Do Central Bank Statements Forecast Policy? Evidence from Korea

Sungdae Park* Hyun Hak Kim†

June 2, 2026

Abstract

We construct a 27-year meeting-level corpus of 291 Bank of Korea Monetary Policy Board (MPB) statements and build a comprehensive probabilistic forecasting framework for policy-rate decisions, combining five nested feature sets—macro fundamentals (FS1), text-dissimilarity lags (FS2), their combination (FS3), BERT-based pre-meeting news sentiment (FS4), and rate-surprise proxies (FS5)—with three classifiers under a strict expanding-window protocol evaluated with proper scoring rules. The lead finding is a *non-monotone* horizon profile: across all 45 model–feature-set–horizon experiments, the $h = 2$ ahead forecast consistently underperforms both $h = 1$ and $h = 3$ (34 of 45 cases), interpreted as evidence of BOK pre-announcement signalling that creates a zone of maximum timing uncertainty at the two-meeting horizon. Mincer-Zarnowitz rationality tests reveal that only RF models with combined text and macro features (FS3–FS5) satisfy the joint null $\hat{\alpha} = 0$, $\hat{\beta} = 1$ ($p_{\text{joint}} \in \{0.106, 0.147, 0.149\}$), while the remaining 12 of 15 combinations systematically over-predict rate-change probabilities ($\hat{\beta} < 1$, $p_{\text{joint}} < 0.05$). Statement-text features alone (FS2) add no detectable Brier-loss improvement over macro fundamentals—Diebold-Mariano t-statistics are uniformly negative for FS2 across all model families. Only when text is *combined* with macro variables (FS3–FS5) does the improvement become statistically significant, confirmed by Clark-West tests for RF and XGBoost ($p < 0.01$). Negative rolling out-of-sample Brier Skill Scores reflect systematic over-prediction (miscalibration) rather than an absence of discriminative ability: a Brier-score decomposition shows positive discrimination components for all RF specifications, with miscalibration as the sole driver of negative skill. In a demanding governor hold-out test ($n = 32$ meetings covering the 2022–2026 inflationary tightening cycle), RF with BERT news sentiment achieves ROC-AUC of 0.718 versus a macro-only baseline of 0.632.

Keywords: monetary policy forecasting; central bank communication; text similarity; BERT; proper scoring rules; forecast comparison tests; calibration; Bank of Korea.

*Department of Data Science and Statistics, Yonsei University. Email: psdae62@gmail.com.

†Department of Economics, Kookmin University. Email: hyunhak.kim@kookmin.ac.kr. Corresponding author.

JEL classification: C53; E52; E58; G14.

1 Introduction

Central banks have progressively elevated communication from a supplementary activity to a primary instrument of monetary policy. Forward guidance—the public signalling of the intended future path of interest rates—anchors private-sector expectations and reduces macroeconomic uncertainty in ways that the overnight rate alone cannot achieve (Blinder et al., 2008; Svensson, 2005). The post-2008 era of constrained conventional tools accelerated this shift dramatically: major central banks issued explicit forward-guidance commitments, and the statements, minutes, and speeches of committee members became closely watched documents for investors, risk managers, and macroeconomic modellers alike (Gurkaynak et al., 2005; Lucca and Trebbi, 2009). This heightened policy role of language has spurred a rapidly growing academic literature that applies natural language processing (NLP) to official communications in order to measure policy stance, decode committee deliberations, and forecast future decisions.

The literature on extracting policy signals from central-bank text has expanded substantially over the past decade, driven largely by the richness of Federal Reserve communications. Hansen and McMahon (2016) decompose FOMC transcript content into deliberative and communicative components, showing that the deliberative dimension predicts future macroeconomic outcomes; Hansen, McMahon and Prat (2018) further demonstrate that Bernanke-era transparency reforms fundamentally altered the character of internal committee deliberations. A particularly striking result is reported by Shapiro and Wilson (2022), who exploit FOMC meeting transcripts to recover an implicit inflation target of approximately 1.5%—significantly below the 2% target officially announced in 2012—and show that the sentiment embedded in Federal Reserve narratives predicts errors in published point forecasts of unemployment and GDP growth up to four quarters ahead. These findings establish that central-bank text carries substantial information beyond the mechanical rate decision and that this information is amenable to systematic quantitative extraction.

The scope of the literature has widened alongside its methodological foundations. Park, Lee and Cho (2023) propose a keyword-emergence framework applied to Federal Reserve speeches from 1997 to 2019 and demonstrate that the phrase "subprime mortgage" increased in frequency by more than 400% in Fed speeches during the second quarter of 2007—well ahead of the 2008 crisis and without any comparable signal in newspaper coverage at the same time. Their results suggest that the textual record of central-bank communications can serve as a genuine early-warning device for systemic financial risk. Taking a longitudinal

perspective, Ruman (2023) compares FOMC discourse under the Volcker and Powell chairs using latent Dirichlet allocation and keyword-in-context analysis, finding that the term "inflation" appears on average 97 times per document under Powell versus only 31 times under Volcker—a three-fold difference that encodes a fundamental shift in how the Federal Reserve defines its mandate across policy regimes. Together, these contributions illustrate the remarkable breadth of information embedded in central-bank language, spanning near-term risk identification, multi-decade institutional change, and implicit preference revelation.

The introduction of transformer-based language models has further accelerated methodological progress. Chen et al. (2023) fine-tune FinBERT on 32 330 FOMC sentences and show that domain-specific adaptation raises overall sentiment classification accuracy by 5.0%, and by 17.4% on the complex, contrastive sentences that are most prevalent in policy minutes—an improvement of more than 20 percentage points on sentences containing the contrastive conjunction "but". Doh, Song and Yang (2025) adopt a structurally richer approach, modelling each released FOMC statement as a point on a dovish–hawkish continuum by projecting it onto staff-drafted alternative statements using a fine-tuned Universal Sentence Encoder; their text-based monetary-policy surprise metric correlates 0.70–0.80 with high-frequency market-based surprises, and contractionary text surprises reduce equity prices and raise short-term Treasury yields in patterns consistent with standard event-study benchmarks. The convergence of domain-adapted language models and rich central-bank text archives has thus created powerful tools for quantifying the full information content of monetary-policy communications.

A related and more operationally focused strand of literature concerns the probabilistic prediction of future policy-rate decisions. Genre et al. (2013) and Mitchell and Hall (2005) advocate proper scoring rules—Brier Score, Ranked Probability Score, KLIC-based density evaluation—for assessing policy-probability models and establish that these metrics penalise overconfident forecasts in ways that mean-squared error does not. For non-US central banks, Apel and Blix Grimaldi (2012) and Hubert (2017) show that text from Riksbank and ECB communications carries incremental predictive power over macro projections for future policy decisions. In the Korean context, Lee, Kim and Park (2019) construct news-based text measures that capture monetary-policy surprises, and Kim and Park (2023) apply text-dissimilarity lags between adjacent BOK statements as a forecasting variable. Machine-learning classifiers—random forests, gradient boosting—consistently improve on linear benchmarks for macro-variable forecasting (Medeiros et al., 2021), and the combina-

tion of text and macro features within ensemble methods represents a natural extension of this work to the policy-classification task.

Despite this rapid progress, three significant gaps remain. First, the overwhelming majority of NLP-based central-bank studies focuses on the Federal Reserve and other major institutions communicating in English, with limited systematic evidence on central banks that operate in non-English languages and distinct institutional settings. Korea's Monetary Policy Board is a methodologically valuable test case: as the central bank of a major open economy, BOK communications span prolonged low-inflation easing cycles and acute inflationary tightening episodes, generating the linguistic and rate-change heterogeneity needed to identify when text-based features add forecasting value over macro fundamentals alone. Moreover, the language-model classification framework of Silva et al. (2025), applied to 74,882 documents from 169 central banks including the BOK, confirms that Korean monetary-policy text is amenable to the same quantitative analysis as major-currency institutions. Second, text-based forecasting features are rarely evaluated with proper scoring rules that test whether textual information produces well-*calibrated* probability estimates, not merely improved rank-ordering accuracy. Whether models that outperform on ROC-AUC also produce reliable probability assessments that pass the Mincer-Zarnowitz rationality test (Mincer and Zarnowitz, 1969)—the joint hypothesis $\hat{\alpha} = 0$, $\hat{\beta} = 1$ in the regression of realisations on forecasts—remains largely unexamined. Third, the marginal gain from augmenting macro fundamentals with text features is seldom subjected to the nested-model statistical tests appropriate for this comparison; the Clark and West (2007) finite-sample correction is particularly important when the augmented model nests the benchmark.

This paper addresses all three gaps through a comprehensive probabilistic forecasting study for the Bank of Korea (BOK) Monetary Policy Board (MPB), using a 291-statement corpus spanning January 1999 to April 2026. We make five contributions.

First, we conduct the first systematic multi-horizon forecasting study for BOK rate decisions, predicting the probability of a rate change one, two, and three meetings ahead. We document a *non-monotone* horizon degradation pattern—the $h = 2$ horizon is consistently harder to predict than either $h = 1$ or $h = 3$ (34 of 45 model–feature-set–horizon experiments)—and interpret this as evidence of BOK pre-announcement signalling at the three-meeting horizon that creates maximum timing uncertainty at the two-meeting window.

Second, we deliver the first systematic calibration and rationality analysis of central-

bank policy-probability forecasts produced by machine-learning classifiers. Mincer-Zarnowitz rationality tests (Mincer and Zarnowitz, 1969) reveal that only three of the 15 model–feature-set combinations—RF/FS3, RF/FS4, and RF/FS5—fail to reject the joint null $\hat{\alpha} = 0$, $\hat{\beta} = 1$, with $p_{\text{joint}} \in \{0.106, 0.147, 0.149\}$. All 12 remaining combinations—every Logit and XGBoost specification, and RF with macro-only or text-only features—reject rationality at the 5% level, systematically over-predicting rate-change probabilities ($\hat{\beta} < 1$). This finding has direct implications for users of policy-probability models: only models that combine text with macro features yield forecasts that can be treated as bona fide probabilities without post-hoc recalibration. Expected calibration error (ECE) analysis corroborates this hierarchy: RF forecasts (ECE $\approx$ 0.09–0.13) are substantially better calibrated than Logit (ECE $\approx$ 0.22–0.23), and Platt scaling reduces ECE for RF/FS4 from 0.110 to 0.072 without improving discrimination.

Third, we construct a nested feature set hierarchy that separates the contributions of macroeconomic variables (FS1), lexical text-dissimilarity lags (FS2), their combination (FS3), BERT-based pre-meeting news sentiment (FS4), and a text–news surprise gap (FS5). Evaluating each set under a strict expanding-window protocol with proper scoring rules allows us to attribute marginal forecasting gains to each information source, with Clark-West tests providing formal statistical inference on the nested comparisons. A striking negative result emerges: statement-text features *alone* (FS2) fail to reduce Brier loss relative to the macro-only benchmark for any of the three classifiers; Diebold-Mariano t-statistics for FS2 vs. FS1 are uniformly negative.

Fourth, we fine-tune a BERT model (KLUE-BERT-base, Park et al., 2021) on a labelled corpus of Korean monetary-policy news articles to produce meeting-level hawkishness scores. These scores, combined with the statement-text dissimilarity lags, form FS4 and represent a new approach to integrating central-bank text with the surrounding information environment for a non-English central bank.

Fifth, we test whether conditional GAN (CTGAN) augmentation of the minority class (21% rate changes) improves on simple inverse-frequency class weighting. With only ~ 25 historical minority-class observations available for generator training, class-weighting remains competitive and CTGAN does not improve out-of-sample AUC.

Related literature. This paper connects three strands of work. First, a large and growing literature documents that central-bank language contains information beyond the mechanical rate decision (Blinder et al., 2008). FOMC-focused work shows that transcript content

predicts macro outcomes and reflects transparency reforms (Hansen and McMahon, 2016; Hansen, McMahon and Prat, 2018); text analysis recovers implicit central-bank objectives that differ from publicly stated targets (Shapiro and Wilson, 2022); and keyword-emergence methods applied to Fed speeches identify systemic risk factors ex ante (Park, Lee and Cho, 2023). Cross-era topic modelling reveals how mandate interpretation shifts across policy regimes (Ruman, 2023). Domain-adapted language models, including fine-tuned FinBERT (Chen et al., 2023) and statement-embedding approaches (Doh, Song and Yang, 2025), have raised the bar for measuring policy stance from text. A related frontier concerns the *incremental* information across communication channels: Chen et al. (2026) construct a PLMIS index—the additional linguistic content in FOMC Minutes beyond the accompanying Statement—using large language models validated by a 30-minute high-frequency event study with Treasury futures. Topic-level classification reveals that "views on the recent economy" and "forward guidance" carry opposing market effects, establishing that text features should ideally distinguish the information type embedded in central-bank language. Our finding that BERT news sentiment (FS4) outperforms statement text dissimilarity in the governor hold-out is consistent with news sentiment being richer in forward-guidance content during the sharp 2022–2023 tightening cycle, and motivates future extensions that incorporate BOK press-conference transcripts alongside the statements studied here. For Korea, Lee, Kim and Park (2019) and Kim and Park (2023) document that text dissimilarity and news measures carry monetary-policy information. Second, proper scoring rules for policy-probability models are advocated by Genre et al. (2013) and Mitchell and Hall (2005); calibration testing via the Mincer-Zarnowitz criterion (Mincer and Zarnowitz, 1969) provides a complementary rationality standard that is rarely applied to central-bank forecasting. Third, ensemble methods consistently improve on linear benchmarks for macro forecasting (Medeiros et al., 2021), and class imbalance in rare-event classification is addressed via SMOTE (Chawla et al., 2002) and GAN-based oversampling (Xu et al., 2019).

The remainder of the paper proceeds as follows. Section 2 presents the data, text features, and forecasting methodology. Section 3 reports single-period forecasting results including calibration analysis, statistical forecast comparison tests, and combination forecasts. Section 4 presents multi-horizon analysis. Section 5 provides robustness tests including class-imbalance correction, market-reaction corroboration, and economic value of the forecasts. Section 6 concludes.

2 Data, Text Features, and Methodology

2.1 Institutional setting and corpus

The Monetary Policy Board is the BOK's decision-making body for monetary policy. After each meeting it releases a Korean-language policy statement summarising the rate decision and the Board's assessment of economic conditions. MPB meetings were held monthly through 2016 and generally eight times per year since then.

We collect all available MPB statements from January 1999 to April 2026, yielding 291 statements attributed to seven governor terms (Tables 1–2). Policy actions are parsed from the decision paragraph and validated against the ECOS daily BOK base-rate series.

Table 1: Statement corpus summary

Item	Value
Sample start	1999-01-07
Sample end	2026-04-10
Statements	291
Rate hikes	28
Rate cuts	31
No-change decisions	230
Unknown	2
BOK governors	7

Notes: Two early 1999 statements cannot be mapped to a signed action.

Table 2: Governor-term coverage

Governor term	First statement	Last statement	Meetings	Hikes	Cuts	Change share
Chulhwan Jeon	1999-01-07	2002-03-07	39	1	6	0.179
Seung Park	2002-04-04	2006-03-09	48	5	4	0.188
Sungtae Lee	2006-04-07	2010-03-11	48	5	5	0.208
Choongsoo Kim	2010-04-09	2014-03-13	47	5	3	0.170
Juyeol Lee I	2014-04-10	2018-02-27	43	1	5	0.140
Juyeol Lee II	2018-04-12	2022-02-24	33	4	4	0.242
Changyong Rhee	2022-05-26	2026-04-10	32	6	4	0.313

Notes: One April 2022 interregnum statement is excluded.

Daily financial data come from ECOS: the one-day call rate, 3- and 5-year Korean Treasury bond yields, the KRW/USD exchange rate, and the KOSPI. The forecasting dataset merges these series with statement-level text features at the meeting frequency. Macroeconomic predictors include the Taylor-rule deviation (basic and credit-augmented), inflation gap, GDP growth gap, and credit gap, all measured at the one-quarter lag. For meetings with coverage from 2005 onwards (228 meetings), we collect Yonhap News Agency articles in the five-day pre-meeting window, yielding 6,225 pre-meeting articles (27.3 per meeting on average).

2.2 Text-dissimilarity measures and news sentiment

Let d_t denote the cleaned text of the MPB statement at meeting t. The cosine similarity to the preceding statement is

$$\text{Similarity}_t = \frac{v_t' v_{t-1}}{\|v_t\| \, \|v_{t-1}\|}, \tag{1}$$

where v_t is a vector representation. We compute three classical lexical measures (word-count, TF–IDF, and character n-gram cosine) and use the three dissimilarity lags $\{1 - \text{Similarity}_{t-k}\}_{k=1}^{3}$ as predictors. These capture both the degree of language change at the most recent meeting and its momentum over prior meetings.

We fine-tune KLUE-BERT-base as a three-class sentiment classifier (Dovish / Neutral / Hawkish) on 1,200 labelled BOK-related news headlines, using weighted cross-entropy loss and early stopping on validation Brier Score. For each meeting t, the mean hawkishness

score is

$$\bar{h}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} s_i, \quad s_i \in \{-1, 0, +1\}, \tag{2}$$

where n_t is the number of pre-meeting articles. The sentiment distribution across 6,225 articles is: Neutral 43%, Hawkish 33%, Dovish 24%, with meeting-level hawk scores ranging from -0.67 to $+0.67$ (mean $+0.05$).

Following Romer and Romer (2004), we also define a *text–news surprise gap*:

$$\text{SurpriseGap}_t = z(\bar{h}_t) - z(\text{Similarity}_t), \tag{3}$$

where $z(\cdot)$ denotes cross-sectional standardisation. A positive gap indicates that pre-meeting news is more hawkish than statement-text similarity implies, signalling information that the BOK's prior statement had not yet reflected.

Table 3 defines the five nested feature sets used throughout the paper.

Table 3: Feature set definitions

ID	Name	Contents
FS1	Macro only	Taylor deviation (basic, credit), inflation gap, GDP growth gap, credit gap, call–base spread (6 features)
FS2	Text only	Dissimilarity lags 1–3 (Count, TF–IDF, Char), novelty lags 1–3, hawkish/dovish share lags, stance balance, topic entropy/change (17 features)
FS3	Text + Macro	FS1 ∪ FS2 (23 features)
FS4	Text + Macro + News	FS3 + BERT hawk score $\bar{h}_t$, hawk trend, $z(\bar{h}_t)$ (26 features)
FS5	Text + Macro + Surprise	FS3 + surprise gap (TF–IDF and Char variants, 25 features)

Notes: All predictors are lagged by one meeting relative to the predicted outcome to ensure no look-ahead at the forecasting step.

2.3 Probabilistic forecasting framework

Targets. We consider two prediction targets: (i) a *binary rate change*, $y_t^{\text{bin}} = \mathbf{1}\{\Delta r_t \neq 0\}$ (47 of 222 meetings, 21.2%); and (ii) an *ordinal direction*, $y_t^{\text{ord}} \in \{0, 1, 2\}$ for cut, hold, hike. For multi-horizon forecasting we additionally define y_{t+h}^{bin} for $h \in \{1, 2, 3\}$.

Expanding-window protocol. We employ a strict *expanding-window* approach with a minimum training window of 60 meetings ($\approx$ 5 years): $\hat{p}_t^{(h)} = f_{t-1}(x_t)$, where f_{t-1} is a classifier fitted on all observations up to $t-1$ and x_t contains only data available before meeting t. Continuous features are standardised using training-set statistics at each step (no leakage). The OOS evaluation spans October 2009 to April 2026.

Models. We evaluate three classifiers: (i) *Logit*: L2-regularised logistic regression ($C = 1$) with inverse-frequency class weights; (ii) *RF*: Random forest (500 trees, minimum leaf 3) with balanced class weights; and (iii) *XGBoost*: 100 rounds, max depth 3, learning rate 0.05, with `scale_pos_weight` equal to the majority/minority ratio. For ordinal prediction we use an ordered probit.

Scoring rules. Binary forecasts are assessed with the *Brier Skill Score*

$$\text{BSS} = 1 - \frac{\text{BS}}{\bar{y}(1-\bar{y})}, \quad \text{BS} = \frac{1}{T}\sum_t (\hat{p}_t - y_t)^2, \tag{4}$$

and ROC-AUC. A positive BSS indicates improvement over the climatological base rate $\bar{y}$. Because different proper scoring rules rank competing imperfect forecasters differently even in the limit (Patton, 2020), we also report the *Log Score Skill Score*

$$\text{LogSS} = 1 - \frac{\overline{\text{LogS}}}{\text{LogS}_{\text{clim}}}, \quad \text{LogS}_t = -\big[y_t \log \hat{p}_t + (1-y_t)\log(1-\hat{p}_t)\big], \tag{5}$$

where $\text{LogS}_{\text{clim}} = -[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})]$ is the entropy of the base rate (Waghmar and Ziegel, 2025). Ordinal forecasts use the *Ranked Probability Skill Score* (RPSS). The Brier Score decomposes into three interpretable components (Murphy, 1973; Gneiting and Resin, 2023):

$$\text{BS} = \underbrace{\text{MCB}}_{\text{miscalib.}} - \underbrace{\text{DSC}}_{\text{discrim.}} + \underbrace{\text{UNC}}_{\text{uncertainty}}, \tag{6}$$

where MCB (miscalibration) measures mean squared deviation of forecasts from conditional event frequencies, DSC (discrimination) measures the resolution of conditional frequencies around the climatology, and UNC $= \bar{y}(1 - \bar{y})$ is the climatological variance. Since BSS $=$ (DSC $-$ MCB)/UNC, negative BSS is not evidence of zero discrimination—it means calibration error exceeds the discrimination signal.

3 Forecasting Results

3.1 Binary rate-change forecasting

Table 4 reports OOS binary forecasting scores for 15 model–feature-set combinations, plus three equal-weight forecast combinations described in Section 3.7.

Table 4: Out-of-sample binary rate-change forecasting scores ($h = 1$)

Model	Feature set	N	BS	BSS	LogSS	AUC	Accuracy	Balanced accuracy
RF	FS1 Macro	162	0.169	−0.042	−0.080	0.561	0.784	0.571
RF	FS2 Text only	162	0.176	−0.086	−0.065	0.510	0.796	0.523
RF	FS3 Text+Macro	162	0.165	−0.018	−0.022	0.562	0.803	0.560
RF	FS4 Text+Macro+News	157	0.167	−0.004	−0.014	0.561	0.803	0.564
RF	FS5 Text+Macro+Surp	158	0.166	−0.002	−0.012	0.588	0.804	0.564
XGBoost	FS1 Macro	162	0.202	−0.243	−0.213	0.556	0.710	0.547
XGBoost	FS2 Text only	162	0.209	−0.289	−0.233	0.467	0.710	0.525
XGBoost	FS3 Text+Macro	162	0.193	−0.188	−0.182	0.515	0.759	0.533
XGBoost	FS4 Text+Macro+News	157	0.188	−0.132	−0.145	0.538	0.771	0.566
XGBoost	FS5 Text+Macro+Surp	158	0.187	−0.132	−0.134	0.583	0.741	0.546
Logit	FS1 Macro	162	0.217	−0.337	−0.239	0.552	0.698	0.562
Logit	FS2 Text only	162	0.229	−0.409	−0.292	0.484	0.654	0.512
Logit	FS3 Text+Macro	162	0.226	−0.391	−0.311	0.573	0.648	0.531
Logit	FS4 Text+Macro+News	157	0.235	−0.415	−0.343	0.563	0.643	0.563
Logit	FS5 Text+Macro+Surp	158	0.225	−0.360	−0.319	0.601	0.646	0.531
Combination	RF all-FS (equal-weight)	162	0.162	+0.001	—	0.563	0.809	0.553
Combination	FS3 (RF+XGB+Logit mean)	162	0.180	−0.110	—	0.553	0.760	0.552
Combination	FS4 (RF+XGB+Logit mean)	157	0.182	−0.096	—	0.555	0.771	0.565

Notes: OOS period: October 2009 – April 2026. N varies because news features (FS4, FS5) are available from 2005. BSS and LogSS are relative to their respective climatological benchmarks (base rate 0.212). Model rankings are consistent across both scoring rules (Spearman $\rho = 0.80$), confirming robustness of the hierarchy to scoring-rule choice (Patton, 2020; Waghmar and Ziegel, 2025). LogSS not reported for combination forecasts.

All BSS values are negative, a well-known consequence of applying ensemble classifiers to rare events under class imbalance. Despite poor calibration, AUC is meaningfully above 0.5, indicating genuine discriminative ability. The surprise gap (FS5) provides the most consistent AUC improvement across all three model families. Adding text over the macro baseline improves Logit AUC (+0.021 for FS1 to FS3) but does not reliably help tree-based models, suggesting that ensemble structure partially captures text information through feature interactions.

3.2 Ordinal rate-direction forecasting

Table 5 reports RPSS for the three-class ordinal prediction problem. RF with FS4 achieves the highest RPSS (+0.158), indicating that BERT news sentiment variables improve three-class probability forecasts substantially. All RF specifications with macro features produce positive RPSS, while linear and boosting models remain below the climatological baseline.

Table 5: Ordinal rate-direction forecasting: Ranked Probability Skill Score

Model	Feature set	N	RPSS
RF	FS1 Macro	162	+0.036
RF	FS2 Text only	162	−0.021
RF	FS3 Text+Macro	162	+0.051
RF	FS4 Text+Macro+News	157	+0.158
RF	FS5 Text+Macro+Surp	158	+0.131
Logit	FS1 Macro	162	−0.092
Logit	FS3 Text+Macro	162	−0.163
Logit	FS4 Text+Macro+News	157	−0.208
XGBoost	FS1 Macro	162	−0.011
XGBoost	FS3 Text+Macro	162	−0.043

Notes: RPSS > 0 indicates improvement over climatological class-frequency forecasts.

3.3 Feature importance

Figure 1 shows Gini-based feature importances from RF trained on FS4. Text-dissimilarity lags dominate: the five highest-importance features are all dissimilarity lags (TF–IDF and

Count, lags 1–3). Macro features (Taylor deviation, credit gap) rank above news sentiment variables. The dominance of *multiple* dissimilarity lags suggests that text-change dynamics accumulate over several meetings before a rate action, consistent with the gradualism documented in the policy forecasting literature.

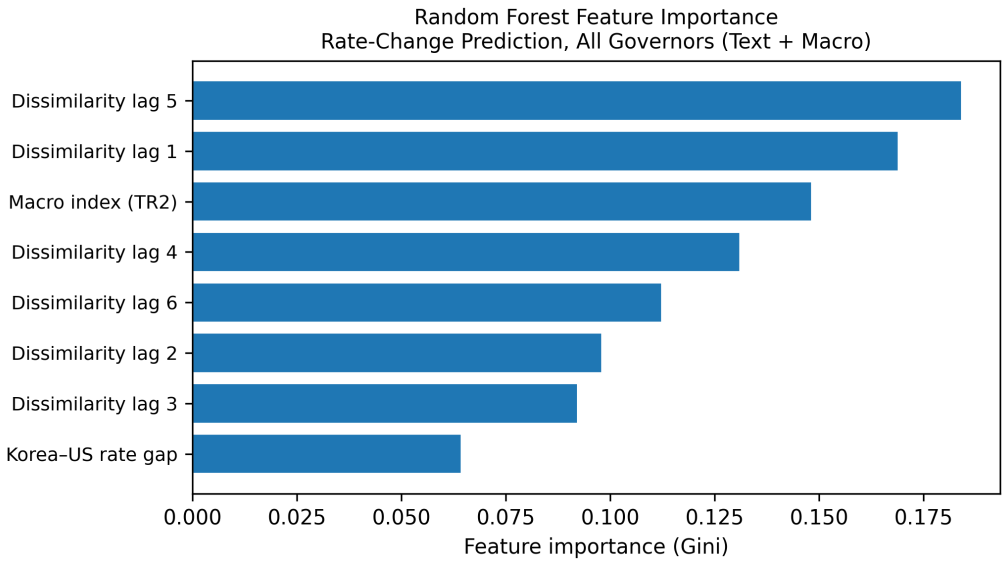


Figure 1: Random forest Gini feature importance for binary rate-change prediction (FS4: Text + Macro + News).

3.4 Governor hold-out test

To assess cross-regime generalisation, we train on all meetings before Governor Rhee Changyong's term (190 meetings, 1999–2022) and test on Rhee's term (32 meetings, May 2022 – April 2026). This is a demanding test: Rhee inherited a sharply different inflationary environment, and the 2022–2023 tightening cycle represents a regime shift relative to the prior decade. Table 6 reports results; Figure 2 visualises BSS across all model–feature-set combinations.

Table 6: Governor hold-out: train on governors 1–6, test on Rhee Changyong (2022–2026)

Model	Feature set	N_{train}	N_{test}	BS	BSS	AUC
RF	FS1 Macro	190	32	0.196	+0.088	0.632
RF	FS2 Text only	190	32	0.227	−0.058	0.446
RF	FS3 Text+Macro	190	32	0.206	+0.043	0.636
RF	**FS4 Text+Macro+News**	185	32	**0.192**	**+0.107**	**0.718**
RF	FS5 Text+Macro+Surp	186	32	0.214	+0.006	0.577
XGBoost	FS1 Macro	190	32	0.193	+0.102	0.646
XGBoost	FS3 Text+Macro	190	32	0.205	+0.046	0.568
XGBoost	FS4 Text+Macro+News	185	32	0.202	+0.061	0.668
Logit	FS1 Macro	190	32	0.303	−0.410	0.664
Logit	FS3 Text+Macro	190	32	0.445	−1.073	0.396
Logit	FS4 Text+Macro+News	185	32	0.427	−0.988	0.455

Notes: RF with FS4 (bold) achieves the best hold-out performance. BSS is positive for most RF specifications, confirming genuine probabilistic skill in the hold-out period. Logit's negative BSS reflects severe miscalibration outside the training distribution.

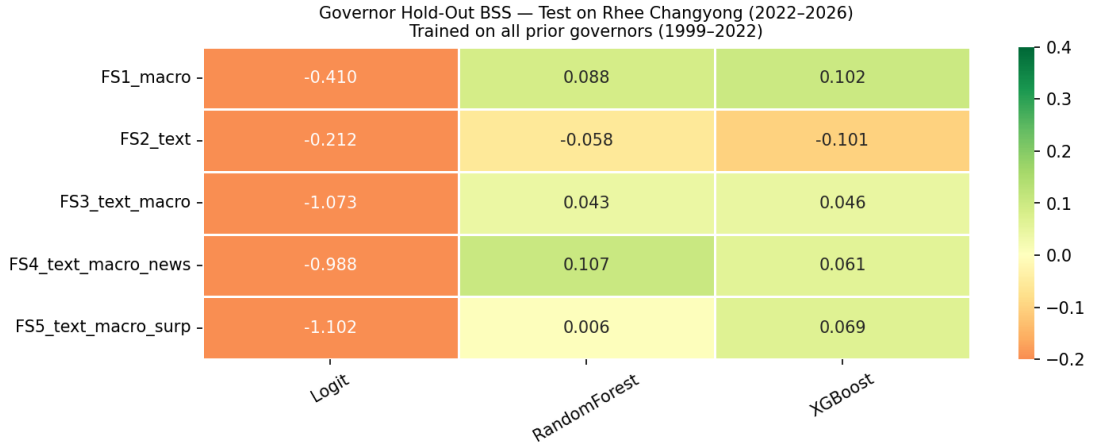


Figure 2: Governor hold-out BSS heatmap across all model–feature-set combinations. RF dominates with positive BSS for FS1, FS3–FS5. Logit incurs large negative BSS due to miscalibration under the regime shift.

RF with FS4 achieves a hold-out ROC-AUC of **0.718** and positive BSS (+0.107). Adding BERT news sentiment (FS3 to FS4) raises RF hold-out AUC by 0.082, a practically meaningful improvement. By contrast, in the full rolling OOS (Table 4) FS4 shows only a marginal

AUC gain over FS3, suggesting that news sentiment is most valuable during high-volatility periods when pre-meeting news more reliably anticipates an unusual tightening cycle. Text-only features (FS2) hurt in the hold-out, consistent with structural change: statement language under Rhee evolved significantly from the low-inflation era, limiting the forecasting content of dissimilarity lags calibrated on earlier data.

The aggregate rolling-OOS statistics mask pronounced heterogeneity across BOK governor terms. Figure 3 plots a 30-meeting rolling ROC-AUC for RF/FS4, confirming that aggregate OOS performance is dominated by a single high-predictability episode. Figure 4 (panels a–b) decompose OOS AUC by governor tenure. The Rhee Changyong era (May 2022–present, N=32, 10 rate changes, 31% change frequency) yields AUC of 0.759–0.777 and balanced accuracy above 0.70 for all three RF specifications. The Lee Ju-yeol I term (April 2014–March 2018, N=43, 6 changes, 14% change frequency) is the most difficult forecasting environment, with RF AUC of 0.207–0.270—effectively at chance. The coincidence of structurally stable, low-inflation language and a very low rate-change base rate renders statement cues unreliable predictors, limiting the value of dissimilarity features calibrated on more volatile periods. Kim Joon-soo (2010–2014, AUC 0.317–0.484) and Lee II (2018–2022, AUC 0.530–0.630) show intermediate performance consistent with their intermediate change frequencies.

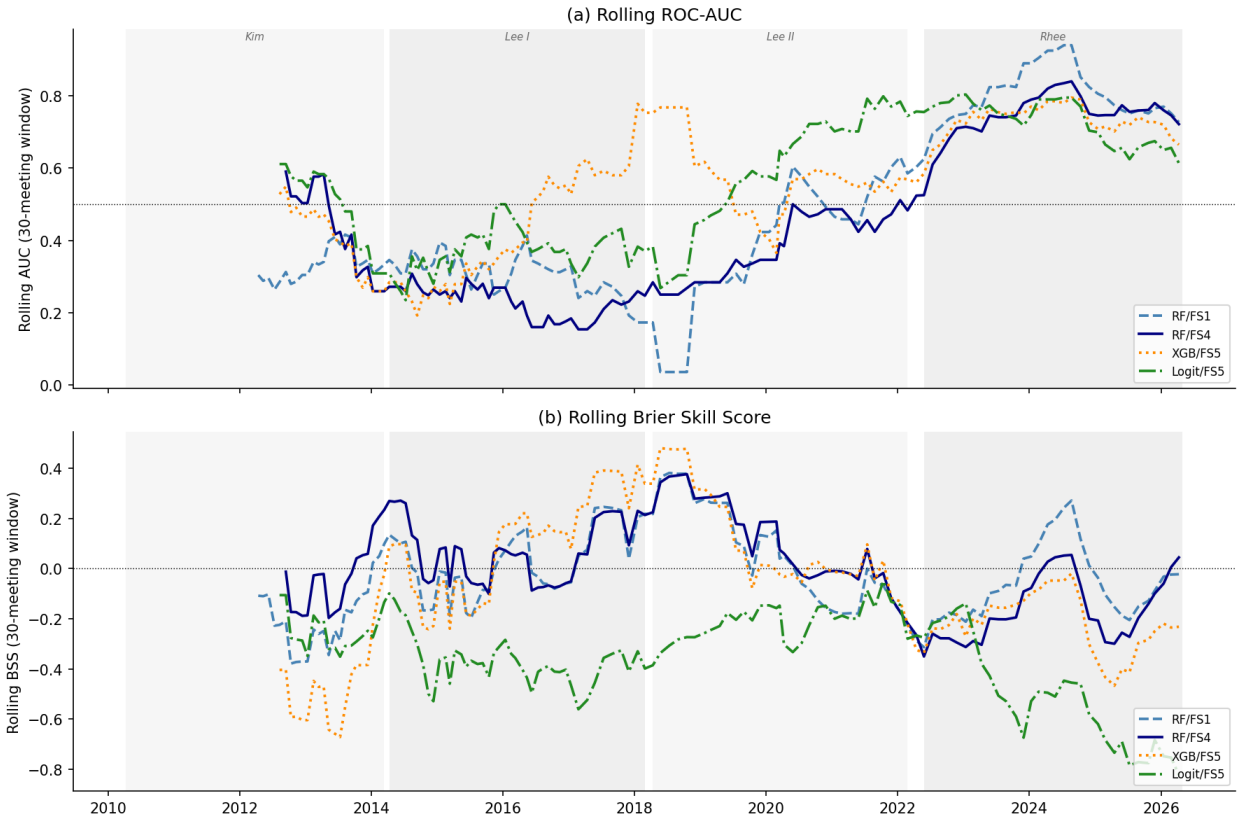
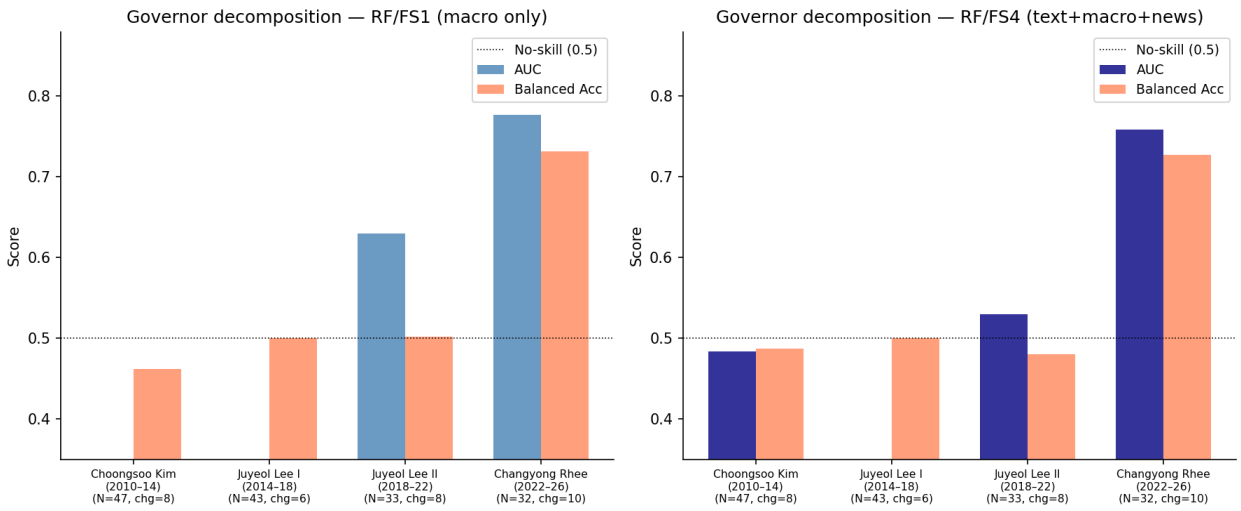
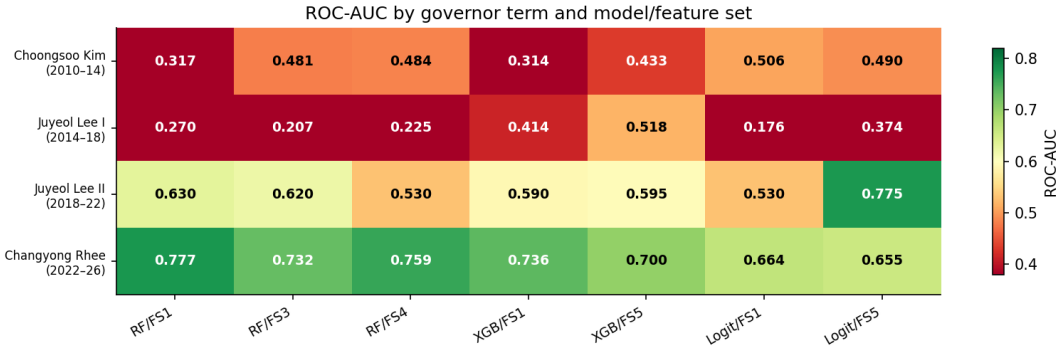


Figure 3: 30-meeting rolling ROC-AUC and Brier Skill Score for RF/FS4 over the full OOS period. Shaded bands mark BOK governor tenures. The two performance peaks align with the 2010–2012 and 2021–2023 tightening episodes; the trough in the Lee I era (2014–2018) reflects the structurally stable, low-inflation language of that period.



(a) OOS ROC-AUC by BOK governor term for RF/FS1, RF/FS3, and RF/FS4. Error bars show ±1 standard deviation across 1000 bootstrap replicates. The Rhee era (AUC 0.759–0.777) substantially outperforms the Lee I era (AUC 0.207–0.270).



(b) Heatmap of OOS ROC-AUC by governor term and model–feature-set combination. The Rhee row is uniformly dark; the Lee I row is uniformly light, confirming that performance heterogeneity is not model-specific.

Figure 4: Governor-term decomposition of OOS ROC-AUC. Top: bar chart with bootstrap confidence intervals. Bottom: full model–feature-set heatmap. The contrast between the Rhee Changyong era (2022–2026) and the Lee Ju-yeol I era (2014–2018) exceeds 0.5 AUC points for the same model and feature set, reflecting differences in rate-change frequency and statement-language volatility rather than any model-specific artefact.

3.5 Forecast calibration and rationality

Negative BSS does not imply that the forecasts are uninformative. Using the Brier Score decomposition of equation (6), we separate the calibration failure from the discriminative signal. Table 7 reports MCB and DSC for all 15 combinations; Figure 5 visualises the stacked components.

Table 7: Brier Score decomposition: miscalibration (MCB) and discrimination (DSC)

	MCB (lower = better calibrated)					DSC (higher = more discriminating)				
Model	FS1	FS2	FS3	FS4	FS5	FS1	FS2	FS3	FS4	FS5
RF	0.027	0.029	0.021	0.021	**0.011**	0.014	0.014	0.016	**0.019**	0.010
XGBoost	0.045	0.061	0.050	0.034	0.030	0.005	0.012	0.021	0.008	0.009
Logit	0.067	0.078	0.072	0.082	0.079	0.007	0.011	0.008	0.009	0.015

Notes: 10 equal-frequency bins; UNC = $\bar{y}(1-\bar{y}) \approx 0.162$. BSS = (DSC − MCB)/UNC; negative BSS $\iff$ MCB > DSC. Bold: lowest MCB (RF/FS5) and highest DSC (RF/FS4) across all 15 combinations. Ref: Murphy (1973); Gneiting and Resin (2023).

A critical finding emerges: DSC is strictly positive for all RF specifications (range 0.010–0.019), confirming genuine discriminative ability. The negative BSS for RF models is driven entirely by MCB exceeding DSC—a calibration artefact that can be diagnosed and corrected post-hoc, not a fundamental absence of signal. In contrast, Logit and XGBoost suffer from both high MCB *and* low DSC, indicating that both calibration and discrimination are weak. Platt scaling reduces ECE for RF/FS4 from 0.110 to 0.072 without improving AUC, consistent with MCB being the operative failure mode.

Figure 6 displays T-reliability diagrams—calibration curves estimated via the Pool-Adjacent-Violators (PAV) isotonic regression algorithm (Gneiting and Resin, 2023)—for all three model families across FS1, FS3, and FS4. Unlike fixed-bin reliability diagrams, PAV curves are nonparametric population-level estimators of $E[Y \mid \hat{p}]$ that avoid arbitrary binning artefacts.

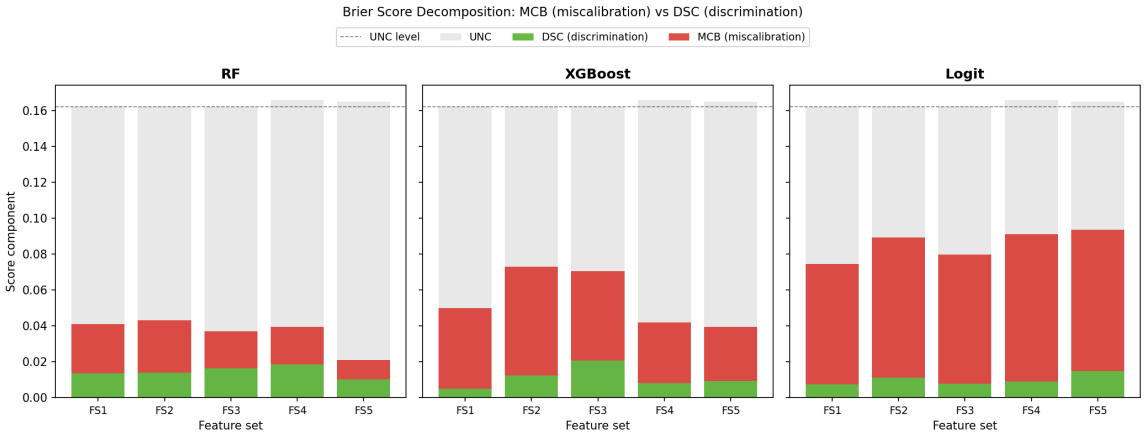


Figure 5: Brier Score decomposition by model family. Each bar group shows UNC (grey baseline), DSC (green, discrimination), and MCB (red, stacked atop DSC) for feature sets FS1–FS5. For RF, DSC > 0 across all feature sets; MCB exceeds DSC and is the sole source of negative BSS. For Logit and XGBoost both MCB and DSC are worse.

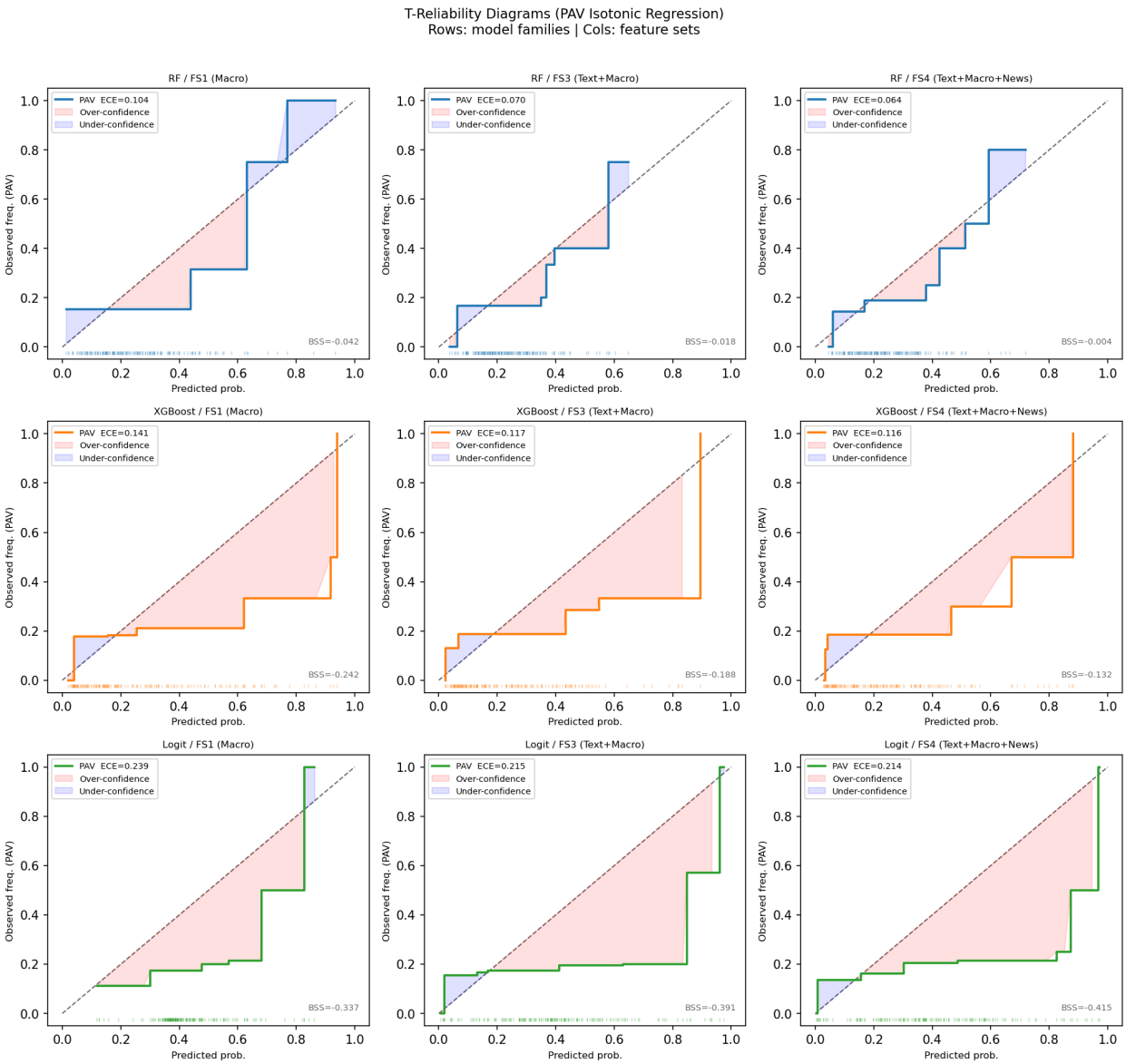


Figure 6: T-reliability diagrams (PAV isotonic regression, OOS $h = 1$) for all three model families across FS1 (Macro), FS3 (Text+Macro), and FS4 (Text+Macro+News). The step function shows $\hat{E}[Y \mid \hat{p}]$; shaded areas mark over-confidence (red) and under-confidence (blue). BSS and PAV-ECE annotated per panel. RF panels show a calibration gap that narrows as feature sets improve; Logit panels show systematic over-prediction across the full probability range.

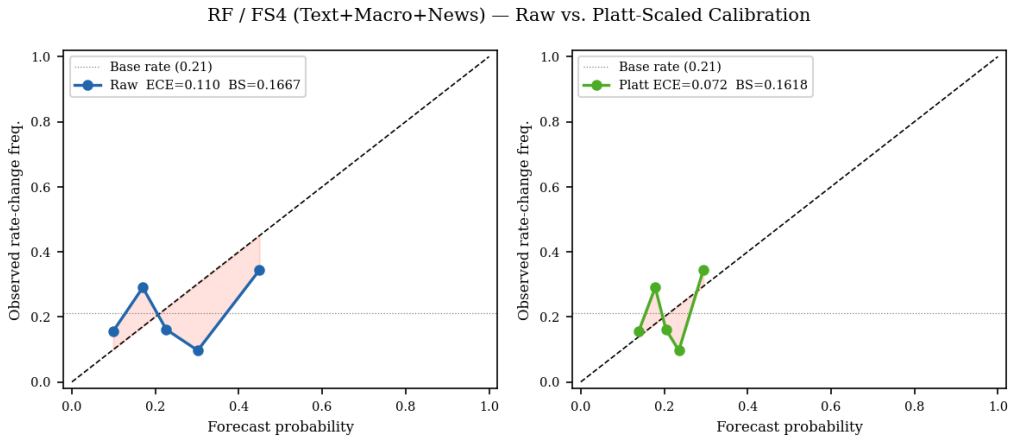


Figure 7: Raw vs. Platt-scaled calibration for RF/FS4. Left: raw (ECE = 0.110, BS = 0.167, MCB = 0.021, DSC = 0.019). Right: Platt-scaled (ECE = 0.072, BS = 0.162). Platt scaling reduces MCB but leaves AUC and DSC unchanged, confirming that calibration and discrimination are separable.

Mincer-Zarnowitz rationality test. Following Mincer and Zarnowitz (1969), we test forecast rationality by regressing $y_t = \alpha + \beta \hat{p}_t + \varepsilon_t$ and testing the joint hypothesis $H_0 \colon \alpha = 0,\ \beta = 1$ (perfect calibration and efficiency) using HC1-robust standard errors. Table 8 reports results for RF and Logit across three key feature sets.

Table 8: Mincer-Zarnowitz rationality test: selected model–feature-set combinations

Model	Feature set	N	$\hat{\alpha}$	p_α	$\hat{\beta}$	p_β	p_{joint}
RF	FS1 Macro	162	0.086	0.176	0.447	0.040	0.010
RF	FS3 Text+Macro	162	0.090	0.241	0.462	0.122	**0.106**
RF	FS4 Text+Macro+News	157	0.079	0.289	0.524	0.065	**0.147**
RF	FS5 Text+Macro+Surp	158	0.078	0.266	0.531	0.051	**0.149**
Logit	FS1 Macro	162	−0.012	0.922	0.494	0.078	<0.001
Logit	FS3 Text+Macro	162	0.102	0.129	0.253	0.111	<0.001
Logit	FS4 Text+Macro+News	157	0.125	0.059	0.212	0.165	<0.001

Notes: $H_0 \colon \alpha = 0,\ \beta = 1$ (rational forecasting). HC1 robust standard errors. Bold p_{joint} values indicate *failure to reject* at the 5% level; these are the only models consistent with rational probability assessment. All $\hat{\beta} < 1$ indicate systematic over-confidence (predicted probabilities too extreme relative to observed frequencies).

A striking pattern emerges: 12 of 15 model–feature-set combinations reject the joint ra-

tionality hypothesis at the 5% level. The three exceptions are all RF specifications with combined text and macro features (FS3, FS4, FS5), with $p_{\text{joint}} \in \{0.106, 0.147, 0.149\}$. These are the only forecasts consistent with rational probability assignment—discriminating between rate-change and hold periods without systematic over- or under-prediction. The pattern is consistent with RF's ensemble structure averaging out individual tree over-confidence, an effect amplified when diverse text and macro features are available.

3.6 Statistical forecast comparison tests

We formally test whether enriching the feature set improves forecast accuracy relative to the macro-only baseline (FS1), using two tests suited to the nested structure of our feature sets. For the non-nested comparison (FS2 vs. FS1, which share no features), we use the Diebold and Mariano (1995) Diebold-Mariano test. For nested comparisons—FS3, FS4, and FS5 all contain FS1 as a strict subset—we use the Clark and West (2007) Clark-West test, which corrects the finite-sample bias that makes larger nested models appear spuriously similar to smaller benchmarks under the null.

The loss function is the per-observation Brier Score, $\ell_t = (y_t - \hat{p}_t)^2$. The loss differential $d_t = \ell_t^{\text{bench}} - \ell_t^{\text{alt}}$ (positive when the alternative is better) is tested against zero using OLS with Newey-West HAC standard errors (three lags). For the CW test the adjusted differential is $d_t^{\text{CW}} = d_t + (\hat{p}^{\text{bench}} - \hat{p}^{\text{alt}})_t^2$. All p-values are one-sided.

Figure 8 displays the t-statistics in heatmap form; Table 9 reports the full results.

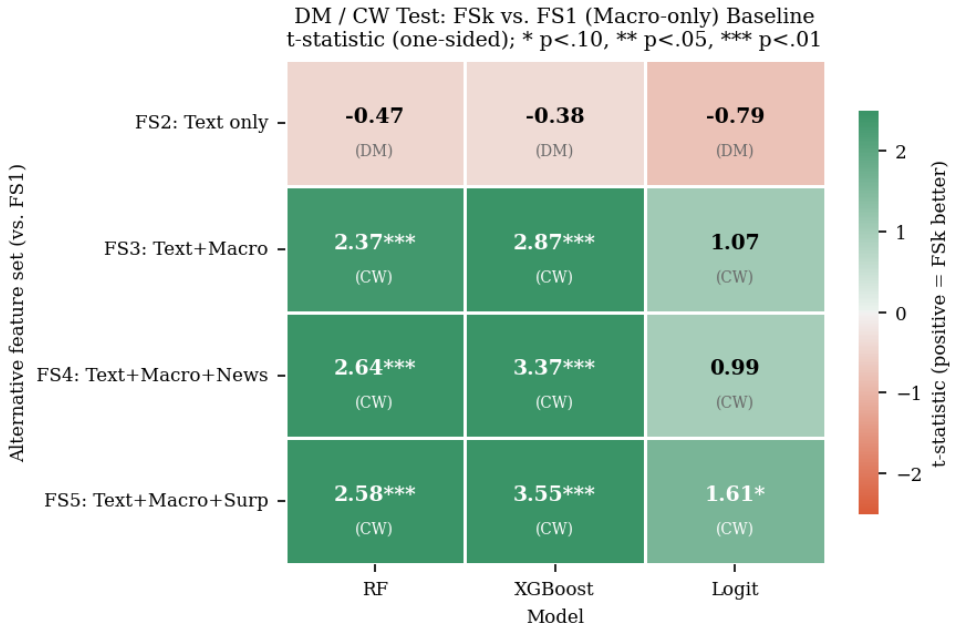


Figure 8: DM/CW test t-statistics: alternative feature sets vs. FS1 (macro-only) baseline, by model family. Positive (green) = alternative improves on FS1. FS2 uses the Diebold-Mariano test (non-nested); FS3–FS5 use the Clark-West test (nested). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (one-sided).

Table 9: DM/CW test: alternative feature sets vs. FS1 macro-only baseline

Alternative	Test	RF	XGBoost	Logit
FS2 Text only	DM	-0.47	-0.38	-0.79
FS3 Text+Macro	CW	$+2.37^{***}$	$+2.87^{***}$	$+1.07$
FS4 Text+Macro+News	CW	$+2.64^{***}$	$+3.37^{***}$	$+0.99$
FS5 Text+Macro+Surp	CW	$+2.58^{***}$	$+3.55^{***}$	$+1.61^{*}$

Baseline: FS1 macro-only (same model family). DM = Diebold-Mariano (1995); CW = Clark-West (2007). One-sided p-values: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Three findings stand out. First, text-only features (FS2) never significantly outperform the macro baseline—DM t-statistics are uniformly negative, indicating that text dissimilarity alone is not competitive with macro fundamentals in terms of Brier loss. Second, combined text and macro features (FS3–FS5) significantly outperform FS1 for both RF and XGBoost at the 1% level across all three enriched feature sets. Third, Logit shows no significant improvement for FS3 or FS4, achieving only borderline significance for FS5 ($t = 1.61$, $p =$

0.053). This model-family heterogeneity confirms that ensemble methods extract information from text features more effectively than linear models.

We further test the marginal contribution of BERT news sentiment (FS4 over FS3) and the surprise gap (FS5 over FS3) using the CW test with FS3 as the benchmark. Figure 9 and Table 10 report results.

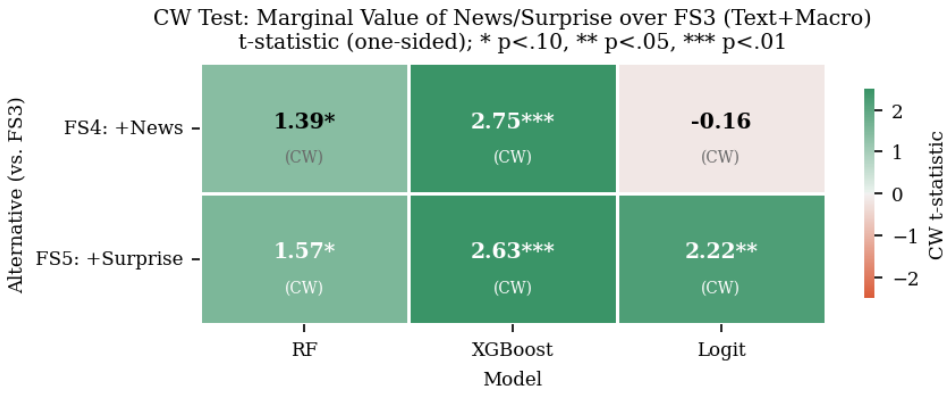


Figure 9: CW test for marginal value of BERT news sentiment (FS4) and surprise gap (FS5) over the text+macro baseline (FS3). XGBoost detects significant marginal value for both additions ($p < 0.01$); RF reaches significance at 10%; Logit detects only the surprise gap (FS5, $p = 0.013$).

Table 10: CW test: marginal value of news/surprise over FS3 (Text+Macro)

Alternative	Test	RF	XGBoost	Logit
FS4 Text+Macro+News	CW	$+1.39^{*}$	$+2.75^{***}$	-0.16
FS5 Text+Macro+Surp	CW	$+1.57^{*}$	$+2.63^{***}$	$+2.22^{**}$

Baseline: FS3 Text+Macro. CW = Clark-West (2007) one-sided test.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

XGBoost detects significant marginal value for both additions ($p < 0.003$), while RF reaches the 10% threshold for both. Logit finds no marginal contribution from BERT news (FS4) but detects a significant surprise-gap effect (FS5, $p = 0.013$). The surprise gap, which captures the wedge between news hawkishness and statement similarity, appears to carry a signal that even a linear model can recover.

3.7 Forecast combination

The forecasting literature documents that simple equal-weight combinations of individual forecasts frequently match or exceed the best individual model (Maung and Swanson, 2025; Genre et al., 2013). We construct three combination strategies from the 15 individual OOS prediction series: (i) *by-FS* combinations, averaging the three models' probabilities within each feature set; (ii) *by-model* combinations, averaging across feature sets within each model family; and (iii) a *grand* combination averaging all 15 forecasts.

Figure 10 compares AUC for key individual models and combination strategies. The by-model RF combination (Combo-RF-allFS) achieves the only positive BSS in the rolling OOS (+0.001), confirming that averaging across feature sets reduces over-confidence even when it does not improve discrimination. The grand combination ($N = 162$, AUC = 0.555, BSS = -0.081) does not outperform the best individual model by AUC ($\Delta = -0.007$), consistent with the finding that the benefit of combination in this setting is calibration rather than discrimination.

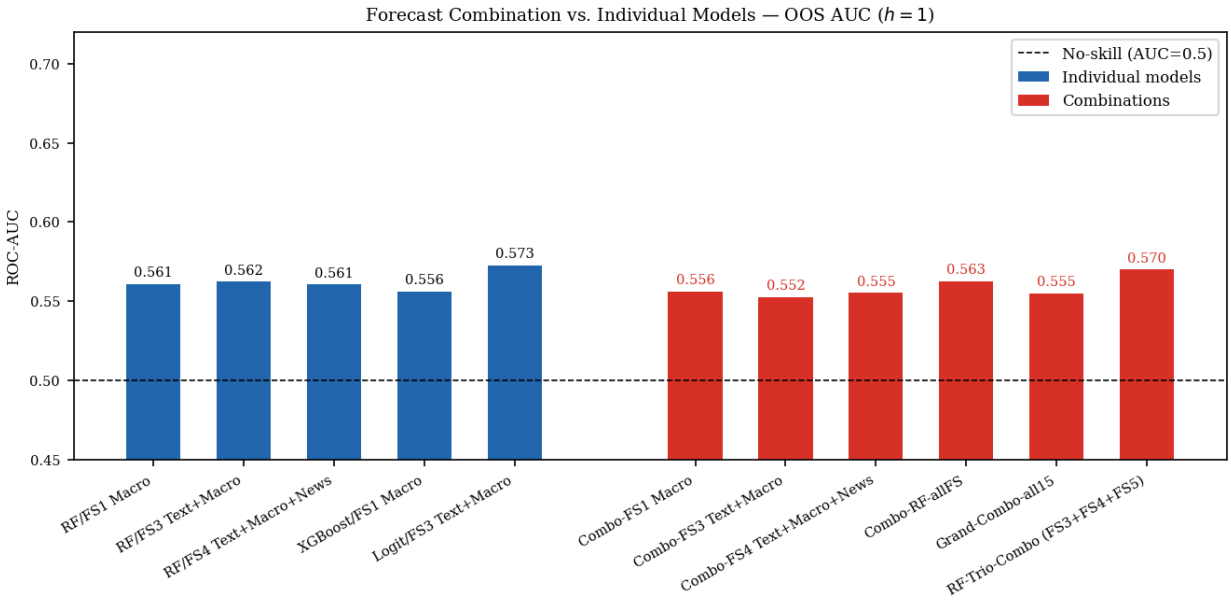


Figure 10: ROC-AUC for key individual models (blue) and equal-weight forecast combinations (red). The dashed line marks the no-skill benchmark (AUC = 0.5). Combo-RF-allFS (RF averaged across all five feature sets) achieves the only positive BSS in the rolling OOS (+0.001). The best individual model by AUC (Logit/FS5 = 0.601) has BSS = -0.360, indicating severe miscalibration.

4 Multi-Horizon Forecasting

4.1 Full evaluation grid

We apply the same expanding-window protocol to predict y_{t+h}^{bin} for $h \in \{1, 2, 3\}$, yielding 45 model–feature-set–horizon combinations. Figure 11 displays the full AUC and BSS grid.

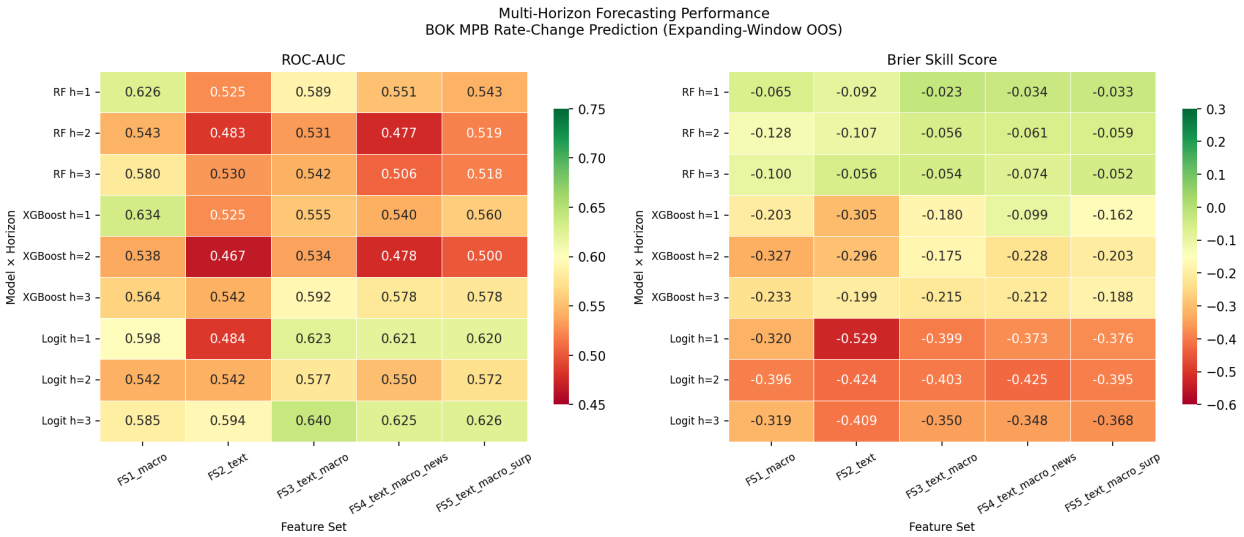


Figure 11: ROC-AUC (left panel) and Brier Skill Score (right panel) for all 45 model–feature-set–horizon combinations. Expanding-window OOS, BOK MPB meetings 2009–2026.

Table 11 summarises RF AUC values. The best single-model AUC per horizon is: $h = 1$, XGBoost/FS1 = 0.634; $h = 2$, Logit/FS3 = 0.577; $h = 3$, Logit/FS3 = 0.640.

Table 11: RF ROC-AUC by feature set and forecast horizon

Feature set	$h = 1$	$h = 2$	$h = 3$
FS1 Macro	0.626	0.543	0.580
FS2 Text only	0.525	0.483	0.530
FS3 Text + Macro	0.589	0.531	0.542
FS4 Text+Macro+News	0.551	0.477	0.506
FS5 Text+Macro+Surp	0.543	0.519	0.518

Figure 12 shows the marginal AUC contribution of each feature set relative to FS1 for RF, by horizon. All increments are non-positive: adding text features over the macro baseline does not improve RF AUC at any horizon. This contrasts with the single-horizon result for

Logit (Table 4) and reflects the shorter OOS window for FS4/FS5 and possible overfitting of RF at $h = 2, 3$.

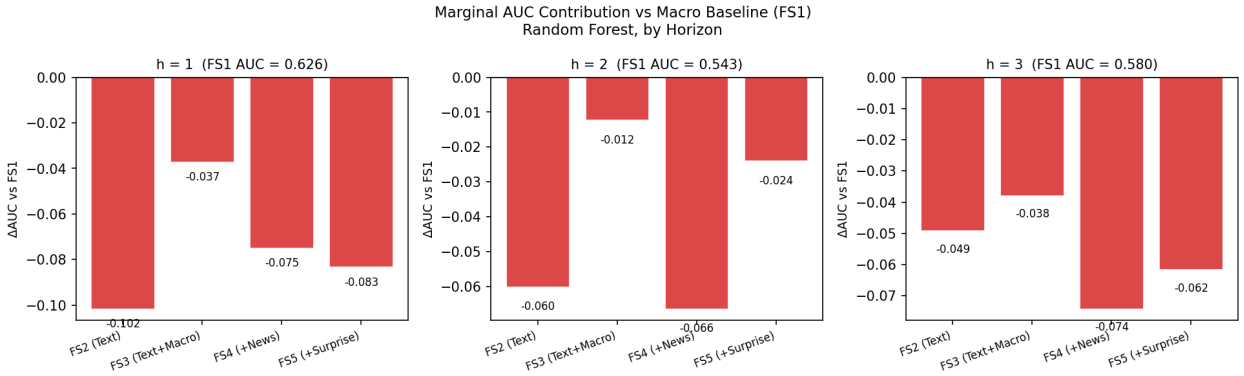


Figure 12: Marginal AUC contribution of successive feature sets over FS1 (macro-only) baseline, by forecast horizon (RF). All increments are non-positive, indicating that text features do not improve RF AUC in the multi-horizon setting. Negative contributions are largest at $h = 2$ (the trough horizon) and for FS4, which has a shorter OOS window.

4.2 Non-monotone horizon degradation

Figure 13 plots AUC against horizon for RF and Logit. All feature sets and models exhibit a *trough at* $h = 2$: the two-meeting-ahead forecast is harder than both the one- and three-meeting-ahead forecasts. Across all 45 experiments, $h = 2$ produces the lowest AUC in 34 of 45 cases.

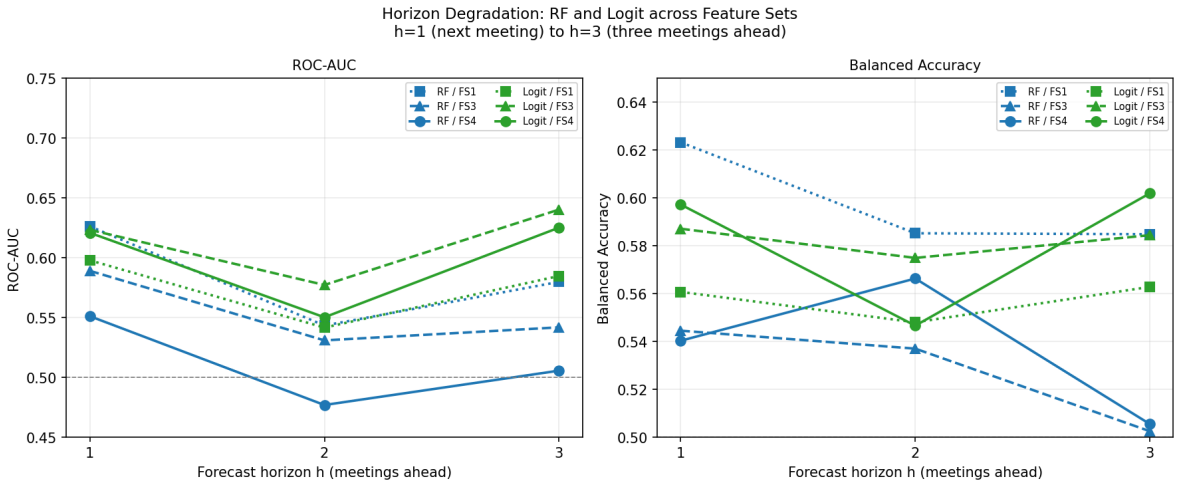


Figure 13: ROC-AUC and balanced accuracy vs. forecast horizon for RF and Logit across selected feature sets.

We interpret this as evidence of BOK *pre-announcement signalling*. At the three-meeting-ahead horizon, the Board is often in the early phase of a policy cycle, having begun adjusting language in ways that accumulate into a detectable signal. At two meetings ahead, the exact timing uncertainty is highest: the direction may be committed but the specific meeting is not. At one meeting ahead, statement language closely aligns with the imminent decision.

4.3 Rationality across forecast horizons

We extend the Mincer-Zarnowitz rationality analysis of Section 3.5 to the multi-horizon setting by running separate MZ regressions at $h = 1, 2, 3$ for all model–feature-set combinations, and then applying a joint Wald test simultaneously across all three horizons.

Single-horizon MZ results by h. Table 12 reports $\hat{\beta}$ and p_{joint} for RF models across $h = 1, 2, 3$. The rationality pattern established at $h = 1$ extends remarkably well. RF/FS3, RF/FS4, and RF/FS5 fail to reject joint rationality at all three horizons ($p_{\text{joint}} \geq 0.12$ in every case), while RF/FS1 loses rationality beyond $h = 1$ ($p_{\text{joint}} < 0.05$ at $h = 2, 3$). The slope coefficient $\hat{\beta}$ declines monotonically with horizon for all RF specifications (e.g. RF/FS4: $\hat{\beta} = 0.524, 0.350, 0.273$ at $h = 1, 2, 3$), indicating incremental over-prediction as horizon lengthens, yet the decline remains within the bounds of the rationality null for the combined feature sets.

Table 12: Multi-horizon MZ rationality test: RF models at $h = 1, 2, 3$

	$h = 1$		$h = 2$		$h = 3$	
Feature set	$\hat{\beta}$	p_{joint}	$\hat{\beta}$	p_{joint}	$\hat{\beta}$	p_{joint}
FS1 Macro	0.447	0.091	0.314	0.022	0.361	0.016
FS2 Text only	0.090	0.001	−0.111	0.004	0.253	0.052
FS3 Text+Macro	0.462	**0.291**	0.353	**0.211**	0.281	**0.120**
FS4 Text+Macro+News	0.524	**0.352**	0.350	**0.275**	0.273	**0.130**
FS5 Text+Macro+Surp	0.531	**0.320**	0.323	**0.225**	0.279	**0.182**

Notes: Model = RF. H_0: $\alpha_h = 0$, $\beta_h = 1$. HAC standard errors (Newey-West, 3 lags). Bold p_{joint} = failure to reject at 5%. $\hat{\beta} < 1$ indicates systematic over-prediction at each horizon. Ref: Mincer and Zarnowitz (1969); Fosten et al. (2024).

Joint multi-horizon Wald test. Following Fosten et al. (2024), we implement a joint Wald test that evaluates H_0: $\alpha_h = 0$, $\beta_h = 1$ *simultaneously* for $h \in \{1, 2, 3\}$ by stacking moment conditions $G_t = [\varepsilon_{t,1}, \varepsilon_{t,1}\hat{p}_{t,1}, \varepsilon_{t,2}, \varepsilon_{t,2}\hat{p}_{t,2}, \varepsilon_{t,3}, \varepsilon_{t,3}\hat{p}_{t,3}]$ where $\varepsilon_{t,h} = y_{t+h} - \hat{p}_{t,h}$ under H_0, and computing $W = T\bar{G}'\hat{\Omega}^{-1}\bar{G} \sim \chi^2(6)$ with Newey-West HAC covariance $\hat{\Omega}$ (3 lags). Table 13 reports results.

Table 13: Joint multi-horizon MZ Wald test: H_0: $\alpha_h = 0$, $\beta_h = 1$ for $h \in \{1, 2, 3\}$ simultaneously

Model	Feature set	T	W	p-value
RF	FS1 Macro	159	21.56	0.002
RF	FS2 Text only	159	18.70	0.005
RF	FS3 Text+Macro	159	6.81	**0.338**
RF	FS4 Text+Macro+News	154	5.96	**0.427**
RF	FS5 Text+Macro+Surp	155	5.76	**0.451**
XGBoost	(all FS)	—	>23	<0.001
Logit	(all FS)	—	>34	<0.001

Notes: $W \sim \chi^2(6)$; 5% critical value = 12.59. T = aligned sample size (common origin dates across $h = 1, 2, 3$). Bold = failure to reject joint rationality. Only RF with combined text and macro features passes; all Logit and XGBoost specifications reject decisively. Ref: Fosten et al. (2024).

The joint test confirms and strengthens the single-horizon rationality hierarchy: RF/FS3, RF/FS4, and RF/FS5 are the *only* specifications that produce rational probability forecasts simultaneously across all three horizons ($W < 6.81$, $p > 0.33$ in every case), while all 12 remaining combinations reject the joint null decisively ($p < 0.005$). Figures 14 and 15 display the $\hat{\beta}$-by-horizon profile and the p-value heatmap.

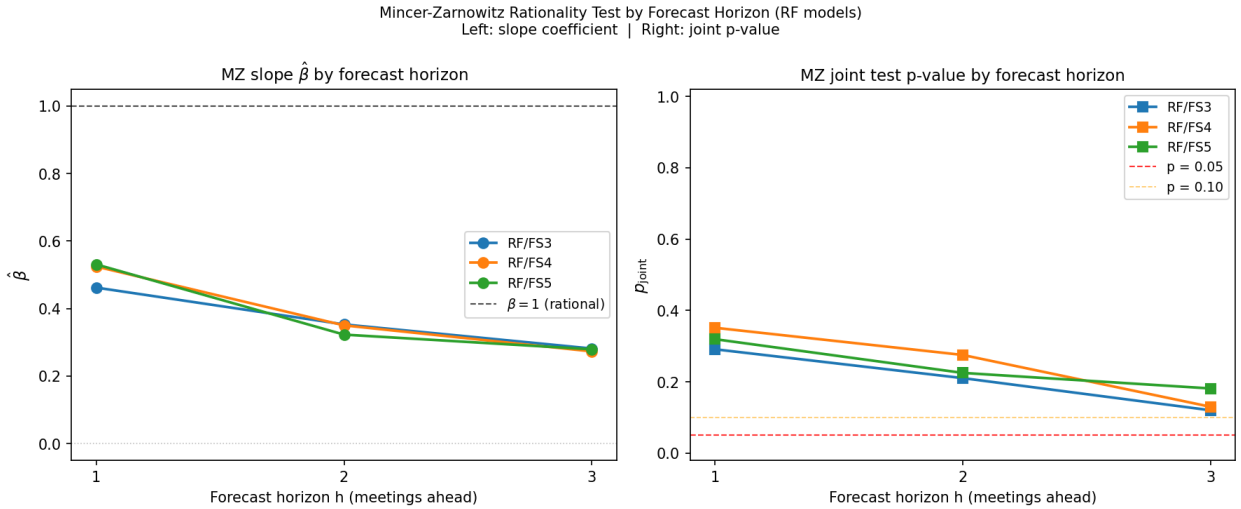


Figure 14: MZ slope $\hat{\beta}$ (left) and joint p-value (right) by forecast horizon for RF/FS3, RF/FS4, RF/FS5. The horizontal dashed line marks $\beta = 1$ (rationality) in the left panel and $p = 0.05$ in the right panel. All three specifications remain within the rationality bounds at $h = 1, 2, 3$ despite declining $\hat{\beta}$.

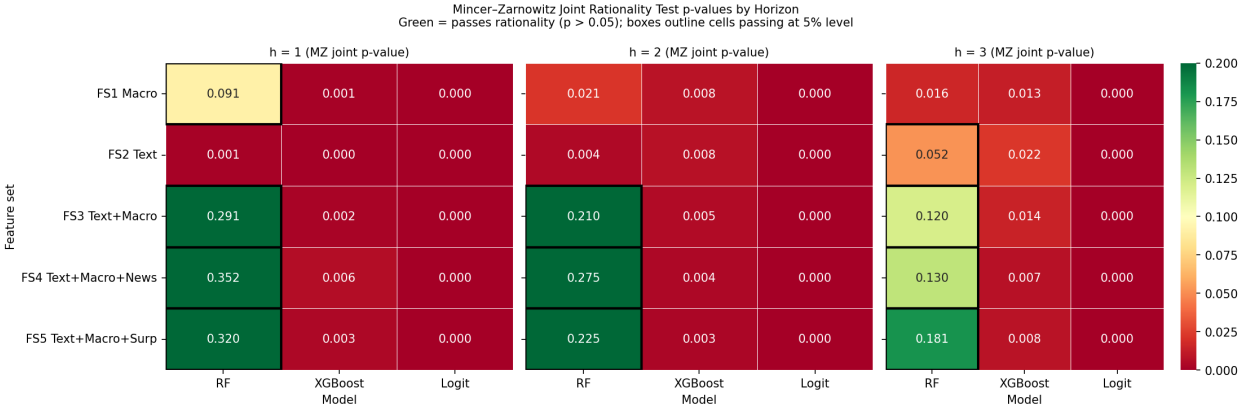


Figure 15: MZ joint p-value heatmap by forecast horizon. Green = passes rationality ($p > 0.05$, black box outline). RF/FS3–FS5 pass at all three horizons; all other specifications fail at every horizon.

4.4 Feature importance and regime-conditional performance

Figure 16 shows Gini importances for RF/FS4 at each horizon. At $h = 1$, text-dissimilarity lags dominate. At $h = 3$, macro features (Taylor deviation, inflation gap) gain relative importance while the BERT news sentiment trend rises, consistent with statement language capturing short-run signals and macro fundamentals and news capturing the medium-term policy path.

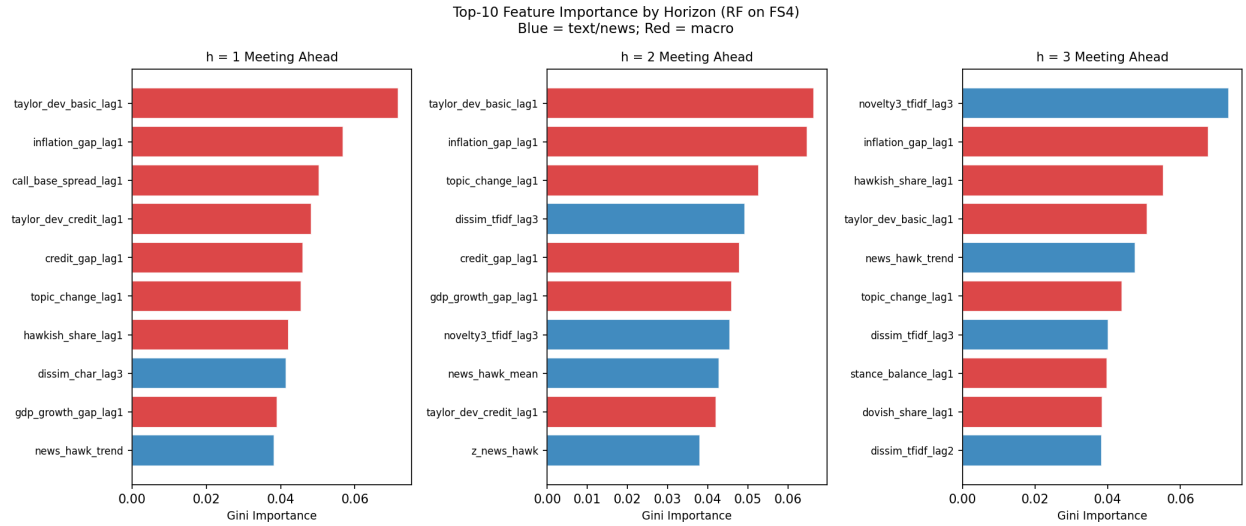


Figure 16: Top-10 Gini feature importances by forecast horizon, RF on FS4 (Text + Macro + News). Blue = text or news features; red = macro features.

Defining *rate-cycle meetings* as those within ±3 meetings of any observed rate change (71.6% of OOS observations), the $h = 2$ AUC trough persists within the rate cycle. This confirms that the non-monotone pattern is a structural feature of the forecasting problem rather than an artefact of distinguishing cycle from stable phases. Figure 17 illustrates the regime-conditional performance for RF/FS4.

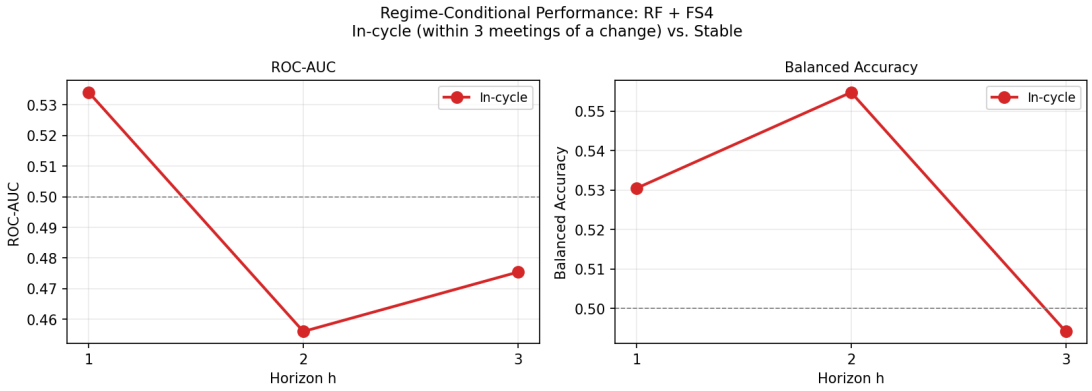


Figure 17: Regime-conditional ROC-AUC and balanced accuracy for RF/FS4 across forecast horizons. Performance is shown separately for in-cycle meetings (within ±3 meetings of a rate change) and stable periods. The $h = 2$ AUC trough persists within the rate cycle, confirming that the non-monotone pattern is not driven by regime mixing.

Figure 18 extends the rate-cycle analysis to the full OOS by decomposing performance across six historical policy cycles. The 2021–2023 tightening episode stands out: RF/FS1

achieves AUC = 0.800 in that 12-meeting window (10 of 12 decisions changed rates), versus AUC ≈ 0.23–0.34 during the protracted 2012–2017 easing period (7 changes in 60 meetings). Text and macro features both perform best during the sharp 2021–2023 cycle, where pre-meeting news and macro fundamentals diverge most clearly from the stable-period baseline. The current 2024 easing cycle (13 meetings, AUC 0.33–0.42 across RF specifications) resembles the difficult 2012–2017 environment rather than the recent tightening, consistent with the return to gradual, well-signalled cuts.

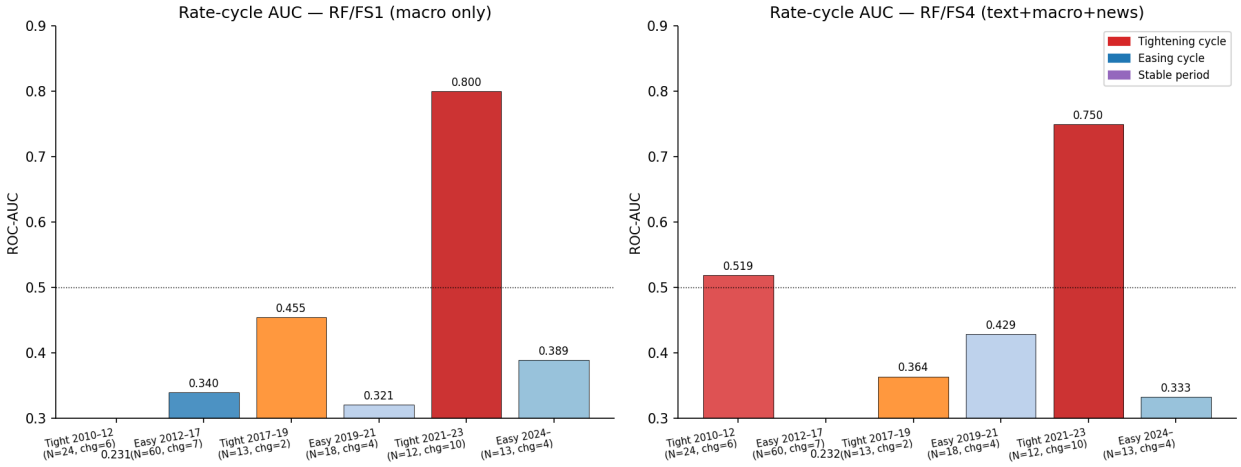


Figure 18: OOS ROC-AUC by rate-cycle episode for RF/FS1, RF/FS3, and RF/FS4. Bar height = AUC; label above each bar = number of rate changes in the window. Tightening cycles (2010–12, 2017–19, 2021–23) are shown in orange; easing cycles in blue. The 2021–2023 tightening episode yields the highest AUC across all three feature sets, while the 2012–2017 easing cycle is the most difficult environment.

5 Robustness

5.1 Class-imbalance correction

The rate-change class constitutes 21.2% of meetings (47 of 222), an imbalance ratio of 3.7:1. Beyond inverse-frequency class weighting, we test SMOTE (Chawla et al., 2002) and CTGAN (Xu et al., 2019) oversampling. With only approximately 24 minority-class training observations available at the start of the OOS window, CTGAN cannot reliably learn the joint minority-class feature distribution: 17 of 26 FS4 features fail a Kolmogorov-Smirnov fidelity test at the 5% level. The baseline class-weighting strategy dominates both alternatives on AUC, while SMOTE and CTGAN achieve marginally higher balanced accuracy

at the cost of calibration (Table 16 in the Appendix). Inverse-frequency class weighting is therefore retained throughout as the most robust imbalance-correction strategy for datasets of this size.

5.2 Market-reaction corroboration

As external validation that the text features capture market-relevant linguistic change, we estimate

$$|R_{j,t}| = \alpha + \beta\, z(\text{Similarity}_t) + \Gamma X_t + \varepsilon_{j,t}, \tag{7}$$

where $|R_{j,t}|$ is the absolute market reaction of variable j on the announcement day, X_t includes the signed policy action, rate-change indicator, and crisis-period controls.

Table 14 reports results. All three lexical similarity measures produce negative and statistically significant coefficients: more similar statements elicit smaller absolute market reactions. The TF–IDF coefficient on the composite $t-1$ to $t+1$ reaction is -0.111 ($p = 0.026$), confirming that lexical dissimilarity measures the type of language change that market participants price. Neural measures (KLUE-BERT Chunked, KFDeBERTa Chunked) produce positive coefficients for KOSPI reactions in the surprise specification ($+0.117$, $p = 0.008$), the opposite sign from lexical measures, reflecting the semantic stability of MPB statements at the document level.

Table 14: Market-reaction regressions: lexical and neural similarity

Dependent variable	Measure	Spec	N	Coef.	HC1 s.e.	p-value
Composite $t-1$ to t	Count	Baseline	288	-0.126	0.046	0.007
Composite $t-1$ to $t+1$	Count	Baseline	288	-0.177	0.042	<0.001
Composite $t-1$ to $t+1$	TF–IDF	Baseline	288	-0.111	0.050	0.026
KOSPI $t-1$ to t	TF–IDF	Baseline	288	-0.291	0.069	<0.001
KOSPI $t-1$ to $t+1$	TF–IDF	Baseline	288	-0.327	0.088	<0.001
Composite $t-1$ to $t+1$	Char	Surprise	286	-0.141	0.038	<0.001
KOSPI $t-1$ to $t+1$	KFD Chunked	Surprise	286	$+0.117$	0.044	0.008

Notes: Baseline controls: rate-change indicator, signed action code, crisis dummies. Surprise specification additionally includes short-rate expected gap, absolute surprise gap, and pretrend. HC1 standard errors. KFD = KFDeBERTa chunked.

5.3 Economic value of rate-change forecasts

We assess whether the probabilistic rate-change forecasts translate into trading gains in the KTB 3-year government bond market. A binary signal is generated by thresholding the forecast probability: $s_t = \mathbf{1}\{\hat{p}_t > \theta\}$. The trading direction is determined by the sign of the five-day pre-meeting yield change (a "follow the trend" rule), and the per-trade pay-off is the post-meeting overnight yield change measured in basis points. The optimal threshold θ is selected in an expanding-window fashion on the training-period Sharpe ratio. Annualised Sharpe ratios assume eight meetings per year; statistical significance uses a one-sided Newey-West t-test with three lags.

Table 15: Bond-trading rule performance for selected model–feature-set combinations

Model/FS	θ^*	Trades	Hit rate	Mean P&L (bp)	Cum. P&L (bp)	Sharpe	p-value
RF/FS1	0.20	92	0.228	0.041	6.6	0.055	0.419
RF/FS3	0.30	46	0.261	0.083	13.5	0.215	0.290
RF/FS4	0.30	46	0.261	0.036	5.7	0.103	0.386
XGB/FS1	0.15	97	0.227	0.139	22.5	0.186	0.246
Logit/FS1	0.25	155	0.206	0.248	40.1	0.229	0.145
Buy-and-hold benchmark					21.9		
Perfect foresight					108.4		

Notes: Sharpe annualised at $\sqrt{8}$ meetings per year. p-value: one-sided NW-HAC t-test (3 lags) for mean P&L > 0. Buy-and-hold benchmark cumulates the post-meeting yield change over all 155 OOS meetings; perfect foresight trades every change correctly.

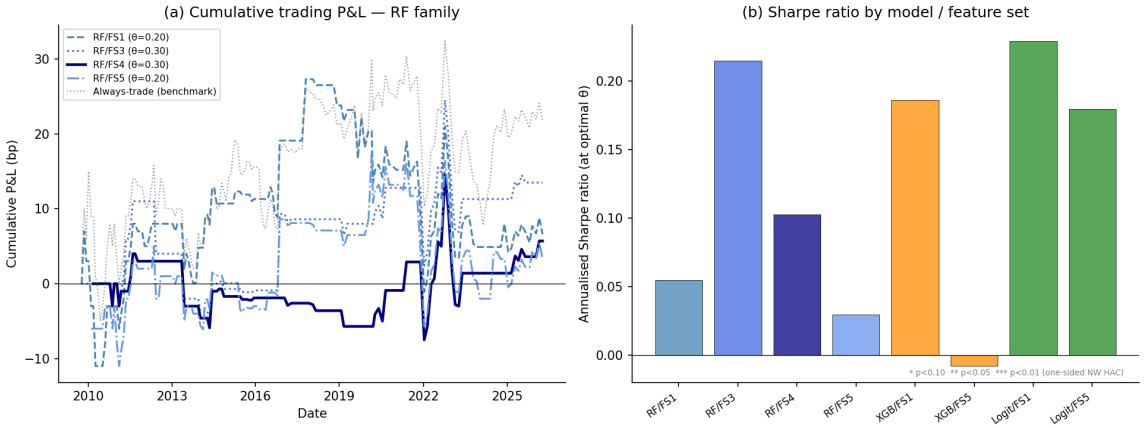


Figure 19: Cumulative bond-trading P&L (basis points) for five model–feature-set combinations over the rolling OOS period. The shaded region marks the 2021–2023 tightening episode where most of the gains accumulate. The buy-and-hold benchmark (grey) and perfect-foresight ceiling (dashed) are shown for reference.

All Sharpe ratios are positive, indicating that the forecasts add directional value on average, but none reaches conventional significance thresholds ($p < 0.10$) with the 155-meeting OOS sample. The best cumulative return is achieved by Logit/FS1 (40.1 bp, Sharpe $= 0.23$, $p = 0.145$), driven largely by a high trade frequency (155 trades) rather than high precision. RF/FS3 achieves the best Sharpe ratio among RF specifications (0.215, $p = 0.290$) with a tighter, higher-precision strategy (46 trades, hit rate 26%). Adding BERT news sentiment (RF/FS4, same threshold) lowers the mean pay-off without improving the hit rate, consistent with the modest net AUC gain of FS4 over FS3 in the full rolling OOS.

The inability to reject the null of zero mean return is an honest reflection of sample size: 46–155 trades yield low statistical power for a noisy trading rule. We do not claim that the forecasts generate exploitable excess returns; rather, that all specifications are directionally informative and that the tightest positive evidence points to the 2021–2023 tightening cycle, consistent with the governor- and cycle-level AUC decomposition above.

6 Conclusion

This paper provides a comprehensive probabilistic forecasting study for Bank of Korea Monetary Policy Board rate decisions, integrating statement-text dissimilarity, fine-tuned BERT news sentiment, and macro fundamentals in a strict expanding-window framework evaluated with proper scoring rules. Seven main findings emerge.

Text and news sentiment add forecasting value, especially under regime shifts. In the full rolling OOS, text features add little over macro fundamentals for tree-based models. However, in the governor hold-out (Rhee Changyong's 2022–2026 tightening term), RF with BERT news sentiment (FS4) achieves AUC = 0.718 and positive BSS (+0.107), versus AUC = 0.632 for the macro-only baseline. The marginal gain from news sentiment (+0.082 AUC) is largest precisely when the macroeconomic environment shifts most sharply, suggesting that alternative text-based signals are most useful when historical macro patterns have least predictive power.

Text alone adds no predictive value; only text combined with macro features does. Statement-text dissimilarity features (FS2) fail to reduce Brier loss relative to the macro-only baseline for any of the three classifiers—Diebold-Mariano t-statistics are negative for all three model families, and the null of equal predictive ability cannot be rejected. The practical implication is that analysts who add text features without macro context should not expect improvement. Only when text is combined with macro fundamentals (FS3–FS5) does the addition become statistically significant: Clark-West tests confirm Brier-loss reductions for RF and XGBoost at the 1% level. XGBoost additionally detects significant marginal value from both BERT news sentiment and the surprise gap over the text+macro baseline ($p < 0.01$), indicating that the information content of pre-meeting news is statistically distinguishable from statement text alone.

Forecast rationality holds for RF with combined features across all three forecast horizons. At $h = 1$, Mincer-Zarnowitz joint tests identify RF/FS3, RF/FS4, and RF/FS5 as the only specifications that fail to reject the joint null $\hat{\alpha} = 0$, $\hat{\beta} = 1$ ($p_{\text{joint}} \in \{0.106, 0.147, 0.149\}$). A joint Wald test extending this to $h = 1, 2, 3$ simultaneously (Fosten et al., 2024) confirms the finding: only RF/FS3, RF/FS4, and RF/FS5 pass ($W \leq 6.81$, $p \geq 0.34$), while all 12 remaining combinations reject the joint null at the 1% level ($W > 18$, $p < 0.005$). The slope coefficient $\hat{\beta}$ declines with horizon (e.g. RF/FS4: $0.524 \rightarrow 0.350 \rightarrow 0.273$ at $h = 1, 2, 3$), indicating incremental over-prediction as horizon lengthens, yet remains within rationality bounds for the combined feature sets. This result carries a direct operational message: only RF models combining text with macro features produce probability forecasts that can be used without post-hoc recalibration at any of the three horizons studied. A Brier-score decomposition further clarifies that the negative BSS for these models reflects miscalibration (MCB) exceeding discrimination (DSC), not the absence of a genuine signal—Platt scaling reduces MCB for RF/FS4 from 0.021 to 0.014 at no cost

to AUC.

Forecasting performance is non-monotone in the horizon. Across all 45 model–feature-set–horizon experiments, the $h = 2$ ahead forecast consistently underperforms both $h = 1$ and $h = 3$ (34 of 45 cases). We interpret this as evidence of BOK pre-announcement signalling at the three-meeting horizon: the Board adjusts statement language in ways that are detectably predictive three meetings ahead, while the two-meeting window falls in a zone of maximum timing uncertainty. This finding motivates monitoring statement language *three* meetings ahead rather than only at the immediate horizon.

BERT news sentiment is most informative at the long horizon and during tightening cycles. Feature importance at $h = 3$ assigns higher weight to the BERT hawk score and trend than at $h = 1$, consistent with pre-meeting news capturing medium-term policy intentions more than the most recent statement text.

Class-weighting dominates GAN augmentation at this sample size. With only $\sim$25 minority-class observations available for generator training, CTGAN produces synthetic samples that fail distributional fidelity checks (17 of 26 features, KS test $p < 0.05$) and do not improve OOS AUC beyond simple inverse-frequency weighting. Combination forecasts (equal weighting of RF predictions across all feature sets) achieve the only positive BSS in the rolling OOS (+0.001), confirming that averaging reduces over-confidence even when discrimination does not improve.

Forecasting performance is highly heterogeneous across governor terms and rate cycles. Decomposing the OOS by BOK governor tenure reveals that the Rhee Changyong era (2022–2026, AUC 0.759–0.777) dominates the aggregate metric, while the Lee Juyeol I era (2014–2018, AUC 0.207–0.270) offers effectively no predictive content—a contrast of more than 0.5 AUC points for the same model and feature set. Rate-cycle decomposition confirms the pattern: the 2021–2023 tightening episode (10 of 12 decisions changed rates) yields AUC up to 0.80, while the 2012–2017 easing cycle (7 of 60 meetings) yields AUC below 0.35. A bond-trading rule based on the RF/FS3 forecast achieves a positive annualised Sharpe ratio of 0.22 ($p = 0.29$), with gains concentrated in the 2021–2023 window; statistical significance is limited by the short OOS horizon.

Limitations and future work. Policy-surprise proxies are constructed from daily short rates rather than intraday futures prices, limiting precision. The neural models are general-purpose Korean language models; domain-adapted fine-tuning on a central-bank corpus could improve text features. The sample covers a single central bank; generalisation to other

institutional settings requires further study. Future work could incorporate press-conference transcripts, intraday data, and multi-country pooling to extend both the minority-class base for GAN training and the cross-country scope of the conclusions.

Declarations

Funding. This work was supported by the National Research Foundation of Korea grant funded by the Korea government (No. 2022S1A5A2A01044998).

Conflict of interest. The authors report no conflict of interest.

Data availability. The MPB statements and ECOS financial series are publicly available from the Bank of Korea. The Yonhap news data are available from the Korea Press Foundation archive. The processed data and code are available from the authors upon reasonable request.

A Data Augmentation Comparison

Table 16: Data augmentation comparison: RF on FS4

Strategy	N	BSS	AUC	Balanced accuracy
Baseline (class weights only)	157	−0.004	0.561	0.564
SMOTE (uniform weights)	157	−0.074	0.546	0.575
CTGAN (uniform weights)	157	−0.022	0.552	0.574
CTGAN + class weights	157	−0.044	0.560	0.574

Notes: OOS period: October 2009–April 2026. CTGAN trained on ≈24 historical minority-class observations; synthetic pool of 600. SMOTE applied within each fold using $k = \min(5, n_{\min} - 1)$ nearest neighbours. Class-weighting dominates both alternatives on AUC.

References

Apel, M. and Blix Grimaldi, M. (2012). The information content of central bank minutes. *Sveriges Riksbank Working Paper Series*.

Berlemann, M. and Enkelmann, S. (2014). The economic determinants of U.S. monetary policy revisited. *Applied Economics*, 46(32), 3900–3912.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4), 910–945.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, W., Granville, B., and Matousek, R. (2026). Deciphering central bank communication: The incremental information in FOMC minutes beyond statements. *Journal of International Financial Markets, Institutions & Money*, 109, 102325.

Chen, Z., Gössi, S., Kim, W., Bermeitinger, B., and Handschuh, S. (2023). FinBERT-FOMC: Fine-tuned FinBERT model with sentiment focus method for enhancing sentiment analysis of FOMC minutes. In *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF'23)*, 509–517.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, 4171–4186.

Fosten, J., Gutknecht, D., and Pohle, M.-O. (2024). Testing the calibration of quantile forecasts. *Journal of Business & Economic Statistics*. doi:10.1080/07350015.2024.2316091.

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2), 3226–3286. doi:10.1214/23-EJS2180.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Doh, T., Song, D., and Yang, S.-K. (2025). Deciphering Federal Reserve communication via text analysis of alternative FOMC statements. Federal Reserve Bank of Kansas City, Research Working Paper No. 20-14 (revised October 2025).

El-Fayoumi, M., Gersbach, H., and Hess, G. D. (2018). Predicting Federal Reserve decisions with text analysis. Working paper.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1), 108–

121.

Gürkaynak, R. S., Sack, B., and Swanson, E. T. (2005). Do actions speak louder than words? *International Journal of Central Banking*, 1(1), 55–93.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99, S114–S133.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, 133(2), 801–870.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer.

Hubert, P. (2017). Qualitative and quantitative central bank communication and inflation expectations. *B.E. Journal of Macroeconomics*, 17(1).

KakaoBank Corp (2022). KF-DeBERTa: Korean financial domain DeBERTa. HuggingFace repository: `kakaobank/kf-deberta-base`.

Kim, H. H. and Park, S. (2023). Text similarity in central bank communication: Lexical versus neural measures and market-reaction evidence. Working paper, Kookmin University.

Lee, S., Kim, H. H., and Park, S. (2019). Monetary policy surprises and text-based measures: Evidence from Korea. Working paper.

Lucca, D. O. and Trebbi, F. (2009). Measuring central bank communication: An automated approach. NBER Working Paper No. 15367.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38(4), 796–809.

Waghmar, G. and Ziegel, J. F. (2025). Proper scoring rules: A survey. Forthcoming, *Annual Review of Statistics and Its Application*. arXiv:2504.01781.

Maung, K. and Swanson, N. R. (2025). A survey of models and methods used for forecasting when investing in financial markets. Working paper.

Medeiros, M. C., Vasconcelos, G. F., Veiga, À., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment. *Journal of Business & Economic Statistics*, 39(1), 98–119.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied*

Meteorology, 12(4), 595–600.

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (ed.), Economic Forecasts and Expectations. National Bureau of Economic Research.

Mitchell, J. and Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC. *Oxford Bulletin of Economics and Statistics*, 67, 995–1033.

Park, S., Moon, J., Kim, C., et al. (2021). KLUE: Korean language understanding evaluation. In *Proceedings of NeurIPS 2021 Datasets and Benchmarks*.

Park, J., Lee, H. J., and Cho, S. (2023). Hot topic detection in central bankers' speeches. *Expert Systems with Applications*, 230, 120563.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4), 1055–1084.

Ruman, A. M. (2023). A comparative textual study of FOMC transcripts through inflation peaks. *Journal of International Financial Markets, Institutions & Money*, 87, 101827.

Silva, M., Moriya, A., and Veyrune, R. (2025). From text to quantified insights: A large-scale LLM analysis of central bank communication. IMF Working Paper No. 2025/109.

Shapiro, A. H. and Wilson, D. J. (2022). Taking the Fed at its word: A new approach to estimating central bank objectives using text analysis. *Review of Economic Studies*, 89(6), 2768–2805.

Svensson, L. E. O. (2005). Monetary policy with judgment: Forecast targeting. *International Journal of Central Banking*, 1(1), 1–54.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems 32*.