentially unchanged (Appendix B.7). Our human validation of the LLM-generated semantic logic is based on 30 documents; while agreement was substantial (Fleiss’ $\kappa = 0.782$) and validity high (83.3%), the estimated 8.5% omitted-pattern rate indicates a modest coverage gap, and larger-scale, recall-oriented evaluation is left to future work. Our RBI-specific validation of the fine-tuned FinBERT (Appendix A.7; 81.3% accuracy, macro- $F_1 = 0.784$, Cohen’s $\kappa = 0.714$ against a human majority vote) confirms adequate transfer to central-bank text but shows residual disagreement concentrated in near-neutral sentiment, so daily-horizon results that lean on fine sentiment gradations should be read with that measurement uncertainty in mind. Propagating the measured SCI generation noise (CV 3–5%) into the downstream interaction model via a $B = 2000$ bootstrap inflates the complexity coefficient’s standard error by only 0.44% (the LLM-induced component is $\approx 9\%$ of the sampling standard error), so our inferences are robust to LLM-generation stochasticity (Appendix A.5).

Composition of the event set. Our 184 event dates are communication-release dates, not exclusively policy-decision dates: 192 of the 337 documents are periodic publications (Bulletins, Macro-

conomic and Monetary Developments, Monetary Policy Reports) rather than rate announcements, and the RBI announced 45 repo-rate changes over the sample period. The event set is therefore diluted relative to a strict announcement study, which is the most likely reason the mean event-window volatility ratio is below unity for NIFTY 50 (≈ 0.97) and India VIX (≈ 0.95) in Figure 1. This affects the interpretation of the *unconditional* event-window level, not the cross-sectional textual associations that are the object of §4.3.2, since those are estimated across event dates rather than against a no-event counterfactual. Restricting the confirmatory family to the policy-decision subset is the natural next check.

Future work. Beyond the firm-, sector-, and control-corpus analyses above, the framework could be extended to bonds, currencies, and commodities; cross-country comparisons would assess generalizability beyond India; real-time tools could operationalize communication-clarity scoring; and quasi-experimental designs could more definitively establish the causal mechanisms underlying the associations we document.

Use of Generative AI

The authors used large language model tools (specifically for grammar correction, proofreading, and improving the readability of the text). All scientific content, analysis, methodology design, and interpretation of results are entirely the authors' own work. The authors take full responsibility and accountability for the content of this paper.

References

- Abhinav Anand, Sankarshan Basu, Xing Huan, and Jalaj Pathak. 2023. [The semantic complexity of financial disclosures](#). *SSRN Electronic Journal*.
- Dogu Araci. 2019. [FinBERT: Financial sentiment analysis with pre-trained language models](#). *arXiv preprint arXiv:1908.10063*.
- Ran Aroussi. 2024. [yfinance: Download market data from yahoo finance](#). Python library, version 0.2.x. Last accessed 2025/02/14.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *arXiv preprint arXiv:2004.05150*.
- Mohammad Reza Besharati and Mohammad Izadi. 2019. [Deciding about semantic complexity of text by DAST model](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, CA.
- Trevor S. Breusch and Adrian R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.
- Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. SenticNet: A publicly available semantic resource for opinion mining. In *Proceedings of AAAI Fall Symposium: Commonsense Knowledge*, pages 14–18.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). *arXiv preprint arXiv:1810.04805*.
- Stephen Hansen and Michael McMahon. 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.
- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. [LayoutLMv3: Pre-training for document ai with unified text and image masking](#). In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4847–4857.
- Christos Ioannidis and Alexandros Kontonikias. 2008. Monetary policy and the stock market: Some international evidence. *Journal of Banking & Finance*, 32(12):2508–2521.
- LlamaIndex Team. 2024. [LlamaParse: A service for parsing complex documents](#). Last accessed 2025/02/14.
- Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.
- A. Edwin Prabu, Indranil Bhattacharyya, and Partha Ray. 2020. Impact of monetary policy on the indian stock market: Does the devil lie in the detail? *Indian Economic Review*, 55(1):1–25.
- Navid Rekabsaz, Mihai Lupu, Allan Hanbury, and Guido Zuccon. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics, pages 1712–1721.

Maryan Rizinski, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov. 2024. **Sentiment analysis in finance: From transformers back to explainable lexicons (XLex)**. *IEEE Access*, 12:7170–7198.

Ankur Sinha and Tanmay Khandait. 2021. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 72(10):1314–1335.

Robert E. Whaley. 1993. Derivatives on market volatility. *The Journal of Derivatives*, 1(1):71–84.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354.

A Supplementary Methodology Details

A.1 Document Corpus and Methodology Overview

Figure 2 presents the detailed six-step methodology pipeline. Table 2 summarizes the distribution of the seven document types comprising the 337-document corpus.

Table 2: Distribution of Document Types

Document Type	Count
Bulletin	131
Macroeconomic and Monetary Developments	54
Monetary Policy Statement	48
Minutes of the Monetary Policy Committee Meeting	42
Monetary Policy Statement, Press Statement by the Governor	32
Minutes of the Meeting of the Technical Advisory Committee	23
Monetary Policy Report	7

A.2 Document Preprocessing Pipeline

A comprehensive eight-stage preprocessing pipeline ensures high-quality text for downstream analysis:

1. **Tokenization:** Word-level tokenization using NLTK (Bird et al., 2009)
2. **Case normalization:** Conversion to lower-case for consistency
3. **Stopword removal:** Filtering common words using NLTK’s English stopwords corpus

Algorithm 1 Semantic Lattice Construction

```

1: Initialize nodes from initial concepts
2:  $queue \leftarrow$  initial nodes
3:  $iterations \leftarrow 0$ 
4: while  $queue \neq \emptyset$  AND  $iterations < max\_iterations$  do
5:   for each rule  $r$  in semantic_rules do
6:     if all left-side symbols of  $r$  present in lattice then
7:       Create new nodes for right-side symbols
8:       Establish parent-child relationships
9:       Add new nodes to  $queue$ 
10:    end if
11:  end for
12:   $iterations \leftarrow iterations + 1$ 
13: end while
14: return Lattice  $G = (V, E)$ 

```

4. **Noise reduction:** Removal of punctuation, special characters, and numerical artifacts
5. **Spell correction:** Using PySpellChecker for typo correction in OCR-extracted text
6. **Part-of-speech tagging:** Identifying grammatical roles using NLTK’s POS tagger
7. **Lemmatization:** Reducing words to base forms using WordNetLemmatizer
8. **Stemming:** Additional root extraction using PorterStemmer for robust matching

A.3 Semantic Lattice Construction Algorithm

Given semantic logic, we construct a directed acyclic graph (DAG) lattice where nodes represent concepts and edges represent derivation relationships. Algorithm 1 outlines the construction process:

Each node is represented by a LatticeNode class containing:

- Unique identifier (UUID)
- Symbol (concept label)
- Parent and child node references
- Applied semantic rule
- Computed complexity value

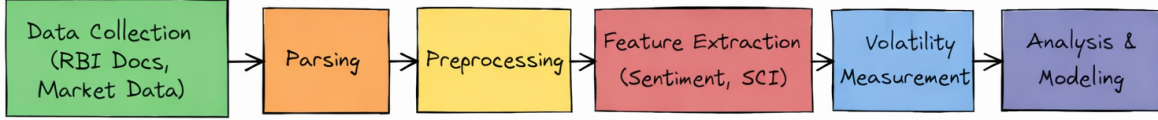


Figure 2: Overview of a detailed methodology pipeline

A.4 SCI Stability Under LLM-Generated Semantic Logic

Since semantic logic generation relies on an LLM (Gemini 2.0 Flash), a natural concern is whether the resulting SCI scores are stable across independent runs. To address this, we conducted a stability experiment on a randomly sampled 10% subset of the corpus (34 documents). For each document, we regenerated the semantic logic (principal intuitions, symbolic representations, and derivation rules) five times using identical prompts but independent LLM calls, then recomputed the full DAST lattice and SCI score for each generation.

Across the 34 documents and 5 runs each (170 total computations), the coefficient of variation of SCI per document ranged from 3% to 5%, indicating that while LLM-generated rules are not perfectly deterministic, the resulting complexity scores exhibit low variance. Rank-order stability was also high: the Spearman rank correlation between any two independent runs exceeded $\rho_S > 0.94$, confirming that relative document ordering by complexity is robust to generation stochasticity. This level of stability is sufficient for the correlation and regression analyses reported in the main paper, where ordinal relationships matter more than exact SCI magnitudes.

We further note that the DAST framework itself has been validated against baselines in prior work: Besharati and Izadi (Besharati and Izadi, 2019) demonstrated that DAST outperforms random baselines with improved precision and competes with alternative complexity methods with lower error percentages, supporting its use as a principled complexity measure for our application.

A.5 Propagating SCI Generation Noise into Downstream Inference

To carry the measured SCI generation noise into downstream inference, we draw $B = 2000$ replicates in which each document’s SCI is multiplied by noise calibrated to the measured coefficient of variation (CV 5%, worst case), and re-estimate the

standardized interaction model on NIFTY BANK immediate-window volatility. The complexity coefficient’s point estimate is $\hat{\beta} = 3.74 \times 10^{-4}$ with a classical standard error of 8.60×10^{-4} ; the additional standard error from LLM stochasticity is 8.08×10^{-5} , i.e. 9.4% of the sampling standard error, raising the combined standard error by 0.44%. Because SCI contributes approximately zero out-of-sample (Appendix B.2), generation noise does not alter the substantive conclusions.

A.6 Human Validation of LLM-Generated Semantic Logic

The reproducibility check above confirms that the Gemini-generated semantic logic is *consistent* across runs, but not that it is *correct*. We therefore conducted a human validation. Three annotators independently reviewed the generated semantic logic for a random sample of 30 documents, judging (i) whether each generated rule-set was valid with respect to the source text (a precision-style criterion) and (ii) which salient patterns present in the text were omitted from the generated logic (a recall-style criterion). Inter-annotator agreement was substantial (Fleiss’ $\kappa = 0.782$). On average 83.3% of generated rule-sets were judged valid (mean per-document validity rate), and annotators flagged an average omitted-pattern rate of 8.5%, corresponding to an estimated coverage of $\approx 91.5\%$ of salient text patterns. These results indicate that the LLM-generated logic is both reliable (high agreement) and largely correct, with a modest, quantified coverage gap, supporting its use as the basis for DAST complexity scoring.

A.7 Human Validation of FinBERT Sentiment Predictions

The fine-tuning corpus consists of general financial text, so we validate the fine-tuned FinBERT directly on central-bank language. Three annotators independently labelled a random sample of 300 sentences drawn from the RBI MPC corpus (30 documents, 10 sentences per document) as pos-

itive, negative, or neutral, blind to the model’s predictions. Inter-annotator agreement was substantial (Fleiss’ $\kappa = 0.761$). Against the majority-vote human reference, the fine-tuned model attains 81.3% accuracy, a macro-averaged F_1 of 0.784, and a model-vs-reference agreement of Cohen’s $\kappa = 0.714$. Most model disagreements occurred between neutral and positive/negative classifications rather than between directly opposing positive and negative classes. These results indicate that the fine-tuned model transfers adequately to RBI communication, while motivating caution around near-neutral sentiment, which we treat as the noisiest region of the scale.

A.8 FinBERT Fine-tuning Hyperparameters

The six fine-tuning datasets comprise:

1. **Financial PhraseBank** (Malo et al., 2014): 4,840 sentences from financial news with expert annotations (positive, negative, neutral)
2. **FiQA** (Maia et al., 2018): Financial opinion mining dataset with 1,173 sentiment-annotated posts
3. **Loughran–McDonald Master Dictionary** (Loughran and McDonald, 2011): 2,355 finance-specific sentiment words with polarity labels
4. **MPQA Subjectivity Lexicon** (Wilson et al., 2005): 8,222 subjectivity clues adapted for financial context
5. **SEntFiN** (Sinha and Khandait, 2021): 10,753 financial news sentences with fine-grained sentiment labels
6. **Senti-DD**: 3,217 economics and finance domain sentences with sentiment annotations

Fine-tuning employed the following configuration:

- Optimizer: AdamW with learning rate 2×10^{-5}
- Batch size: 16 with gradient accumulation over 2 steps
- Epochs: 3 with early stopping (patience=2)
- Loss: Cross-entropy with class weights for imbalanced data

Two of the six sources—the Loughran–McDonald Master Dictionary (2,355 words) and the MPQA Subjectivity Lexicon (8,222 clues)—contribute single-token or short-phrase instances rather than full sentences, accounting for roughly 35% of the $\sim 30,560$ training instances. We include them because they supply finance-specific polarity coverage that the sentence corpora lack, but the resulting mixture is one plausible source of the near-neutral disagreement observed in our human validation (Appendix A.7). We note that this also makes the lexicon comparison of Appendix B.18 conservative: the Loughran–McDonald dictionary is inside the model’s training set, so any advantage the transformer shows over it cannot be attributed to the lexicon being unseen.

A.9 Multi-Method Sentiment Scoring Definitions

To ensure robust sentiment quantification without overfitting to specific aggregation methods, we employ three complementary scoring approaches:

Logit-based Score: Direct use of model output logits for sentiment classes. For a document with K sentences, each receiving logit scores $\mathbf{l}_k = [l_k^{\text{neg}}, l_k^{\text{neu}}, l_k^{\text{pos}}]$, the document score is:

$$S_{\text{logit}} = \frac{1}{K} \sum_{k=1}^K (l_k^{\text{pos}} - l_k^{\text{neg}}) \quad (4)$$

CLS Token Score: Utilizes the final hidden state of the [CLS] token from FinBERT’s last layer as a sentence representation. The CLS embedding $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^{768}$ is projected to sentiment space through the model’s classification head. Document-level aggregation follows Equation 4 structure.

Attention-weighted Score: Extracts attention weights from FinBERT’s multi-head attention mechanism to identify influential tokens. For sentence k with tokens $\{t_1, \dots, t_n\}$ and attention weights $\mathbf{a}_k \in \mathbb{R}^n$ (averaged across heads and layers), the weighted sentiment is:

$$S_{\text{attn}}^k = \sum_{i=1}^n a_{k,i} \cdot s(t_i) \quad (5)$$

where $s(t_i)$ is the token-level sentiment from FinBERT’s embeddings. Document score aggregates weighted sentence sentiments:

$$S_{\text{attn}} = \frac{1}{K} \sum_{k=1}^K S_{\text{attn}}^k \quad (6)$$

Dynamic Aggregated Score: Combines the three methods using confidence-based weighting to avoid arbitrary choices:

$$S_{\text{dynamic}} = \frac{\sum_{m \in M} w_m \cdot S_m}{\sum_{m \in M} w_m} \quad (7)$$

where $M = \{\text{logit}, \text{CLS}, \text{attn}\}$ and weights w_m are derived from prediction confidence (softmax probability of maximum class).

A.10 Attention-Distribution Diagnostic Analysis

We extract and analyze attention weights as model-internal diagnostics of how attention is distributed across tokens. For each document, we compute:

- Mean attention weight: $\mu_{\text{attn}} = \frac{1}{n} \sum_{i=1}^n a_i$
- Attention standard deviation: σ_{attn} indicating focus concentration
- Top-k attended tokens with their sentiment contributions

These quantities characterize the model’s internal attention allocation; we do not interpret attention weights alone as faithful explanations of the classifier’s decisions.

A.11 Attention Concentration vs. Dispersion

The reported `attention_std` (σ_{attn}) is computed from the [CLS] token’s last-layer attention distribution, averaged over heads, for the FinBERT model input. Inputs are truncated to a maximum sequence length of 512 tokens; the attention measures therefore characterize the model-visible input rather than the complete text of longer documents. In the exact-document reconstruction sample, all 295 successfully matched FinBERT inputs contain exactly 512 model-visible tokens. Model-visible token count therefore has zero variance in this sample. A partial correlation controlling for token count is consequently undefined, but, more importantly, variation in `attention_std` within the validation sample cannot be mechanically generated by variation in model-visible sequence length. The attention-concentration analysis therefore compares the shape of attention distributions over fixed-length model inputs rather than comparing sequences of different model-visible lengths. Note also that the two textual pipelines see different inputs: document sentiment aggregates over all K sentences (Equation 4), whereas the

attention statistics come from a single 512-token forward pass. Because every document in the reconstruction sample saturates the 512-token limit, `attention_std` characterises the opening segment of each communication rather than its full text.

Because standard deviation is computed over the attention distribution, larger values indicate greater *concentration*: a small number of tokens receive disproportionately large attention weights. We validate this interpretation using an exact-document recomputation rather than a date-only merge. Reliable document-level matching is obtained for 295 of 335 eligible analysis rows (88.1%); ambiguous dates containing multiple unmatched text files are excluded rather than assigned heuristically. On these 295 matched rows, the recomputed attention standard deviation reproduces the released `attention_std` exactly ($r = 1.000$), and is strongly negatively correlated with length-normalized Shannon entropy ($r = -0.958$, $p < 10^{-150}$).

We then aggregate the correctly matched attention-shape measures to the event-date level and relate them to the standardized minute-level volatility spike. On the resulting matched event subset ($N = 163$), recomputed attention standard deviation is positively associated with both NIFTY 50 ($r = 0.221$, $p = 0.0047$) and NIFTY BANK ($r = 0.242$, $p = 0.0019$), but not India VIX ($r = -0.024$, $p = 0.762$). Normalized entropy exhibits the expected opposite relationship for both equity indices (NIFTY 50: $r = -0.193$, $p = 0.0135$; NIFTY BANK: $r = -0.230$, $p = 0.0031$) and is unrelated to India VIX ($p = 0.709$).

The finite-sample-normalized Gini coefficient is directionally consistent: it is positively associated with NIFTY BANK volatility ($r = 0.168$, $p = 0.0318$), positive but non-significant for NIFTY 50 ($r = 0.117$, $p = 0.137$), and non-significant for India VIX ($r = -0.059$, $p = 0.457$). Taken together, the near-perfect reconstruction of `attention_std`, its strong inverse relationship with entropy, and the opposite-signed equity associations of concentration and entropy confirm that higher `attention_std` represents concentrated rather than dispersed model attention.

We stress that this is a *measurement* result. The market association it was introduced to support does not survive the inference protocol of §3.6: the Pearson equity correlations reverse sign under rank correlation ($\rho = -0.118$ and $\rho = -0.097$), have BCa intervals spanning zero, and vanish under the

time-matched baseline (§4.5, Appendix B.17). The correlations reported above between recomputed concentration, entropy and volatility are Pearson correlations and inherit the same limitation; we retain them because their internal sign consistency is what validates the construct, not because they establish a market effect.

A.12 Minute-Level Volatility Definitions

For the primary intraday analysis, we use the Parkinson high–low volatility estimator computed from one-minute OHLC observations. For minute m ,

$$PV_m = \sqrt{\frac{[\log(H_m/L_m)]^2}{4 \log 2}}, \quad (8)$$

where H_m and L_m denote the minute high and low prices, respectively.

For each trading date d and clock-time window w , we average the minute-level Parkinson values:

$$PV_{d,w} = \frac{1}{|w|} \sum_{m \in w} PV_m. \quad (9)$$

Because exact publication timestamps are not consistently available for the full corpus, the intraday analysis uses standardized clock-time windows rather than claiming exact release-time alignment for every communication:

- Pre-window baseline: 9:15–10:00 AM
- Standardized event window: 10:00–11:00 AM
- Short-term window: 11:00 AM–1:00 PM
- Intraday stabilization: 1:00–3:30 PM

The headline text–volatility correlations use the standardized 10:00–11:00 AM event-window spike relative to the same clock-time interval on non-event trading days. Additional intraday profiles are used to characterize the subsequent normalization of volatility.

A.13 Daily Volatility Metric Definitions

Event-Window Annualized Realized Volatility (RV): For the daily event-window regressions, we compute the sample standard deviation of daily log returns within each post-announcement window and annualize it:

$$RV_i^{[a,b]} = s(\{r_{i,d} : d \in [a,b]\}) \sqrt{252}, \quad (10)$$

where $r_{i,d} = \log(P_{i,d}/P_{i,d-1})$ is the daily log return for event i , and $[a,b] \in \{[0,1], [2,3], [4,5]\}$. These event-window RV measures are distinct from the longer-horizon rolling ARSD measure defined below.

Event-Window Annualized Return (AR): For auxiliary robustness targets, we also compute the annualized event-window return

$$AR_i^{[a,b]} = \exp\left(\sum_{d=a}^b r_{i,d} \frac{252}{n_{i,[a,b]}}\right) - 1, \quad (11)$$

where $n_{i,[a,b]}$ is the number of trading-day returns observed in the event window. AR is a return measure rather than a volatility measure and is used only as an auxiliary robustness target.

Annualized Rolling Standard Deviation (ARSD): Historical volatility over rolling windows:

$$ARSD_t = \sigma_{\text{daily}} \times \sqrt{252} \quad (12)$$

where σ_{daily} is computed over windows of 5, 10, and 21 trading days:

$$\sigma_{\text{daily}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (R_{t-i} - \bar{R})^2} \quad (13)$$

with $R_t = \log(P_t/P_{t-1}) \times 100$ as the daily log-return.

India VIX: Forward-looking implied volatility from NSE options market, representing market expectations of 30-day volatility.

Event Windows: We define three post-MPC event windows for analysis:

- Window 1: Days 0–1 (immediate impact)
- Window 2: Days 2–3 (short-term adjustment)
- Window 3: Days 4–5 (stabilization period)

Additionally, a pre-event window (Days -5 to -1) establishes baseline volatility levels.

A.14 Volatility Differential Metrics

In addition to the volatility spike metric reported in the main paper, we define two further differential measures used in exploratory analysis:

Volatility Decay: Rate of volatility normalization:

$$\text{Decay}_t = V_t^{[0,1]} - V_t^{[4,5]} \quad (14)$$

measuring how quickly markets stabilize after the announcement.

Time to Mean Reversion: Number of trading days for volatility to return within one standard deviation of its 60-day pre-announcement mean:

$$T_{\text{reversion}} = \min\{t : |\text{ARSD}_t - \mu_{\text{pre}}| \leq \sigma_{\text{pre}}\} \quad (15)$$

A.15 Robust Inference for the Confirmatory Family

Every test in the nine-member confirmatory family is reported under four inference procedures.

Pearson correlation, as in prior work, with the caveat that its p -value assumes bivariate normality and is sensitive to leverage.

Spearman rank correlation:

$$\rho_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (16)$$

where d_i is the difference between ranks of paired observations. A discrepancy in sign between r and ρ_S indicates that the linear association is carried by a small number of extreme observations.

BCa bootstrap: bias-corrected and accelerated 95% confidence intervals on Pearson r from $B = 10,000$ paired resamples of the event dates.

Regime-conditional partial correlation: both variables are residualised on a post-March-2020 indicator before correlating, which removes any association induced purely by the common pre/post level shift documented in §4.7.

A.16 Event-Based Window Regression Specifications

We estimate separate regression models for each post-announcement window to identify when document characteristics have maximal impact:

$$\text{RV}_i^{[0,1]} = \beta_0^{(1)} + \beta_1^{(1)} S_i + \beta_2^{(1)} \text{SCI}_i + \epsilon_i^{(1)} \quad (17)$$

$$\text{RV}_i^{[2,3]} = \beta_0^{(2)} + \beta_1^{(2)} S_i + \beta_2^{(2)} \text{SCI}_i + \epsilon_i^{(2)} \quad (18)$$

$$\text{RV}_i^{[4,5]} = \beta_0^{(3)} + \beta_1^{(3)} S_i + \beta_2^{(3)} \text{SCI}_i + \epsilon_i^{(3)} \quad (19)$$

Comparing coefficient magnitudes and significance across windows reveals the temporal dynamics of impact.

A.17 Polynomial Regression

To capture non-linear relationships:

$$V_t = \beta_0 + \sum_{j=1}^p \beta_j S_t^j + \sum_{k=1}^q \gamma_k \text{SCI}_t^k + \epsilon_t \quad (20)$$

with polynomial degrees $p, q \in \{2, 3\}$ selected via cross-validation.

A.18 Time Series Analyses

Autocorrelation Functions: We compute ACF and PACF for post-MPC volatility series:

$$\text{ACF}(k) = \frac{\sum_{t=1}^{T-k} (V_t - \bar{V})(V_{t+k} - \bar{V})}{\sum_{t=1}^T (V_t - \bar{V})^2} \quad (21)$$

to assess persistence of volatility shocks.

Figure 3 presents the ACF and PACF for the India VIX differential volatility series (post-announcement VIX change relative to pre-announcement baseline).

The rapid decline in autocorrelation beyond the first lag indicates that the India VIX differential exhibits little serial persistence across successive event observations. This reduces concern that the corresponding event-level regressions are being driven by long-lived persistence in this dependent variable. We treat this as supporting evidence for the event-level specification rather than as a formal proof that all communication events or all market outcomes are statistically independent.

A.19 Machine Learning Models

To capture complex non-linear relationships and identify important predictors:

Random Forest Regression: Ensemble of decision trees providing feature importance scores:

$$\hat{V} = \frac{1}{B} \sum_{b=1}^B f_b(\mathbf{X}) \quad (22)$$

where $\mathbf{X} = [S, \text{SCI}, \text{attention_mean}, \text{attention_std}, \dots]$ and B is the number of trees.

Gradient Boosting: Sequential ensemble optimizing:

$$\hat{V}_m = \hat{V}_{m-1} + \eta \cdot h_m(\mathbf{X}) \quad (23)$$

where h_m is the m -th weak learner and η is the learning rate. Feature importance from these models provides descriptive rankings indicating which variables receive greater importance within

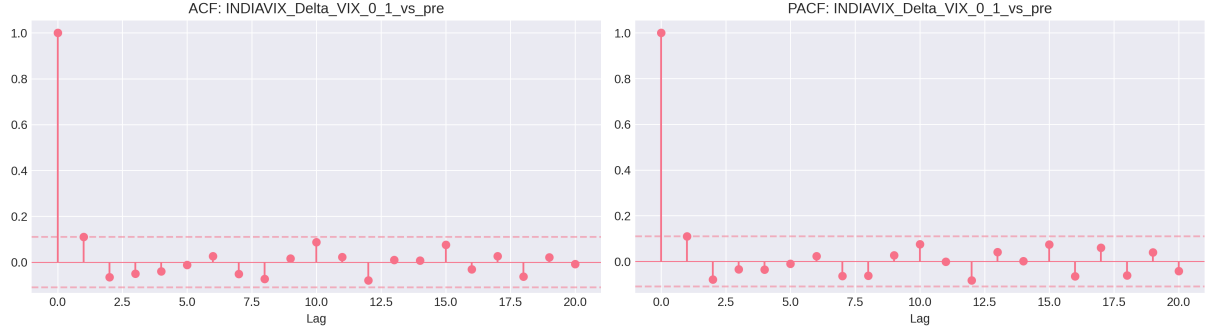


Figure 3: Autocorrelation function (left) and partial autocorrelation function (right) for the India VIX post-event differential series. Dashed red lines indicate 95% confidence bounds. Serial dependence is concentrated at the first lag, with subsequent lags remaining within the confidence bands. The result shows little persistent dependence across successive event observations and supports the use of event-level rather than cumulative specifications for this outcome.

the fitted ensemble models; it does not establish genuine out-of-sample predictive contribution (see the feature-group OOS test, Appendix B.2).

A.20 Robustness Checks

We use a two-sided significance level of $\alpha = 0.05$ throughout. For the nine-test primary minute-level family, we report both Benjamini–Hochberg false-discovery-rate and Bonferroni adjustments (Appendix B.4). Regression inference uses heteroskedasticity-robust standard errors, with HC3 reported for the daily interaction models. BCa bootstrap confidence intervals (Appendix A.15) and Breusch–Pagan diagnostics (Breusch and Pagan, 1979) are used in the corresponding robustness analyses where applicable.

Figure 4 presents Breusch–Pagan test results across all volatility targets.

The single Breusch–Pagan rejection occurs for the NIFTY BANK annualized-return target (Days 0–1). We therefore use heteroskedasticity-robust standard errors for regression inference rather than relying on homoskedastic OLS standard errors.

A.21 Minute-Level Event Analysis Specifications

To capture immediate market reactions with greater temporal precision, we conduct minute-level volatility analysis on MPC announcement days.

A.21.1 Data Alignment

Event-specific publication timestamps are not consistently available across the full sample. We therefore use a standardized 10:00–11:00 AM IST clock-time event window for the minute-level analysis,

with 9:15–10:00 AM representing the preceding market window. The standardized event window should not be interpreted as exact release-time alignment for every communication.

Multiple communications on the same event date. Because multiple RBI communications can map to the same trading date while the market response is common to that date, the primary minute-level analysis is conducted at the event-date level. For dates with multiple associated communications, each textual feature is averaged across the documents on that date before being matched to the common market response. This avoids arbitrarily selecting one communication and yields one textual predictor vector per event date. The resulting primary minute-level sample contains $N = 184$ event dates.

A.21.2 Volatility Spike Identification

For each event date d , index j , and standardized clock-time window w , we define the Parkinson-volatility spike relative to the corresponding non-event baseline as

$$\text{Spike}_{d,j,w}(\%) = 100 \times \frac{PV_{d,j,w} - \overline{PV}_{\text{non-event},j,w}}{\overline{PV}_{\text{non-event},j,w}}, \quad (24)$$

where $\overline{PV}_{\text{non-event},j,w}$ is mean Parkinson volatility for index j in the same clock-time window on non-event trading days. Positive values therefore indicate volatility above the matched clock-time baseline.

A.21.3 Multivariate Content Volatility Model

To assess whether the primary textual associations remain when sentiment, attention concentration,

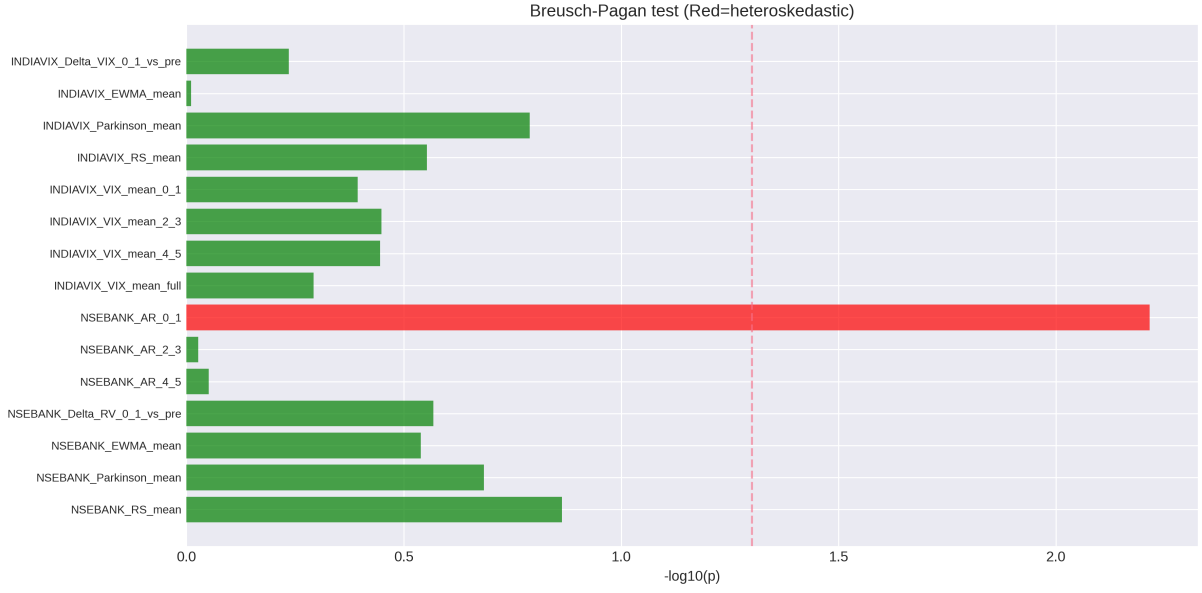


Figure 4: Breusch–Pagan heteroskedasticity test results for sentiment market response regressions across the tested targets. Bars show $-\log_{10}(p)$; the dashed pink line marks the $\alpha = 0.05$ significance threshold. Green bars indicate homoskedastic residuals (no correction needed); the single red bar (NSEBANK AR 0–1) exceeds the threshold, indicating heteroskedastic residuals for which heteroskedasticity-robust standard errors are used for regression inference. All India VIX and broader NIFTY targets pass the homoskedasticity assumption.

and semantic complexity are considered jointly, we estimate the following event-date-level model for each equity index:

$$\text{Spike}_{i,j} = \gamma_0 + \gamma_1 P_i^{\text{neg}} + \gamma_2 A_i + \gamma_3 \text{SCI}_i + \epsilon_{i,j}, \quad (25)$$

where $\text{Spike}_{i,j}$ is the standardized-window Parkinson-volatility spike for event date i and equity index j , P_i^{neg} is the FinBERT negative-sentiment probability, A_i is attention concentration measured by attention_std, and SCI_i is DAST semantic complexity. When multiple communications occur on the same trading date, all textual predictors are averaged at the event-date level before estimation. Predictors are standardized for coefficient comparability, and inference uses HC3 heteroskedasticity-robust standard errors.

For India VIX, where the primary bivariate sentiment association is obtained with the CLS sentiment score, we estimate the analogous specification replacing P_i^{neg} with S_i^{CLS} . The joint models are treated as conditional robustness checks rather than as an expansion of the pre-specified nine-test confirmatory correlation family.

B Supplementary Results

B.1 Descriptive Statistics of Variables

Table 3 presents summary statistics for selected document-level textual features and market-response measures used in the analysis.

For compactness, the daily realized-volatility rows report, for each event, the mean of the Days 0–1, 2–3, and 4–5 annualized RV measures; the window-specific RV variables are used separately in the regression analyses. The NIFTY 50 and NIFTY BANK summaries contain 326 and 304 complete events, respectively.

The wide range of semantic complexity values reflects the diversity of MPC document types in our corpus (Table 2), from brief press statements to comprehensive quarterly bulletins, and the multiplicative aggregation in Equation 1, which maps a modest range of lattice depths onto an extremely wide numerical range. Sentiment measures cluster near neutrality with occasional extreme values during crisis periods. Minute-level volatility spikes exhibit extreme outliers (max values $> 800\%$) corresponding to specific high-impact announcements, and correspondingly negative medians (-8.45% for NIFTY 50, -10.19% for NIFTY BANK, -12.21% for India VIX), with 42.9%, 42.4% and 39.1% of event dates respectively above the matched non-event baseline. This skew is the

Table 3: Descriptive Statistics of Document and Market Variables

Variable	Mean	Std Dev	Min	Max
Document Characteristics				
Semantic Complexity (SCI) [†]	3.98×10^{20}	7.24×10^{21}	5	1.32×10^{23}
Sentiment (Dynamic Aggregated)	-0.056	0.259	-0.773	0.806
Sentiment (Logits)	0.453	0.276	-0.253	0.912
Probability Negative	0.165	0.272	0	0.970
Attention Std	0.00367	0.00107	0	0.00744
Attention Mean [‡]	0.00194	0.00015	0	0.00195
Daily Event-Window Realized Volatility				
NIFTY 50 mean event-window RV (%)	11.26	8.81	0.56	88.18
NIFTY BANK mean event-window RV (%)	16.69	12.31	0.49	113.16
India VIX Change (%)	-0.053	1.51	-5.51	10.67
Minute-Level Spike (%)				
NIFTY 50 (10:00–11:00 AM)	9.45	81.97	-100.00	867.61
NIFTY BANK (10:00–11:00 AM)	14.30	93.76	-100.00	992.77
India VIX (10:00–11:00 AM)	-2.60	41.39	-100.00	182.15

[†]SCI is reported on its raw scale. Because Equation 1 aggregates child complexities multiplicatively, the raw index is extremely heavy-tailed (median = 21, interquartile range 10–38); all regression analyses use an outlier-trimmed SCI variable and correlation analyses are interpreted ordinally. [‡]Mean attention is mechanically $1/512$ for a fully truncated input (99.4% of documents lie within 0.1% of $1/512$); it is reported for transparency and excluded from all model feature sets.

reason the confirmatory family is reported under rank and bootstrap inference as well as Pearson correlation (§3.6). Three rows of the released feature table duplicate an existing (date, document) pair exactly and are collapsed before analysis, leaving 334 unique documents; semantic complexity is available for 332 of them.

Document Type and Volatility: We tested whether document type (Table 2) is associated with volatility outcomes by comparing market-response distributions across the seven document categories using Kruskal–Wallis tests. No statistically significant category-level differences emerge (all $p > 0.15$). Thus, the observed market responses are not readily reducible to coarse document type alone. This result supports analyzing continuous textual characteristics such as sentiment, semantic complexity, and attention concentration rather than treating document category as the primary explanatory dimension.

B.2 Feature-Group Out-of-Sample Contribution

Because impurity/permutation importance is descriptive and can over-credit features that do not generalize, we directly test each feature group’s out-of-sample contribution. For each target we fit a ridge model on all four feature groups (persistence/lagged volatility, sentiment, complexity, attention) under repeated cross-validation, then drop each group and measure the loss in out-of-sample

R^2 ; significance is assessed against a permutation null in which the group is replaced by a same-size shuffled (noise) block. Table 4 reports the results. Only lagged-volatility (persistence) features contribute significantly out of sample (5.5–33.1 percentage points, permutation $p \approx 0.005$ across targets); sentiment, complexity, and attention contribute approximately zero or slightly negative amounts and are non-significant ($p > 0.1$) at every horizon. This supports the substantive claim that volatility persistence dominates, and clarifies that the textual signals documented elsewhere in the paper are in-sample/associational rather than out-of-sample predictive.

B.3 Intraday Volatility Profile: MPC vs. Non-MPC Days

Figure 6 presents minute-level Parkinson volatility profiles on MPC announcement days versus non-MPC trading days for all three indices, providing the raw magnitude comparison underlying the mean-reversion dynamics reported in §4.3.1.

Three cross-index patterns emerge. First, the banking index exhibits the largest absolute MPC-day elevation, consistent with the bank lending channel through which monetary policy most directly transmits (Prabu et al., 2020). Second, the convergence of MPC-day and non-MPC-day profiles by midday across equity indices corroborates the 2-hour mean-reversion window reported in Figure 1. Third, the near-identical MPC and non-

Table 4: Out-of-sample ΔR^2 (%) contributed by each feature group (drop-group test). * denotes permutation $p < 0.05$. Only persistence is significant; textual groups are not.

Target	Persistence	Sentiment	Complexity	Attention
NIFTY BANK RV 0–1	8.2*	−3.4	−0.5	−1.6
NIFTY BANK RV 2–3	33.1*	−1.2	0.4	−1.3
NIFTY BANK RV 4–5	20.9*	−2.0	−0.3	−1.5
NIFTY BANK ΔRV 0–1	5.5*	−3.2	−0.1	−2.0
NIFTY 50 RV 0–1	13.1*	−2.3	−0.6	−1.6
NIFTY 50 RV 2–3	26.2*	−1.4	0.0	−1.6
NIFTY 50 RV 4–5	21.3*	−2.0	−0.4	−1.5
India VIX 0–1	27.5*	−2.1	0.2	−0.5

NIFTY BANK AR 0–1 omitted: the full model has negative OOS R^2 (no fit).

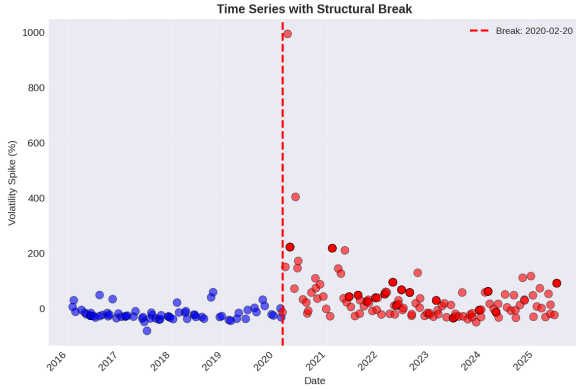


Figure 5: Structural break in sentiment–volatility relationship identified via Chow test. Pre-COVID regime (blue): near-flat relationship between negative sentiment probability and NIFTY BANK volatility, with spikes rarely exceeding 50%. Post-COVID regime (red): dramatically elevated volatility baseline with regular spikes of 100–400%. Under this full-sample-baseline construction the level shift is large and highly significant while the slope change is *not* separately identified ($p = 0.228$, 0.256); the slope steepening is identified only once the baseline is time-matched (§4.7, Appendix B.17). Break date February 20, 2020 (pre-specified COVID date; Chow $F = 7.52$, $p < 10^{-3}$ for NIFTY BANK and $F = 9.91$, $p < 10^{-4}$ for NIFTY 50); a Quandt–Andrews test estimates the break endogenously at March 11, 2020. This non-stationarity explains why pooled models yield low R^2 .

MPC profiles for India VIX indicate that implied-volatility turbulence is more closely related to sentiment content than to the announcement-day profile itself, consistent with the CLS-based sentiment association reported in Table 1.

B.4 Multiple-Comparison Correction

We treat the three pre-specified minute-level channels (negative-sentiment probability, attention concentration, CLS sentiment) $\times$ three indices as the confirmatory family ($m = 9$) and apply Bonfer-

roni and Benjamini–Hochberg (BH) FDR corrections (Table 5). All five channels significant at the raw level survive BH-FDR at $q = 0.05$; the two strongest (attention concentration NIFTY BANK and CLS India VIX) also survive Bonferroni. Multiplicity correction controls the error rate across the family but says nothing about whether an individual correlation is leverage-driven: as §4.5 shows, the Bonferroni-surviving attention result does not withstand rank or bootstrap inference, whereas the CLS result does. The two adjustments should therefore be read together with Table 1, not in place of it.

For completeness we also corrected the full daily feature $\times$ target grid ($m = 36$): no daily-horizon correlation survives either correction (smallest raw $p = 0.014$), consistent with our finding that daily associations are weak, conditional, and non-additive (§4.2, §4.4).

B.5 Joint Multivariate Event-Date Models

Because negative sentiment and attention concentration are both associated with equity volatility in the primary bivariate analysis, we additionally test whether the two measurements provide conditionally separable associations. At the corrected event-date level, negative-sentiment probability and attention_std are moderately correlated ($r = 0.451$, $p = 1.34 \times 10^{-10}$, $N = 184$). Attention concentration is, however, essentially uncorrelated with semantic complexity ($r = 0.013$, $p = 0.867$, $N = 181$), while negative sentiment and semantic complexity are only weakly correlated ($r = 0.129$, $p = 0.083$, $N = 181$). The CLS score is also essentially uncorrelated with attention concentration ($r = 0.013$, $p = 0.858$, $N = 184$).

Table 6 reports standardized coefficients from HC3-robust event-date regressions. Semantic-complexity availability leaves $N = 181$ complete

Table 5: Confirmatory minute-level family after event-date aggregation: raw and multiple-testing-adjusted p -values ($N = 184$ event dates). All five raw-significant associations survive Benjamini–Hochberg FDR; two also survive Bonferroni correction.

Channel	Index	r	p	p_{BH}	p_{Bonf}
Prob. Negative	NIFTY 50	0.194	0.0084	0.0181	0.0759
Prob. Negative	NIFTY BANK	0.189	0.0101	0.0181	0.0907
Prob. Negative	India VIX	0.077	0.2972	0.3809	1.000
Attention Std	NIFTY 50	0.198	0.0071	0.0181	0.0635
Attention Std	NIFTY BANK	0.218	0.0029	0.0131	0.0261
Attention Std	India VIX	−0.041	0.5811	0.5811	1.000
CLS (Norm)	NIFTY 50	−0.094	0.2063	0.3094	1.000
CLS (Norm)	NIFTY BANK	−0.071	0.3386	0.3809	1.000
CLS (Norm)	India VIX	−0.224	0.0022	0.0131	0.0202

event dates for these specifications. This subset is defined by SCI availability and is distinct from the equally sized data-quality-filtered sample of Appendix B.17, which excludes three dates for zero-volatility and attention-extraction failures; the two coincide in size but not in membership.

For NIFTY 50, negative sentiment retains a nominally significant conditional association, whereas attention concentration and semantic complexity do not. For NIFTY BANK, neither negative sentiment nor attention concentration is individually significant in the joint model despite both displaying positive coefficients. The attention coefficients are far from significance in both equity models, which is consistent with the bivariate diagnosis in §4.5: the nominal attention correlation is not a monotone association and does not survive robust inference. The polarity association, by contrast, is nominally significant for NIFTY 50 and retains the expected sign for NIFTY BANK.

For India VIX, the CLS sentiment association remains highly significant after joint adjustment, strengthening the evidence that this sentiment representation is associated with event-window turbulence in the implied-volatility index. The negative semantic-complexity coefficient is also statistically significant in the joint model; because this conditional association was not one of the pre-specified primary minute-level channels, we report it as exploratory rather than as a new confirmatory finding.

B.6 Endogenous Break-Point Estimation

We estimate the candidate break endogenously using the Quandt–Andrews $\sup-F$ statistic, sweeping candidate break points over the central 70% of the date-ordered event sample (15% trimming) for the regression of the minute-level volatility spike on negative-sentiment probability. After correcting

same-date textual aggregation, the maximum remains March 11, 2020 for both indices, with 69 observations before and 115 after the candidate break. For NIFTY 50, $\sup-F = 10.14$ exceeds the Andrews (1993) 10% reference critical value (≈ 9.84 for the stated specification) but not the 5% value (≈ 11.79). For NIFTY BANK, $\sup-F = 7.64$ remains below conventional reference thresholds. The fact that the independently searched maximum occurs on the same date for both equity indices provides cross-index evidence about the location of the regime transition, while the statistical strength of the single-break test differs by index. A complementary Chow test at the pre-specified February 20, 2020 boundary strongly rejects parameter stability for both NIFTY 50 ($F = 9.91$, $p = 8.3 \times 10^{-5}$) and NIFTY BANK ($F = 7.52$, $p = 7.3 \times 10^{-4}$). We therefore interpret the combined evidence as supporting COVID-period regime dependence rather than requiring an instantaneous common structural break exactly on one trading date.

B.7 Robustness of the Composite Sentiment Score

We examined the confidence-weighted composite along three axes. (i) **Convergent validity**: the three scoring methods are only modestly correlated (mean pairwise Pearson $r = 0.27$; logit vs. CLS $r = 0.37$), and the learned weights are dominated by the logit method ($\alpha_{\text{logit}} = 0.76 \pm 0.04$), so the composite is in practice close to the logit score ($r = 0.99$). (ii) **Weight calibration**: the per-document confidence does not track inter-method agreement ($r = -0.03$, $p = 0.65$), so we do not interpret the weights as calibrated uncertainty. (iii) **Robustness to weighting**: replacing the confidence-weighted composite with a simple equal-weight

Table 6: Joint event-date-level regressions of standardized 10:00–11:00 AM Parkinson-volatility spikes on textual characteristics. Coefficients are standardized; p -values use HC3 heteroskedasticity-robust standard errors. The equity specifications include negative-sentiment probability, attention concentration, and semantic complexity. The India VIX specification replaces negative-sentiment probability with the CLS sentiment score. All models use $N = 181$ complete event dates.

Index	Predictor	Std. β	HC3 p	Model R^2
NIFTY 50	Prob. Negative	0.131	0.0492	0.0516
	Attention Std	0.135	0.5458	
	Semantic Complexity	0.006	0.9661	
NIFTY BANK	Prob. Negative	0.102	0.1275	0.0618
	Attention Std	0.169	0.4556	
	Semantic Complexity	0.071	0.6824	
India VIX	CLS Score (Norm)	-0.228	0.0010	0.0708
	Attention Std	-0.033	0.7103	
	Semantic Complexity	-0.136	0.0009	

average yields nearly identical associations with volatility (mean $|r| = 0.031$ vs. 0.033 ; mean absolute difference 0.020). We therefore present the composite as an exploratory feature whose downstream conclusions do not depend on the weighting scheme.

B.8 Temporal Co-evolution of Sentiment and Market Volatility

Figure 7 presents the full time-series evolution of dynamically aggregated FinBERT sentiment alongside India VIX over the 2010–2025 study period.

The figure visually confirms three properties of the sentiment–volatility relationship: (1) sentiment provides episodic rather than continuous explanatory power, with alignment concentrated during stress periods; (2) the COVID-19 VIX spike (peaking above 45) dwarfs all prior volatility episodes, underscoring the structural break documented in §4.7; and (3) post-2020 sentiment exhibits higher-frequency oscillations with larger amplitude, consistent with the regime shift toward heightened communication sensitivity.

B.9 Minute-Level Text Volatility Correlations

Table 1 in the main text reports the event-date-level confirmatory family. When multiple communications map to the same trading date, textual features are averaged at the event-date level before being matched to the common market response; the resulting primary sample contains $N = 184$ event dates.

The event-date correlations reveal a cross-market separation among the textual channels: negative sentiment is associated with realized equity volatility, whereas CLS-based sentiment is associated

with India VIX turbulence. Effect sizes are modest in absolute terms ($r^2 \approx 3.6$ – 5.0% for the surviving associations), indicating that textual characteristics explain a meaningful but limited portion of event-window variation—a magnitude that is typical for event-level financial text and that we do not interpret as economically large. All five raw-significant associations survive Benjamini–Hochberg correction (Table 5), but only the polarity and CLS associations survive the rank, bootstrap and regime-conditional criteria of §3.6.

B.10 Sentiment-Innovation Robustness

To test whether the minute-level sentiment associations are driven merely by the predictable persistence of sentiment across successive communications, we orthogonalize each sentiment series against its own first lag. Specifically, we fit a document-ordered AR(1) model and use the residual as the sentiment innovation. We then repeat the event-date-level correlations using these innovations in place of the original sentiment levels.

The results are essentially unchanged. For negative-sentiment probability, the level correlation for NIFTY 50 is $r = 0.194$ ($p = 0.0084$), while the innovation correlation is $r = 0.194$ ($p = 0.0082$). For NIFTY BANK, the corresponding values are $r = 0.189$ ($p = 0.0101$) and $r = 0.189$ ($p = 0.0102$). CLS sentiment remains specifically associated with India VIX: the level correlation is $r = -0.224$ ($p = 0.0022$), and the innovation correlation remains $r = -0.208$ ($p = 0.0046$). The equity CLS associations and the negative-sentiment India VIX association remain non-significant. Thus, the principal polarity and implied-volatility associations are not explained by

simple AR(1) persistence in the sentiment series; the innovation component carries essentially the same cross-sectional signal as the level.

Implied-volatility level check. The minute-level India VIX target measures the intraday variability of the implied-volatility index rather than its level, so we additionally correlate each textual feature with the daily India VIX level change over the Days 0–1 window relative to the pre-event baseline. None of the three is significant: CLS sentiment $r = 0.035$ ($p = 0.54$), negative-sentiment probability $r = -0.043$ ($p = 0.45$), attention concentration $r = -0.032$ ($p = 0.57$), with $N = 314$. The uncertainty channel therefore does *not* extend from turbulence in implied volatility to the level of implied volatility, and we restrict the claim accordingly.

B.11 Machine Learning Models: Feature Importance and Non-linear Effects

Figures 8 and 9 present comprehensive visualizations of feature importance rankings and polynomial regression fits for non-linear effect detection.

B.11.1 Feature Importance Comparison: Random Forest vs. XGBoost

The six-panel layout enables direct cross-model comparison: Random Forest importance scores (left column) are computed via mean decrease in impurity, while XGBoost scores (right column) are computed via gain-based importance. Despite different calculation methods, the top-5 features remain largely consistent within each index. For NIFTY50, inflation surprise proxy and NSEI Parkinson mean consistently rank highest in the descriptive importance rankings. For INDIIVIX, NSEI realized volatility metrics dominate, reflecting the derivative nature of VIX as an index constructed from NSE options. For NIFTYBANK, banking-sector-specific volatility lags (NSEBANK EWMA, NSEBANK Parkinson) far exceed those of the general market or content features, consistent with sectoral momentum effects.

Among content features, sentiment logits rank above semantic complexity across the six models in the descriptive rankings. This is consistent with an in-sample time-scale pattern but does not establish differential out-of-sample predictive contribution. Attention-based features (attention max, attention std, attention mean) show variable importance across indices, ranking moderately for equity indices and lower for INDIIVIX. These are

in-sample impurity/gain rankings and, given the robustness results of §4.5, should not be read as support for an attention channel. We further note that attention_mean is mechanically constant for fully truncated inputs (Table 3) and is retained in this figure only for comparability with the original feature set.

B.11.2 Non-linear Effects: Polynomial Regression Fits

The polynomial fits were estimated using regularized polynomial regression, with model complexity evaluated out of sample. Probability-negative and semantic-complexity curves remain close to their linear fits, whereas attention_std exhibits visibly stronger curvature for the equity indices. NIFTY 50 peaks near attention_std ≈ 0.0035 and NIFTY BANK near ≈ 0.004 , while India VIX exhibits the opposite curvature with a minimum around the same intermediate range.

This cross-market reversal is specific to the attention-distribution variable rather than appearing generically across all textual predictors. We report it descriptively only. Because the underlying bivariate attention association does not survive rank or bootstrap inference (§4.5), and because the polynomial specifications do not improve out-of-sample prediction, we do not interpret the curvature as evidence of a salience mechanism and do not treat the turning points as stable thresholds.

B.12 Residual Analysis: Content Effects Beyond Event Windows

To test whether textual features carry explanatory power beyond event-driven volatility, we regress each volatility target on event indicators and lagged volatility controls, then examine whether sentiment and complexity correlate with the residuals. Figure 11 presents two complementary views.

The uniformly red bars in the left panel confirm that content features cannot explain volatility residuals after accounting for event-driven and persistence-based effects. The right panel’s clustering near the origin (all correlations within ± 0.06) reinforces this conclusion for both sentiment and complexity independently. This result is important for two reasons: (1) it validates that the content effects documented in §4.3 and §4.4 operate *within* event windows rather than persisting beyond them; and (2) it confirms that the modest R^2 values in our main models are not artifacts of omitted persistent content effects but reflect the genuinely transient

Table 7: In-sample content-feature permutation importance for NIFTY BANK RV (Days 0–1), shown descriptively. Out-of-sample group contributions appear in Appendix B.2.

Feature	Importance Score	Feature Type
Sentiment (Logits)	0.260	Current
Sentiment (Lag 1)	0.200	Temporal
Semantic Complexity	0.190	Current
Sentiment Change	0.180	Temporal
Complexity Change	0.095	Temporal
Complexity (Lag 1)	0.075	Temporal

Full feature importance chart appears below in Figure 13.

nature of communication-driven volatility.

B.13 Daily Volatility Analysis: Interaction Effects and Feature Importance

Table 8 reports the event-window interaction-model comparison, while Figure 13 provides the descriptive feature-importance ranking.

The interaction results in Table 8 complement the descriptive feature-importance visualization in Figure 13. The feature-importance figure facilitates direct comparison of the fitted model’s content-feature rankings; these rankings are descriptive and should not be interpreted as evidence of out-of-sample predictive contribution. The prominence of both current and lagged sentiment features, combined with the elevated importance of semantic complexity at daily horizons, is consistent with the in-sample observation that textual signals operate through time-dependent mechanisms distinct from immediate sentiment-driven reactions.

Stratification and temporal-feature interpretation. The high-sentiment/high-complexity volatility elevation is consistent with two possible interpretations: high sentiment intensity may proxy informational intensity, while high complexity may proxy linguistic density requiring extended processing; both interpretations could contribute to greater response variability. Temporal features (lagged sentiment, sentiment change, complexity change) collectively account for 50% of total in-sample importance, suggesting, descriptively, that *changes* relative to prior MPC statements receive substantial importance within the fitted model—a hawkish pivot from dovishness could, on this reading, elicit stronger reactions than sustained hawkishness.

B.14 DAST vs. Classical Readability Baselines

To assess whether DAST adds value over simpler, well-established complexity measures, we computed fourteen classical readability and surface-complexity metrics (Flesch Reading Ease, Flesch–Kincaid Grade, Gunning Fog, SMOG, Coleman–Liau, ARI, Dale–Chall, Linsear–Write, difficult-word count, word and sentence counts, average sentence length, average syllables per word, and type token ratio) on the same MPC corpus, and compared them to DAST along three dimensions: redundancy, incremental out-of-sample predictive value, and feature importance.

DAST is not redundant with classical readability: across fourteen standard indices its maximum absolute correlation is $|r| = 0.21$ (Table 9), confirming it captures hierarchical semantic structure that syllable- and sentence-length-based formulas do not. In a random forest combining DAST with all fourteen readability features, DAST is the single most important complexity feature for the three difference-based targets (NIFTY BANK AR 0–1, NIFTY BANK ΔRV 0–1, NIFTY 50 ΔRV 0–1; note that AR is a return rather than a volatility measure, Appendix A.13) and ranks in the top four for five of eight targets (Table 11), indicating it carries announcement-relevant signal beyond any individual surface metric. We emphasize, consistent with §B.15, that at the event level no complexity measure—DAST or readability—is a strong standalone out-of-sample predictor of volatility; complexity operates conditionally rather than as an additive main effect. DAST’s contribution is therefore as a distinct, interpretable construct rather than as a superior linear predictor.

B.14.1 Explicit Document-Length Control

Because DAST could in principle covary with the amount of text available for semantic-rule generation, we additionally tested document length directly. Using document-level matching between each analysis row and its parsed source document ($N = 290$), DAST is only weakly negatively associated with length (Pearson $r = -0.162$, $p = 0.0058$; Spearman $\rho = -0.229$, $p < 0.001$). The correlation with $\log(1 + \text{word count})$ is similarly weak ($r = -0.188$, $p = 0.0013$).

We then re-estimated the sentiment complexity interaction after adding $\log(1 + \text{word count})$ as an explicit covariate. The interaction coefficient retains its original sign for all nine market-response targets, but none is significant under HC3-

Table 8: Exploratory sentiment $\times$ complexity interaction models for NIFTY BANK event-window realized volatility. R^2 values are in-sample percentages; ΔR^2 is the interaction-model gain over the corresponding base model. HC3 p -values refer to the interaction coefficient. The final two columns report the interaction gain and HC3 p -value after additionally controlling for $\log(1 + \text{WordCount})$.

Event Window	Base R^2	+Int. R^2	ΔR^2	HC3 p	$\Delta R^2 + \text{Length}$	HC3 $p + \text{Length}$
Days 0–1 (Immediate)	0.128	0.153	0.026	0.777	0.010	0.856
Days 2–3 (Short-term)	1.019	1.285	0.266	0.208	0.181	0.353
Days 4–5 (Medium-term)	0.234	1.044	0.810	0.093	0.780	0.113

Table 9: DAST vs. fourteen classical readability and surface-complexity metrics. Correlations are weak throughout. Reported p -values are two-sided tests on the Pearson coefficient. The three length-based rows (word count, sentence count, difficult words) are computed on the $N = 290$ subset for which reliable document-level text matching is available; the remaining eleven readability indices are computed on all $N = 332$ documents with a DAST score, which is why p -values are not monotone in $|r|$ across rows.

Readability metric	Pearson r	Spearman ρ	p
Type token ratio	0.211	0.240	< 0.001
Avg. syllables/word	0.204	0.218	< 0.001
Coleman–Liau	0.187	0.202	0.001
Sentence count	−0.162	−0.239	0.003
Difficult words	−0.161	−0.235	0.004
Word count	−0.162	−0.229	0.006
Flesch Reading Ease	−0.154	−0.131	0.005
Dale–Chall	0.146	0.178	0.008
ARI	0.133	0.139	0.016
Gunning Fog	0.117	0.087	0.035
Flesch–Kincaid Grade	0.105	0.085	0.058
Linsear–Write	−0.102	−0.158	0.065
SMOG	0.098	0.072	0.076
Avg. sentence length	0.006	−0.015	0.919

robust inference after length control (minimum $p = 0.113$). Likewise, no target exhibits a significant repeated- k -fold out-of-sample improvement under the permutation null (minimum permutation $p = 0.289$). The mean length-controlled repeated- k -fold ΔR^2_{OOS} is -0.41 percentage points, and the mean forward-chaining ΔR^2_{OOS} is -2.23 percentage points. Thus, controlling explicitly for document length does not alter the substantive conclusion: the sentiment complexity interaction is at most a small exploratory in-sample pattern and does not provide robust predictive gain.

B.15 Out-of-Sample Validation of the Interaction Model

Because the interaction gains in Table 8 are small and estimated in-sample, we tested whether they reflect genuine predictive signal or in-sample overfitting. For each market-response target we compared the base model (S, C) against the interaction model ($S, C, S \times C$) using three out-of-sample schemes—repeated 5-fold cross-validation, Group-KFold by event date (so documents from one date never span train and test), and forward-chaining

time-series splits—plus a permutation null for the OOS ΔR^2 and HC3 robust standard errors for the interaction coefficient. Outliers in semantic complexity ($|z| > 5$) were removed beforehand, leaving 326 events.

The interaction term produces small in-sample fit gains, but does not yield a statistically significant out-of-sample improvement on any tested market-response target (Table 12). In the original specification, the mean repeated- k -fold ΔR^2_{OOS} is -0.37% and the mean forward-chaining ΔR^2_{OOS} is -1.35% . No target shows a significant OOS improvement under the permutation null, and no interaction coefficient is significant under robust standard errors. The explicit document-length robustness check leads to the same conclusion (Appendix B.14.1). We therefore interpret the interaction as an exploratory in-sample pattern rather than a predictive result.

B.16 GARCH + Policy-Rate Baseline

We fit an expanding-window GARCH(1,1) to daily index log-returns and, at each MPC date, take the one-step-ahead conditional-volatility forecast us-

Table 10: Sentiment $\times$ complexity interaction after controlling for $\log(1 + \text{word count})$. ΔR_{in}^2 is the incremental in-sample fit from adding the interaction to a model already containing sentiment, SCI, and document length. ΔR_{OOS}^2 is the analogous repeated- k -fold out-of-sample change. Permutation p tests the conditional OOS interaction gain.

Target	N	$\beta_{S \times \text{SCI}}$	HC3 p	ΔR_{in}^2 (%)	ΔR_{OOS}^2 (%)	Perm. p
NIFTY BANK RV 0-1	275	0.000290	0.856	0.010	-0.966	0.678
NIFTY BANK RV 2-3	270	0.001237	0.353	0.181	-0.154	0.399
NIFTY BANK RV 4-5	270	0.002489	0.113	0.780	0.071	0.289
NIFTY BANK AR 0-1	276	41.653915	0.314	0.455	-0.557	0.857
NIFTY BANK Δ RV 0-1	272	0.000217	0.896	0.005	-0.631	0.628
NIFTY 50 RV 0-1	290	-0.001230	0.247	0.392	-0.332	0.618
NIFTY 50 RV 2-3	290	0.000180	0.842	0.006	-0.375	0.369
NIFTY 50 RV 4-5	290	0.000344	0.681	0.032	-0.408	0.369
India VIX 0-1	290	0.024450	0.541	0.063	-0.357	0.389

Table 11: DAST permutation-importance rank among 15 complexity features (DAST + 14 readability) in a random forest per target. DAST is the single most important complexity feature for the three difference-based targets (NIFTY BANK AR 0-1, NIFTY BANK Δ RV 0-1, NIFTY 50 Δ RV 0-1; note that AR is a return rather than a volatility measure, Appendix A.13).

Volatility target	DAST rank (of 15)
NIFTY BANK AR 0-1	1
NIFTY BANK Δ RV 0-1	1
NIFTY 50 Δ RV 0-1	1
NIFTY 50 RV 0-1	3
NIFTY BANK RV 0-1	4
NIFTY BANK RV 2-3	8
India VIX 0-1	13
NIFTY BANK RV 4-5	15

ing only prior information; the announced repo-rate change (basis points, taken from the RBI’s official repo-rate record—45 changes over 2010–2025—aligned to each release within a ± 3 -day window) enters as the policy regressor. We compare three nested models for the standardized 10:00–11:00 AM event-window volatility spike under forward-chaining (TimeSeriesSplit) cross-validation: M0 (GARCH only), M1 (GARCH + policy-rate change and hike/cut indicators), and M2 (GARCH + rate + text).

Table 13 reports out-of-sample R^2 for each model, the incremental contribution of text, and an in-sample nested F -test. Event-level spikes are difficult to predict out of sample—all three models have negative OOS R^2 —and adding text reduces OOS R^2 on every index (mean $\Delta R_{\text{OOS}}^2 = -8.3\%$), with paired squared-loss differences non-significant across both rate encodings ($p = 0.28$ – 0.73). In-sample, the nested F -test favours the text-augmented model for India VIX ($F = 2.33$, $p = 0.035$) and NIFTY BANK ($F = 3.29$, $p = 0.005$) but not NIFTY 50 ($F = 1.67$, $p = 0.13$);

this in-sample result is robust to the alternative carried-forward policy-rate-change encoding (F -test $p \leq 0.03$ across indices). This reframes text from “differs from zero” to “incremental over an econometric baseline,” and corroborates the feature-group OOS null (Appendix B.2).

B.17 Data-Quality Filter and Time-Matched Baseline

Data-quality filter. Two issues affect a small number of event dates. One event date records zero Parkinson volatility across the entire 10:00–10:59 window on all three indices, which indicates a data gap rather than a quiet hour and produces a mechanical -100% spike; two further event dates carry documents whose FinBERT attention extraction failed (attention_std = 0). Excluding these three dates leaves $N = 181$ event dates. All minute-window counts are otherwise complete (minimum 54 of 60 minutes present; no date falls below 30). Conclusions are unchanged by the filter, so the main text reports the unfiltered $N = 184$ sample and this appendix reports the filtered one.

Time-matched baseline. Equation 24 normalises each event window by the mean non-event Parkinson volatility in the same clock-time window, computed over the whole sample. That single scalar absorbs the low-frequency level of market volatility, so an event observed in a calm year is scored against a baseline that includes turbulent years. We therefore recompute every spike against a *local* baseline formed from the $k = 20$ non-event trading days nearest in calendar time to each event, and re-run the confirmatory family. Tables 14 and 15 report both constructions on the filtered sample.

Three associations are robust to both baseline constructions: negative-sentiment probability with NIFTY 50 and NIFTY BANK volatility, and CLS sentiment with India VIX turbulence. Under the

Table 12: Out-of-sample validation of the sentiment $\times$ complexity interaction. ΔR^2 is interaction minus base. In-sample gains are small, out-of-sample changes are predominantly negative, and no target exhibits a statistically significant out-of-sample improvement.

Target	In-sample ΔR^2 (%)	k -fold OOS ΔR^2 (%)	Time-series OOS ΔR^2 (%)	Perm. p	Robust p
NIFTY BANK RV 0–1	0.00	−0.86	−2.07	0.83	0.97
NIFTY BANK RV 2–3	0.04	−0.28	−1.35	0.51	0.63
NIFTY BANK RV 4–5	0.72	−0.07	0.24	0.11	0.11
NIFTY BANK AR 0–1	0.36	−0.43	−0.53	0.82	0.31
NIFTY BANK Δ RV 0–1	0.01	−0.55	−2.44	0.60	0.87
NIFTY 50 RV 0–1	0.42	−0.18	−1.90	0.19	0.20
NIFTY 50 RV 2–3	0.00	−0.29	−1.34	0.37	0.87
NIFTY 50 RV 4–5	0.00	−0.36	−1.22	0.42	0.92
India VIX 0–1	0.03	−0.29	−1.57	0.26	0.63
Mean	0.18	−0.37	−1.35	-	-

Table 13: GARCH+rate baseline vs text-augmented model (standardized 10:00–11:00 AM event-window volatility spike, forward-chaining OOS). $\Delta R^2_{\text{text}} = \text{M2} - \text{M1}$.

Index	M0	M1	M2	ΔR^2_{text}	nested F (p)
India VIX	−0.035	−0.045	−0.128	−0.083	2.33 (0.035)
NIFTY 50	−0.201	−0.024	−0.120	−0.096	1.67 (0.131)
NIFTY BANK	−0.470	−0.115	−0.185	−0.071	3.29 (0.005)

M0: GARCH only; M1: +policy-rate; M2: +text.

Paired squared-loss $p = 0.28$ – 0.73 (n.s., both encodings). $n = 184$ events, 82 folds.

time-matched baseline, CLS sentiment with NIFTY 50 additionally clears the full inference protocol. Because this fourth association is baseline-dependent, we treat it as a sensitivity result rather than as a core cross-baseline finding. Attention concentration does not survive under either construction.

Regime decomposition under the time-matched baseline. Estimating the regime-interaction specification of §4.7 on the filtered sample with the local baseline gives, for NIFTY 50, $\hat{\alpha}_1 = 13.46$ (HC3 $p = 0.0010$) and $\hat{\beta}_1 = 53.35$ (HC3 $p = 0.0006$), and for NIFTY BANK $\hat{\alpha}_1 = 14.32$ ($p = 0.0015$) and $\hat{\beta}_1 = 73.43$ ($p = 0.0001$); heteroskedasticity-robust Wald tests of joint stability give $F = 21.68$ and $F = 20.35$ respectively ($p < 10^{-4}$). The unfiltered $N = 184$ estimates quoted in §4.7 are within 10% of these ($\hat{\alpha}_1 = 14.7$ vs. 13.46; 15.6 vs. 14.32) and identical in sign and significance; the slope estimates agree to within 1%.

Sensitivity to the baseline window. Table 16 reports the confirmatory correlations for $k \in \{10, 20, 40, 60\}$. The polarity association is essentially flat in k ($r = 0.29$ – 0.34 on both equity indices), the CLS–India VIX association peaks near $k = 20$ but remains $|r| > 0.18$ throughout, and attention concentration remains non-significant in rank terms at every k , with the Pearson and Spear-

man estimates of opposite sign across the whole range. The choice $k = 20$ is therefore not pivotal.

B.18 Lexicon Baseline for Sentiment

The fine-tuned transformer is only worth its cost if it improves on a dictionary count. We compute Loughran–McDonald (Loughran and McDonald, 2011) negative-word share and net polarity on the parsed document text, aggregate them to the event date, and evaluate them against the same minute-level target under the time-matched baseline. Reliable text matching is available for 295 of 334 documents, giving a common evaluation sample of 164 event dates on which both the FinBERT and lexicon features are defined. Table 17 reports the comparison.

The fine-tuned model dominates on the equity targets, where neither lexicon feature is significant, and on India VIX turbulence, where LM polarity is nominally significant but weaker than the CLS score. The two negativity measures are also close to uncorrelated with each other ($r = 0.090$, $p = 0.25$; $\rho = -0.102$, $p = 0.20$), so the transformer is not recovering lexicon counts by another route. Given the near-neutral concentration of the model’s residual disagreement with human annotators (Appendix A.7), we read this as evidence that the contextual model separates hedged central-

Table 14: Confirmatory family on the data-quality-filtered sample ($N = 181$), full-sample baseline. Columns as in Table 1.

Feature	Index	r	p	95% BCa CI	ρ	ρp	Partial $r(p)$	Robust
Prob. Negative	NIFTY 50	0.192	0.0093	[0.065, 0.302]	0.186	0.012	0.149 (0.045)	✓
	NIFTY BANK	0.188	0.0111	[0.067, 0.301]	0.167	0.024	0.149 (0.045)	✓
	India VIX	0.077	0.3004	[-0.016, 0.201]	0.022	0.771	0.090 (0.226)	
Attention Std	NIFTY 50	0.212	0.0041	[-0.076, 0.544]	-0.119	0.111	0.342 (< 0.001)	
	NIFTY BANK	0.233	0.0016	[-0.060, 0.552]	-0.097	0.192	0.349 (< 0.001)	
	India VIX	-0.012	0.8695	[-0.135, 0.230]	-0.010	0.898	-0.035 (0.638)	
CLS Score (Norm)	NIFTY 50	-0.084	0.2591	[-0.180, 0.056]	-0.053	0.476	-0.116 (0.118)	
	NIFTY BANK	-0.062	0.4060	[-0.155, 0.071]	-0.039	0.603	-0.089 (0.234)	
	India VIX	-0.208	0.0049	[-0.340, -0.070]	-0.289	< 0.001	-0.203 (0.006)	✓

Table 15: Confirmatory family on the data-quality-filtered sample ($N = 181$) with the time-matched baseline ($k = 20$ nearest non-event trading days). The polarity and CLS associations strengthen; the attention association disappears.

Feature	Index	r	p	95% BCa CI	ρ	ρp	Partial $r(p)$	Robust
Prob. Negative	NIFTY 50	0.292	0.0001	[0.127, 0.476]	0.173	0.019	0.257 (0.0005)	✓
	NIFTY BANK	0.340	< 0.0001	[0.178, 0.519]	0.202	0.006	0.309 (< 0.001)	✓
	India VIX	0.121	0.1042	[0.002, 0.253]	-0.006	0.938	0.091 (0.221)	
Attention Std	NIFTY 50	0.086	0.2495	[-0.088, 0.304]	-0.095	0.203	0.200 (0.0067)	
	NIFTY BANK	0.115	0.1231	[-0.078, 0.352]	-0.064	0.392	0.229 (0.0019)	
	India VIX	-0.073	0.3261	[-0.218, 0.210]	-0.155	0.036	-0.015 (0.845)	
CLS Score (Norm)	NIFTY 50	-0.175	0.0185	[-0.296, -0.042]	-0.170	0.022	-0.211 (0.0042)	✓
	NIFTY BANK	-0.124	0.0967	[-0.247, 0.033]	-0.128	0.085	-0.156 (0.035)	
	India VIX	-0.263	0.0003	[-0.380, -0.126]	-0.299	< 0.001	-0.286 (0.0001)	✓

bank negativity from mere negative-word density, which a dictionary cannot do.

B.19 Lattice Shape and the SCI Construct

Because SCI is generated rather than measured, we report the shape of the lattices it is computed from. Across the 332 generated lattices, the number of derivation rules averages 8.8 (SD 6.7; range 3–62), the number of lattice nodes averages 20.2 (SD 18.3; range 5–249), and the number of edges averages 11.0 (SD 11.8; range 0–78). Median SCI is 21 with an interquartile range of 10–38; on a $\log_{10}$ scale the distribution has mean 1.40 and SD 1.24.

Two features of this distribution qualify the interpretation of DAST in our setting. First, 30.1% of lattices contain no edges at all, so Equation 1 returns 1.0 at every node and SCI equals the node count exactly for roughly a third of the corpus. Second, across all documents SCI is close to a monotone transform of lattice size: Spearman $\rho = 0.980$ with node count, 0.941 with edge count and 0.767 with rule count. Since all our analyses use SCI ordinally (Appendix A.4), the operative content of the index is therefore the amount of semantic structure the generation step extracts, not a fine-grained

gradation of hierarchical depth.

This does not make SCI redundant. It remains largely orthogonal to fourteen classical readability indices ($|r|_{\max} = 0.21$; Appendix B.14) and ranks first among fifteen complexity features for three announcement-change targets (Table 11). But it does mean that “semantic complexity” here should be read as generated-structure volume, and that the multiplicative form of Equation 1 contributes numerical range rather than additional discriminative information.

C Institutional and Socioeconomic Context

The statistical associations documented in the main paper operate within a specific institutional ecosystem. This appendix discusses possible channels through which RBI communications may be associated with market volatility, and identifies the socioeconomic and institutional factors that mediate this relationship.

Table 16: Sensitivity of the confirmatory correlations to the number of non-event days k used for the time-matched baseline ($N = 181$). Pearson r with Spearman ρ in parentheses.

Feature	Index	$k = 10$	$k = 20$	$k = 40$	$k = 60$
Prob. Negative	NIFTY 50	0.308 (0.174)	0.292 (0.173)	0.302 (0.207)	0.293 (0.198)
	NIFTY BANK	0.337 (0.175)	0.340 (0.202)	0.335 (0.219)	0.326 (0.219)
	India VIX	0.206 (0.089)	0.121 (−0.006)	0.140 (0.057)	0.141 (0.056)
Attention Std	NIFTY 50	0.068 (−0.106)	0.086 (−0.095)	0.143 (−0.068)	0.156 (−0.060)
	NIFTY BANK	0.093 (−0.078)	0.115 (−0.064)	0.158 (−0.049)	0.171 (−0.037)
	India VIX	0.054 (−0.078)	−0.073 (−0.155)	−0.031 (−0.111)	−0.033 (−0.108)
CLS Score (Norm)	NIFTY 50	−0.185 (−0.171)	−0.175 (−0.170)	−0.133 (−0.120)	−0.127 (−0.127)
	NIFTY BANK	−0.124 (−0.142)	−0.124 (−0.128)	−0.097 (−0.107)	−0.096 (−0.111)
	India VIX	−0.186 (−0.203)	−0.263 (−0.299)	−0.205 (−0.260)	−0.194 (−0.250)

Table 17: FinBERT versus Loughran–McDonald on an identical sample ($N = 164$ event dates, time-matched baseline). p_{BH} applies Benjamini–Hochberg across the twelve tests shown.

Feature	Index	r	p	p_{BH}
Prob. Negative	NIFTY 50	0.292	0.0001	0.0009
	NIFTY BANK	0.326	< 0.0001	0.0003
	India VIX	0.102	0.1921	0.3294
CLS (Norm)	NIFTY 50	−0.175	0.0248	0.0595
	NIFTY BANK	−0.109	0.1662	0.3294
	India VIX	−0.268	0.0005	0.0021
LM negative share	NIFTY 50	−0.011	0.8866	0.8866
	NIFTY BANK	−0.074	0.3436	0.4123
	India VIX	0.086	0.2757	0.3677
LM polarity	NIFTY 50	−0.092	0.2392	0.3588
	NIFTY BANK	−0.038	0.6266	0.6835
	India VIX	−0.191	0.0141	0.0423

C.1 Possible Information Transmission Channels

RBI monetary-policy communications reach markets through a layered chain: official documents are published on the RBI website; algorithmic news-parsing systems operated by institutional traders can extract salient headlines rapidly; financial media relay summaries; and retail investors subsequently encounter interpreted versions through brokerage applications and other media channels. Each layer introduces potential for signal distortion—algorithmic parsers may react to keyword sentiment without contextual nuance, while media summarization may oversimplify complex conditional guidance.

The *polarity channel* provides one plausible interpretation of the observed association: negative language may prompt automated or discretionary risk-reduction responses, including portfolio rebalancing and hedging. Such a mechanism is consistent with the observed correlation between negative sentiment and equity volatility spikes within the announcement hour (§4.3), but these trading re-

sponses are not directly measured in our data. The stronger association for NIFTY BANK may reflect the greater direct exposure of banking stocks to monetary-policy changes and sector-specific portfolio flows, although these mechanisms should be interpreted as plausible explanations rather than directly identified causal pathways.

A second plausible interpretation—a *salience* mechanism, in which attention concentrated on a few competing phrases produces more heterogeneous interpretations of the net policy signal—is *not* supported by our data and we do not advance it. The nominal attention association reverses sign under rank correlation, has a bootstrap interval spanning zero, is not significant in the joint equity models, and disappears under a time-matched baseline (§4.5). We record the hypothesis here only because it motivated including the feature in the pre-specified family, and because the discrepancy between its Pearson and rank behaviour is itself informative about how attention-derived features should be evaluated.

The *implied-volatility turbulence* association

with CLS sentiment is the second pattern that does survive. A plausible reading is that a more positive overall document representation coincides with a less contested repricing of option-implied uncertainty within the announcement hour. As with the polarity channel, the microstructure mechanism is not directly observed in our data, and the association is with the variability of India VIX rather than with its level (Appendix B.10).

C.2 Why Effects are Incremental Rather Than Dominant

Several institutional features may explain why textual signals exhibit only modest incremental *in-sample* explanatory importance (cumulative content contribution of 15–25% in feature importance, §4.8), rather than dominant or out-of-sample predictive power:

1. **Pre-announcement information leakage:** Market participants form expectations about MPC decisions through macroeconomic data releases, RBI Governor speeches, and media commentary in the weeks preceding announcements. By the time the MPC document is released, a substantial portion of its informational content has already been priced in through the “expectations channel, leaving the *surprise component* deviations from consensus as one plausible source of announcement-day volatility.
2. **Algorithmic speed vs. linguistic complexity:** High-frequency trading algorithms react to easily parsed features (rate decision magnitude, keyword counts) within milliseconds, while the semantic complexity and nuanced sentiment that our framework captures require human-timescale cognitive processing. This could create a temporal separation between rapidly processed simple signals and more slowly interpreted complex signals, a pattern consistent with the observed time-scale split between sentiment (minute-level) and complexity (daily) documented in §4.4.
3. **Concurrent macroeconomic noise:** MPC announcements do not occur in isolation. Global risk events (US Federal Reserve decisions, geopolitical tensions, commodity price shocks), domestic data releases (GDP, CPI, trade balance), and firm-level earnings announcements all contribute to daily volatility,

diluting the marginal contribution of textual features in any statistical model.

4. **Market microstructure effects:** India’s equity markets feature circuit breakers, price bands, and periodic call auctions that mechanically constrain volatility responses regardless of information content. These institutional guardrails limit the maximum volatility any single information event can generate, placing an upper bound on achievable R^2 for text-based prediction.

C.3 The COVID-19 Regime Shift in Institutional Context

The structural break identified around March 11, 2020 (§4.7) has a clear institutional interpretation. Pre-COVID, MPC communications operated in a low-volatility regime where monetary policy was broadly predictable and forward guidance was well anchored. The pandemic introduced three simultaneous changes: (i) *unprecedented policy actions* (emergency rate cuts, moratorium announcements, liquidity facilities) that had no historical precedent for market participants to calibrate against; (ii) a *surge in retail investor participation*—new demat accounts opened in India grew from approximately 4 million in FY2019 to over 15 million in FY2021—bringing less sophisticated participants who may be more sensitive to communication tone; and (iii) *heightened global uncertainty* that amplified the marginal information content of any domestic policy signal. Together, these factors provide plausible institutional context for the observed post-COVID regime shift: for both equity indices the event-window volatility differential shifts upward, and once the baseline is time-matched the fitted sentiment–volatility slope also becomes substantially and significantly steeper (§4.7). These institutional mechanisms remain plausible interpretations rather than directly identified causal contributions.

C.4 Implications for Central Bank Communication Design

These empirical patterns have several implications for communication design and evaluation. First, the event-window analysis shows that the largest volatility elevation is temporally concentrated and followed by rapid normalization, making the immediate communication window particularly important for monitoring market interpretation. Second, the daily sentiment complexity models pro-

vide weaker but suggestive evidence that linguistically dense communication may condition how sentiment is absorbed over subsequent days; because this interaction is not robust out of sample, we treat it as a design hypothesis rather than a validated intervention rule. Third, our results caution against a specific evaluation practice. An attention-derived “salience” feature passed a pre-specified, multiplicity-corrected screen and would, under the inference standard common in this literature, have been reported as a channel; it does not survive rank or bootstrap inference. Communication-monitoring tools built on attention-derived features should therefore be validated with distribution-free inference before being used to guide disclosure design. The present observational results establish market associations rather than the causal effect of any particular redesign.

These considerations are particularly relevant for emerging markets like India, where the central bank simultaneously serves as monetary authority, banking regulator, and foreign exchange manager, making its communications inherently multi-dimensional and increasing the potential for interpretive ambiguity.

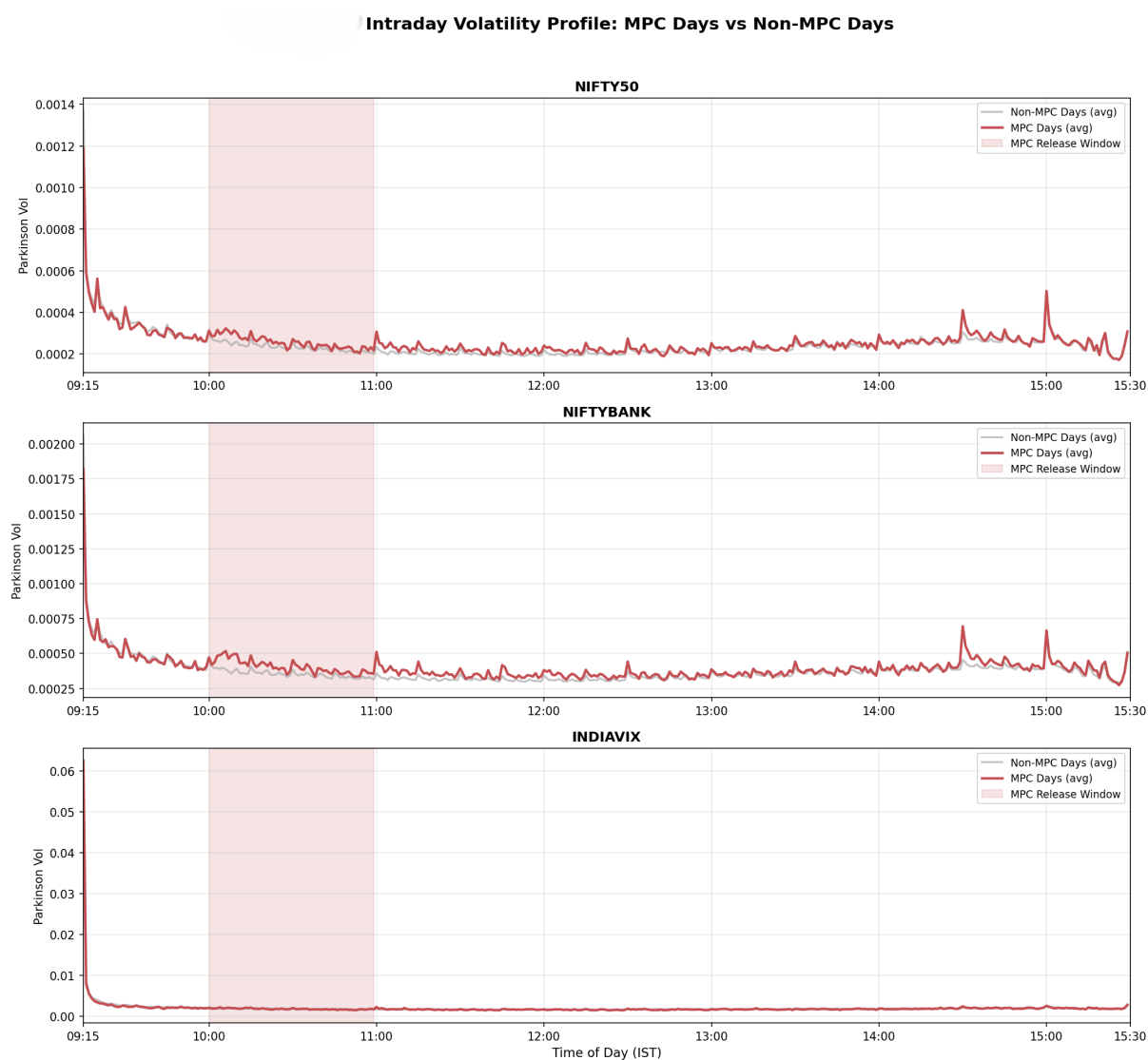


Figure 6: Intraday Parkinson volatility profiles: MPC days (red) vs. non-MPC days (grey) for NIFTY50 (top), NIFTYBANK (middle), and INDIIVIX (bottom). The shaded pink region marks the standardized event window (10:00–11:00 AM IST). NIFTY50 event-day volatility rises into the standardized event window but stays close to the non-MPC average throughout (see Figure 1 for the ratio view). NIFTYBANK shows the sharpest MPC-day elevation during 10:00–11:00, converging to the non-MPC baseline by approximately 12:00 PM. INDIIVIX displays negligible MPC-day divergence from baseline, consistent with implied-volatility turbulence being more closely related to sentiment content than to the announcement-day profile itself. All three indices exhibit a common opening-auction spike (9:15 AM) unrelated to MPC effects.

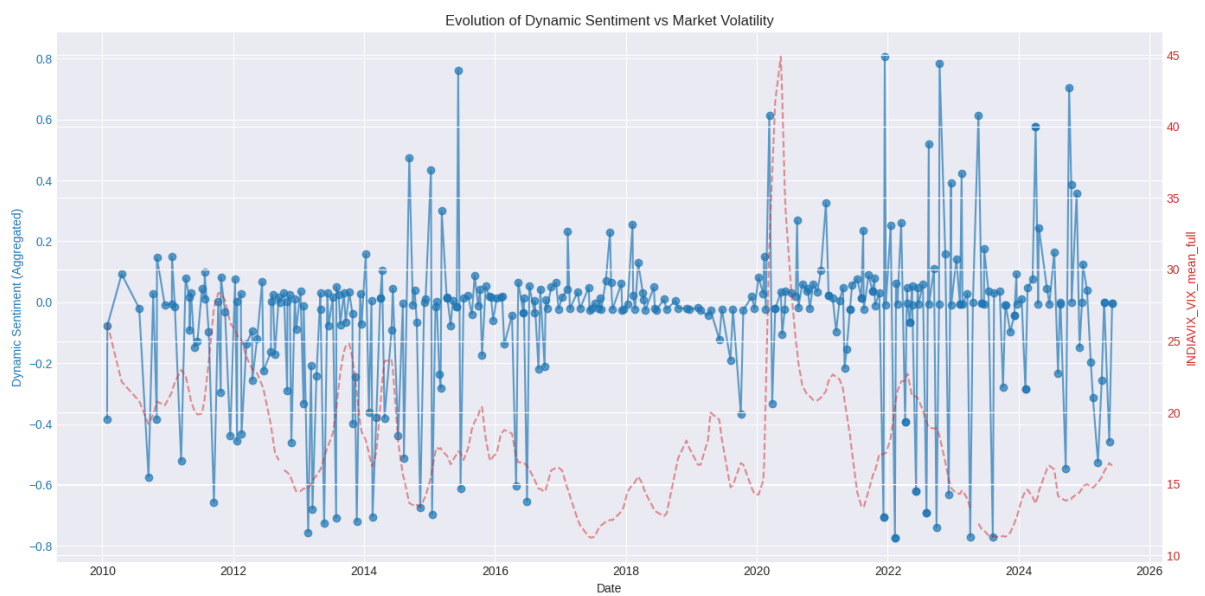


Figure 7: Temporal co-evolution of dynamic sentiment (blue, left axis) and India VIX (red dashed, right axis), 2010–2025. Sentiment oscillates around neutrality, with no sustained trend, while VIX spikes dramatically at the onset of COVID-19 (March 2020). Alignment between sentiment troughs and VIX peaks is episodic rather than continuous, with the most pronounced co-movement occurring during stress periods (2011 eurozone crisis, 2013 taper tantrum, 2020 COVID-19, 2022 rate-hiking cycle). Outside these windows, sentiment and VIX evolve largely independently, explaining the weak unconditional correlation reported in §4.

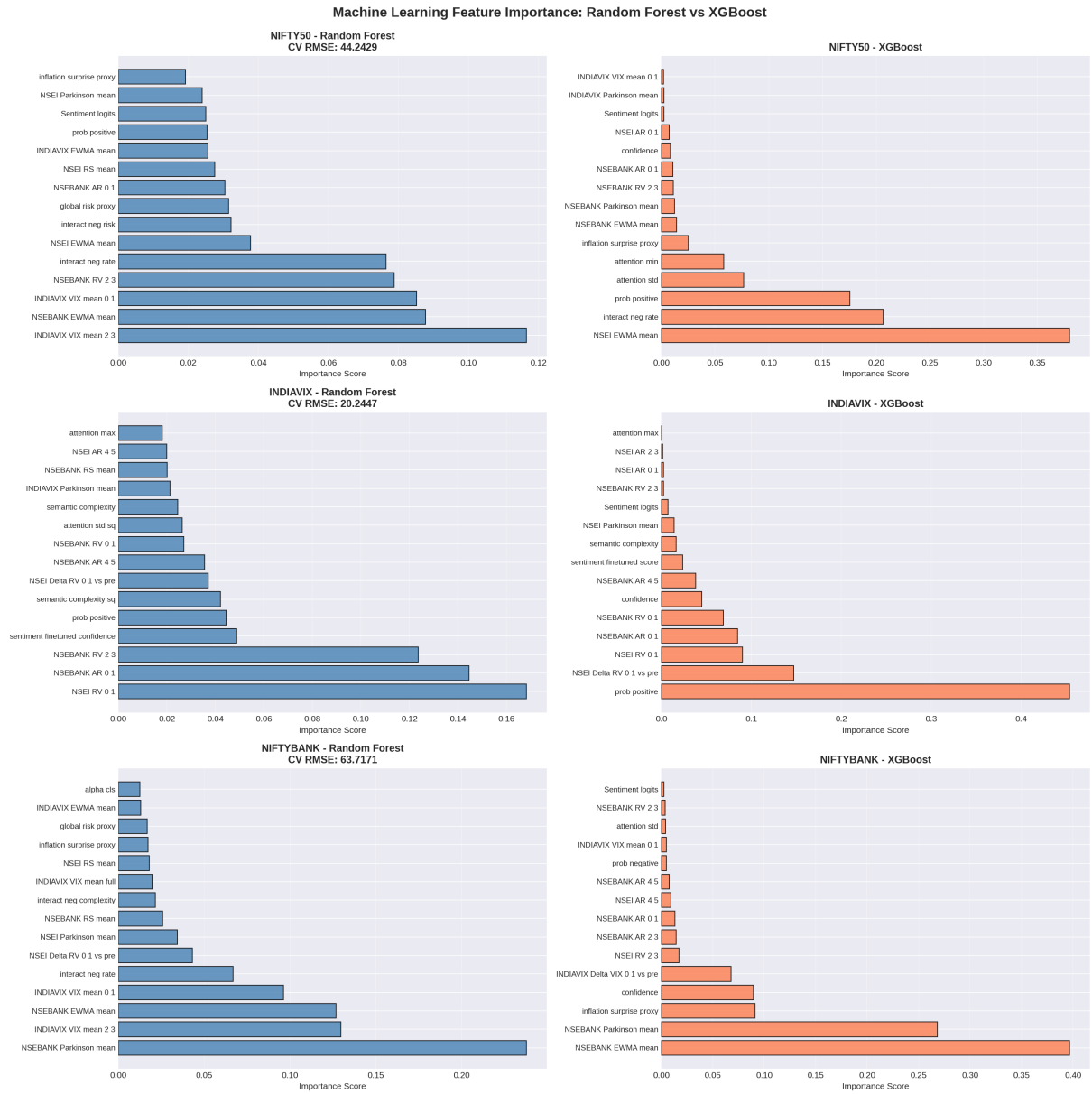


Figure 8: Feature importance comparison across Random Forest and XGBoost models for minute-level volatility spike prediction. Six panels show importance rankings for three indices (NIFTY50, INDIIVIX, NIFTYBANK) using two ensemble methods. Lagged volatility proxies (Parkinson range, EWMA mean, realized volatility metrics) consistently rank highest across all models, with importance scores ranging from 0.10 to 0.35. Content-derived features (sentiment logits, semantic complexity, attention metrics) appear in lower rankings with importance scores of 0.02–0.06. The rankings are consistent across Random Forest (blue) and XGBoost (orange) and stable across random seeds; we treat them as descriptive. Cross-validated RMSE values are reported in panel titles. Key observation: for NIFTYBANK, lagged NSEBANK EWMA mean and Parkinson mean are the two strongest predictors (importance > 0.20), while sentiment and complexity features rank positions 8–15. This descriptive pattern is consistent with the separate out-of-sample result that volatility persistence dominates; textual features receive smaller fitted-model importance.

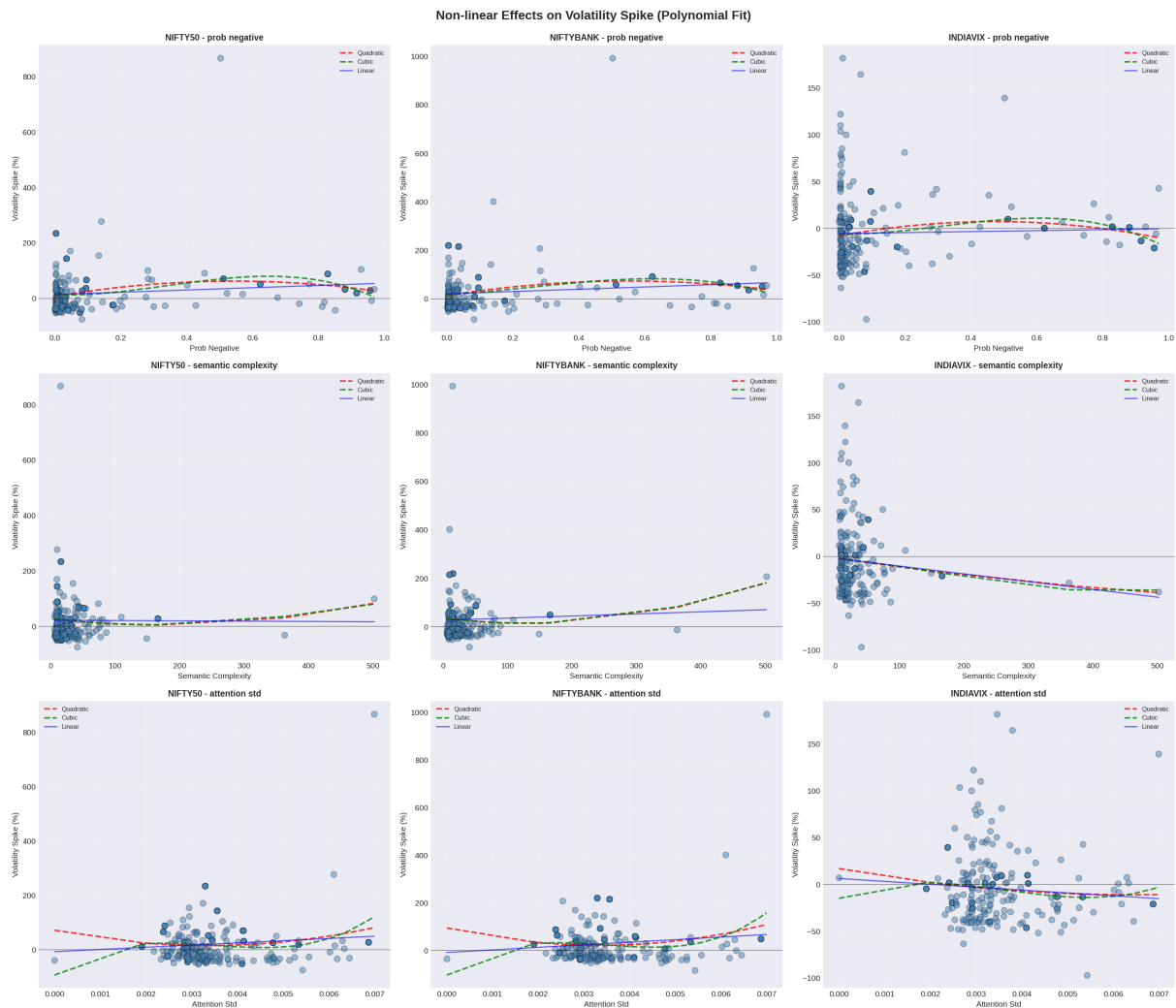


Figure 9: Non-linear effects of textual features on volatility spikes via polynomial regression. Nine panels show scatter plots with fitted curves for three textual predictors (probability negative, semantic complexity, attention std) across three indices. Each panel overlays linear (blue), quadratic (red dashed), and cubic (green dashed) fits. **Top row (prob_negative):** Near-linear relationships across all indices; polynomial fits provide minimal improvement over linear, indicating no threshold or saturation effects for sentiment polarity. **Middle row (semantic_complexity):** Similarly, linear patterns; complexity effects do not exhibit non-monotonic behavior within the observed range (0–500). **Bottom row (attention_std):** An apparent inverted-U shape for NIFTY50 and NIFTYBANK, with the fitted curve peaking near $\text{attention_std} \approx 0.003\text{--}0.004$, and the opposite shape for INDIAVIX. These curves are shown for completeness only. The underlying bivariate attention–volatility association does not survive rank or bootstrap inference (§4.5), polynomial expansions do not improve out-of-sample prediction, and the apparent curvature is driven by the same small set of high-leverage event dates; we therefore draw no mechanistic conclusion from it.

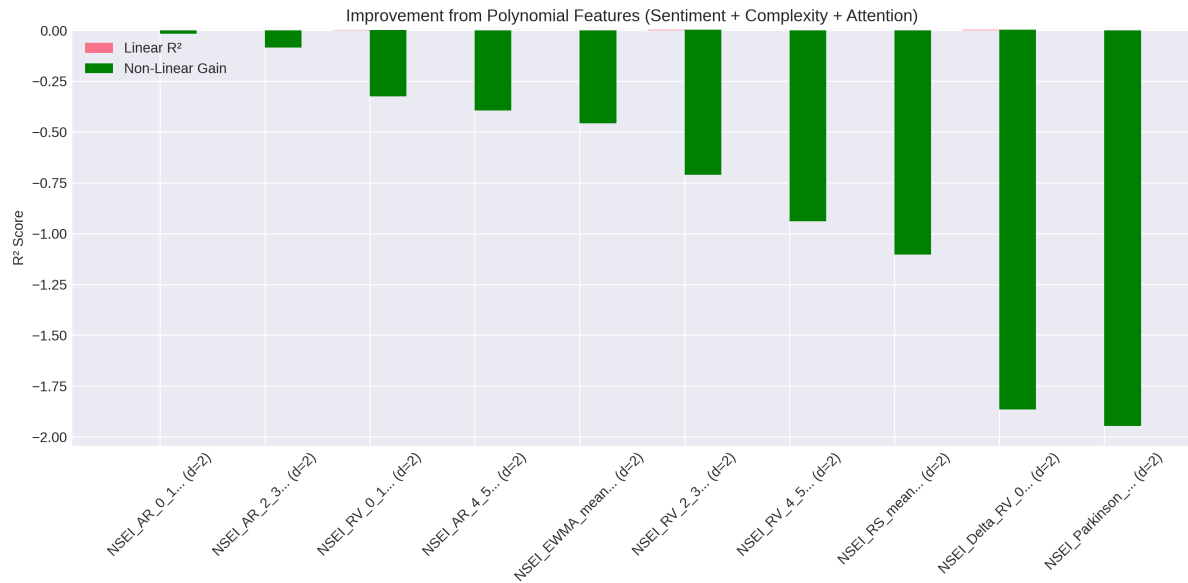


Figure 10: Cross-validated performance (RMSE) for linear vs. polynomial feature models. Polynomial expansions consistently degrade out-of-sample prediction across all volatility targets despite improving in-sample R^2 . This overfitting pattern motivates using parsimonious interaction terms rather than high-capacity nonlinear expansions, particularly given limited sample size and regime non-stationarity.

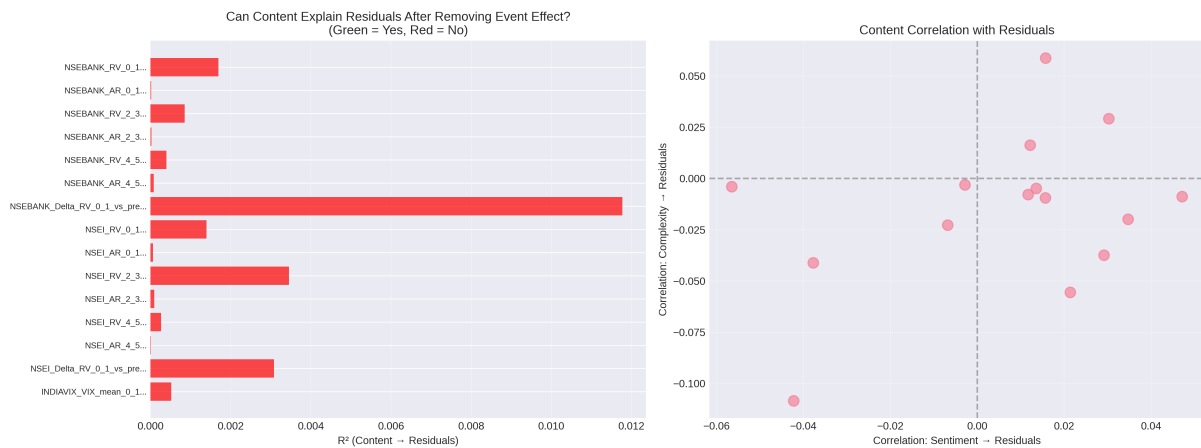


Figure 11: Residual analysis after removing event effects. **Left panel:** R^2 of content features regressed on residual volatility. All bars are red (non-significant), with R^2 values below 1.2% across all targets. NSEBANK Delta RV residuals show the highest (but still negligible) content sensitivity. **Right panel:** Scatter of sentiment-residual correlations (x-axis) vs. complexity-residual correlations (y-axis) across all volatility targets. Points cluster near the origin, indicating that neither sentiment nor complexity shows a systematic association with residual volatility.

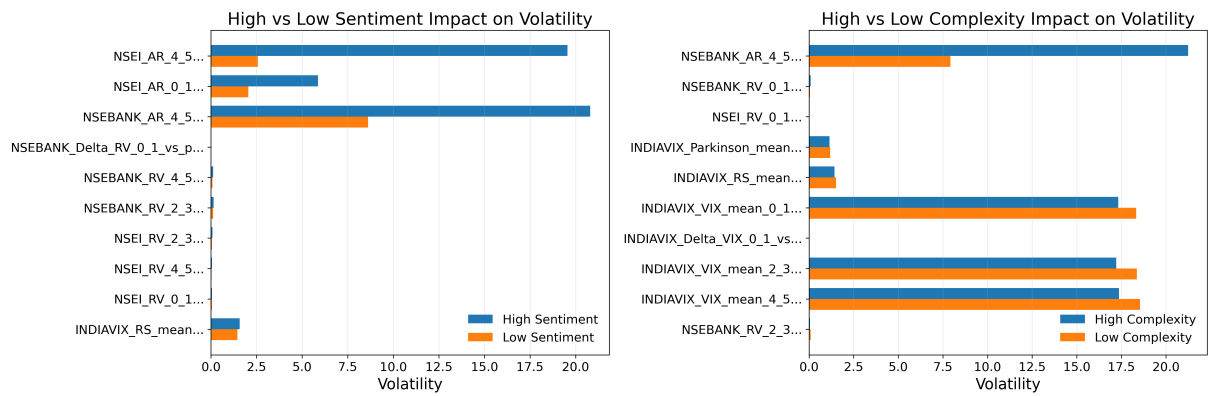


Figure 12: Volatility distributions stratified by sentiment intensity and semantic complexity (median split). High-sentiment documents (left panel) and high-complexity documents (right panel) exhibit elevated volatility across multiple metrics, though patterns are not universal across all measures.

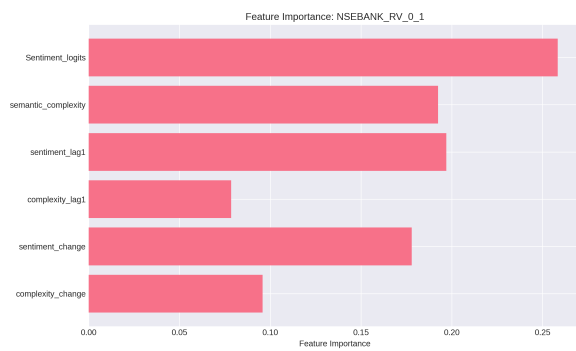


Figure 13: Feature importance scores from Random Forest regression predicting NIFTY BANK realized volatility (Days 0–1). The analysis focuses solely on content-derived features, excluding lagged volatility proxies. Sentiment logits emerge as the most important textual feature (0.26), followed by lagged sentiment (0.20) and semantic complexity (0.19). Semantic complexity ranks higher in-sample at the daily than the minute horizon; however, this is an in-sample importance pattern only—its out-of-sample contribution is not statistically significant (Appendix B.2).