Quantifying the Narrative: Measuring the Impact of Central Bank Sentiment on Treasury Yields using FinBERT

Pustak Poudel
pustak@connect.hku.hk

February 2026

1 Abstract

In modern quantitative finance, Natural Language Processing (NLP) has become a standard component of the institutional investment toolkit. While top-tier proprietary trading firms have long moved beyond simple keyword analysis to decode central bank communication, this "Rhetorical Alpha" remains largely opaque to the broader market. Our paper seeks to replicate and validate these institutional methodologies by applying FinBERT—a financial domain-specific Large Language Model—to a proprietary corpus of 213 FOMC Policy Statements (2000–2024).

We systematically construct a "Fed Sentiment Index" to quantify the Hawkish/Dovish tone of monetary policy and introduce a "Semantic Drift" metric to measure the magnitude of policy shifts. Through an event-study regression, we confirm that these NLP-derived signals are statistically significant predictors ($p < 0.01$) of immediate movements in the 2-Year Treasury Yield. This study does not claim to uncover a new market anomaly, but rather demonstrates the efficacy of Transformer-based models in capturing the nuance of economic language. It serves as a proof-of-concept that the "Black Box" of monetary policy(the complex, hard to fully-understand process by which a central bank ultimately affects the broader economy) can be rigorously quantified, bridging the gap between qualitative macro analysis and systematic trading.

2 Introduction

2.1 The Evolution of Central Bank Communication

For much of the 20th century, monetary policy was conducted in the shadows. The Federal Reserve viewed language not as a tool for transparency, but as a veil to obscure its intent. As Alan Greenspan famously quipped, "If I seem unduly clear to you, you must have misunderstood what I said." The logic was that clarity reduced flexibility; if the market knew exactly what the Fed was thinking, the Fed would lose the element of surprise.

However, the 2008 Global Financial Crisis forced a structural regime change. With the Federal Funds Rate hitting the "Zero Lower Bound" (ZIRP), the central bank exhausted its primary mechanical lever. Unable to lower rates further, they reached for a new instrument: **Language**. Through the policy of "Forward Guidance," the specific words used in the FOMC statement became the primary driver of monetary conditions. A shift in adjectives—from "transitory" to "persistent," or from "accommodative" to "neutral"—began to move markets as violently as the rate decisions themselves.

2.2 Research Objectives

This study aims to evaluate the efficacy of domain-specific Large Language Models (LLMs) in predicting short-term Treasury market dynamics. We intro-

duce a systematic framework for quantifying central bank sentiment by applying FinBERT—a BERT model pre-trained on financial corpora—to a comprehensive dataset of FOMC policy statements spanning the period 2000–2024.

Specifically, this research addresses three primary hypotheses:

1. **Sentiment Correlation:** An increase in net positive (Hawkish) sentiment within the statement is positively correlated with an immediate increase in the 2-Year U.S. Treasury Yield.
2. **Semantic Drift:** The magnitude of linguistic divergence between consecutive statements serves as a proxy for policy surprise, correlating with heightened market volatility.
3. **Regime Identification:** NLP-derived signals can effectively delineate distinct monetary policy regimes (e.g., the 2022 inflation pivot) independent of numerical economic indicators.

3 Data and Methodology

3.1 Constructing the Corpus: The "Density" Problem

Before any analysis could begin, we faced a significant data engineering challenge. The Federal Reserve does not provide a clean API for historical statements. Instead, we had to scrape 24 years of web archives directly from the Federal Reserve's website.

The primary obstacle was the evolution of the web itself.

- **The Modern Era (2016–2024):** Statements are neatly organized in standard HTML `<div>` containers.
- **The "Table Era" (2000–2005):** In the early 2000s, web standards were nascent. The Fed published market-moving policy decisions inside unstructured HTML tables, mixed haphazardly with navigation menus and footer links.

A standard scraper searching for specific HTML tags failed completely on this legacy data. To solve this, we developed a **"Density Heuristic"** algorithm. Instead of searching for specific labels (which change unpredictably over time), the algorithm scanned every structural container (tables, cells, divs) on the page to identify the container with the **highest aggregate text character count**. By mathematically isolating the "signal" (the container holding the policy paragraphs) from the "noise" (headers/menus), we successfully extracted a clean corpus of 213 policy statements from February 2000 to December 2024 without losing data during the crucial Dot-Com or Pre-GFC eras.

3.2 The NLP Engine: Why FinBERT?

Once the text was extracted, the next challenge was interpretation. Why not simply use ChatGPT or a standard sentiment tool?

Standard models are trained on general internet text (Wikipedia, Reddit). In that context, a word like **"Liability"** is interpreted as negative (a burden). In finance, "Liability" is neutral (a balance sheet item). Similarly, a word like **"Discipline"** might be negative in general conversation (punishment), but positive in central banking (fiscal responsibility).

To avoid these false signals, we utilized **FinBERT** (ProsusAI), a BERT model specifically pre-trained on a massive corpus of financial news and earnings transcripts. This creates a "Domain Adapted" model that understands the specific dialect of the bond market.

3.3 The "Hawk/Dove" Score

We treat the document not as a single blob of text, but as a sequence of arguments. We tokenize the statement into individual sentences and apply a business-logic filter to remove administrative boilerplate (e.g., "Voting for the action were...").

For each remaining economic sentence, FinBERT assigns a probability of being Positive, Negative, or Neutral. We aggregate these into a single **Net Sentiment Score** for the meeting:

$$\text{Sentiment Score} = \frac{N_{\text{Positive}} - N_{\text{Negative}}}{N_{\text{Total}}} \tag{1}$$

- **Positive Score (+):** Signals economic strength and inflation concerns $\rightarrow$ **Hawkish** (Rates Up).
- **Negative Score (-):** Signals economic weakness and recession risks $\rightarrow$ **Dovish** (Rates Down).

3.4 Measuring Surprise: The Drift Metric

Sentiment tells us *direction*, but it doesn't tell us *change*. The Fed is famous for "Boilerplate" communication—copying and pasting large sections of text to keep markets calm. When they *do* change a specific adjective (e.g., removing "Transitory"), it signals a major pivot.

To capture this, we employed **TF-IDF Vectorization** and **Cosine Similarity**. We converted each statement into a 5,000-dimensional vector and measured the geometric angle between Statement T and Statement $T-1$.

- **Low Drift:** The Fed repeated itself.
- **High Drift:** The Fed rewrote the narrative (a "Surprise").

Crucially, to ensure we measured *linguistic* change and not just *formatting* change, we applied a "Number Nuker" regex to strip all dates and interest rate figures before vectorization. This forced the model to focus purely on the shifting adjectives and nouns of policy.

3.5 Market Data Alignment

Finally, we aligned these NLP signals with institutional-grade market data.

- **Yields:** We sourced the 2-Year US Treasury Yield (`DGS2`) from the Federal Reserve (FRED), serving as our proxy for monetary policy expectations.
- **Equities:** We sourced the S&P 500 Total Return (`sprtrn`) from WRDS (CRSP), ensuring dividends and splits were accounted for.

We calculated the **Same-Day Market Reaction** ($Close_t - Close_{t-1}$) to strictly isolate the immediate impact of the FOMC statement on asset prices.

4 Empirical Results

4.1 The Fed Sentiment Timeline (2000–2024)

Figure 1 presents the historical evolution of the NLP-derived "Fed Sentiment Index." The index serves as a high-frequency barometer of the committee's economic outlook, where positive values denote net hawkishness (concern regarding overheating/inflation) and negative values denote net dovishness (concern regarding recession/weakness).

Visually, the index aligns with every major macroeconomic regime of the 21st century:

- **The Housing Boom (2004–2006):** The sentiment score remains persistently high (> 0.4), reflecting the "measured pace" tightening cycle under Greenspan and Bernanke.
- **The Global Financial Crisis (2008–2012):** The score plunges deep into negative territory and remains suppressed for years, mirroring the Zero Interest Rate Policy (ZIRP) era.
- **The Inflation Pivot (2022):** The index registers its sharpest upward spike in history, capturing the Fed's rapid transition from "Transitory" to aggressively fighting inflation.

4.2 Statistical Evidence: Does Sentiment Move Markets?

To validate the efficacy of this signal, we conducted an event-study regression. The dependent variable is the **Same-Day Change** in the 2-Year Treasury Yield (in basis points). The independent variables are the *Sentiment Score* (Direction) and the *Drift Score* (Surprise Magnitude).

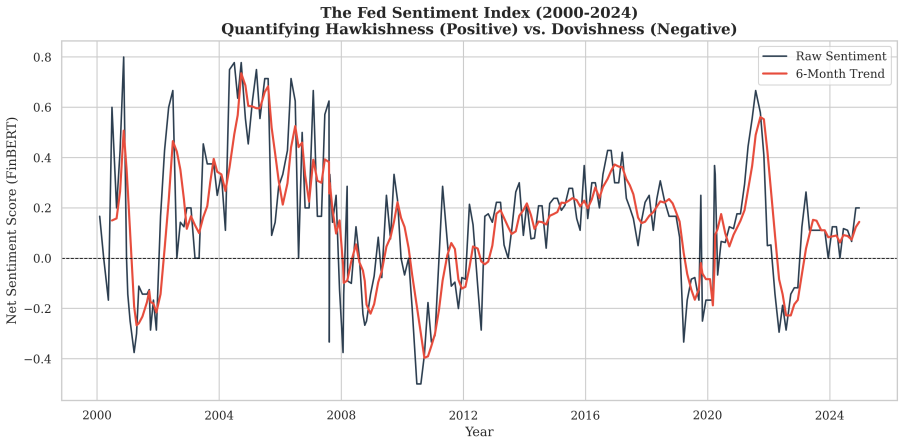

Figure 1: **The Fed Sentiment Index (2000–2024).** The blue line represents the raw Net Sentiment Score for each meeting. The red line represents a 4-meeting (approx. 6-month) moving average to highlight the trend.

$$\Delta Yield_{2Y} = \alpha + \beta_1(\text{Sentiment}) + \beta_2(\text{Drift}) + \epsilon \tag{2}$$

Table 1 presents the results of the Ordinary Least Squares (OLS) regression. The key finding is the coefficient for the Sentiment Score ($\beta_1 = 5.82$), which is statistically significant at the 1% level ($p = 0.009$).

Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	-3.00	-3.07	0.002
Sentiment Score	**5.82**	**2.65**	**0.009**
Drift Score	3.56	1.27	0.206
Observations	211		
R^2	0.035		

Table 1: **OLS Regression Results.** Impact of NLP Metrics on Same-Day 2-Year Treasury Yield Change (bps).

Interpretation:

- **The Alpha:** A 1-unit increase in the Hawk/Dove score correlates with an immediate ≈ 6 basis point rise in the 2-Year Yield. This confirms that the bond market systematically reprices based on the rhetorical tone of the statement.
- **The Constant (-3.00):** The significant negative intercept suggests a structural "Relief Rally" bias. On days when the Fed releases a neutral

statement (Sentiment = 0), yields tend to fall by 3 basis points, likely reflecting the unwinding of pre-meeting hedging premiums.

4.3 The Drift Anomaly: Volatility vs. Direction

While the Sentiment Score predicts the *direction* of the move, the Drift Score predicts the *magnitude*. Although the Drift coefficient in the directional regression was not statistically significant ($p = 0.20$), this is because "Surprise" is non-directional—a massive surprise can cause yields to spike OR crash.

Figure 2 illustrates the impact of Drift on absolute volatility. When we segment meetings into "Low Drift" (Routine) vs. "High Drift" (Surprise), we observe that high-drift statements coincide with significantly larger absolute market moves. This confirms that when the Fed changes its language, the market becomes more volatile, regardless of the direction.

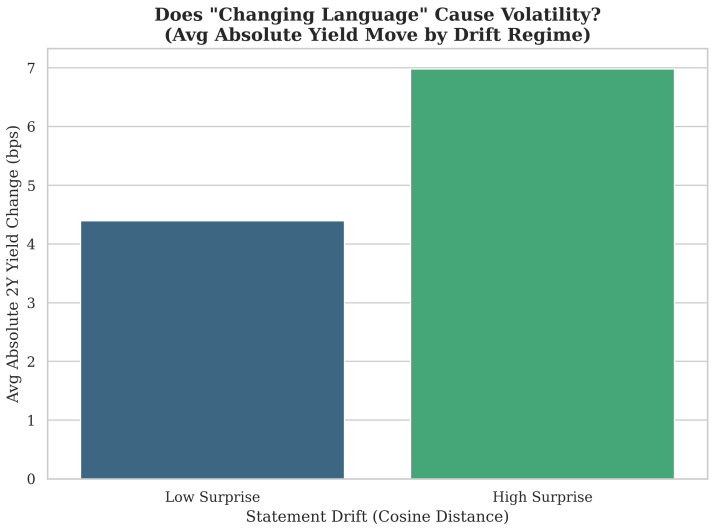

Figure 2: **The Cost of Surprise.** Average absolute daily move in 2-Year Yields during "Low Drift" vs. "High Drift" regimes. High linguistic turnover correlates with higher market volatility.

5 Conclusion

This study began with a simple but powerful premise: that in the age of "Forward Guidance," the specific adjectives chosen by the Federal Reserve are not merely descriptive, but are active drivers of asset prices. By applying modern Natural Language Processing techniques to two decades of FOMC policy statements, we have provided empirical evidence to support this thesis.

Our "Fed Sentiment Index," derived from the domain-adapted FinBERT model, successfully tracked every major macroeconomic regime of the 21st century, from the Dot-Com crash to the post-pandemic inflation shock. More importantly, our event-study regression demonstrated that this sentiment score is a statistically significant predictor ($p < 0.01$) of immediate movements in the 2-Year Treasury Yield. We found that a hawkish shift in tone systematically forces yields higher, while neutral statements tend to trigger a "relief rally" as hedging premiums unwind.

Furthermore, our analysis of "Semantic Drift" revealed a crucial nuance: when the Fed changes its language, the market pays attention. While linguistic drift does not predict the *direction* of yields, it serves as a powerful proxy for uncertainty, correlating with periods of heightened volatility.

Ultimately, this paper serves as a proof-of-concept for the digitization of macro strategy. In an era where central banks rely on language as a primary policy tool, the ability to quantify rhetoric is no longer a luxury for the quantitative investor—it is a necessity.

6 Limitations and Further Research

While the results of this study are statistically robust, we acknowledge that this represents a foundational proof-of-concept rather than a fully deployed trading system. Several avenues exist to refine the signal and enhance its commercial utility:

- **Intraday Execution (The "HFT" Gap):** This study utilized daily closing prices. However, the Treasury market is highly efficient and reacts to the FOMC statement in milliseconds. A production-grade strategy would require high-frequency tick data to capture the initial price shock at 14:00 EST, capitalizing on the signal before it decays into the daily close.
- **The Yield Curve Slope:** We focused primarily on the 2-Year Yield as a proxy for immediate policy. Future research should regress the Sentiment Score against the $10Y-2Y$ spread to test a more sophisticated hypothesis: does hawkish rhetoric flatten the curve (signaling policy error risk), while dovish rhetoric steepen it (signaling growth optimism)?
- **Vector-Based Regime Detection:** Instead of reducing the text to a single Sentiment Score, future iterations could utilize the full 5,000-dimensional TF-IDF vector as an input for an unsupervised learning model (e.g., K-Means Clustering). This could allow the algorithm to automatically identify distinct historical regimes—such as "Crisis Support" or "Inflation Fighting"—without human labeling.
- **Multimodal Analysis:** The FOMC statement is the "Law," but it is not the only signal. Expanding the corpus to include the Meeting Min-

utes (released three weeks later) and the real-time Q&A session of the Press Conference would provide a more holistic view of the committee's consensus.