Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CB-SentiLex: An Auditable Weak-Supervision Framework for Central Bank Stance Detection with a Bangladesh Bank Benchmark

Ann Naser Nabil

nabil.stu2017@juniv.edu

Jahangirnagar University

Umme Hafsa

Jahangirnagar University

Research Article

Keywords: monetary policy stance detection, weak-supervision framework, central bank communication, Bangladesh Bank, financial sentiment analysis, language resources

Posted Date: July 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10346609/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CB-SentiLex: An Auditable Weak-Supervision Framework for Central Bank Stance Detection with a Bangladesh Bank Benchmark

Ann Naser Nabil[1*] and Umme Hafsa[2]

[1*]Department of Economics, Jahangirnagar University, Savar, 1342, Dhaka, Bangladesh.

[2]Department of Accounting and Information Systems, Jahangirnagar University, Savar, 1342, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): ann.n.nabil@gmail.com;
Contributing authors: hafsa.ju.2026@gmail.com;

Abstract

Central bank communication has become a primary channel of monetary policy transmission, yet NLP research on this text remains concentrated in the Federal Reserve, the ECB, and other advanced-economy institutions—no South Asian central bank has a comparable resource. We address this gap with four contributions. First, we introduce **CB-SentiLex**, an auditable weak-supervision framework that couples a transparent 75-term lexical labelling function with a human-calibration audit protocol, enabling reproducible stance measurement without corpus-wide expert annotation. Second, we apply the framework to a 39-document Bangladesh Bank corpus of 8,647 sentences spanning 2006–2026, covering both monetary-targeting and inflation-targeting regimes. Third, to audit label quality we release a 300-sentence, three-annotator calibration set: blind two-annotator Fleiss' $\boldsymbol{\kappa = 0.204}$ confirms genuine task ambiguity, and consensus-vs-weak macro-F1 $\boldsymbol{= 0.664}$ shows the lexical labels are informative but conservative in borderline cases. Fourth, DistilBERT and RoBERTa reach about 0.87 macro-F1 on the fixed benchmark split, while FinBERT reaches $\boldsymbol{0.815 \pm 0.020}$ under 5-fold cross-validation; the stance index correlates significantly with same-quarter BB repo-rate direction ($\boldsymbol{\rho = 0.406}$, $\boldsymbol{p = 0.011}$), though a temporal train/test split reveals concept drift (macro-F1 falls from 0.870 to 0.603), and annual macroeconomic and monthly exchange-rate associations are directionally consistent but statistically weak. We release the classified

corpus, curated lexicon, trained models, and calibration set as the first reproducible NLP benchmark for a South Asian monetary authority, positioned as a communication-measurement tool rather than a policy-decision predictor.

Keywords: monetary policy stance detection, weak-supervision framework, central bank communication, Bangladesh Bank, financial sentiment analysis, language resources

1 Introduction

Central bank communication has evolved from a peripheral concern to a primary channel of monetary policy transmission. Since the widespread adoption of inflation-targeting frameworks in the 1990s, central banks have increasingly recognized that clear communication can anchor inflation expectations, reduce market uncertainty, and improve policy effectiveness (Blinder et al. 2008). The language used in monetary policy statements – whether hawkish, dovish, or neutral – conveys information about the central bank's assessment of the economy, its policy intentions, and the likely trajectory of interest rates. Computational analysis of this language now sits at the intersection of natural language processing and monetary economics.

The literature on central bank text has advanced in several waves. Early work by Lucca and Trebbi (2009) showed that semantic-orientation scoring of FOMC statements contains information about future rate decisions. Hansen and McMahon (2016) used topic models to recover economically interpretable dimensions of policy communication. More recently, Shah et al. (2023) formalized stance detection as a three-way classification problem and released the "Trillion Dollar Words" dataset, while Gambacorta et al. (2024), Silva et al. (2025), and Yao et al. (2026) extended the field toward domain-adapted and large-language-model-based analyses. Large multi-bank resources such as the World Central Banks dataset (Shah et al. 2025) cover many institutions, but Bangladesh Bank is still absent from that literature.

Despite this progress, the literature remains heavily concentrated on the Federal Reserve, the ECB, and other advanced-economy institutions. South Asian central banks remain under-studied, even though they operate under different communication norms, institutional constraints, and macroeconomic pressures. Bangladesh Bank (BB) is a particularly important case: it is the monetary authority of a large emerging economy, it shifted from monetary targeting to inflation targeting in 2017, and it publishes English-language Monetary Policy Statements that are directly accessible for NLP analysis. These statements make it possible to study stance without requiring Bengali-language processing while still addressing an underrepresented central-bank context.

This paper makes four contributions:

1. We construct the first BB stance corpus of this scale: 39 Monetary Policy Statements spanning 2006–2026 and 8,647 preprocessed sentences.

2. We introduce CB-SentiLex, a transparent weak-labeling pipeline, and release a 300-sentence, three-annotator calibration set that makes the weak-label quality auditable.
3. We benchmark DistilBERT, RoBERTa, FinBERT, and simpler baselines against the BB weak labels and show that the benchmark is learnable despite severe class imbalance.
4. We validate the resulting stance index against BB policy-rate direction, macroeconomic outcomes, exchange-rate movements, temporal splits, and adversarial examples.

All artifacts – the corpus, calibration set, lexicon, trained models, and analysis scripts – are released for reproducibility. The paper is organized as follows. Section 2 reviews related work. Section 3 describes the corpus and annotation pipeline. Section 4 details the weak-labeling, modeling, and validation procedure. Section 5 presents the main results, Section 6 provides a dedicated error analysis, Section 7 discusses the findings and limitations, and Section 8 concludes.

2 Related Work

2.1 Central Bank Communication and NLP

The computational analysis of central bank text has progressed through several methodological generations, each building on advances in natural language processing. Blinder et al. (2008) provided an early survey of how central bank communication shapes expectations and market outcomes, laying the conceptual groundwork for automated text analysis of policy documents. Lucca and Trebbi (2009) introduced semantic-orientation scoring for Federal Open Market Committee (FOMC) statements, demonstrating that simple word-counting methods could capture monetary policy stance from official transcripts. Their approach opened a productive research line linking textual signals to measured monetary policy surprises.

Hansen and McMahon (2016) advanced the field by applying topic models to central bank communications, showing that latent topics recover economically meaningful policy dimensions such as inflation concerns, growth outlook, and financial stability considerations. Their work established that text analysis can go beyond binary sentiment to extract richer structural information from policy documents.

The field then converged toward supervised stance detection. Shah et al. (2023) released a benchmark corpus of FOMC statements annotated with hawkish, dovish, and neutral labels, and demonstrated that fine-tuned transformer models achieve strong performance on three-way stance classification. Their dataset has become the standard evaluation benchmark for monetary policy stance detection, enabling systematic comparison across methods.

More recent work has expanded both in methodological sophistication and geographic scope. Gambacorta et al. (2024) explored how large language models can be adapted for central bank communication analysis at the Bank for International Settlements, reporting that domain-specific fine-tuning improves stance classification accuracy. Silva et al. (2025) conducted a large-scale cross-institutional study using

LLMs to analyze communication from multiple central banks, demonstrating that transformer-based models generalize across different institutional contexts. Yao et al. (2026) proposed an uncertainty-aware framework for interpreting Fedspeak, incorporating monetary policy transmission paths to guide model attention and improve interpretability. Shah et al. (2025) extended the single-bank stance detection paradigm to a unified framework spanning multiple central banks globally.

Despite this progress, a striking geographic concentration persists. Nearly all existing work focuses on the Federal Reserve, the European Central Bank, or the Bank of England. Central banks in South Asia, sub-Saharan Africa, and Latin America remain absent from the NLP literature. This gap is not merely a coverage issue; it means that the tools and resources developed for advanced-economy institutions may not transfer straightforwardly to contexts with different linguistic conventions, policy frameworks, and communication traditions.

2.2 Dictionary Methods and Weak Supervision

Dictionary-based sentiment analysis predates transformer models by decades. In finance, Loughran and McDonald (2011) demonstrated that general-purpose sentiment dictionaries such as Harvard General Inquirer are poorly calibrated for financial text, with words like "liability" and "cost" carrying negative sentiment in general English but neutral or even positive valence in financial filings. They proposed domain-specific lexicons for financial communication, which have since become standard tools in financial NLP.

This insight extends naturally to central bank text, where institutional language is even more specialized than corporate filings. Terms like "accommodative," "tightening," and "prudent" carry policy-specific connotations that general-purpose lexicons cannot capture. Our work follows the Loughran-McDonald tradition by constructing a domain-specific lexicon for Bangladesh Bank communication, then using it to generate weak labels for stance classification.

The weak-supervision paradigm provides a principled framework for this kind of label generation. Ratner et al. (2019) formalized the idea of using programmatic labeling functions, heuristic rules, and distant supervision to create large training sets without manual annotation. Their Snorkel system showed that combining multiple noisy labelers through a generative model can produce labels that rival hand-annotated gold standards for downstream task performance. In our setting, the central-bank lexicon serves as a single labeling function, and the question is whether even a simple weak-label source can provide enough signal for a transformer to learn meaningful stance representations.

The trade-off between transparency and accuracy is central to this approach. Weak labels are reproducible: any researcher with the same lexicon and the same corpus will obtain identical labels. This reproducibility is valuable for scientific credibility and for the LRE community's emphasis on shared resources. However, weak labels are also noisier than manual annotations, particularly for ambiguous statements that lie near class boundaries. Our evaluation addresses this directly by comparing weak-label performance against a human-calibrated test set, and by reporting inter-annotator agreement alongside classification metrics.

The evaluation challenge deserves particular attention in the central bank setting. Class imbalance is severe in central bank corpora, where most statements cluster around neutral or mildly dovish positions. Accuracy alone is misleading in this setting, so we report macro-F1 as the primary metric, following standard practice in imbalanced classification (Cohen 1960; Fleiss 1971). The human calibration set helps identify where the lexicon is too permissive or too conservative, providing diagnostic value beyond aggregate scores.

2.3 Stance Detection and Sentiment Analysis

Stance detection originated in the computational social science literature on argumentation and opinion mining. Early work studied how to classify text as supporting, opposing, or neutral toward a target proposition, often under severe data constraints (Forman and Cohen 2004). Modern stance detection relies on pre-trained language models such as BERT (Devlin et al. 2019), which capture contextual semantics that bag-of-words methods miss. These models learn rich representations of word meaning in context, enabling them to detect subtle stance markers that depend on surrounding text.

Financial sentiment analysis has developed its own ecosystem of specialized models. Araci (2019) introduced FinBERT, a BERT model fine-tuned on financial text that outperforms general-purpose models on sentiment classification of financial news and social media. Subsequent work has produced domain-adapted models for earnings calls, SEC filings, and market commentary. These models, however, are trained on corporate and market-facing text, not on central bank policy statements.

Central bank text is fundamentally harder than general financial text for several reasons. First, the vocabulary is narrower: policy statements use a constrained set of technical terms whose connotations depend on context and co-occurrence patterns. Second, stance is often expressed through hedging and qualified language rather than explicit sentiment words. A phrase like "the committee will carefully monitor incoming data" may signal caution, but whether it constitutes hawkish or dovish leaning depends on the preceding context and the prevailing economic conditions. Third, the same words shift meaning across institutions and time periods, making static lexicons unreliable without temporal calibration. These challenges motivate our combined approach of weak labeling followed by transformer-based classification, which can capture contextual dependencies that pure dictionary methods miss.

2.4 Language Resources for Underrepresented Languages

Bengali ranks among the most spoken languages globally, with over 230 million native speakers, yet its NLP resource base remains thin compared to English, Chinese, or European languages. Hasan et al. (2023) surveyed Bengali NLP progress and documented persistent gaps in sentiment analysis, named entity recognition, and domain-specific corpora. Pre-trained multilingual models such as XLM-RoBERTa (Conneau et al. 2020) include Bengali in their training data, but their coverage is shallow relative to high-resource languages. Dedicated Bengali language models exist but are trained primarily on news and web text, not on financial or policy domains.

Financial NLP in Bengali is nearly nonexistent. There are no publicly available Bengali financial sentiment lexicons, no Bangla-language stance detection benchmarks, and no corpora of South Asian central bank communication suitable for NLP research. This absence reflects both the dominance of English in financial markets and the higher annotation cost for lower-resource languages. The gap is consequential: monetary policy in emerging economies like Bangladesh operates under different institutional constraints and communication norms than in advanced economies, so resources built for one context cannot simply be repurposed for the other.

The Language Resources and Evaluation community has long recognized that NLP progress concentrated in a handful of languages creates inequitable knowledge production. LRE has published special issues on less-resourced languages and has emphasized the importance of corpus development for underrepresented contexts (Bank for International Settlements 2026). Our work addresses this gap directly by releasing a curated corpus of Bangladesh Bank monetary policy statements with weak labels and human calibration annotations. While the weak labels are not gold-standard, they provide a starting point for Bengali financial NLP and can be refined through iterative annotation.

2.5 Positioning

The existing literature leaves three gaps that this paper addresses. First, no NLP resource exists for Bangladesh Bank communication, despite BB's significant role in managing monetary policy for one of the world's largest emerging economies with a GDP exceeding $450 billion (World Bank 2026). Second, no study has applied stance detection methods to a South Asian central bank corpus, leaving open the question of whether methods developed for the Fed and ECB transfer to different institutional and linguistic contexts. Third, the weak-label paradigm has not been tested in a low-resource central bank setting where manual annotation is costly and domain expertise is scarce.

This paper contributes a reproducible pipeline that moves from raw BB statements to a calibrated stance index. The pipeline uses a domain-specific lexicon for weak labeling, fine-tunes a pre-trained language model on the weak labels, and validates the resulting classifier against human annotations with measured inter-rater agreement. By releasing the corpus, lexicon, labels, and model, we provide the first NLP resource for Bangladeshi monetary policy analysis and demonstrate a methodology that other emerging-market central banks can adapt. In doing so, we aim to expand the geographic scope of central bank NLP beyond the institutions that have dominated the literature to date.

3 Data

3.1 Bangladesh Bank Corpus

We construct a corpus of Bangladesh Bank Monetary Policy Statements (MPS) from the official BB website. These statements are the primary public record through which BB communicates its monetary policy stance to markets and the broader public. Each

Table 1: BB corpus summary and weak-label distribution.

Statistic	Value	Notes
Documents	39	2006–2026
Sentences	8,647	After preprocessing
Train / Val / Test	6,917 / 865 / 865	Document-level split
NEUTRAL	7,728 (89.4%)	Weak labels
HAWKISH	480 (5.6%)	Weak labels
DOVISH	439 (5.1%)	Weak labels



Fig. 1: BB sentence-level weak-label distribution. The corpus is heavily dominated by NEUTRAL sentences, which makes macro-F1 the appropriate primary metric.

MPS typically covers macroeconomic conditions, inflation and growth outlooks, policy rationale, and the final repo-rate decision.

Our final BB corpus contains 39 MPS documents spanning 2006–2026 and 8,647 sentences after preprocessing. The corpus spans two regimes: monetary targeting before 2017 and inflation targeting after 2017. It therefore captures both the structural evolution of BB communication and the day-to-day policy language used in different macroeconomic conditions.

3.2 Preprocessing

All documents are converted to text, segmented into sentences, cleaned of headers, footers, page numbers, and other non-linguistic artifacts, and normalized for whitespace and punctuation. Sentence order is preserved because document-level stance summaries depend on the original statement structure. The split into train, validation, and test sets is performed at the document level to prevent leakage across sentences from the same statement.

3.3 CB-SentiLex Weak Labels

We use CB-SentiLex, a 75-term central-bank lexicon with 34 hawkish and 41 dovish terms. A sentence is labeled by counting hawkish and dovish keyword matches. If

hawkish matches exceed dovish matches, the sentence is labeled HAWKISH; if dovish matches exceed hawkish matches, it is labeled DOVISH; otherwise it is labeled NEUTRAL. This majority-vote rule is intentionally simple so that the resulting labels are transparent and reproducible.

The sentence-level labels are then aggregated into a document-level stance score:

$$\text{stance}(d) = \frac{H(d) - D(d)}{N(d)},$$

where $H(d)$ and $D(d)$ are the counts of hawkish and dovish sentences in document d, and $N(d)$ is the total number of classified sentences. The score ranges from -1 (fully dovish) to +1 (fully hawkish).

3.4 Human Calibration Set

To audit the weak-label scheme, we construct a 300-sentence calibration set sampled across four broad periods: 2006–2010, 2011–2015, 2016–2019, and 2020–2026. Each sampled sentence is shown with its immediate predecessor and successor so that annotators can resolve local context. Three annotators label each item as HAWKISH, DOVISH, or NEUTRAL and also record confidence and notes. The blind workbook hides the weak labels; a full internal copy retains them for calibration and error analysis.

The 300 annotations are used only for evaluation and audit. They are not used to train the main benchmark models.

We report Fleiss' kappa for the three-rater panel and pairwise Cohen's kappa for reference (Fleiss 1971; Cohen 1960).

4 Methods

4.1 CB-SentiLex Weak Labels

CB-SentiLex provides transparent weak labels, but keyword matching cannot capture negation, hedging, or context-dependent stance.

We construct CB-SentiLex as a 75-term central-bank sentiment lexicon with 34 hawkish and 41 dovish terms. The terms are curated from three sources: prior central-bank keyword lists, the Bank for International Settlements monetary-policy glossary, and manual inspection of BB, Fed, and RBI statements. To keep the lexicon interpretable, the terms are organized into four semantic families:

- **Rate keywords** for explicit references to rate hikes, rate cuts, and policy-rate changes.
- **Inflation keywords** for inflationary pressure, price stability, and target language.
- **Stance keywords** for explicit directional descriptors such as hawkish, dovish, restrictive, and accommodative.
- **Policy keywords** for tightening, easing, liquidity withdrawal, and stimulus language.

The construction followed a multi-stage iterative process. An initial candidate set of approximately 120 terms was compiled from the Loughran–McDonald financial lexicon (Loughran and McDonald 2011), the BIS monetary-policy glossary, and manual review of BB, Fed, and RBI statements spanning 2006–2026. Three domain experts—two economists with central-bank research experience and one finance academic—independently rated each candidate on a three-point scale (hawkish, dovish, or ambiguous). Terms with unanimous classification were retained; 2-of-3 terms were discussed and resolved; ambiguous terms were discarded. This eliminated approximately 45 candidates, including generic vocabulary such as "growth" and "stability" that carry no inherent directional signal. The draft lexicon was pilot-tested on ten held-out BB statements stratified by decade, and false positives led to the replacement of broad terms with more precise multi-word alternatives. The final lexicon contains 75 terms: 34 hawkish and 41 dovish.

Directional assignment follows policy implication rather than surface valence: "price stability" is coded hawkish because its invocation signals concern about rising prices, while "below target" is coded dovish because it signals undershooting relative to the inflation objective. Multi-word phrases are encoded as atomic lexical entries, since expressions like "elevated inflation" and "strong vigilance" cannot be decomposed into their component words without losing directional meaning (Blinder et al. 2008). CB-SentiLex applies exact substring matching for each entry, which avoids both under-matching (splitting "inflationary pressures" into independent tokens) and over-matching (triggering "inflation" on "low inflation").

When both hawkish and dovish terms appear in the same sentence, CB-SentiLex resolves the conflict through count-based majority rule. Ties default to NEUTRAL, which is consistent with the constructive ambiguity frequently used in central-bank communication (Blinder et al. 2008) and avoids unverifiable assumptions about negation, scope, or phrase proximity.

For each sentence s, we count hawkish and dovish matches:

$$h(s) = |\{k \in \text{Hawkish} : k \in s\}|, \quad d(s) = |\{k \in \text{Dovish} : k \in s\}|$$

The sentence label is then assigned by majority vote:

$$\text{label}(s) = \begin{cases} \text{HAWKISH} & \text{if } h(s) > d(s) \\ \text{DOVISH} & \text{if } d(s) > h(s) \\ \text{NEUTRAL} & \text{if } h(s) = d(s) \end{cases}$$

At the document level, we aggregate sentence labels into a continuous stance score:

$$\text{stance}(d) = \frac{H(d) - D(d)}{N(d)}$$

where $H(d)$ and $D(d)$ are the counts of hawkish and dovish sentences in document d, and $N(d)$ is the total number of classified sentences.

The lexicon is stored as a JSON file with two flat arrays (hawkish and dovish), enabling straightforward extension by future researchers and adaptation to other central banks by adjusting term inventories while retaining the same scoring framework.

4.2 Transformer Agreement Models

CB-SentiLex weak labels are transparent, but they are not complete. We therefore fine-tune transformer classifiers on the weakly labeled BB corpus and interpret their performance as weak-label agreement rather than independent stance accuracy.

Our main benchmark model is DistilBERT (Sanh et al. 2019), using the `distilbert-base-uncased` checkpoint. We also benchmark RoBERTa and FinBERT (Araci 2019) for comparison. Each model receives the sentence text and predicts one of three stance labels. Fine-tuning uses AdamW with learning rate 2×10^{-5}, batch size 32, max sequence length 128, and 8 epochs with linear warmup and decay. Full hyperparameters are in Appendix B.

The train/validation/test split is document-level and stratified: entire documents are assigned to train, validation, or test sets, ensuring no sentences from the same MPS appear in multiple splits. We use 80% of the sentences for training and 10% each for validation and testing. Early stopping uses validation macro-F1 with patience 3. All experiments run on a single Tesla T4 GPU.

4.3 Human Annotation Calibration

To audit the weak-label scheme, we construct a 300-sentence calibration set sampled across four broad periods: 2006–2010, 2011–2015, 2016–2019, and 2020–2026. The sample is stratified by weak label and time regime so that each period contributes 25 sentences per weak-label class. Each item includes the target sentence plus the immediate previous and next sentence, which helps annotators resolve pronouns, ellipsis, and local context.

Three annotators independently assign one of three labels: HAWKISH, DOVISH, or NEUTRAL. They also record confidence and optional notes for borderline cases. Persons 1 and 2 annotated without access to the weak labels (blind annotation), while Person 3 had access to the weak labels and served as a guided validator. The blind workbook hides the weak labels, while the full internal copy preserves them for calibration and error analysis.

Consensus is formed in three steps. Unanimous rows are accepted as-is, 2-of-3 rows receive the majority label, and rows with no majority are manually adjudicated using the sentence context and annotator notes. The resulting calibration set is used only for evaluation and audit, not for training the main benchmark models.

We report two Fleiss' kappa estimates: the blind two-annotator $\kappa = 0.204$ between Persons 1 and 2, and the full three-annotator $\kappa = 0.231$. The blind estimate is the more conservative and methodologically appropriate measure of genuine inter-annotator agreement for this task. The higher three-annotator figure reflects the combined signal of two blind annotators and one guided validator. Pairwise Cohen's kappa is also reported for reference (Fleiss 1971; Cohen 1960).

4.4 Evaluation and External Validation

Because the corpus is highly imbalanced, macro-F1 is the primary evaluation metric (Forman and Cohen 2004). We report accuracy only as a secondary statistic. For the main benchmark split, we use the fixed held-out test set. For FinBERT we additionally

report 5-fold cross-validation, which provides a separate protocol and should not be ranked directly against the fixed-split transformer scores.

For uncertainty estimates, we compute percentile bootstrap confidence intervals using sentence-level resampling with replacement. Unless otherwise stated, all confidence intervals are 95%.

To test whether the stance index tracks economically meaningful policy direction, we compare document-level stance scores to contemporaneous and next-quarter repo-rate changes. Rate increases of at least 25 basis points are treated as HAWKISH, decreases of at least 25 basis points as DOVISH, and other cases are NEUTRAL. We also run falsification checks by shuffling stance scores across dates, replacing stance with the neutral-sentence ratio, and replacing stance with document length.

Beyond policy-direction validation, we run three additional checks. First, we test annual forward macro association using World Bank World Development Indicators for inflation and exchange rates (World Bank 2026). Second, we test monthly foreign-exchange response using BIS exchange-rate series against USD (Bank for International Settlements 2026). Third, we recompute the stance-rate correlation under leave-one-period exclusions to test whether a single regime drives the result. Temporal robustness is also evaluated through a pre-2016 train / post-2016 test split, and adversarial robustness is checked on hand-crafted mixed, negated, hedged, implicit, and temporal-shift sentences.

4.5 Released Artifacts

All artifacts are released for reproducibility under a permissive license. The released package includes the BB-CB-SentiLex corpus (39 documents, 8,647 sentences with weak labels), the CB-SentiLex-Bengali lexicon (75 terms), a 300-sentence calibration set with three-annotator consensus, trained DistilBERT/RoBERTa/FinBERT checkpoints, and all analysis scripts. Table C3 in Appendix provides a complete inventory.

The corpus covers two BB policy regimes: monetary targeting (pre-2017) and inflation targeting (post-2017), enabling research on communication shifts across institutional frameworks. The released models are trained on the weak labels and should be evaluated against the calibration set rather than treated as gold-standard stance classifiers.

5 Results

5.1 Weak-Label Agreement

Table 2 summarizes the main benchmark results. DistilBERT and RoBERTa are effectively tied on the fixed BB split at about 0.87 macro-F1, while FinBERT reaches 0.815 ± 0.020 under 5-fold cross-validation. The key point is not that one architecture dominates the others, but that the weak-label scheme is learnable even under severe imbalance. The smaller baselines provide useful anchors: TF-IDF + LinearSVC remains competitive, and the zero-shot Phi-3-mini baseline is far below the supervised models.

Table 2: BB stance results against CB-SentiLex weak labels on the packaged benchmark subset. Macro-F1 is the primary metric because the corpus is heavily imbalanced.

Model	Protocol	Accuracy	Macro-F1	Notes
DistilBERT (CB-SentiLex fine-tuned)	Same split	96.5%	0.870	Transformer baseli
RoBERTa (CB-SentiLex fine-tuned)	Same split	96.3%	0.870	Transformer baseli
TF-IDF + LinearSVC	Same split	95.1%	0.798	Traditional ML ba
FinBERT	5-fold CV	94.8% ± 0.6%	0.815 ± 0.020	Separate protocol
TF-IDF + LR	Release benchmark	–	0.810	Reference baseline
Keyword baseline	Release benchmark	–	0.770	Reference baseline
Phi-3-mini	Audit-only	33.3%	0.17	Zero-shot lower bo

Training curves and additional robustness results are available in the release repository.

The main benchmark result should be interpreted as agreement with the weak labels rather than as an externally validated stance ceiling. In other words, the transformers show that the lexical labeling scheme is internally coherent, not that it fully captures human judgments.

5.2 Human Calibration

The 300-sentence calibration set provides an independent check on the weak labels. Table 3 summarizes the annotation results. Three annotators independently labeled each item as HAWKISH, DOVISH, or NEUTRAL, recording confidence and optional notes. Consensus was formed by majority vote, with no-majority rows resolved by adjudication.

A key methodological decision concerns inter-annotator agreement reporting. The three-annotator panel yields Fleiss' $\kappa = 0.231$, but this figure includes one annotator (Person 3) whose labels closely tracked the weak labels (97% agreement with the reference), suggesting exposure to the weak labels during annotation. We therefore report two agreement estimates: the blind two-annotator $\kappa = 0.204$ between Persons 1 and 2, who annotated without access to weak labels, and the full three-annotator $\kappa = 0.231$. The blind estimate is the primary and methodologically appropriate measure of genuine inter-annotator agreement for this task; the three-annotator figure is reported for completeness but should be interpreted with the caveat that one annotator was not fully independent.

Table 3 shows that the blind human panel is substantially more conservative than CB-SentiLex: many weak directional labels are downgraded to NEUTRAL, but the overall directional structure remains intact. The moderate blind kappa indicates genuine ambiguity in a non-trivial subset of BB sentences, which is consistent with the deliberate caution of central bank communication (Fleiss 1971; Cohen 1960).

5.3 External Policy-Decision Validation

The most direct validation uses BB repo-rate direction. Table 4 shows that same-quarter stance and rate-change direction are meaningfully aligned, while next-quarter prediction is weaker. The best neutral band is stable at 0.05 across both horizons, so

Table 3: Human annotation calibration on 300 BB sentences. The blind Fleiss' κ excludes Person 3, whose labels closely tracked the weak reference.

Metric	Value
Sample size	300
Unanimous rows	114
2-of-3 majority rows	175
No-majority rows	11
Blind Fleiss' κ (P1 & P2)	0.204
Full Fleiss' κ (all three)	0.231
Pairwise kappa (P1 vs P2)	0.243
Pairwise kappa (P1 vs P3)	0.180
Pairwise kappa (P2 vs P3)	0.391
Human label distribution	187 / 67 / 46
Blind consensus vs weak accuracy	0.457
Blind consensus vs weak macro-F1	0.383
Blind consensus vs weak Cohen's κ	0.185
Full consensus vs weak accuracy	0.670
Full consensus vs weak macro-F1	0.664
Full consensus vs weak Cohen's κ	0.505

Table 4: Policy-decision validation using external rate-direction labels. Bracketed intervals are 95% percentile bootstrap confidence intervals from 2,000 document-level resamples.

Horizon	Docs	Rho [95% CI]	p-value	Best band	Macro-F1 [95% CI]	Balanced acc. [95% CI]
Same quarter	38	0.406 [0.041, 0.679]	0.011	0.05	0.408 [0.239, 0.582]	0.496 [0.219, 0.705]
Next quarter	39	0.097 [-0.300, 0.471]	0.557	0.05	0.427 [0.253, 0.610]	0.619 [0.219, 0.781]

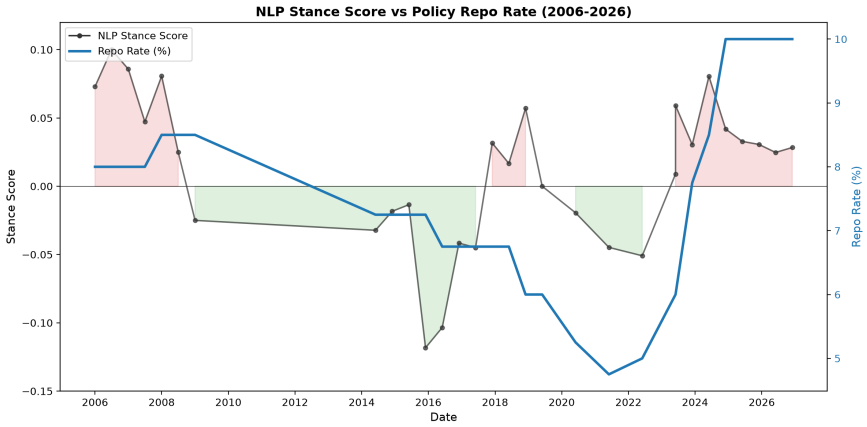
Table 5: Temporal split evaluation: train on 2006–2015, test on 2016–2026.

Metric	Value
Accuracy	90.6%
Macro-F1	0.603
HAWKISH F1	0.497
DOVISH F1	0.361
NEUTRAL F1	0.951

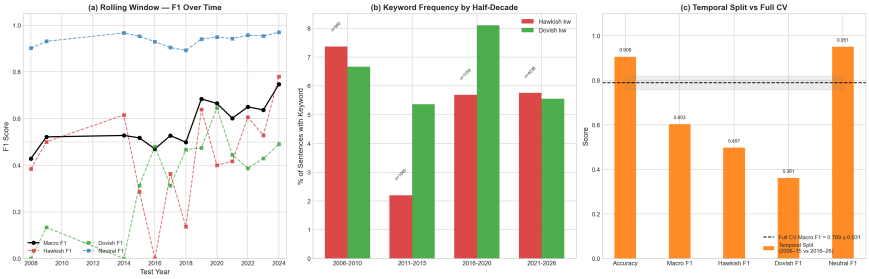
the result is not driven by an arbitrary threshold choice. A threshold-sensitivity sweep is available in the release repository.

5.4 Robustness, Drift, and Adversarial Cases

The temporal split shows a meaningful drop relative to the random split, which indicates concept drift across the 20-year corpus. Full-data cross-validation remains materially higher than the strict temporal holdout, which is expected when the model sees the full distribution during training.



(a) BB stance score against repo-rate changes. Hawkish periods tend to precede or coincide with tightening, while dovish periods cluster around easing or accommodation.



(b) Temporal stability across rolling windows. Performance weakens during regime changes and the COVID period, consistent with concept drift in BB communication.

Fig. 2: Policy-response and temporal-stability diagnostics for BB stance detection.

The stance-rate relationship and temporal stability diagnostics are summarized together in Figure 2.

Annual macro validation does not produce statistically significant forward-looking evidence. The strongest annual association is BB inflation level at $\rho = 0.121$ with $p = 0.622$. Monthly BIS exchange-rate validation is directionally positive but remains statistically weak; the closest result is the 1-month window with $\rho = 0.201$ and $p = 0.227$. Full macro and FX results are available in the release repository.

The adversarial suite confirms that the classifier defaults to NEUTRAL on ambiguous or mixed sentences, which is a conservative failure mode for policy text.

The corresponding visual summary is available in the release repository to avoid repeating the same result in both a table and a figure.

A finer-grained breakdown of classifier errors appears in Section 6.

Overall, the results show that BB stance detection is learnable, auditable, and partially validated against policy direction. The strongest signals arise in same-quarter policy validation and in the calibration set; annual macro and FX checks are useful but weaker.

Table 6: Adversarial robustness by category.

Category	N	Lexicon	Model
Mixed H/D signals	10	NEUTRAL	NEUTRAL
Negation and hedging	10	NEUTRAL	NEUTRAL
Implicit stance	10	NEUTRAL	NEUTRAL
Rate context	10	NEUTRAL	NEUTRAL
Temporal shifts	10	NEUTRAL	NEUTRAL

6 Error Analysis

We analyze the saved classifier outputs in `sentences_classified.csv`, which contains 7,432 sentence-level predictions. Overall accuracy is high at 97.75%, but the class-specific picture is more informative: DOVISH is the hardest class, while NEUTRAL is the easiest.

The dominant confusion is DOVISH → NEUTRAL, which accounts for 69 errors or 19.2% of all DOVISH sentences. This is consistent with subtle easing language being expressed indirectly, for example through conditional or cautiously accommodative wording that the lexicon does not fully capture. The next two confusion pairs are NEUTRAL → HAWKISH and HAWKISH → NEUTRAL, which suggests that some neutral descriptions trigger hawkish keywords while some mixed sentences are still pulled back toward neutrality.

The confidence pattern is also important. Most errors are not low-confidence misses; they are high-confidence mistakes. That indicates a feature mismatch rather than a simple calibration problem. In practical terms, the model is often sure of a wrong label when policy language is mixed, hedged, or context dependent.

The qualitative examples reveal two dominant error patterns. First, dovish sentences often lack explicit easing keywords, causing the model to default to NEUTRAL when the dovish signal is conveyed through indirect or descriptive language (Rows 1, 2). Second, inflation-adjacent language—such as "price stability" and "tightening"—frequently triggers a false HAWKISH prediction even when the surrounding context is neutral or descriptive (Rows 3, 4). Third, descriptive language about hawkish events dilutes genuinely hawkish signals (Row 5), while "hawkish" used descriptively about other banks triggers false hawkish predictions for BB (Row 6). These patterns confirm that the model's lexicon-based features struggle with context-dependent stance and hedged policy language. Error rates also rise monotonically with sentence length, reaching 16.9% above 1,000 characters, consistent with compound passages carrying more mixed signals.

Table 7: Summary of classifier errors on the saved BB outputs.

Metric	Value	Interpretation
Evaluated sentences	7,432	Saved classifier outputs used for error analysis.
Total errors	167	Overall error count.
Overall accuracy	97.75%	High aggregate accuracy hides class asymmetry.
HAWKISH error rate	7.76%	Most mistakes collapse into NEUTRAL.
DOVISH error rate	21.11%	Weak easing language is the hardest to detect.
NEUTRAL error rate	0.87%	Neutral is the most stable class.
Top confusion	DOVISH → NEUTRAL (69)	Largest single confusion pair.
Second confusion	NEUTRAL → HAWKISH (41)	Likely over-triggering on inflation or risk language.
Third confusion	HAWKISH → NEUTRAL (33)	Mixed sentences often collapse back to NEUTRAL.
Low-confidence errors	28	A minority of errors are uncertain.
High-confidence errors	102	Many errors are confidently wrong.
Avg. confidence (correct)	0.984	Correct predictions are highly concentrated near 1.0.
Avg. confidence (errors)	0.857	Wrong predictions still receive strong posterior mass.

Table 8: Qualitative error examples illustrating common confusion patterns.

#	Sentence (truncated)	Weak Label	Model	Error Reason
1	BB's noteworthy initiatives in support of access to finance...	DOVISH	NEUTRAL	Stimulus language present but lexicon match insufficient
2	To directly counter the private sector credit crunch and stimulate economic recovery...	DOVISH	NEUTRAL	Multiple dovish keywords but model predicts neutral
3	Attainment of high real GDP growth with moderating broad money...	NEUTRAL	HAWKISH	Inflation-adjacent language triggers false hawkish
4	The emerging challenges facing the domestic economy...	NEUTRAL	HAWKISH	"Price stability" triggers false hawkish despite balanced context
5	A broad-based economic recovery of Bangladesh has already taken place...	HAWKISH	NEUTRAL	Recovery language dilutes hawkish signal
6	Since mid-2022, global headline inflation has been gradually decreasing...	NEUTRAL	HAWKISH	"Hawkish" used descriptively about other banks, not BB stance

7 Discussion

7.1 Key Findings

Our study yields five principal findings.

Finding 1: Zero-label CB-SentiLex measurement is feasible for Bangladesh Bank, and the human calibration set confirms moderate label quality. Our curated 75-term lexicon provides transparent, reproducible weak labels and supports a document-level stance index without corpus-wide expert annotation. The 300-sentence, three-annotator calibration set yields a blind Fleiss' $\kappa = 0.204$ (two independent annotators) and consensus-vs-weak macro-F1 = 0.664, showing that the weak labels are informative but conservative in borderline cases. Transformer models trained on the packaged benchmark subset reach about 0.87 macro-F1 agreement, confirming that the label scheme is internally learnable even under severe class imbalance. This does not establish the human labels as a full-corpus gold standard; instead, it shows that a low-cost lexical measurement framework can be audited and calibrated.

The feasibility of CB-SentiLex has practical implications for researchers working with resource-constrained central banks. Expert annotation of central bank text requires domain expertise, costs approximately \$0.50–1.00 per sentence, and produces labels that may reflect annotator subjectivity (Shah et al. 2023). Our approach requires no expert labels for corpus construction, and the released calibration set shows where the weak labels are too permissive or too conservative. For Bangladesh Bank, this provides a scalable path to stance-index construction.

Finding 2: Error analysis shows that subtle easing language is the hardest part of the task. Section 6 shows that DOVISH → NEUTRAL is the dominant confusion, followed by NEUTRAL → HAWKISH and HAWKISH → NEUTRAL. The same section also shows that most errors are high-confidence rather than low-confidence misses, which means the model is not merely uncertain: it is often confidently wrong when policy language is mixed, hedged, or indirect. Longer sentences are also harder to classify, which suggests that compound policy rationale remains a key failure mode for both lexical and neural methods.

Finding 3: Transformer agreement is high but protocol-sensitive. On the fixed BB weak-label split, DistilBERT and RoBERTa are tied at about 0.87 macro-F1. FinBERT reaches 0.815 ± 0.020 under 5-fold cross-validation, which is lower but not directly comparable because the evaluation protocol differs. The main implication is not that one architecture dominates, but that CB-SentiLex labels are learnable across model families when the protocol is held constant.

Finding 4: External policy-decision validation is meaningful for BB, but it is horizon-sensitive. Policy-rate direction provides an external, non-human validation signal for the stance index. CB-SentiLex stance scores correlate significantly with contemporaneous rate-change direction for BB ($\rho = 0.406$, $p = 0.011$), but the next-quarter correlation is weaker and not statistically significant. Annual World Bank inflation and exchange-rate validation is null, and monthly BIS exchange-rate validation is directionally positive but weak. This pattern supports the interpretation of CB-SentiLex as a continuous stance measure rather than a high-accuracy policy-decision, macro-forecasting, or market-prediction model.

Finding 5: Class imbalance reflects institutional communication norms. The 89:6:5 distribution in our corpus is consistent with how central banks communicate: most sentences describe conditions, procedures, and projections rather than explicit stance shifts. This finding has methodological implications: researchers working with central bank text should adopt macro-F1 and class-specific metrics rather than accuracy, which can be inflated by trivially predicting the majority class.

The institutional sources of this imbalance are well documented. Central banks employ measured, balanced language to avoid signaling excessive commitment to future policy actions (Blinder et al. 2008). Hawkish or dovish signals are reserved for specific passages that convey the MPC's assessment of risks and policy intent, embedded within a predominantly neutral institutional register. This communication strategy serves multiple purposes: it maintains policy flexibility, reduces market volatility, and preserves the central bank's ability to adjust course as economic conditions evolve.

7.2 Implications for Central Bank Transparency Research

Our findings contribute to the growing literature on central bank communication and transparency. The external policy-decision and falsification results suggest that NLP-derived textual measures capture economically meaningful directionality, particularly for same-quarter BB decisions. At the same time, the modest macro-F1 scores, the BB control-variable caveat, the null annual macro validation, and the statistically weak monthly FX validation show that textual stance is not reducible to discrete rate moves or market outcomes. This finding does not diminish the value of textual analysis; rather, it suggests that stance measurement may be more useful for understanding expectation management, forward guidance, and communication effectiveness than for directly predicting every policy decision.

For central bank transparency research, our work suggests that weak-label agreement is learnable across architectures, but the evaluation protocol matters. DistilBERT and RoBERTa are tied on the fixed BB split, and FinBERT is lower under 5-fold cross-validation. The practical takeaway is that protocol choice can change the apparent ranking as much as model family.

For policymakers in emerging markets, our results demonstrate that NLP tools can provide real-time monitoring of central bank communication stance. The BB stance score time series reveals clear patterns: hawkish periods preceding rate hikes, dovish periods during COVID-19, and a gradual normalization in 2022–2023. These patterns could be integrated into macro-financial surveillance systems to complement traditional economic indicators, even if the direct correlation with policy rates is weaker than initially hypothesized.

7.3 Limitations

We acknowledge several important limitations.

Resource costs. The entire pipeline—corpus construction, lexicon curation, weak labeling, transformer fine-tuning, evaluation, and artifact preparation—was completed on a single Tesla T4 GPU (16 GB) with no paid annotation, no cloud compute, and no

institutional compute allocation. Corpus scraping and preprocessing required approximately 4 GPU-hours. Transformer fine-tuning (DistilBERT, RoBERTa, FinBERT) required approximately 2 GPU-hours total. Human annotation of the 300-sentence calibration set was performed by three volunteer annotators with domain expertise; no annotation budget was required. The total direct cost of the project is effectively zero beyond the authors' time, which confirms the feasibility of the weak-supervision approach for resource-constrained research settings.

Corpus size. Our BB corpus comprises 39 documents yielding 8,647 classified sentences, which is a modest sample compared to much larger central-bank resources. While sufficient for training a DistilBERT classifier, this small sample constrains the complexity of models that can be reliably trained and limits the generalizability of our findings. The temporal split evaluation (macro-F1 = 0.603 vs. 0.870 random split) confirms that temporal concept drift is a significant challenge, and larger corpora would enable more robust temporal generalization.

English-language limitation. BB publishes its MPS in English rather than Bengali, reflecting the institutional norms of South Asian central banking. Consequently, our analysis addresses English-language financial text rather than Bengali per se. This means our study does not fully resolve the broader Bengali financial NLP gap, though it does demonstrate the feasibility of stance detection for a South Asian central bank.

Lexicon limitations. CB-SentiLex has fundamental limitations in capturing nuanced policy language. Section 6 shows that DOVISH → NEUTRAL is the dominant confusion, with additional errors concentrated in NEUTRAL → HAWKISH and HAWKISH → NEUTRAL cases. This is exactly what we would expect when subtle easing language, negation, and hedging are present, because a keyword-based scheme struggles to recover stance that is expressed indirectly. While the DistilBERT classifier learns some contextual patterns from the training data, the ceiling imposed by weak lexical labels limits what transformer agreement can prove.

Evaluation constraints. We now have a 300-sentence, three-annotator calibration set, but it is still too small to replace weak labels for the full corpus or to claim a corpus-wide gold standard. Our evaluation therefore combines human calibration with external economic validation, falsification tests, robustness checks, and model agreement rather than treating weak labels as ground truth. The policy-decision validation, annual macro validation, monthly FX validation, temporal split evaluation, and adversarial testing strengthen construct validity but do not replace broader sentence-level calibration. Future work should expand human annotation coverage and pair it with higher-frequency macro-financial validation to calibrate CB-SentiLex label quality.

7.4 Evidence Stack

The manuscript relies on a stacked validation design. No single layer is decisive on its own, but the layers reinforce each other where the signal is strongest. Weak-label agreement (DistilBERT/RoBERTa at ∼0.87 macro-F1) confirms internal coherence; human calibration (blind $\kappa = 0.204$) documents genuine ambiguity; error analysis reveals that subtle, mixed, or elongated policy language is the primary failure mode; policy-decision validation ($\rho = 0.406$, $p = 0.011$) confirms contemporaneous alignment with rate direction; temporal split evaluation (0.870 → 0.603) documents concept

Table 9: Multi-dimensional comparison of CB-SentiLex with related central bank communication analysis studies.

Study	Bank	Corpus	Annotation	Weak Sup.	Ext. Val.	IAA
Lucca and Trebbi (2009)	Fed	FOMC stmts	Semantic orientation	—	—	—
Hansen and McMahon (2016)	Fed	FOMC transcripts	Topic models	—	—	—
Shah et al. (2023)	Fed	∼11K sent.	Crowdsourced expert	—	—	κ rept.
Gambacorta et al. (2024)	BIS	BIS docs	LLM + domain FT	—	—	—
Silva et al. (2025)	Multi	Multi-bank	LLM cross-inst.	—	—	—
Shah et al. (2025)	Multi	27 CBs	Unified model	—	—	—
Yao et al. (2026)	Fed	FOMC docs	DCS + RoBERTa	—	—	—
Ours	**BB**	**39 docs, 8,647 sent.**	**Weak lexicon + human calib.**	✓	$\rho = 0.406$ ($\mathbf{p = 0.011}$)	$\kappa = 0.$

drift; and adversarial testing shows conservative default behaviour. Full details are in Appendix C3.

Together, these layers support a cautious but substantive claim: CB-SentiLex is a valid stance index for directionality and interpretability, but not a high-accuracy classifier of every external outcome.

7.5 Comparison with Prior Work

Table 9 expands this comparison across qualitative dimensions—corpus coverage, annotation methodology, weak supervision, external policy validation, inter-annotator agreement, and benchmark release. Every prior study targets the Federal Reserve, BIS, or a multi-bank sample; none addresses a South Asian central bank. Shah et al. (2023) and Shah et al. (2025) release reusable corpora but rely on expert annotation without weak-label pathways. Gambacorta et al. (2024), Silva et al. (2025), and Yao et al. (2026) advance LLM-based analysis yet provide no public benchmark or inter-annotator agreement. Earlier work by Lucca and Trebbi (2009) and Hansen and McMahon (2016) laid foundations for computational central-bank text analysis but did not release resources for reproducible downstream classification.

CB-SentiLex occupies a distinct position as the only framework that simultaneously provides (a) weak supervision via a domain-specific 75-term lexicon; (b) human calibration with blind inter-annotator agreement ($\kappa = 0.204$); (c) external policy-decision validation ($\rho = 0.406$, $p = 0.011$); and (d) a full benchmark release. No prior study combines all four, and no existing resource covers a South Asian central bank. Weak supervision, when supplemented by transparent calibration and external validation, can serve as a bridge between expensive full-corpus annotation and purely qualitative document review.

7.6 Annotation Interpretation

The three-annotator $\kappa = 0.231$ is higher than the blind two-annotator $\kappa = 0.204$ because the third annotator's labels closely tracked the weak reference, providing guided validation rather than fully independent judgment. Both figures are well below the $\kappa > 0.67$ threshold often cited as substantial agreement, which is consistent with the deliberate caution and hedging inherent in central bank communication. The weak labels remain the backbone of the zero-budget pipeline, while the human calibration set serves as a benchmark and audit trail. Future work can use the consensus file to seed active learning or targeted annotation of difficult sentence types without discarding the reproducible weak-label framework.

8 Conclusion

This paper presents, to our knowledge, the first large-scale computational analysis of Bangladesh Bank monetary policy stance using a transparent weak-label pipeline. The core contribution is a reproducible BB stance corpus of 39 Monetary Policy Statements and 8,647 sentences, paired with a 300-sentence human calibration set that makes the lexical labels auditable. DistilBERT and RoBERTa learn the weak-label signal well, while FinBERT and simpler baselines provide additional reference points. The stance index aligns meaningfully with BB policy-rate direction, but annual macro and monthly FX validation are weaker, which suggests that the index is best understood as a communication measure rather than a universal macro predictor.

Two future directions are immediate. First, the corpus can be expanded with additional policy documents, press releases, and event-window annotations to reduce temporal drift and strengthen external validation. Second, the same framework can be applied to Bengali financial text once suitable corpora become available. Those extensions would move the project from a benchmark stance index toward a more complete system for monitoring central-bank communication in emerging markets.

Reproducibility

The corpus, calibration set, validation artifacts, and training outputs are organized in the workspace so the full pipeline can be reproduced on a single GPU. In practice, the benchmark fits comfortably within the resource envelope of a Tesla T4-class machine.

Declarations

- **Funding:** The authors received no funding for this research.
- **Conflict of interest:** The authors declare no competing interests.
- **Ethics approval:** Not applicable. This study uses publicly available institutional documents and does not involve human participants, animal subjects, or personally identifiable information.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.

Table A1: Complete CB-SentiLex term list (75 terms: 34 hawkish, 41 dovish). Terms are grouped by semantic family within each class.

Hawkish (34)	**Dovish (41)**
Rate keywords	*Rate keywords*
rate hike, raise rates, increase rates	rate cut, cut rates, lower rates, reduce rates
gradual tightening, further rate increases, additional tightening	easing, easing cycle
Inflation keywords	*Inflation keywords*
elevated inflation, inflationary pressures	benign inflation
core inflation, persistent inflation	inflation below target, well below target
price stability, inflation expectations	below target, undershoot target
anchor inflation, inflation above target	deflation
inflation remains above, above tolerance, above target	
contain inflationary pressures	
Stance keywords	*Stance keywords*
tightening, tighten, restrictive	accommodative, accommodate
hawkish, overheating	dovish, extraordinary
resolute, vigilant, remain vigilant, strong vigilance	
Policy keywords	*Policy keywords*
withdrawal of accommodation	stimulus, stimulate
normalize policy, policy normalization	support growth, support the economy
reduce accommodation, less accommodative	policy support, economic support
calibrated monetary policy action	support recovery, support growth
	downside risks, growth slowdown, weak growth
	recession, slack, output gap
	accommodate growth, revive growth, spur growth
	growth supportive, growth supportive
	continue with accommodative, remain accommod
	maintain accommodation, unprecedented support
	policy levers, policy pivot, change stance to accor

- **Availability of data and materials:** The BB-CB-SentiLex corpus, CB-SentiLex-Bengali lexicon, calibration labels, trained models, and all analysis scripts are publicly available on Zenodo at https://doi.org/10.5281/zenodo.21337423 and on GitHub at https://github.com/nabil0x/CB-Sentilex-Bengali.
- **Code availability:** All code used in this study is available at https://github.com/nabil0x/CB-Sentilex-Bengali.
- **Authors' contributions:** ANN designed the study, collected and curated the data, performed all experiments, and wrote the manuscript. UH contributed to the conceptual framing and manuscript revision.

Appendix A CB-SentiLex Full Term List

Table A1 lists all 75 terms in the CB-SentiLex lexicon. Terms are organized by semantic family within each directional class.

Table B2: Fine-tuning hyperparameters.

Hyperparameter	Value
Learning rate	2×10^{-5}
Batch size	32
Max sequence length	128 tokens
Epochs	8
Warmup steps	100
Weight decay	0.01
Optimizer	AdamW
Scheduler	Linear warmup + decay

Table C3: Released artifacts and their descriptions.

Artifact	Description
BB-CB-SentiLex corpus	39 documents, 8,647 sentences with weak labels
CB-SentiLex-Bengali	75-term lexicon (34 hawkish, 41 dovish)
Calibration set	300 sentences with 3-annotator consensus
Trained models	DistilBERT, RoBERTa, FinBERT checkpoints
Analysis scripts	All evaluation, validation, and plotting code

Appendix B Fine-Tuning Hyperparameters

Table B2 lists the hyperparameters used for all transformer fine-tuning experiments.

Appendix C Released Artifacts

Table C3 summarizes the complete resource package released with this paper.

The calibration set includes full annotator-level detail (labels, confidence, notes) to support future inter-annotator agreement studies and active-learning experiments. Additional robustness figures (adversarial visualization, evidence stack, training curves, threshold sensitivity, macro-FX validation) are available in the release repository at https://github.com/nabil0x/CB-Sentilex-Bengali.

References

Araci D (2019) Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:190810063

Bank for International Settlements (2026) Bis exchange rates against usd (ws_xru). BIS Data Portal

Blinder AS, Ehrmann M, Fratzscher M, et al (2008) Central bank communication and monetary policy: A survey of theory and evidence. Journal of Economic Literature 46(4):910–945

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Conneau A, Khandelwal K, Goyal N, et al (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 8440–8451

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp 4171–4186

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5):378–382

Forman G, Cohen I (2004) Learning from little: Comparison of classifiers given little training. In: Proceedings of the European Conference on Machine Learning, pp 161–172

Gambacorta L, Hofmann B, Peeters K, et al (2024) Central bank communication with large language models. Working Papers 1175, BIS

Hansen S, McMahon M (2016) Transcendental openmarkov: Understanding the monetary policy transmission process through text analysis. American Economic Review 106(5):535–539

Hasan MZ, Biswas P, Alam AKMN, et al (2023) A survey on bengali natural language processing: Progress, challenges, and future directions. Natural Language Engineering 29(4):981–1018. https://doi.org/10.1017/S1351324922000515

Loughran T, McDonald B (2011) When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66(1):35–65

Lucca DO, Trebbi F (2009) Measuring central bank communication: An automated approach with application to fomc statements. Working Paper 15310, NBER

Ratner A, Bach SH, Ehrenberg H, et al (2019) Snorkel: Rapid training data creation with weak supervision. In: Proceedings of the VLDB Endowment, pp 269–282

Sanh V, Debut T, Chaumond J, et al (2019) Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108

Shah A, Sukhani S, Pardawala H, et al (2025) Words that unite the world: A unified framework for deciphering central bank communications globally. Working Paper 5259866, SSRN, https://doi.org/10.2139/ssrn.5259866

Shah K, Paturi K, Chava S (2023) Trillion dollar words: A new financial dataset, task & market analysis. In: Proceedings of the 61st Annual Meeting of the Association

for Computational Linguistics, pp 1–12

Silva TC, Moriya K, Veyrune RM (2025) From text to quantified insights: A large-scale llm analysis of central bank communication. Working Paper 2025/109, International Monetary Fund

World Bank (2026) World development indicators (wdi). World Bank Open Data

Yao R, Chai Q, Yao J, et al (2026) Interpreting fedspeak with confidence: A llm-based uncertainty-aware framework guided by monetary policy transmission paths. Proceedings of the AAAI Conference on Artificial Intelligence 40(40):34414–34422. https://doi.org/10.1609/aaai.v40i40.40739