Enhancing Monetary Policy Forecasting Through Sentiment Analysis of Central Bank Communications

Group 10:
Gaia Iori, Michael Ladaa, Luca Milani, Matteo Roda, Sofia Villa

Abstract

This paper investigates how central bank communications contain predictive signals for monetary policy decisions beyond traditional economic indicators. Focusing on Federal Reserve and European Central Bank speeches between 1999 and 2025, we developed a RoBERTa-based classifier and aspect-based sentiment analysis to predict interest rate changes. By integrating the best-performing classifier into a SARIMAX time series model alongside macroeconomic variables, we achieved a 32.8% RMSE reduction compared to the baseline model for ECB rate forecasting. Our findings reveal that ECB communications are more suitable to NLP analysis than Fed speeches, leading to significant improvements in interest rate forecasting accuracy.

1 Introduction

Central banks increasingly rely on communication as a powerful tool, given its ability to influence financial markets and improve the predictability of monetary policy decisions (Blinder et al., 2008). Traditional methods for forecasting interest rates typically draw on macroeconomic indicators, employing Taylor rules (Giannone et al., 2004), time series models (Johannsen and Mertens, 2021), or structural economic models (European Central Bank et al., 2024). However, few studies incorporate comprehensive Natural Language Processing (NLP) analyses of central bank communication. To that end, we conducted a joint NLP analysis of speeches from both the European Central Bank (ECB) and the Federal Reserve (Fed), covering the period from 1999 to 2025. This combined approach is motivated by Granger causality tests indicating bidirectional influence between the two institutions, as well as clear signs of convergence in speech timing, tone, and policy decisions during financial crises (Appendix A). We employ two NLP-based strategies to build a model that predicts whether a central bank speech will be followed by an interest rate hike or cut. First, we fine-tune an end-to-end RoBERTa classifier and benchmark its performance against a logistic regression baseline using TF-IDF vectorization. Second, we implement Aspect-Based Sentiment Analysis (ABSA), applying FinBERT to extract sentiment on specific economic themes identified through Latent Dirichlet Allocation (LDA). We then incorporate these NLP-derived features into a SARIMAX model, alongside conventional macroeconomic indicators, to evaluate whether the inclusion of speech sentiment improves forecasting performance. The paper proceeds as follows: we first describe the data collection and preprocessing steps, then we detail the classification models and present results from integrating these outputs into a SARIMAX model for interest rate forecasting, and lastly we discuss the limitations of our approach and propose directions for future research.

2 Method

2.1 Data Collection and Cleaning

We collected speeches from the Federal Reserve and the European Central Bank (ECB) covering the period from 1999 to 2025. Speeches from 1999 to 2024 were sourced from a Hugging Face dataset[1], while those from 2025 were obtained directly from the ECB [2] and Fed [3] websites. In total, the dataset has 4,818 speeches: 53% from the ECB and 47% from the Fed. We also downloaded the corresponding interest rate data [4] [5] and merged them with speech datasets by aligning each speech with the prevailing interest rate of the country at the time

[1] https://huggingface.co/datasets/istat-ai/ECB-FED-speeches
[2] https://www.ecb.europa.eu/press/key/html/downloads.en.html
[3] https://www.federalreserve.gov/newsevents/speeches.htm
[4] https://fred.stlouisfed.org/series/FEDFUNDS
[5] https://data.ecb.europa.eu/data/datasets

it was delivered. Additionally, we calculated the time period from each speech to the next interest rate change and classified the speech's impact (cut or rise) based on whether the subsequent interest rate was lower or higher than the current rate.

2.2 Classifier

We built models for both binary (cut or rise) and 5-class classification (immediate cut, immediate rise, upcoming cut, upcoming rise, or stable), using two approaches: a TF-IDF-based model (HULK) and a fine-tuned transformer (RoBERTa). For both methods, we trained on 14/15 of the dataset and validated on the remaining 1/15, using two strategies: randomized sampling and chronological (time-based) splits to assess robustness over time. The baseline used TF-IDF features with character n-grams (2 to 6, word-boundary based), class balancing, and selection of the top 30,000 features via 5-fold cross-validated grid search. Randomized training produced strong binary classification results (macro F1 > 0.7) and moderate 5-class performance (macro F1 > 0.3), while time-aware performance dropped significantly (macro F1 ≈ 0.14), likely due to concept drift. RoBERTa was fine-tuned with a maximum input length of 128 tokens to reduce training time and GPU usage, a learning rate of 1e-5 (halved if validation loss did not improve in the following epoch), dropout (0.2) to randomly deactivate 20% of the model, weight decay at 0.05 to penalize large weights, and early stopping (patience 7). We trained for up to 15 epochs (binary) and 100 epochs (multiclass). Randomized training yielded excellent results (macro F1: 0.95 binary, 0.86 multiclass), but time-aware models suffered similarly to HULK (macro F1 ≈ 0.14 binary, below 0.10 multiclass), confirming a consistent challenge with temporal generalization. Table 1 summarizes F1 scores.

Model	Random	Time-aware
TF-IDF Binary	> 0.7	≈ 0.14
TF-IDF 5-class	> 0.3	≈ 0.14
RoBERTa Binary	0.95	≈ 0.14
RoBERTa 5-class	0.86	< 0.10

Table 1: Macro F1 scores for baseline (HULK) and RoBERTa models on binary and 5-class classification tasks, comparing randomized and time-aware splits.

2.3 Aspect Based Sentiment Analysis

To further investigate our dataset, we applied an aspect-based sentiment analysis on various topics discussed in the speeches. We enhanced our preprocessing by creating domain-specific stopwords and incorporating economic bigrams to better capture relevant financial expressions. We then vectorized the cleaned text with stricter frequency filtering and applied Latent Dirichlet Allocation (LDA) to uncover focused topics. This yielded four interpretable topics: 1) Inflation, 2) Financial Stability, 3) Microeconomic and Business, and 4) Post-crisis reforms (Appendix B). For each topic, we extracted representative keywords and computed the topic distribution for each speech. Next, we employed the FinBERT model to assess the sentiment expressed towards each topic in every speech.We excluded neutral predictions to better reflect directional sentiment. The resulting aspect-level sentiment scores were then combined with the topic distribution weights to produce weighted sentiment indicators for each speech. These features were then integrated and used to train a classifier, using Random Forest and Logistic Regression, and evaluated using time-aware cross-validation. The aspect-based sentiment features yielded moderate F1-scores, with the Random Forest model achieving an average F1-score of 0.554. Topic 4, which centers on post-crisis reforms and institutional evolution, emerged as a key predictor of interest rate changes, suggesting that speeches emphasizing structural policy shifts and institutional credibility provide valuable forward guidance on monetary tightening decisions.

2.4 Model selection

Model selection focused on three key dimensions: transformer-based vs. classical models, binary vs. multiclass classification, and random vs. time-aware data splitting. RoBERTa consistently outperformed HULK, achieving a macro F1-score above 0.95 in the binary task with balanced precision and recall, while HULK reached only 0.71, struggling especially with "cut" recall and "rise" accuracy. RoBERTa also outperformed in the 5-class task, so we adopted it as our core model. The multiclass setup faced challenges: categories like "immediate," "upcoming," and "stable" were arbitrarily defined to balance classes, but imbalance and differing temporal patterns between ECB and Fed speeches added noise. Therefore, we chose

the more robust binary classification. Data splitting was critical. Random splits leaked future information into training, causing unrealistically high performance. Time-aware splits, while more realistic, led to lower results, reflecting significant language evolution in central bank speeches. This highlights the need for adaptive models such as continual learning or periodic fine-tuning. Despite leakage risks, we used RoBERTa with random splits to assess maximum predictive potential, though future work should prioritize time-aware approaches. We also tested Aspect-Based Sentiment Analysis (ABSA) with LDA topic modeling and a Random Forest classifier, achieving a moderate F1 of 0.554. While topics were economically meaningful, ABSA lacked the predictive power of direct RoBERTa classification and was therefore excluded from the final SARIMAX integration.

2.5 SARIMAX Integration

In the first phase of our study, we developed a SARIMAX model to forecast interest rate decisions by the ECB and the Fed [6], leveraging seven macroeconomic indicators: inflation [7], unemployment [8], real GDP growth [9], yield spreads [10], consumer sentiment [11], 5-year inflation expectations[12], and oil prices[13]. Covering data from 1999 to 2025, we ensured temporal alignment and data quality through preprocessing steps, including deduplication and standardization. As SARIMAX requires stationary inputs, we applied differencing to non-stationary series, validated using Augmented Dickey-Fuller tests. Our choice of SARIMAX over ARIMA or ARCH was driven by its ability to integrate exogenous variables and capture asymmetric market responses to policy signals. We observed persistent rate change momentum up to three months and improved sensitivity to hawkish versus dovish tones using SARIMAX's moving average component. This phase provided a robust baseline for interest rate forecasting grounded in traditional macroeconomic signals. In the second phase, we extended this model by incorporating sentiment features derived from central bank speeches using a fine-tuned RoBERTa model, enhancing predictive power and offering deeper insight into how policymakers' communication drives market expectations and rate shifts.

[6] https://fred.stlouisfed.org

[7] https://www.ecb.europa.eu/stats/macroeconomic_and_sectoral/hicp/html/index.en.html

[8] https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/table_hist_unem.en.html

[9] https://ec.europa.eu/eurostat/databrowser/view/namq_10_gdp/default/table?lang=en

[10] https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/euro_area_yield_curves/html/index.en.html

[11] https://ec.europa.eu/eurostat/databrowser/view/ei_bssi_m_r2/default/table?lang=en

[12] https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html

[13] https://fred.stlouisfed.org/series/DCOILBRENTEU

3 Results

Integrating RoBERTa-based sentiment features significantly enhanced SARIMAX forecasting performance, with the ECB model achieving a 32.8% RMSE reduction (from 8.53 to 5.64). The RoBERTa sentiment coefficient was statistically significant (–0.0939, $p < 0.001$), alongside other key predictors including real GDP growth and consumer sentiment. Improved log-likelihood, AIC, and BIC scores further validated the enhanced model fit. The Fed model showed more modest gains, with RMSE changing from 13.9 to 14.2, though log-likelihood improvements and reduced AIC indicated better overall model coherence. The inclusion of sentiment features also increased the significance of other macroeconomic predictors. These contrasting results reflect fundamental differences in institutional communication styles. ECB speeches are more detailed and explicit in expressing policy direction, embedding linguistic signals that sentiment models can effectively capture. Fed communications appear more standardized and intentionally vague to preserve policy flexibility, limiting the linguistic variability needed for strong sentiment extraction. This communication strategy explains the smaller performance gains in the Fed model and supports our hypothesis that transparent central bank communications are more amenable to NLP-enhanced forecasting techniques.

4 Related work

The literature on interest rate forecasting is extensive, but it primarily focuses on quantitative indicators. Only in recent years significant attention has been given to the predictive role of central bank communication, following earlier work emphasizing its growing influence on shaping market expectations and signaling future policy moves

(Blinder et al., 2008). Most studies analyzing central bank discourse employ topic modeling techniques. For instance, Carboni et al. (2020) used Latent Dirichlet Allocation (LDA) to show that topics related to the core tasks and functions of central banks are consistently present in Fed and ECB speeches, while emerging challenges leave small measurable traces. Similarly, Petropoulos and Siakoulis (2021) applied LDA to extract relevant topics and performed sentiment analysis using custom dictionaries to forecast financial market turbulence, specifically S&P500 and VIX volatility, via an XGBoost model. Their findings confirm that central bank speeches are a valuable source of predictive information. In the context of monetary policy forecasting, Su et al. (2025) investigated communications from the People's Bank of China, combining BERTopic and Word2Vec to model topics. The study then used TF-IDF to extract signals of expansionary or tightening stances and constructed variables to enhance a forecasting model based on the Taylor rule. This study revealed that BERTopic produced more coherent and stable topics than LDA and that Taylor rule–based models can outperform SARIMAX in certain settings. Chortareas et al. (2025) also employed LDA to extract topics from Bank of England (BoE), ECB, and Fed communications, and used the Loughran and McDonald (LM) dictionary to quantify the tone of these topics. Their analysis linked sentiment to macroeconomic outcomes such as unemployment and stock market returns. We attempted to apply the LM dictionary in our sentiment analysis pipeline, but found it insufficiently expressive for our use case. Nonetheless, Chortareas et al. (2025) reported findings similar to ours, particularly in observing that the Fed's communications exhibit weaker signaling effects compared to the ECB. Despite the growing body of research on central bank communications, we found no prior studies that apply NLP techniques on ECB and Fed speeches specifically to forecast interest rate changes.

5 Conclusion

This study demonstrates that integrating RoBERTa-based sentiment analysis of central bank speeches into SARIMAX models significantly improves interest rate forecasts. For the ECB, our approach reduced RMSE by 32.8%, highlighting the strong predictive power of its more transparent communication style. Although improvements for the Fed were smaller, the inclusion of sentiment features still enhanced overall model coherence. These findings underscore that central bank speeches contain valuable and timely signals that complement traditional economic indicators, showcasing the potential of NLP techniques to advance monetary policy forecasting. In contrast, our exploration of Aspect-Based Sentiment Analysis (ABSA), which produced economically interpretable topics via LDA, revealed that simpler, end-to-end transformer models tend to be more effective for this task. This suggests that the complex linguistic patterns predictive of monetary policy decisions are better captured through direct fine-tuning of transformer models, rather than by decomposing text into topic-sentiment components.

5.1 Limitations

One key limitation of this study is temporal instability. While RoBERTa performed well under randomized cross-validation, its predictive accuracy declined in time-aware settings. This suggests difficulty in generalizing across economic cycles and evolving policy regimes. Furthermore, the analysis assumes a close link between central bank speeches and subsequent interest rate decisions, but such decisions are influenced by a broader set of factors. The different results for the Federal Reserve in the times series model, also highlight potential differences in communication styles, which may limit the generalization of the proposed methods across central banks.

5.2 Future improvements

Future research could extend this analysis to the communications of other central banks, enabling broader coverage and allowing for a comparative assessment of policy communication styles. Additionally, ABSA could benefit from using dynamic topic modeling techniques such as BERTopic, which may capture more nuanced and contextual aspects than traditional LDA or manually defined topics. Sentiment analysis could also focus on specific economic indicators such as inflation or employment by using specific dictionaries. This strategy would allow these sentiment signals to be incorporated as predictive features in interest rate forecasting models. Finally, while this study employs SARIMAX to integrate NLP-based signals into time series forecasting, alternative approaches such as machine learning models could be explored to evaluate potential gains in predictive accuracy.

References

Alan S. Blinder, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen. 2008. Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–45.

Marika Carboni, Vincenzo Farina, and Daniele A. Previati. 2020. *ECB and FED Governors' Speeches: A Topic Modeling Analysis (2007–2019)*, pages 9–25. Springer International Publishing, Cham.

Georgios Chortareas, Fotis Papailias, and Linda Shuku. 2025. Does central bank talk matter for forecasting? evidence from speeches of the boe, ecb, and fed. *International Journal of Finance & Economics*, n/a(n/a).

European Central Bank, M. Ciccarelli, M. Darracq Pariès, and R. Priftis. 2024. *ECB macroeconometric models for forecasting and policy analysis – Development, current practices and prospective challenges*. Publications Office of the European Union.

Domenico Giannone, Lucrezia Reichlin, and Luca Sala. 2004. Monetary policy in real time. *NBER Macroeconomics Annual*, 19:161–200.

Benjamin K. Johannsen and Elmar Mertens. 2021. A time-series model of interest rates with the effective lower bound. *Journal of Money, Credit and Banking*, 53(5):1005–1046.

Anastasios Petropoulos and Vasilis Siakoulis. 2021. Can central bank speeches predict financial market turbulence? evidence from an adaptive nlp sentiment index analysis using xgboost machine learning technique. *Central Bank Review*, 21(4):141–153.

Shiwei Su, Ahmad Hassan Ahmad, Justine Wood, and Songbo Jia. 2025. Monetary policy analysis using natural language processing: Evaluating the people's bank of china's minutes and report summary with the taylor rule. *Economic Modelling*, 149:107121.

Appendices

A Macroeconomic and Trend Analysis

Our macroeconomic analysis reveals that interest rate changes tend to cluster around periods of crisis and recovery, underscoring the cyclical nature of monetary policy. Notably, the ECB exhibits greater rate change volatility than the Fed, suggesting a more reactive or less constrained policy regime in the Eurozone. Granger causality tests at a lag of 4 show statistically significant bidirectional influence: the Fed leads ECB decisions ($p = 0.0028$) and vice versa ($p = 0.0245$), implying a complex interdependence in transatlantic policy setting. In terms of timing, the Fed tends to release speeches closer to decision dates (within 15–30 days), while ECB communications often precede decisions by over 90 days, reflecting differing communication strategies. We observe a negative correlation (Pearson $r = -0.578$) between ECB speech frequency and the magnitude of rate changes, suggesting that more frequent communication may act as a signaling mechanism to avoid abrupt policy shifts. However, the timing of speeches alone is a weak predictor of rate change size ($\rho \approx 0.2$), indicating that speech content likely carries more predictive power than cadence. Crisis periods mark a convergence in speech and policy timing, with speeches clustering around decisions and rate cuts dominating the response. Seasonal trends show heightened speech and rate activity in March–June and October–November, with dips in August and December. These patterns suggest informal cycles in policymaking and communication, though no statistically significant seasonal effect links speech frequency directly to the size of rate changes. Additionally, we find no dominant effect from individual speakers, reinforcing that structural and macroeconomic factors, rather than personality or communication style, drive policy discourse and decisions.

B Topic Modelling (LDA)

Topic 1: Inflation and core monetary policy (e.g., inflation, rate, price): low in 2009, peaks again from 2021–2025, coinciding with the global return of inflationary pressures and renewed monetary tightening in the post-pandemic economy.
Topic 2: Financial stability (e.g., credit, liquidity, crisis). Its sharp peak in 2009 aligns closely with the global financial crisis, a period when central banks focused intensely on providing liquidity, backstopping credit markets, and managing systemic risks.
Topic 3: Real economy and microeconomic focus (e.g., firm, payment, community). Interestingly, this theme remains remarkably constant across time, suggesting that supporting economic activity at the micro or sectoral level has been a consistent strategic concern for central banks, regardless of macroeconomic shocks.
Topic 4: Post-crisis reforms and institutional evolution (e.g., reform, currency, price stability). This topic is most prominent from 2009 to 2013, when speeches emphasized long-term policy redesign and stabilization measures in the wake of the financial crisis. Its gradual decline in subsequent years reflects the transition toward newer economic

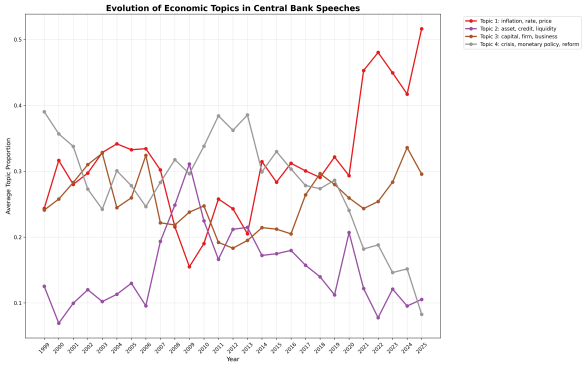

Figure 1: Topic evolution over time as detected by LDA.

priorities.

As shown in Figure 1, thematic trajectories demonstrate how central bank discourse evolves in response to macroeconomic conditions and policy priorities, with a pivot from stabilization post-crisis to inflation control in the 2020s.