Can We Reliably Predict the Fed's Next Move? A Multi-Modal Approach to U.S. Monetary Policy Forecasting

Fiona Xiao Jingyi and Lili Liu*, National University of Singapore

***Abstract*—Forecasting central bank policy decisions remains a critical challenge for financial institutions, investors, and policymakers, given the profound influence of monetary actions on market dynamics and macroeconomic conditions. In particular, anticipating changes in the U.S. federal funds rate is essential for effective risk management and the formulation of informed trading strategies. Yet, traditional approaches relying solely on structured economic indicators may fall short in capturing the forward-looking nuances conveyed through central bank communications.**

This study examines whether predictive performance can be improved by integrating structured macroeconomic data with unstructured textual signals from Federal Reserve communications. We adopt a multi-modal modeling framework to compare the effectiveness of traditional machine learning classifiers, transformer-based language models, and deep learning architectures across both unimodal and hybrid configurations.

Empirical results demonstrate that hybrid models consistently outperform unimodal baselines. The strongest performance is achieved by integrating TF-IDF representations of FOMC texts with structured economic features in a gradient boosting (XGBoost) classifier, achieving a test AUC of 0.83. In contrast, deep learning approaches using FinBERT-derived sentiment probabilities provide marginal improvements in ranking but underperform in classification accuracy, particularly in the presence of class imbalance. SHAP analysis further reveals that interpretable, sparse features better reflect policy-relevant signals in formal monetary texts.

These findings suggest that, in the context of financial policy prediction, simplicity and interpretability can be powerful assets. Our results offer practical insights for scholars, policymakers, and practitioners alike, highlighting the value of hybrid, transparent models for navigating the evolving landscape of central bank decision-making.

***Index Terms*—Federal Reserve, Monetary Policy, Forecasting, Multi-Modal Learning, Sentiment Analysis, Machine Learning**

I. Introduction

Forecasting interest rate decisions by the U.S. Federal Reserve is a core challenge in financial and macroeconomic analysis. These decisions influence asset prices, guide investor expectations, and underpin overall economic stability. Historically, such forecasts have relied heavily on structured macroeconomic indicators—such as inflation rates, employment levels, and GDP growth—closely aligned with the Fed's dual mandate of price stability and full employment.

In the aftermath of the Global Financial Crisis (GFC), the Federal Reserve has placed growing emphasis on forward guidance as a key component of monetary policy communication [1], [2]. Formal channels—such as statements, speeches, meeting minutes, and press conferences—now serve not merely as reflections of policy decisions but as instruments to shape expectations and convey strategic intent [3]. This development reflects a broader methodological shift: from purely rule-based, data-driven forecasting toward frameworks that integrate both structured economic indicators and unstructured textual signals.

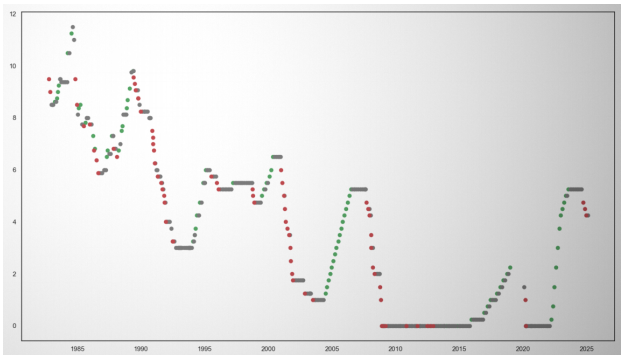

Fig. 1. Federal Funds Rate trajectory (Jan 1980 – Jan 2025).

This development raises an important research question:

> *Can combining macroeconomic indicators with unstructured Fed communication improve the predictive accuracy and interpretability of interest rate forecasts?*

As shown in Figure 1, U.S. interest rates have exhibited marked volatility over the past decades, driven by complex macroeconomic and geopolitical dynamics. Anticipating these movements remains a high-stakes task for market participants and policymakers alike.

To address this challenge, we introduce a multi-modal forecasting framework that integrates structured macroeconomic indicators with sentiment signals derived from Federal Reserve communications. Our main contributions are as follows:

- We develop a hybrid model that synthesizes sentiment probabilities with curated economic features to capture both quantitative fundamentals and qualitative policy narratives.
- We perform a rigorous comparative evaluation across multiple model architectures, quantifying the individual and combined predictive strengths of each data modality.
- We apply SHAP-based interpretability to clarify the role of textual and numerical features, enhancing transparency and supporting meaningful economic insight.

By jointly modeling structured and unstructured inputs, our approach advances monetary policy forecasting in both accuracy and interpretability. The proposed system achieves a test AUC of 0.83, while offering decision-relevant explanations aligned with the practical needs of economists and central bank observers.

II. Literature Review

The majority of existing research has focused on one data modality in isolation—either structured or unstructured. This has limited the potential to uncover richer insights from the interplay of data types. In what follows, we provide a structured review across both dimensions and highlight recent efforts toward multi-modal approaches.

A. Structured Economic Indicators in Policy Forecasting

Rule-based models, such as the Taylor Rule [10], represent early structured approaches to policy forecasting. These models relate policy rates to deviations in inflation and output gaps. Building on these foundations, econometric frameworks like vector autoregressions (VARs) [15] have been widely used to capture dynamic interdependencies among macroeconomic variables. Furthermore, GARCH-type models [19] have offered valuable tools for estimating and forecasting volatility in financial time series. Despite their rigor, these methods often lack the ability to incorporate qualitative or forward-looking signals embedded in central bank communication.

B. Central Bank Communications and Textual Analysis

In parallel, a growing body of work investigates the predictive utility of central bank discourse. Early efforts used basic textual metrics or manual annotations to assess tone and emphasis. The use of natural language processing (NLP) has since expanded, allowing for more systematic analysis. Texts such as meeting minutes, statements, speeches, and press conference transcripts offer rich linguistic cues that can influence market expectations and policy interpretation [2], [3]. The evolution of computational linguistics has made it possible to extract patterns, narratives, and sentiment embedded within these documents [1].

C. Sentiment Analysis Techniques in Central Bank Communication

Lexicon-based sentiment analysis, particularly using the Loughran–McDonald dictionary [20], was among the first attempts to quantify tone in financial texts. This approach categorizes words as positive, negative, uncertain, or litigious based on financial context. Later empirical studies [18], [22] showed that shifts in central bank tone could predict changes in interest rate expectations and asset prices. More advanced methods integrate domain-specific lexicons with traditional NLP pipelines, enhancing robustness.

D. Deep Learning and the Emergence of Contextual Language Models

With the advent of deep learning, transformer-based models like BERT and FinBERT have substantially improved contextual understanding in text data [21]. FinBERT, fine-tuned on financial corpora, excels in extracting nuanced sentiment and semantic features from central bank documents. These models outperform earlier techniques in capturing subtleties such as policy uncertainty, indirect signaling, and sentiment asymmetry. However, most studies employing these models treat text in isolation, without integrating structured macroeconomic signals.

E. Toward Multimodal Prediction

The integration of structured and unstructured data sources is still nascent. Some recent studies attempt to fuse time series data with text embeddings, using concatenation or attention mechanisms [4]. Yet, comprehensive frameworks for monetary policy prediction remain limited. Bridging this gap requires not only technical innovation but also interpretability to ensure insights are meaningful to economists and policymakers. Our approach contributes to this space by unifying sentiment features with macroeconomic indicators, enabling a more holistic and explainable forecast.

III. Description of Dataset

This section outlines the data sources used in our multimodal forecasting framework. The dataset comprises both structured macroeconomic indicators and unstructured textual data from official Federal Reserve communications. The target variable is constructed from historical interest rate policy decisions made by the Federal Open Market Committee (FOMC).

A. Structured Dataset: Economic Indicators

The structured component consists of key macroeconomic indicators commonly referenced in monetary policy analysis. These include:

- Inflation measures: Consumer Price Index (CPI), Personal Consumption Expenditures (PCE) price index
- Labor market indicators: Unemployment rate, Nonfarm Payroll Employment (NFP)
- Housing metrics: Housing Starts (HOUST), Home Price Index (HPI)
- Interest rate spreads: 10-Year Treasury Yield minus 3-Month Treasury Bill (10Y–3M spread)

All data are retrieved from the Federal Reserve Economic Data (FRED) database, maintained by the Federal Reserve Bank of St. Louis [8]. Features are transformed into monthly and year-over-year differences where appropriate and standardized to ensure comparability across input dimensions.

B. Unstructured Dataset: Federal Reserve Communications

The unstructured dataset includes official textual releases by the Board of Governors of the Federal Reserve System. These documents span from January 2011 to January 2025 and cover five key types of communication:

- FOMC Statements
- Meeting Minutes
- Speeches by Federal Reserve officials
- Testimonies before Congress
- Press Conference transcripts (presconf), which represent the prepared remarks delivered by the Federal Reserve Chair

Text preprocessing involves tokenization, stopword removal, and lemmatization. Sentiment features are extracted using two approaches: (1) Term Frequency–Inverse Document Frequency (TF–IDF) vectors combined with Loughran–McDonald sentiment scores [20], and (2) sentiment class probabilities generated by FinBERT, a transformer-based language model fine-tuned for financial text classification [21].

C. Target Variable

The prediction task is framed as a three-class classification problem, with the target variable representing the direction of FOMC interest rate decisions. Each policy action is labeled as one of the following:

- *Raise* – indicating an increase in the federal funds target rate
- *Hold* – indicating no change in the target rate
- *Lower* – indicating a decrease in the target rate

Labels are derived from official post-meeting rate announcements and validated against historical FOMC decision records published by the Federal Reserve [**?**]. This formulation enables a structured evaluation of how well each data modality contributes to the forecast of monetary policy shifts.

IV. Baseline Models

This section presents the construction, training, and evaluation of baseline models that establish performance benchmarks for monetary policy classification. The methodology spans four key components: data preprocessing, feature engineering, model selection, and evaluation design.

A. Data Preprocessing and Feature Engineering

We begin by assembling a unified dataset that integrates structured macroeconomic indicators with unstructured textual sentiment from Federal Reserve communications. Macroeconomic variables—such as inflation, employment, and GDP growth—are standardized to ensure temporal coherence and comparability across features.

To capture linguistic signals, we extract sentiment scores from monetary policy documents using FinBERT, a domain-specific transformer model fine-tuned for financial text. Sentiment probabilities (positive, negative, and neutral) are aggregated at both the document and policy decision levels. In parallel, TF–IDF representations are constructed for each communication, and Loughran–McDonald sentiment scores are appended to enrich interpretability. Collectively, these features form a comprehensive pipeline capturing both economic fundamentals and narrative tone.

B. Model Architecture

We benchmark several supervised learning algorithms to establish comparative performance baselines. These include Logistic Regression, Random Forest, Extra Trees, and Gradient Boosting classifiers. For text-based inputs, we consider two representations: (i) traditional TF–IDF embeddings enhanced with sentiment scores, and (ii) FinBERT-derived sentiment probabilities. For hybrid models, we concatenate these representations with structured macroeconomic variables to assess joint predictive strength. Among these, Gradient Boosting consistently demonstrated superior capability in modeling complex and nonlinear feature interactions.

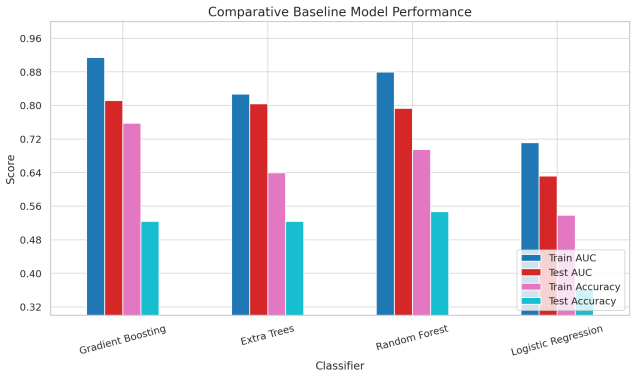

Fig. 2. Baseline comparison of classifier performance in terms of ROC AUC and accuracy. Gradient Boosting consistently outperforms other models on both training and test sets. Logistic Regression exhibits underfitting, while tree-based models demonstrate stronger predictive capacity.

C. Experimental Design

To ensure rigorous and generalizable evaluation, we adopt a stratified 5-fold cross-validation strategy. This preserves label distribution across splits and mitigates overfitting. Performance is assessed using both threshold-independent metrics (ROC AUC) and threshold-dependent metrics (accuracy, precision, recall, and F1-score). Given the class imbalance in the target variable—particularly underrepresentation of "Hike" and "Cut" decisions—we incorporate the Synthetic Minority Oversampling Technique (SMOTE) and class-weighted loss functions during training. This strategy reflects real-world challenges in imbalanced policy classification.

D. Model Tuning and Class Imbalance Handling

To optimize model performance, we employ a two-stage tuning strategy. First, we perform randomized hyperparameter search to efficiently navigate the parameter space, followed by fine-tuning using grid search. For the Gradient Boosting classifier, the optimal configuration includes:

- `n_estimators = 10`
- `learning_rate = 0.01`
- `max_depth = 4`
- `max_features = 'sqrt'`
- `min_samples_leaf = 10`
- `min_samples_split = 10`

Four classifiers are benchmarked under this framework. Each model is trained using stratified cross-validation, and its generalization capacity is assessed on a held-out test set. The class imbalance problem is further addressed by integrating SMOTE, which synthetically augments minority class instances to promote equitable learning.

E. Performance Summary and Insights

As illustrated in Figure 2, gradient Boosting delivers the most robust results across both AUC and accuracy metrics. It effectively captures complex feature interactions and remains resilient under cross-validation. Extra Trees and Random Forest models perform moderately well but show tendencies toward overfitting. Logistic Regression, while interpretable,

lags in generalization and fails to capture the nuances of multimodal inputs.

Our tuned Gradient Boosting model achieves excellent in-sample performance (Train AUC: 0.9139; Accuracy: 75.7%) and respectable generalization on the test set (Test AUC: 0.8116; Accuracy: 52.3%). Notably, after applying SMOTE to mitigate class imbalance, the model's AUC improved, though accuracy declined slightly due to the introduction of synthetic variance.

In conclusion, Gradient Boosting emerges as a strong candidate for further development within our multimodal prediction framework. Its balance between predictive power and robustness makes it well-suited for capturing the intricacies of monetary policy classification, driven by both data and discourse.

V. Text-Only Models

This section investigates the standalone predictive power of unstructured textual data drawn from official Federal Reserve communications. Specifically, it evaluates whether policy-related texts—such as FOMC statements, meeting minutes, press conferences, and speeches—contain sufficient informational signals to forecast U.S. interest rate decisions without the aid of traditional economic indicators.

We adopt a rigorous pipeline consisting of natural language preprocessing, exploratory linguistic and sentiment analyses, and the application of both traditional machine learning classifiers and transformer-based deep learning models. The objective is to assess the quality and limitations of textual signals in capturing monetary policy intent.

A. Data Preprocessing

To prepare the unstructured text data for modeling, a standardised cleaning process was applied across all document types. This included case normalization, removal of extraneous characters, punctuation filtering, and stopword removal. Importantly, lemmatization was excluded to preserve domain-specific terminology critical in financial discourse.

Each document was aligned with the Federal Reserve's corresponding rate decision using the FOMC calendar. In cases of overlapping or compound documents, chunking strategies were employed to handle input length constraints for transformer models.

B. Exploratory Data Analysis

We conducted exploratory linguistic analyses to uncover patterns across different document types and decision categories. Distributions of word counts were visualized to inform model constraints (e.g., BERT's token limits), and top-words analysis revealed vocabulary clusters associated with each decision type—offering early evidence of tone divergence between hawkish, neutral, and dovish communications.

We next analyze dominant terms associated with each decision class: *Lower*, *Hold*, and *Raise*. As shown in Figure 3, references to financial instability dominate the language of "Lower" decisions, while "Raise" statements emphasize inflation and labor markets. "Hold" decisions reflect a neutral, status-quo tone.

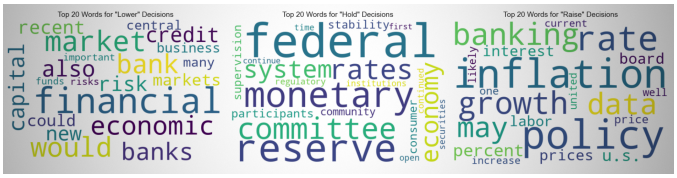

Fig. 3. Top words per decision class. Themes vary significantly across monetary policy stances.

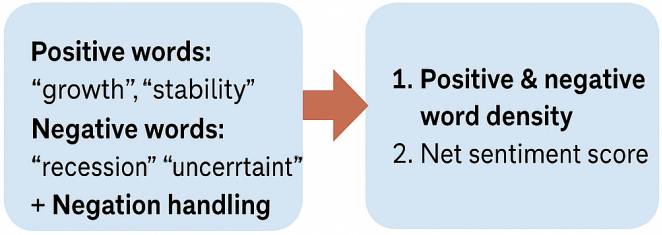

Fig. 4. Rule-based sentiment pipeline using Loughran–McDonald dictionary.

C. Sentiment Analysis

To quantify linguistic tone, we applied both rule-based and model-based sentiment methods.

1) Dictionary-Based Sentiment Analysis: Using the Loughran–McDonald (LM) sentiment lexicon [20], we computed sentiment scores from positive and negative word counts, normalized by document length. A negation-aware scoring mechanism was implemented to reduce semantic errors. We extracted three metrics: positive density, negative density, and net sentiment. Temporal and categorical analyses showed that sentiment shifts generally align with known macroeconomic stress periods, validating the LM dictionary's utility in formal policy text. Figure 4 summarizes this extraction pipeline.

Figure 5 presents the evolution of net sentiment over time. Notably, sentiment declines tend to precede major recession periods, reinforcing its potential as a leading indicator.

As seen in Figure 6, sentiment polarity varies by document type. Statements show a slightly positive skew, whereas speeches and testimonies present more diverse and often negative sentiment. Press conferences (*presconf*) exhibit a narrow, neutral range, aligning with their explanatory tone. These distinctions offer valuable inputs for forecasting interest rate decisions.

2) Transformer-Based Sentiment Analysis: We also employed FinBERT [21], a domain-specific transformer model trained on financial corpora, to classify sentiment across the document corpus. Predictions were aggregated from chunk-level softmax probabilities to derive document- and decision-level sentiment summaries. While FinBERT returned mostly neutral classifications, its probabilistic outputs were retained as features in downstream models. FinBERT's strength lies in contextual understanding, offering a probabilistic counterpoint to the interpretable yet rigid dictionary-based scores.

As shown in Figure 7, most communications are labeled

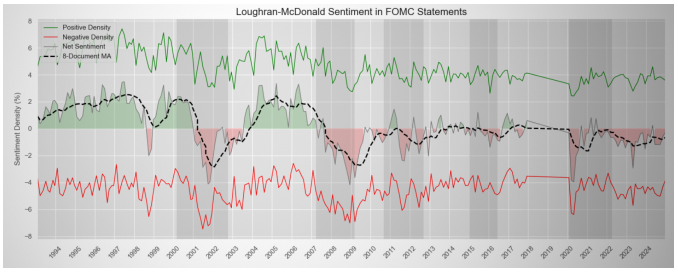

Fig. 5. Net sentiment trends over time. Dips align with U.S. recession periods (shaded).

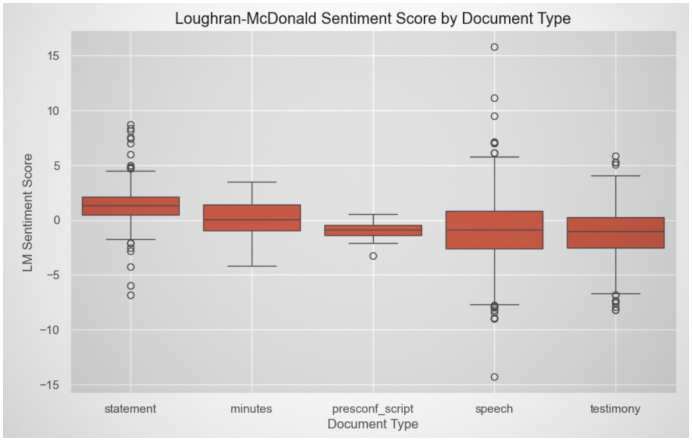

Fig. 6. Distributions of LM sentiment scores by document type.

"neutral," consistent with the Fed's measured communication style. FinBERT's sentiment outputs mirror the measured and neutral tone commonly found in central bank communications, leading to limited variation across sentiment categories. Although less effective in isolation, these scores serve as a valuable probabilistic complement to rule-based methods, motivating the exploration of FinBERT as a direct classifier to capture nuanced financial language better.

D. Model Training and Evaluation

Two categories of models were developed using only textual inputs to forecast interest rate decisions.

- **Traditional classifiers** (Logistic Regression, Naïve Bayes, Random Forest, Extra Trees, Gradient Boosting) were trained on TF-IDF representations of cleaned texts. Despite some variation in accuracy, overall AUC scores remained modest, suggesting that shallow models struggle to extract deep context from policy language.
- **FinBERT fine-tuned classifier** was trained to directly predict the rate decision class. Document chunking and weighted cross-entropy loss were applied to address token constraints and class imbalance. The best checkpoint achieved a validation ROC AUC of 0.69 and accuracy of 0.67, outperforming traditional models but still showing a persistent bias toward the majority "Hold" class.

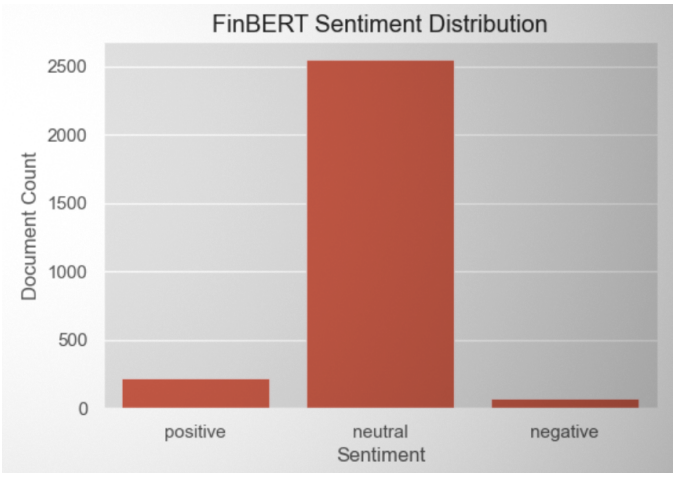

Fig. 7. FinBERT sentiment classification across central bank documents. Majority are labeled as neutral.

TABLE I
TEXT-ONLY MODELS PREDICT INTEREST RATE DECISIONS

Model	Test ROC AUC	Test Accuracy
TF-IDF + Logistic Regression	0.6290	0.4959
TF-IDF + Gradient Boosting	0.6759	0.5190
FinBERT (fine-tuned)	**0.6869**	**0.6690**

E. Evaluation Insights

Although textual models offer some predictive capacity—especially with FinBERT's contextual strengths—the performance ceiling appears constrained by the formal, neutral tone of Fed communications. Discriminative signals are subtle and often insufficient when used in isolation. As summarized in Table I, the standalone use of unstructured text falls short of capturing the full complexity of interest rate decisions. These limitations motivate the integration of structured economic data into a multi-modal learning framework.

VI. Multi-Modal Models

To explore how structured economic data (ED) and unstructured data can complement each other, we develop three multi-modal modeling frameworks. Each framework combines macroeconomic indicators with a different form of textual representation. Our goal is to understand how different combinations influence the accuracy and interpretability of monetary policy forecasts.

1) Method 1: ED + TF-IDF and LM Sentiment Features + XGBoost: The first framework merges structured economic features with two sets of textual inputs: (1) 500-dimensional TF-IDF vectors derived from Federal Reserve communications, and (2) 50 sentiment features based on the Loughran–McDonald (LM) financial dictionary.

Algorithm 1 outlines the process for combining TF-IDF and Loughran–McDonald (LM) sentiment features into a unified feature matrix for model training. The input consists of a document corpus and the LM sentiment lexicon, which includes predefined lists of positive, negative, uncertain, and

Algorithm 1 Combine TF-IDF and LM Sentiment Features

Require: Document corpus $D = \{d_1, d_2, \ldots, d_n\}$
Require: LM sentiment lexicon $L = \{L_{\text{pos}}, L_{\text{neg}}, L_{\text{unc}}, L_{\text{lit}}\}$
Preprocess each document: remove stopwords, lowercase, tokenize
for each document d_i in D **do**
  Compute TF-IDF vector $T_i \leftarrow \text{TFIDF}(d_i)$
  Initialize sentiment counts: $s_{\text{pos}}, s_{\text{neg}}, s_{\text{unc}}, s_{\text{lit}} \leftarrow 0$
  for each token t in d_i **do**
    if $t \in L_{\text{pos}}$ **then**
      $s_{\text{pos}} \leftarrow s_{\text{pos}} + 1$
    end if
    if $t \in L_{\text{neg}}$ **then**
      $s_{\text{neg}} \leftarrow s_{\text{neg}} + 1$
    end if
    if $t \in L_{\text{unc}}$ **then**
      $s_{\text{unc}} \leftarrow s_{\text{unc}} + 1$
    end if
    if $t \in L_{\text{lit}}$ **then**
      $s_{\text{lit}} \leftarrow s_{\text{lit}} + 1$
    end if
  end for
  Normalize sentiment scores by total tokens if desired
  Form LM sentiment vector $L_i \leftarrow [s_{\text{pos}}, s_{\text{neg}}, s_{\text{unc}}, s_{\text{lit}}]$
  Concatenate: $F_i \leftarrow [T_i, L_i]$
end for
return Feature matrix $F = \{F_1, F_2, \ldots, F_n\}$ for model training

litigious words. Each document is first preprocessed through tokenization, stopword removal, and lowercasing. Next, a TF-IDF vector is computed to capture the importance of terms in the corpus. Simultaneously, sentiment counts are tallied by matching tokens against the LM lexicon categories. These counts are optionally normalized and then concatenated with the TF-IDF vector to form a combined feature representation. The resulting feature matrix integrates both frequency-based textual relevance and domain-specific sentiment signals, enabling richer input for downstream classification or prediction tasks.

We use these combined features to train an XGBoost classifier. Figure 8 shows the SHAP summary plot for the XGBoost model using economic and textual features. The most influential variable is the deviation from policy inertia (`Inertia_diff`), which captures intentional shifts in monetary policy behavior. Bond market expectations (`10YUST_diff_prev`) and consumer sentiment (`UMich_diff_prev`) also exhibit strong contributions, especially in identifying "Lower" rate decisions. Housing-related indicators (`HOUST_diff_year`, `HPI_diff_year`) reflect broader macroeconomic conditions and play a notable role. Additionally, textual features such as `tfidf_basis` and `tfidf_difficult`, extracted from FOMC statements, capture subtle narrative tones and help improve both prediction accuracy and model interpretability. This distribution of feature importance supports the value of combining structured data

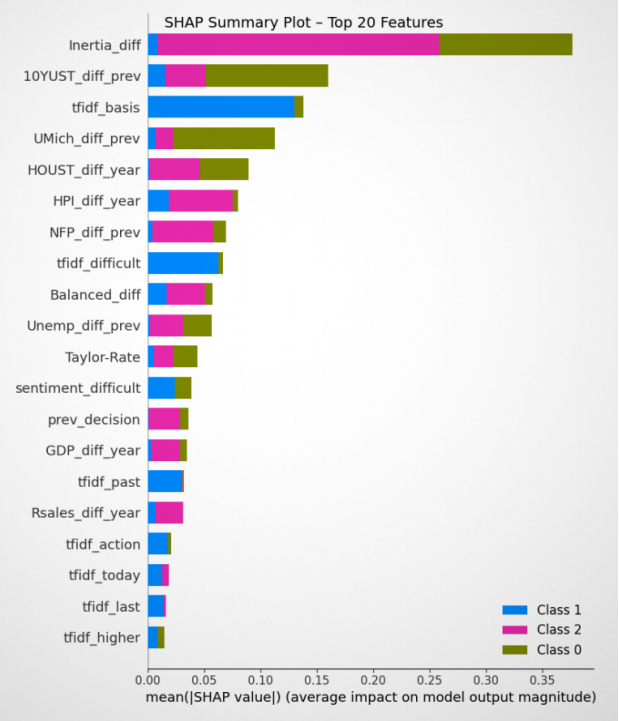

Fig. 8. Method 1: SHAP summary plot highlighting the top features influencing the XGBoost model's predictions

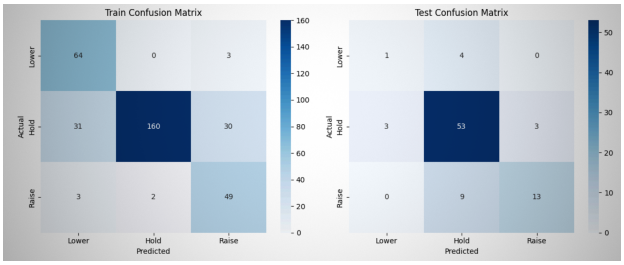

Fig. 9. Method 1: ED + TF-IDF & LM Sentiment + XGBoost

with domain-specific textual cues.

Figure 9 shows the confusion matrix for Method 1. The model performs well across all three decision categories—"Raise," "Hold," and "Lower." Most predictions fall along the diagonal, indicating a high level of agreement with actual FOMC outcomes. While some confusion exists between "Hold" and the adjacent classes, the overall balance suggests that the combined feature set captures relevant policy signals effectively. The model is particularly accurate in identifying the most frequent "Hold" decisions, without severely misclassifying the less common classes.

This model performs strongly, achieving a test AUC of **0.8304** and an accuracy of **77.91%**. Its balance of simplicity and interpretability makes it a promising baseline.

2) Method 2: ED + FinBERT Sentiment Probabilities + XGBoost: In the second modeling framework, we integrated structured economic indicators with sentiment classification probabilities generated by a fine-tuned FinBERT model. While the `TF-IDF` approach captures surface-level word frequencies, this transformer-based method was designed to extract

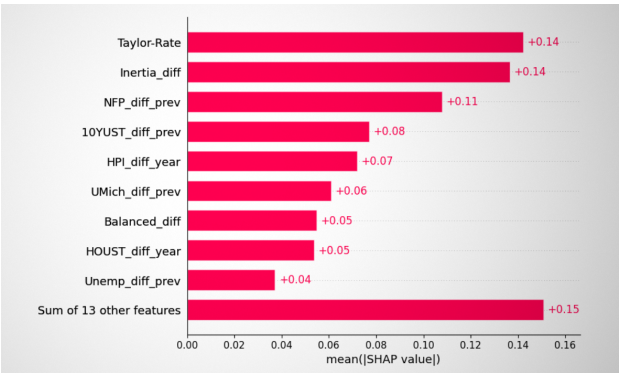

Fig. 10. Method 2: SHAP Summary Plot

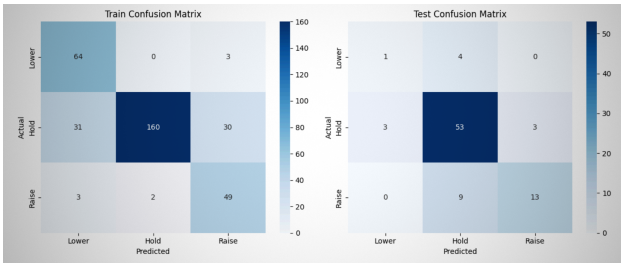

Fig. 11. Method 2: ED + FinBERT Sentiment + XGBoost

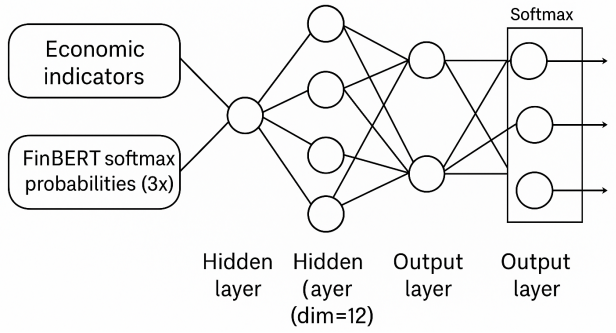

Fig. 12. Feedforward Neural Network (FNN) Architecture

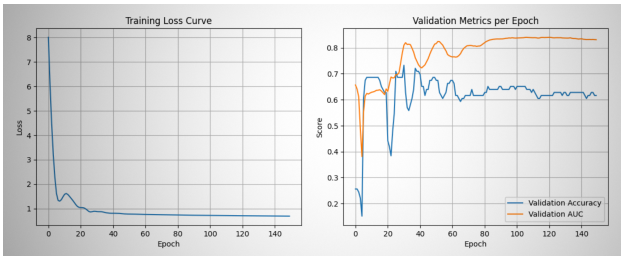

Fig. 13. FNN Training Curve: Loss, AUC, and Accuracy

deeper contextual meaning from Federal Reserve communications. The goal was to evaluate whether FinBERT's domain-specific sentiment scores could enhance predictive performance when combined with macroeconomic data.

We trained an `XGBoost` classifier on the combined feature set. As illustrated in the SHAP summary plot (Figure 10), macroeconomic variables remained the most influential predictors. These included the Taylor Rule-implied interest rate (`Taylor_Rate`), the deviation from policy inertia (`Inertia_diff`), labor market momentum (`NFP_diff_prev`), bond yield changes (`10YUST_diff_prev`), and consumer sentiment (`UMich_diff_prev`).

By contrast, the FinBERT-derived sentiment probabilities contributed modestly to the model's decision process. This suggests that although `FinBERT` captures rich contextual nuance, the relatively shallow architecture of `XGBoost` may not fully exploit its representational depth. These results highlight a trade-off between model simplicity and the complexity of text-based features in financial policy prediction tasks.

Figure 11 presents the confusion matrix for Method 2. The model shows reasonable performance in predicting "Hold" decisions but struggles with minority classes, especially "Raise." Many instances of "Raise" are incorrectly predicted as "Hold," highlighting the challenge posed by class imbalance. This pattern also suggests that FinBERT sentiment probabilities, while capturing general tone, may lack the precision needed to differentiate more subtle policy shifts. Compared to Method 1, this approach trades off some classification accuracy for sentiment-based contextual richness.

The model achieves a test AUC of **0.7960** and an accuracy of **59.30%**. Performance declines suggest that sentiment probabilities from FinBERT may smooth over subtle textual distinctions. The effect is more noticeable under class imbalance.

3) Method 3: ED + FinBERT Sentiment Probabilities + FNN: The third framework tests a deep learning setup. We use a feedforward neural network (FNN) with two hidden layers (64 and 32 units). Input features include economic indicators and FinBERT sentiment probabilities.

This framework achieves the highest test AUC at **0.8404**, suggesting strong ranking ability. However, its accuracy is lower at **61.63%**. The FNN struggles with calibration and misclassifies less common classes like "Raise" and "Lower." Sensitivity to class imbalance may explain this behavior.

4) Comparative Performance and Insights: Figure 15 compares all three hybrid models. The FNN shows the highest AUC but lower overall accuracy. In contrast, the XGBoost model with TF-IDF and LM sentiment offers both solid performance and strong interpretability.

Several insights emerge. First, sparse, transparent text features may capture formal monetary language better than dense transformer outputs. Second, simpler models like XGBoost remain competitive—particularly when paired with carefully selected and engineered inputs. Finally, interpretability tools like SHAP help validate economic meaning and ensure the model aligns with domain intuition.

In sum, we find that hybrid models hold great promise. While deep learning offers flexibility, structured approaches built on interpretable features may be more practical for financial policy forecasting, especially when clarity and trust are essential.

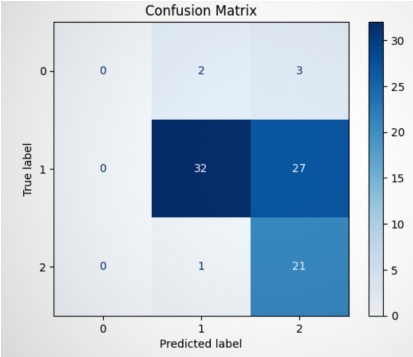

Fig. 14. Method 3: ED + FinBERT Sentiment + FNN

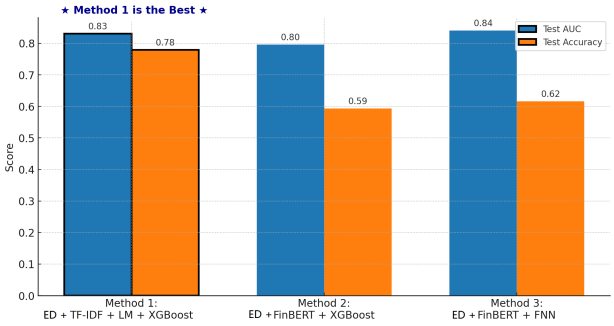

Fig. 15. Comparative Performance of Hybrid Models

VII. Conclusion

A. Key Insights

This study offers a data-driven perspective on forecasting U.S. monetary policy using multi-modal machine learning models. We examined how combining structured economic indicators with unstructured central bank communication can enhance predictive performance and interpretability. Several key insights emerged:

- **Integrating structured and unstructured inputs** consistently improved performance over single-modality models. This highlights the complementary value of quantitative fundamentals and qualitative narratives in policy forecasting.
- **TF-IDF features**, when combined with Loughran–McDonald sentiment scores, captured meaningful linguistic cues from FOMC documents. These sparse, interpretable signals outperformed transformer-based sentiment embeddings in both accuracy and clarity.
- **Shallow models like XGBoost**, when paired with thoughtfully engineered hybrid features, achieved the best balance of performance, robustness, and interpretability. They also managed class imbalance more effectively than deep neural networks, especially in settings with limited or skewed data.

B. Limitations

We remain mindful of the limitations of our current approach. One major challenge is class imbalance, particularly the small number of "Lower" rate decisions. While we applied class-weighting and SMOTE to address this, synthetic oversampling introduced variance and did not significantly improve generalization.

Additionally, FinBERT sentiment probabilities, while rich in context, lacked the granularity needed for fine-grained prediction. Their compressed format limited both precision and transparency in downstream tasks.

C. Future Directions

There are several promising directions for future work:

1) Develop targeted or ensemble classifiers to improve recall on minority classes and reduce model bias.
2) Explore lighter, fine-tuned transformer models adapted specifically to monetary policy language.
3) Incorporate external signals—such as market-based expectations or global economic trends—to broaden the model's situational awareness.

In closing, we believe this work provides a modest contribution to the evolving field of data-driven policy analysis. Our findings highlight the potential of hybrid, interpretable frameworks to support deeper understanding of central bank decision-making. While challenges remain, we hope this research encourages further exploration at the intersection of machine learning, economics, and institutional communication.

VIII. Acknowledgments

We gratefully acknowledge the National University of Singapore (NUS) for its generous support and collaborative environment, which played a pivotal role in enabling this research. In particular, we thank the NUS School of Computing for funding this work through the Graduate Project Supervision Fund (SF). Their guidance, resources, and academic ecosystem have been instrumental in shaping and advancing the direction of this study.

References

[1] S. G. Cecchetti, M. S. Mohanty, and F. Zampolli, "Monetary Policy in the Next Recession?," *CEPR Policy Insight*, no. 103, 2020.
[2] A. S. Blinder, M. Ehrmann, M. Fratzscher, J. de Haan, and D.-J. Jansen, "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence," *Journal of Economic Literature*, vol. 46, no. 4, pp. 910–945, 2008.
[3] R. Fortes and T. Le Guenedal, "Tracking ECB's Communication: Perspectives and Implications for Financial Markets," *Banque de France Bulletin*, no. 231, 2020.
[4] J. Wong and L. Liu, "Portfolio Optimization through a Multi-Modal Deep Reinforcement Learning Framework," *Authorea Preprints*, 2025.
[5] J. B. Taylor, "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, vol. 39, pp. 195–214, 1993.
[6] T. Bollerslev, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.
[7] T. Loughran and B. McDonald, "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.
[8] Federal Reserve Bank of St. Louis, "Federal Reserve Economic Data (FRED)," *Online Resource*, 2025. [Online]. Available: https://fred.stlouisfed.org/
[9] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models," *arXiv preprint arXiv:1908.10063*, 2019.

[10] J. B. Taylor, "Discretion versus policy rules in practice," *Carnegie-Rochester Conference Series on Public Policy*, vol. 39, pp. 195–214, 1993.
[11] M. Apel and M. Grimaldi, "Monetary policy decision-making, market expectations and commitment," *Journal of Monetary Economics*, vol. 59, no. 6, pp. 601–621, 2012.
[12] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.
[13] S. Hansen and M. McMahon, "Transparency and Deliberation in Monetary Policy," *Econometrica*, vol. 86, no. 2, pp. 499–530, 2018.
[14] R. F. Engle, "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.
[15] M. Apel and M. Grimaldi, "Monetary policy decision-making, market expectations and commitment," *Journal of Monetary Economics*, vol. 59, no. 6, pp. 601–621, 2012.
[16] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.
[17] S. Hansen and M. McMahon, "Transparency and Deliberation in Monetary Policy," *Econometrica*, vol. 86, no. 2, pp. 499–530, 2018.
[18] N. Jegadeesh and D. Wu, "Word power: A new approach for content analysis," *Journal of Financial Economics*, vol. 117, no. 2, pp. 371–394, 2015.
[19] T. Bollerslev, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.
[20] T. Loughran and B. McDonald, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.
[21] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint arXiv:1908.10063*, 2019.
[22] S. Hansen and M. McMahon, "Transparency and Deliberation in Monetary Policy," *Econometrica*, vol. 86, no. 2, pp. 499–530, 2018.