ISSN: 2993-8643

Engineering: Open Access

Research Article

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*

Department of Information Systems and Analytics, School of Computing, National University of Singapore

***Corresponding Author**
Lili Liu, Department of Information Systems and Analytics, School of Computing, National University of Singapore.

Submitted: 2025, Jun 13; **Accepted:** 2025, Jul 21; **Published:** 2025, Aug 06

Citation: F. J. Xiao, L. Liu (2025). Can We Reliably Predict the Fed's Next Move? A Multi-Modal Approach to U.S. Monetary Policy Forecasting. *Eng OA, 3*(8), 01-09.

Abstract

Forecasting central bank policy decisions remains a persistent challenge for investors, financial institutions, and policymakers due to the wide-reaching impact of monetary actions. In particular, anticipating shifts in the U.S. Federal Funds Rate is vital for risk management and trading strategies. Traditional methods relying only on structured macroeconomic indicators often fall short in capturing the forward-looking cues embedded in central bank communications.

This study examines whether predictive accuracy can be enhanced by integrating structured data with unstructured textual signals from Federal Reserve communications. We adopt a multi-modal framework, comparing traditional machine learning models, transformer-based language models, and deep learning architectures in both unimodal and hybrid settings.

Our results show that hybrid models consistently outperform unimodal baselines. The best performance is achieved by combining TF–IDF features of Federal Reserve communications with economic indicators in an XGBoost classifier, reaching a test AUC of 0.83. FinBERT-based sentiment features marginally improve ranking but perform worse in classification, especially under class imbalance. SHAP analysis reveals that sparse, interpretable features more effectively capture policy-relevant signals.

These findings underscore the importance of integrating textual and structured signals transparently. For monetary policy forecasting, simpler hybrid models can offer both accuracy and interpretability, delivering actionable insights for researchers and decision-makers.

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 Federal Open Market Committee (FOMC) meeting statements, 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 supports a broader methodological shift in policy forecasting: from purely rule-based, data-driven prediction toward frameworks that integrate both structured economic indicators and unstructured textual signals.

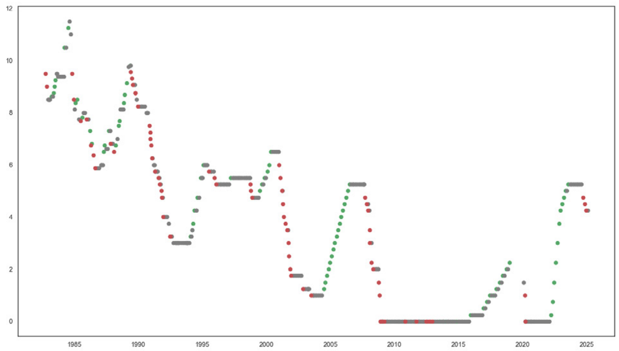

Figure 1: Federal Funds Rate Trajectory (Jan 1980 – Jan 2025).

We hence raise an important research question:

> *Can combining macroeconomic indicators with unstructured Federal Reserve communications improve the predictive accuracy and interpretability of interest rate forecasts?*

As shown in Figure 1, U.S. interest rates have experienced significant fluctuations over time, reflecting complex macroeconomic and geopolitical forces. Anticipating these policy shifts 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 signals derived from Federal Reserve communications. Our main contributions are as follows:

- We develop a hybrid model that synthesizes relevant economic features, linguistic cues, and sentiment probabilities 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.

2. Literature Review

The majority of existing research has focused on one data modality–either structured or unstructured data–in isolation. 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 modeling.

2.1 Structured Economic Indicators in Policy Forecasting

Rule-based models, such as using the Taylor Rule, represent early structured approaches to policy forecasting [4]. These models relate policy rates to deviations in inflation and output gaps. Building on these foundations, econometric frameworks like vector autoregressions (VARs) have been widely used to capture dynamic interdependencies among macroeconomic variables [5]. Despite their rigor, these methods often lack the incorporation of qualitative or forward-looking signals embedded in central bank communication. Their predictive power alone has been questioned during periods of heightened uncertainty.

2.2 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 evolved, allowing for more systematic analysis. Texts such as FOMC 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 such documents [1].

2.3 Sentiment Analysis Techniques in Central Bank Communication

Lexicon-based sentiment analysis, particularly using the Loughran-McDonald (LM) dictionary, was among the first techniques used to quantify tone in financial texts [6]. This approach categorizes words using predefined lists; the LM dictionary in particular categorizes words as positive, negative, uncertain, or litigious using a dictionary catered to financial texts. Later empirical studies demonstrated that shifts in central bank tone–as captured by such dictionaries–can predict changes in interest rate expectations and asset prices [7,8]. More advanced methods integrate domain-specific lexicons with broader NLP pipelines, enhancing robustness of analysis.

2.4 Deep Learning and the Emergence of Contextual Language Models

With the advent of deep learning, transformer-based models like BERT have substantially improved models' contextual understanding in interpreting text data [9]. FinBERT, fine-tuned on financial corpora, excels in extracting nuanced sentiment and semantic features from financial text. These models outperform traditional approaches in detecting subtle signals such as policy uncertainty, indirect messaging, and shifts in sentiment polarity. However, most studies employing these models treat text in isolation, without integrating structured macroeconomic signals.

2.5 Toward Multimodal Prediction

The integration of structured and unstructured data sources is still relatively unexplored. Some recent studies attempt to fuse time series data with text embeddings, using concatenation or attention mechanisms [10]. 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 combining sentiment features with macroeconomic indicators, enabling a more holistic and explainable policy forecast.

3. 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.

3.1 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)
- Labor market indicators: Unemployment rate, Nonfarm Payrolls (NFP)
- Growth metrics: Real Gross Domestic Product (GDP), Retail Sales
- Housing metrics: Housing Starts (HOUST), Home Price Index (HPI)
- Consumer sentiment: University of Michigan Consumer Sentiment

All data was retrieved from the Federal Reserve Economic Data (FRED) database, maintained by the Federal Reserve Bank of St. Louis [8]. Features were transformed using temporal differencing, policy rule-based composites, and moving average smoothing, with final selection based on correlation strength with target and distinctiveness.

3.2 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 between February 1993 and January 2025, and cover five key types of communication:

- FOMC Statements
- FOMC Meeting Minutes
- Speeches by Federal Reserve officials
- Testimonies by Federal Reserve officials
- Press conference transcripts from the Fed Chair

Text preprocessing involved tokenization, stopword removal, and lemmatization. Sentiment features were extracted using two approaches: (1) Term Frequency–Inverse Document Frequency (TF–IDF) vectors combined with Loughran-McDonald sentiment scores, and (2) sentiment class probabilities generated by FinBERT, a transformer-based language model fine-tuned for financial text classification [7,9].

3.3 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 reduction 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.

4. 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.

4.1 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 aligned using their dates of release and the official FOMC calendar 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 document, and Loughran-McDonald sentiment scores are added to enrich interpretability. Collectively, these features form a comprehensive representation capturing both economic fundamentals and narrative tone.

In the initial stage of model development, we focus exclusively on structured economic factors to establish a performance benchmark against which later models that incorporate unstructured communications data can be compared.

4.2 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. Among these, Gradient Boosting consistently demonstrated superior capability in modeling complex and nonlinear feature interactions.

4.3 Experimental Design

To ensure realistic and temporally-consistent evaluation, we conduct time series cross-validation with out-of-sample evaluation, which respects the chronological order of observations and avoids data leakage. This is chosen over a rolling-window approach given the small size of our dataset. Performance is assessed using both threshold-independent (ROC AUC) and threshold-dependent (accuracy, precision, recall) metrics. Given the imbalanced nature of the target variable–particularly the underrepresentation of "Hike" and "Cut" decisions–ROC AUC is emphasized as the primary evaluation metric due to its robustness to skewed class distributions.

4.4 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, both with time series cross-validation. 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. To further address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied post-tuning to the training set. This strategy synthetically augments minority class instances to promote equitable learning, and reflects real-world challenges in imbalanced policy classification.

4.5 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 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.

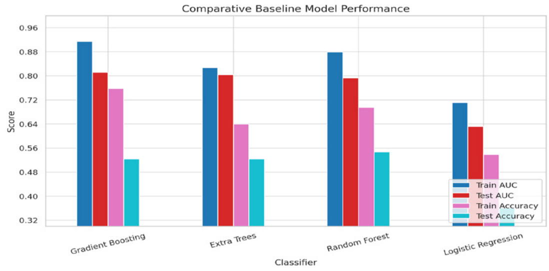

Figure 2: Baseline Comparison of Classifier Performance in Terms of ROC AUC and Accuracy Gradient Boosting consistently outperforms other models on both training and test sets. Logistic Regression exhibits underfitting, while tree-based models demonstrate stronger predictive capacity.

In conclusion, Gradient Boosting emerges as a strong candidate for further development within our multimodal prediction framework in subsequent stages. Its balance between predictive power and robustness makes it well-suited for capturing the nuances of monetary policy classification.

5. 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.

5.1 Data Preprocessing

To prepare the unstructured text data for modeling, a standardized cleaning process was applied across all document types. This included case normalization, removal of extraneous characters, punctuation filtering, and stop word 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 documents, chunking strategies were employed to handle input length constraints for transformer models.

5.2 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 input token limits),

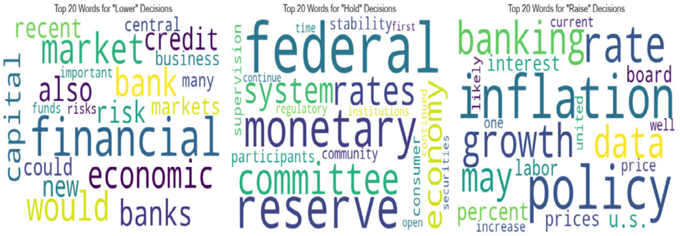

Figure 3: Dominant Terms Per Decision Class. Themes vary significantly across monetary policy decisions.

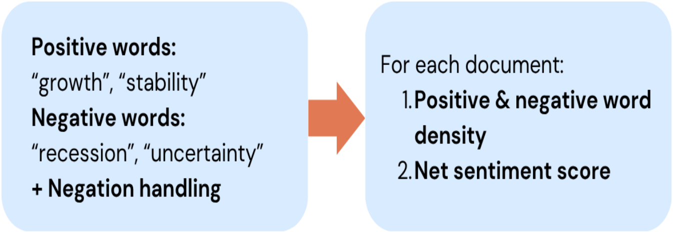

Figure 4: Rule-Based Sentiment Pipeline Using Loughran-McDonald Dictionary

and analysis of dominant terms revealed vocabulary clusters associated with each decision type–offering initial evidence of tone divergence between hawkish, neutral, and dovish communications. As shown in Figure 3, references to financial instability dominate the language of “Lower” decisions, while “Raise” statements emphasize inflationary concerns. “Hold” decisions reflect a neutral, status-quo tone.

5.3 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 sentiment lexicon, we computed sentiment scores from positive and negative word counts, normalized by document length [7]. A negation-aware scoring mechanism was implemented to reduce semantic errors. We extracted three metrics: positive density, negative density, and net sentiment. Figure 4 summarizes this extraction pipeline.

Temporal and categorical analyses showed that sentiment shifts generally align with known periods of economic stress, validating the LM dictionary’s utility in policy text interpretation. Figure 5 presents the evolution of net sentiment over time. Notably, sentiment declines tend to precede major recession periods, suggesting its potential as a leading indicator.

As seen in Figure 6, sentiment polarity varies by document type. FOMC statements, minutes, and press conference transcripts (*presconf*) exhibit a more narrow range, reflecting their formal and policy-focused nature. Speeches and testimonies present a more diverse range of sentiment, reflecting the rhetorical flexibility of such formats, which are often tailored to different audiences and economic narratives. These distinctions offer valuable inputs in later modeling stages.

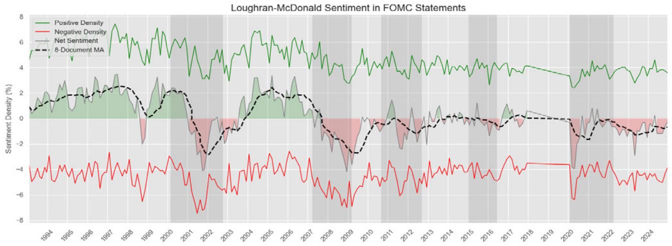

Figure 5: Net Sentiment Trends Over Time. Dips align with U.S. recession periods (shaded).

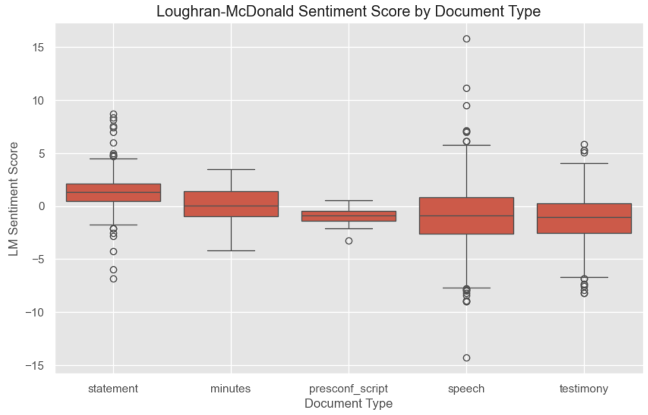

Figure 6: Distributions of LM sentiment Scores by Document Type

2) Transformer-Based Sentiment Analysis: We also employed FinBERT, a domain-specific transformer model trained on financial corpora, to classify sentiment across the document corpora [9]. 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 rigid though interpretable dictionary-based scores.

As shown in Figure 7, FinBERT’s sentiment outputs are predominantly labeled “neutral,” reflecting the Fed’s characteristically-measured communication style. 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 in the modeling stage.

5.4 Model Training and Evaluation

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

- **Traditional classifiers** (Logistic Regression, Na¨ıve 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 alone.
- **FinBERT classifier** was fine-tuned 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

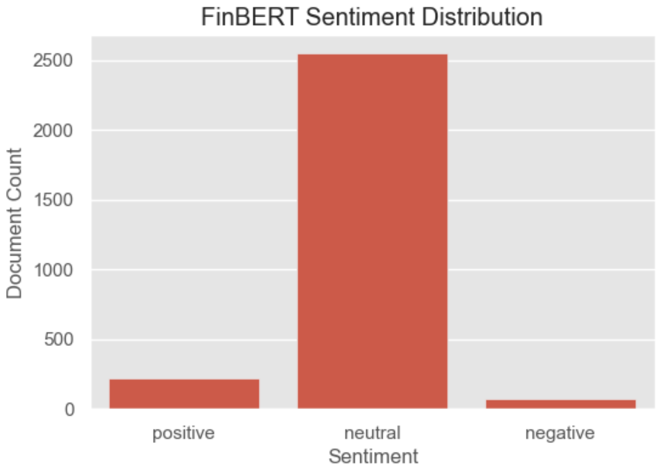

Figure 7: FinBERT sentiment classification across central bank documents. Majority documents are labeled as neutral.

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**

Table I: Text-Only Models Comparative Performance

of 0.67, outperforming traditional models though still showing a persistent bias toward the majority “Hold” class.

5.5 Evaluation Insights

Although textual models offer some predictive capacity–especially with FinBERT’s contextual strengths–their performance 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.

6. Multi-Modal Models

To explore how structured economic data (ED) and unstructured data can complement each other, we developed three multi-modal modeling frameworks. Each framework combines macroeconomic indicators with a distinctive approach to integrating textual data. We aimed to understand how different data modalities and modeling methods influence the performance and interpretability of monetary policy forecasts.

6.1 Method 1: ED + TF–IDF and LM Sentiment Features + X G Boost

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 Combine TF–IDF and LM Sentiment Features

```
Require: Document corpus $D = \{d_1, d_2, \ldots, d_n\}$
Require: LM sentiment lexicon $L = \{L_{\text{pos}}, L_{\text{neg}}, L_{\text{negate}}\}$
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: $s_{\text{pos}}, s_{\text{neg}} \leftarrow 0$
    for each token $t_j$ in $d_i$ do
        if $t_j \in L_{\text{pos}}$ and $t_{j-1} \notin L_{\text{negate}}$ then
            $s_{\text{pos}} \leftarrow s_{\text{pos}} + 1$
        end if
        if $t_j \in L_{\text{neg}}$ then
            $s_{\text{neg}} \leftarrow s_{\text{neg}} + 1$
        end if
    end for
    Normalize sentiment scores by total tokens
    Form LM sentiment vector $LM_i \leftarrow [s_{\text{pos}}, s_{\text{neg}}]$
    Concatenate: $F_i \leftarrow [T_i, LM_i]$
end for
return Feature matrix $F = \{F_1, F_2, \ldots, F_n\}$ for model training
```

Algorithm 1 outlines the process of combining TF–IDF and 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 and negative words, as well as a custom list of negation terms. Each document is first preprocessed through tokenization, stop-word 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 normalized and 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 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 Fed’s Inertial Rule (`Inertia_diff`), which captures intentional shifts in monetary policy. Bond market expectations (`10YUST_diff_prev`) and consumer sentiment (`UMich_diff_prev`) also exhibit strong contributions, especially in identifying minority “Lower” rate decisions Housing-related indicators (`HOUST_diff_year`, `HPI_diff_year`) reflect broader macroeconomic conditions and play a notable role as well. Additionally, textual features extracted from FOMC communications, such as `tfidf_basis` and `tfidf_difficult`, capture subtle narrative indicators and help improve both predictive ability and performance interpretability. The distribution of feature importance supports the value of combining structured data

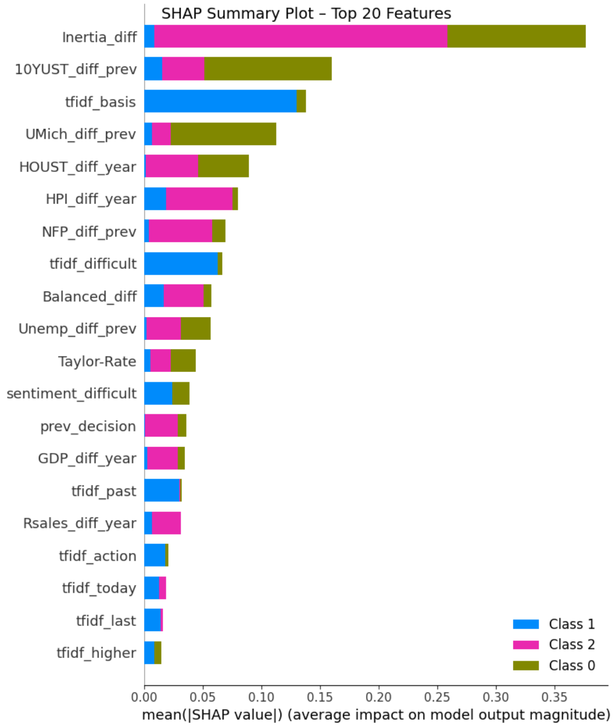

Figure 8: Method 1: SHAP Summary Plot Highlighting the Top Features Influencing the XGBoost Model's Predictions

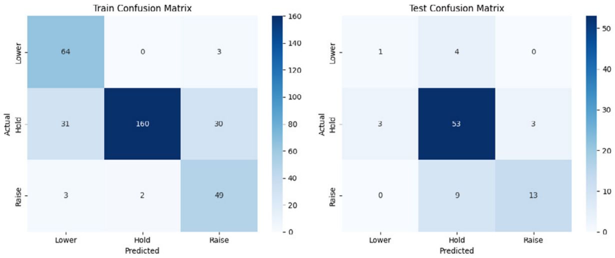

Figure 9: Method 1: ED + TF–IDF & LM Sentiment + XGBoost

with domain-specific textual cues.

Figure 9 shows the confusion matrix for Method 1. The model performs well across all three decision categories. Most model predictions fall along the diagonal, indicating a high level of agreement with actual FOMC outcomes. While some confusion exists between "Hold" and the other two 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 for future multi-modal model enhancements.

6.2. 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 frequency of relevant terms, this transformer-based method was designed to extract deeper contextual meaning from Federal Reserve communications. The goal was to evaluate whether FinBERT's context-aware sentiment scores could enhance predictive performance when combined with macroeconomic data.

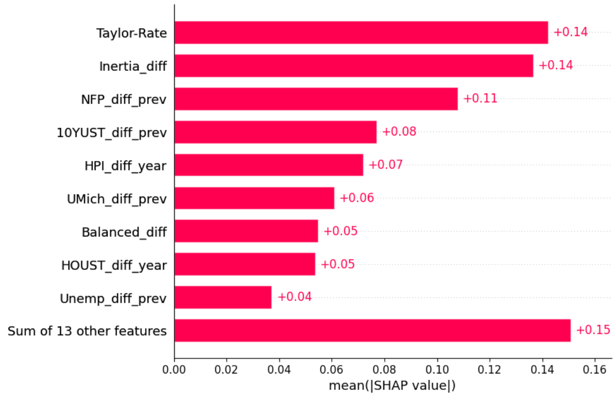

Figure 10: Method 2: SHAP Summary Plot

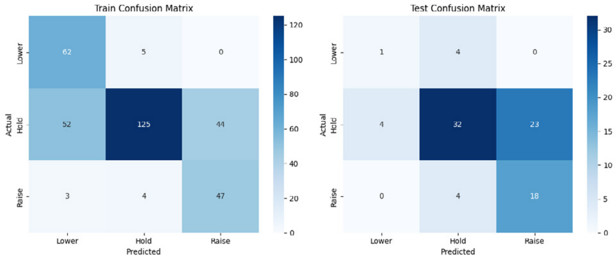

Figure 11: Method 2: ED + FinBERT Sentiment + XGBoost

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 (`Taylor_Rate`), the deviation from the Inertial rule (`Inertia_diff`), labor market momentum (`NFP_diff_prev`), long-term Treasury yield changes (`10YUST_diff_prev`), and shifts in consumer sentiment (`UMich_diff_prev`).

By contrast, the FinBERT-derived sentiment probabilities contributed modestly to the model's decision process. This suggests that while FinBERT captures contextual nuance, the relatively shallow architecture of XGBoost may be insufficient to fully leverage FinBERT's representational depth. These findings underscore a trade-off between model simplicity and the ability to extract meaningful insights from complex textual features in policy forecasting.

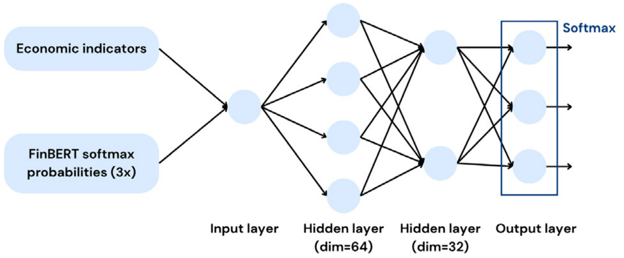

Figure 12: Feedforward Neural Network (FNN) Architecture

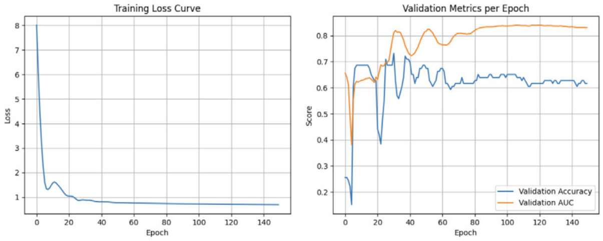

Figure 13: FNN Training Curve: Loss, AUC, and Accuracy

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 forgoes some classification accuracy for sentiment-based contextual awareness.

The model achieves a test AUC of 0.7960 and an accuracy of 59.30%. Performance declines from Method 1 suggest that sentiment probabilities from FinBERT may smooth over subtle textual distinctions. The effect is likely exacerbated under class imbalance.

6.3 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), as illustrated in Figure 12. 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 than the first method's at **61.63%**. The FNN struggles with calibration (as shown in Figure 13) and misclassifies minority classes "Raise" and "Lower" (Figure 14), which may be explained by a lack of sensitivity to class imbalance. Attempts to mitigate class imbalance in the neural network using focal loss were unable to alleviate this behaviour, which may be attributed to the limitations of a small dataset.

6.4 Comparative Performance and Insights

Figure 15 compares all three hybrid models. The FNN (Method 3) shows the highest AUC but lower overall accuracy. In contrast, the XGBoost model with TF–IDF and LM sentiment (Method 1) offers both solid performance and strong interpretability.

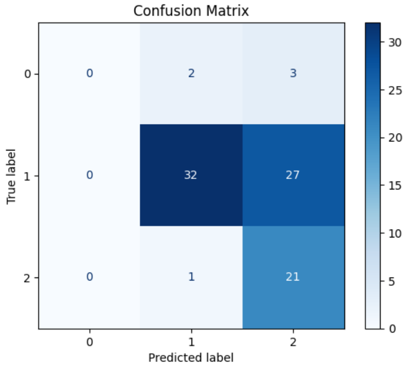

Figure 14: Method 3: ED + FinBERT Sentiment + FNN

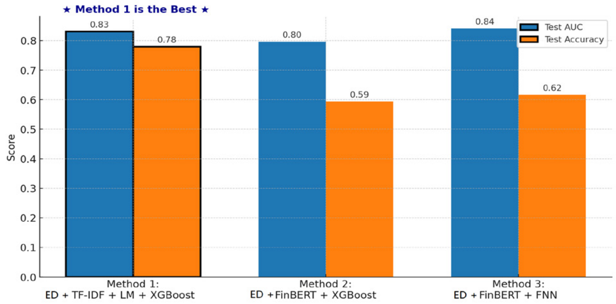

Figure 15: Comparative Performance of Hybrid Models

Several insights emerge. First, sparse, transparent text features may better capture formal monetary language compared to dense transformer outputs. Second, simpler models such as 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 modeling holds great promise. While deep learning offers flexibility, structured approaches utilizing interpretable features may be more practical for financial policy forecasting, especially where interpretability of predictions is essential.

7. Conclusion

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 communications 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 communications documents. These sparse, interpretable signals outperformed transformer-based sentiment embeddings in both model accuracy and interpretability.
- **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 given limited and imbalanced data.

We remain mindful of the limitations of our current approach. One major challenge was 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 precise prediction. Their compressed format limited both accuracy and interpretability in the modeling stage.

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 models' 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 the evolving landscape of policy forecasting [11-19].

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. Cecchetti, S. G., Feroli, M., Kashyap, A. K., Mann, C. L., & Schoenholtz, K. L. (2020). Monetary policy in the next recession?.
2. Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D. J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of economic literature, 46*(4), 910-945.
3. Fortes, R., & Le Guenedal, T. (2020). Tracking ECB's communication: Perspectives and Implications for Financial Markets.
4. Taylor, J. B. (1993, December). Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy* (Vol. 39, pp. 195-214). North-Holland.
5. Apel, M. and Grimaldi, M. "Monetary policy decision-making, market expectations and commitment," Journal of Monetary Economics, vol. 59, no. 6, pp. 601–621, 2012.
6. 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.
7. Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of financial economics, 110*(3), 712-729.
8. Hansen, S. and McMahon, M. "Transparency and Deliberation in Monetary Policy," Econometrica, vol. 86, no. 2, pp. 499–530, 2018.
9. Araci, D. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv preprint arXiv:1908.10063, 2019.
10. Wong, J. and Liu, L. "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.
11. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics, 31*(3), 307-327.
12. 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.
13. Federal Reserve Bank of St. Louis, "Federal Reserve Economic Data (FRED)," Online Resource, 2025.
14. Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063.*
15. Apel, M. and Grimaldi, M. "Monetary policy decision-making, market expectations and commitment," Journal of Monetary Economics, vol. 59, no. 6, pp. 601–621, 2012.
16. Hansen, S. and McMahon, M. "Transparency and Deliberation in Monetary Policy," Econometrica, vol. 86, no. 2, pp. 499–530, 2018.
17. Engle, R. F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation," Econometrica, vol. 50, no. 4, pp. 987–1007, 1982.
18. Hansen, S. and McMahon, M. "Transparency and Deliberation in Monetary Policy," Econometrica, vol. 86, no. 2, pp. 499–530, 2018.
19. Engle, R. F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation," Econometrica, vol. 50, no. 4, pp. 987–1007, 1982.