Predicting Stock Market Performance from Federal Reserve Governors' Speeches: Data Mining and NLP-Driven Event Study

Prashant Gavit
Computer Engineering Department
San Jose State University
prashat.gavit@sjsu.edu
017463277

Saim Sheikh
Computer Engineering Department
San Jose State University
saim.sheikh@sjsu.edu
01490171

Zach Kuo
Computer Engineering Department
San Jose State University
zach.kuo@sjsu.edu
018327569

***Abstract*—This study looks at how linguistic features in Federal Reserve Governors speeches affect short term U.S. stock market performance. The analysis draws on the Federal Reserve Governors Speeches dataset from 1996 to 2020, along with market data pulled from Yahoo Finance. Natural language processing techniques and data mining methods help quantify aspects like sentiment, uncertainty, and policy tone. Text features come from FinBERT models and financial lexicons. These features connect to market movements via event study methods, regression analysis, and machine learning classification. The goal centers on seeing if certain rhetorical patterns, such as hawkish or dovish language, can predict shifts in market indices and how long those reactions last. We use SHAP (Shapley Additive Explanations) values, an explainable AI method that attributes model predictions to individual input features using concepts from cooperative game theory, to provide interpretability by pointing out the terms that most influence market responses. Overall, the findings offer evidence on how central bank communication shapes investor expectations. They also suggest a framework for gauging market reactions to upcoming Federal Reserve statements.**

I. Introduction

Financial markets respond strongly to signals from central banks. The United States Federal Reserve stands out in this regard. Its communications shape investor expectations and asset prices in key ways. Tools like interest rate changes, quantitative easing, balance sheet tweaks, and forward guidance help the Federal Reserve guide economic activity. Even prior to actual policy shifts, the wording in central bank statements can spark notable changes in equity and bond markets. A change in tone, say from accommodative to restrictive, or the mention of inflationary pressures or slowing growth, can reshape traders' views on future policy. This often leads to increased short term volatility.

In today's markets, filled with algorithmic and high frequency trading, monetary policy talk acts as a live data feed. Institutional investors and hedge funds keep close watch on speeches, press conferences, and official releases. Many use natural language processing systems to parse tone and sentiment for clues on market direction. Some models go further. They examine non verbal elements like facial expressions, posture, or vocal stress in Federal Reserve briefings. Such details might hint at whether news will prove positive or negative for the economy. All this highlights the rich information value in policy decisions and their presentation.

Academic work on the linguistic side of Federal Reserve communications has been somewhat sparse so far. Much prior research centers on the effects of clear cut rate decisions or meeting minutes. Yet the subtle phrasing in governors' speeches, given outside formal meetings, reveals much about the bank's outlook and focus areas. Recent progress in text mining and machine learning makes it possible to measure this qualitative data systematically. Researchers can now model how elements like sentiment, uncertainty, or thematic stress affect later market responses.

This study fills that methodological gap by systematically quantifying the relationship between Federal Reserve communication and market dynamics. We examine how nuanced tone and phrasing in Federal Reserve Governors' speeches affect short-term stock market performance, using a comprehensive dataset spanning 1996–2020 — a period that encompasses multiple economic cycles and policy regimes. Our approach combines advanced NLP methods, including FinBERT sentiment analysis and financial lexicons, with techniques to identify hawkish and dovish policy signals from speech content to predict the impact of policy on financial markets.

The derived linguistic features undergo rigorous evaluation against both same-day and multi-day market returns across major stock indices through event study methodology and predictive modeling frameworks. Rather than establishing simple correlations, our investigation identifies specific linguistic patterns that reliably predict positive or negative market movements, with SHAP [11] (Shapley Additive Explanations) analysis, discussed further in Section V, providing interpretable insights into feature importance and model decision-making processes.

The research contributes in two main ways. First, it builds an empirical setup for measuring the info in central bank language within financial markets. Second, it gives predictive views on how speeches could sway investor sentiment and market paths ahead. In the end, the results deepen knowledge of how communication drives market behavior. They show text analytics as a useful add on to standard economic forecasting.

II. Related Work

The application of natural language processing and textual analysis to financial market prediction has emerged as a transformative research domain, fundamentally reshaping how researchers and practitioners approach market forecasting. This evolution represents a paradigm shift from traditional quantitative models that rely exclusively on numerical market indicators toward more comprehensive approaches that integrate the rich information content embedded in human communication. The growing recognition that financial markets are fundamentally driven by information flow and sentiment has catalyzed extensive research into diverse textual data sources, each offering unique perspectives on market dynamics and investor behavior.

For the purposes of this literature review, we organize contemporary research in this field into three complementary streams, each focusing on distinct types of textual data sources and their applications to financial prediction. This categorization facilitates our understanding of the multifaceted nature of textual information in financial markets and highlights the varying methodological approaches required to extract meaningful predictive signals from different communication channels.

A. *Financial Forecasting Using Social Media Data*

The integration of social media data, particularly Twitter content, into financial forecasting models has emerged as a transformative research area in computational finance. Social media platforms serve as real-time barometers of public sentiment and investor behavior, providing a rich, complementary data source to traditional financial indicators. This evolving body of research demonstrates how microblogging data can significantly enhance market prediction accuracy through increasingly sophisticated methodological approaches.

The foundational work in this domain established the viability of combining Twitter-derived sentiment indicators with conventional forecasting methods. Oliveira et al. pioneered this approach by demonstrating that Twitter sentiment scores and posting volume, when integrated with traditional investor surveys through advanced filtering techniques such as Kalman Filters, can substantially improve forecasts of critical market variables including returns, trading volume, and volatility [1]. Their comprehensive evaluation revealed that these hybrid approaches consistently outperform autoregressive baseline models, establishing that social media data provides unique and valuable predictive signals. The fundamental appeal of microblogging data stems from its accessibility, cost-effectiveness, and ability to capture investor sentiment with daily granularity and near real-time availability—advantages that traditional survey methods cannot match.

The advent of deep learning architectures has fundamentally revolutionized this research domain. Wu et al. introduced a groundbreaking cross-modal attention-based hybrid recurrent neural network (CH-RNN) that seamlessly integrates daily Twitter content with historical stock price sequences [2]. Their innovative architecture employs sophisticated attention mechanisms to dynamically align social media information with corresponding stock trends, resulting in measurably improved accuracy for price movement prediction. The key innovation lies in the model's ability to focus selectively on the most relevant social media signals while filtering out noise—a critical capability given the inherently noisy nature of social media streams.

Building upon these foundational sentiment-based approaches, subsequent research has developed increasingly refined scoring methodologies. Guo and Li introduced the Twitter Sentiment Score (TSS) model, which incorporates novel baseline correlation techniques to achieve remarkable predictive performance for major market indices, particularly the FTSE 100 [3]. Their approach demonstrated the remarkable ability to anticipate market direction up to 30 hours in advance with approximately 67% accuracy. Perhaps most intriguingly, the TSS model exhibited pronounced asymmetric performance characteristics, showing significantly higher accuracy for predicting upward market movements compared to downward trends. This asymmetry provides valuable insights into how social sentiment translates differently to bullish versus bearish market conditions, suggesting that positive sentiment may be a more reliable predictor than negative sentiment.

Addressing the persistent challenges of noise and data sparsity inherent in social media data, the most recent advances have focused on self-supervised learning paradigms. Soun et al. developed SLOT (Self-supervised Learning of Tweets), which represents a significant methodological advancement by learning joint embeddings for tweets and stocks within a unified semantic space through self-supervised objectives [4]. This approach captures complex multi-level relationships between companies based on social media discourse patterns, making the model remarkably robust to unreliable or sparse tweet data. By leveraging information across all available tweets rather than focusing solely on direct stock mentions, SLOT successfully uncovers latent semantic connections and achieved state-of-the-art accuracy in stock movement prediction. This work exemplifies the growing importance of sophisticated representation learning techniques for extracting meaningful signals from inherently noisy social media text, setting new standards for the field.

B. Financial News-Based Prediction Models

Financial news has evolved into a critical data source for market prediction, representing a fundamental shift from purely quantitative analysis to incorporating rich textual information that captures market psychology and investor behavior. Unlike traditional financial indicators such as price, volume, and volatility, news articles provide contextual narratives that explain market movements and reveal underlying causal mechanisms. These textual sources capture breaking developments, earnings announcements, regulatory changes, and macroeconomic events that often serve as catalysts for significant market reactions. The challenge lies in systematically extracting actionable insights from the massive, unstructured streams of financial news generated daily across global markets.

The emergence of sophisticated natural language processing (NLP) and deep learning techniques has revolutionized the field's ability to process and interpret financial news at scale. Modern approaches go beyond simple keyword matching or basic sentiment analysis, instead employing advanced techniques such as event detection, semantic embeddings, and contextual understanding to capture nuanced market signals. This technological evolution has enabled the development of predictive models that can identify subtle linguistic patterns associated with market movements, detect emerging trends from news narratives, and quantify the impact of specific events on asset prices. The integration of these capabilities represents a paradigm shift toward news-informed predictive modeling that bridges the traditional gap between human market interpretation and algorithmic decision-making.

A pioneering contribution in this domain came from Ding et al., who developed an innovative event-driven framework that fundamentally changed how financial news is processed for market prediction [5]. Their approach moves beyond traditional sentiment analysis by focusing on structured event extraction, treating news articles as collections of discrete, semantically meaningful events rather than unstructured text. The methodology employs a sophisticated neural tensor network architecture to transform extracted events into dense vector representations that capture complex semantic relationships between market entities, companies, and economic events. These event embeddings are subsequently processed through a convolutional neural network (CNN) designed to model both immediate and longer-term market dependencies. The framework's effectiveness was demonstrated through comprehensive evaluations on the S&P 500 index and individual stock predictions, achieving approximately 6% improvement over state-of-the-art baseline methods. Importantly, market simulation experiments validated the practical utility of their approach, showing superior profitability compared to existing systems. This work established event-driven modeling as a crucial paradigm for linking unstructured news content to structured market outcomes.

Extending the integration of structured knowledge with deep learning, Liu et al. introduced a hybrid approach that addresses the fundamental challenge of data sparsity in financial news analysis [6]. Their methodology represents a significant advancement by combining traditional text-based feature extraction with structured knowledge representation through knowledge graphs. The core innovation lies in the integration of TransE knowledge graph embeddings with convolutional neural networks, creating a unified framework that enhances semantic understanding of financial news headlines while reducing the impact of limited training data. This hybrid architecture captures both explicit textual information and implicit relationships between financial entities, market sectors, and economic concepts. The approach demonstrates remarkable versatility by seamlessly integrating multiple data modalities—combining news-derived features with daily trading data and technical indicators before evaluation through both traditional machine learning methods (Support Vector Machines) and modern deep learning architectures (Long Short-Term Memory networks). Applied to Apple's stock performance within the S&P 500 context, the framework achieved impressive results: 97.66% accuracy in news sentiment classification and 55.44% accuracy in directional price movement prediction, consistently outperforming traditional machine learning baselines. This work exemplifies the growing importance of hybrid modeling approaches that leverage both the rich contextual information in news text and the structured relationships encoded in knowledge graphs, providing more robust and interpretable decision support for financial market participants.

C. Alternative Textual Data Sources for Financial Prediction

While social media and financial news represent prominent textual data sources for market prediction, the research frontier has expanded to encompass a diverse array of unconventional textual sources that offer unique insights into market dynamics. These alternative data sources—ranging from corporate communications and regulatory speeches to earnings call transcripts—provide rich contextual information that traditional quantitative models often overlook. The exploration of these diverse textual sources reflects a fundamental recognition that financial markets are influenced by a complex ecosystem of communications, where different stakeholders contribute distinct types of information that collectively shape investor sentiment and market behavior. Unlike the broad, democratized nature of social media or the event-focused characteristics of financial news, alternative textual sources often provide more structured, authoritative, and contextually rich information that can offer deeper insights into market fundamentals and future expectations.

One prominent direction examines the linguistic content of earnings calls as a predictive signal of financial analysts' behavior. Keith and Stent analyzed how the pragmatics and semantics of analysts' questions during earnings calls influence subsequent changes in price targets [7]. By extracting a set of twenty pragmatic features and embedding-based representations, their regression and classification models demonstrated that linguistic cues moderately predict analysts' post-call decisions, even after accounting for external factors such as market conditions and private communications. This work highlights the subtle role of conversation tone, information asymmetry, and dialogue structure in financial decision-making.

Complementing corporate communication analysis, researchers have recognized central bank speeches as particularly valuable sources of macroeconomic sentiment and policy signals. Petropoulos and Siakoulis advanced this field by developing an adaptive NLP sentiment index specifically designed to quantify central bankers' expectations regarding economic stability [8]. Their methodology combines the robustness of XGBoost machine learning techniques with sophisticated text mining approaches, creating sentiment indices derived from both predefined financial dictionaries and corpus-specific vocabularies built from historical speech patterns. The resulting framework demonstrated remarkable predictive capability for financial market turbulence, establishing that linguistic sentiment embedded in policy communications can serve as an effective early warning system for market volatility. This research underscores the growing recognition that qualitative macroeconomic narratives, when systematically analyzed, can provide crucial insights for anticipating systemic risks and market disruptions.

Extending beyond official institutional communications, the field has embraced event-driven trading strategies that exploit news-based corporate event detection as sophisticated predictive mechanisms. Zhou et al. introduced a groundbreaking bi-level classification framework specifically designed to detect and analyze corporate events from PRNewswire articles for stock movement prediction [9]. Their innovative architecture combines token-level and article-level detection mechanisms, creating a hierarchical understanding of event significance and market impact. The framework's effectiveness was demonstrated through comprehensive evaluations showing consistent outperformance of baseline sentiment-based models across multiple metrics, including excess market returns and win rates. This research represents a paradigm shift from broad sentiment analysis toward fine-grained event recognition, illustrating that specific event detection—rather than general sentiment classification—can more accurately capture the underlying causal mechanisms that drive price dynamics in modern financial markets.

D. Literature Synthesis and Research Direction

The literature demonstrates that textual analysis can effectively predict financial market movements across various data sources, from social media sentiment to financial news events. Building on these findings, we believe that Federal Reserve governors' speeches similarly contain linguistic signals that influence market behavior. While Petropoulos and Siakoulis examined central bank communications for broad market turbulence prediction, limited research has focused specifically on the short-term market impact of Federal Reserve speeches.

This work investigates whether linguistic features extracted from Federal Reserve governors' speeches can predict short-term stock market movements. We develop a methodology that combines established NLP techniques with domain-specific features to quantify the relationship between monetary policy communication and market reactions.

III. Problem Statement

The Federal Reserve exerts a powerful influence over financial markets, shaping investor expectations and economic outlooks long before formal policy actions are implemented. While interest rate changes and official policy decisions are well-studied, the *language* and *tone* used in Federal Reserve Governors' speeches also play a crucial, yet less quantifiable, role in influencing market sentiment. Subtle linguistic cues—whether a speech adopts a "hawkish" (restrictive) or "dovish" (accommodative) stance—can drive investor perceptions of future monetary policy, impacting asset prices even in the absence of concrete policy changes.

However, despite the growing recognition of communication as a policy tool, there remains a gap in systematically measuring how the sentiment and rhetoric in these speeches correlate with short-term stock market performance. Traditional econometric studies primarily analyze event outcomes such as interest rate announcements or meeting minutes, leaving the *linguistic dimension* of monetary signaling underexplored.

This project aims to address that gap by developing a data-driven framework that objectively quantifies and models the relationship between the tone of Federal Reserve speeches and subsequent market behavior. By applying natural language processing (NLP) techniques, sentiment analysis models such as FinBERT, and event study methodologies, we seek to answer the following central research question:

> *To what extent does the sentiment and tone expressed in Federal Reserve Governors' speeches predict short-term movements in U.S. stock market indices?*

Our goal is to move beyond anecdotal or qualitative interpretations of central bank communication and instead provide empirical evidence on how linguistic factors—such as sentiment polarity, uncertainty, and policy tone—translate into measurable market reactions. The outcome of this study will not only shed light on how investors interpret monetary policy communication but also contribute a predictive framework for assessing future market responses to Federal Reserve statements.

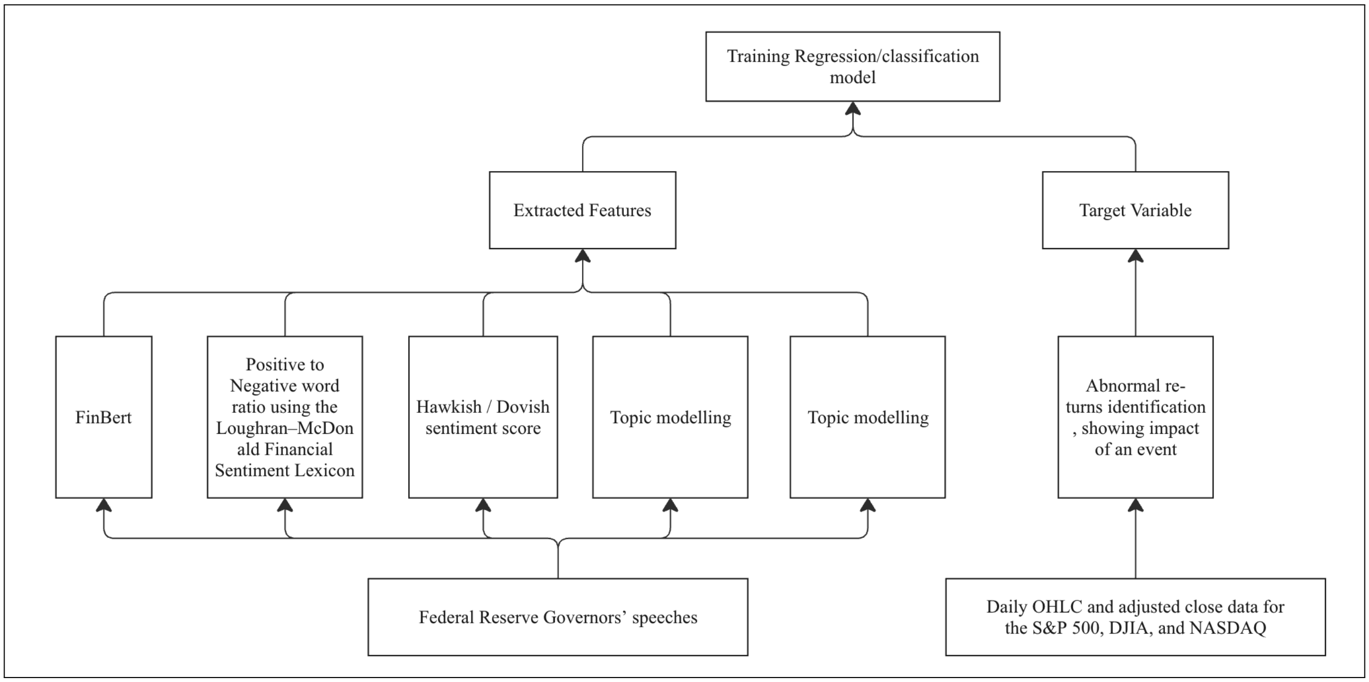

Fig. 1. Comprehensive Data Mining Pipeline for Federal Reserve Speech Analysis. The methodology indicates five key stages: (1) Data Collection of Federal Reserve speeches and market data, (2) Feature Extraction using FinBERT, financial lexicons, and hawkish-dovish indicators, (3) Event Study Analysis and Machine Learning modeling

IV. Proposed Solution

A. Overview

We propose a data mining pipeline that integrates textual analysis and event study methods to assess how Federal Reserve speeches affect short-term U.S. stock market movements. Figure 1 illustrates our comprehensive methodology framework:

1) **Data Collection:**
   - **Text Data:** Federal Reserve Governors' speeches (1996–2020) from the official Fed website and publicly available datasets.
   - **Market Data:** Daily OHLC and adjusted close data for the S&P 500, DJIA, and NASDAQ obtained via the Yahoo Finance API.
2) **Preprocessing:**
   - **Text Cleaning and Tokenization**: Removing punctuation, lowercasing, and splitting text into tokens (words).
   - **Lemmatization and Stopword Removal**: Reducing words to their root form (e.g., "running" → "run") and removing high-frequency, low-information words (e.g., "the", "and").
   - **Metadata Extraction**: Capturing attributes such as the speech date, speaker, title, and topic category for downstream alignment and analysis.
   - **Temporal Alignment**: Matching each speech to the corresponding trading day (t=0) and subsequent days (e.g., t+1) to define event windows, enabling measurement of the market's immediate and delayed response.
3) **Feature Extraction:**
   - **Sentiment Scores:** We use FinBERT (ProsusAI/finbert), a transformer-based language model fine-tuned from BERT (Bidirectional Encoder Representations from Transformers) specifically for financial text [10]. Unlike generic sentiment models, FinBERT is trained on financial corpora such as analyst reports and earnings calls, enabling it to correctly interpret domain-specific language—for example, recognizing that "lower interest rates" may carry positive implications for equities. It outputs probabilities for positive, negative, and neutral sentiment classes, which we aggregate to form numerical sentiment scores for each speech.
   - **Uncertainty/Polarity:** The Loughran–McDonald financial sentiment lexicon is a domain-specific dictionary that redefines conventional sentiment categories for the financial domain. For instance, while general-purpose lexicons might label "liability" or "capital" as neutral or positive, the Loughran–McDonald lexicon correctly identifies "liability" as negative in a financial context. We use this lexicon to compute counts or ratios of words associated with uncertainty, negativity, and constraining tone, which serve as features representing market-relevant sentiment nuances.
   - **Hawkish–Dovish Index:** We build a custom dictionary distinguishing hawkish (restrictive or inflation-fighting) from dovish (accommodative or growth-supporting) lan-

guage. This captures the Fed's policy stance, where hawkish tones may signal rate hikes (potentially negative for equities), while dovish tones may suggest easing (potentially positive).

- **Topic Modeling:** We apply Latent Dirichlet Allocation (LDA) or BERTopic to extract key policy themes (e.g., inflation, employment, financial stability). LDA models documents as probabilistic mixtures of topics, whereas BERTopic leverages contextual embeddings from transformer models, offering semantically richer clusters.
- **Linguistic Complexity:** We measure readability and syntactic complexity using metrics such as average sentence length, type-token ratio, and Flesch–Kincaid readability scores. Research suggests that more complex or ambiguous communication can heighten market uncertainty and volatility.

4) **Event Study and Modeling:** We compute abnormal returns as the deviation between the observed market return and the expected return from a benchmark model (e.g., market model or mean-adjusted return). Cumulative abnormal returns (CAR) aggregate AR values over short windows (e.g., t = 0, +1, +3) to measure the net effect of speeches. We then use OLS regression and machine learning classifiers (Logistic Regression, Random Forest, XGBoost) to link linguistic features with market direction.

5) **Interpretability:**
To make the predictive models explainable, we apply SHAP (Shapley Additive Explanations), an explainable AI framework based on cooperative game theory. SHAP assigns each feature a "Shapley value" representing its marginal contribution to a model's prediction [11]. This allows us to:

- Quantify how much each linguistic feature (e.g., sentiment score, hawkish index, topic proportion) influences the direction of predicted market movement.
- Visualize global and local feature importance (e.g., summary plots, dependence plots).

By combining predictive performance with interpretability, SHAP ensures the analysis remains both transparent and economically meaningful.

B. Algorithmic and Model Details

In our analysis we build both regression and classification models to link linguistic features from speeches to subsequent market returns and direction. Models are trained on engineered features (FinBERT embeddings, lexicon counts, hawkish/dovish scores, topic proportions, readability and volatility features) and evaluated using time-aware validation.

1) Regression Analysis

We start with ordinary least squares (OLS) as a baseline and extend to regularized and robust variants to improve generalization and handle multicollinearity:

$$R_t = \alpha + \beta_1 S_t + \beta_2 U_t + \beta_3 D_t + \epsilon_t \tag{1}$$

where R_t denotes the market return, S_t the sentiment score, U_t the uncertainty measure, and D_t the dovish index.

2) Machine Learning Classification Framework

For market direction prediction, we implement a comprehensive suite of machine learning algorithms, treating the problem as a binary classification task where the target variable $y_t \in \{0, 1\}$ indicates negative (0) or positive (1) market movements. Each algorithm brings unique strengths to capture different aspects of the relationship between linguistic features and market behavior.

Support Vector Machine (SVM): SVM finds the optimal hyperplane that maximizes the margin between classes in a high-dimensional feature space. For non-linearly separable data, SVM employs kernel functions to map input features into higher-dimensional spaces:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{2}$$

where $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function (RBF, polynomial, or linear), α_i are Lagrange multipliers, and b is the bias term. Key hyperparameters include the regularization parameter C and kernel-specific parameters (e.g., γ for RBF kernel).

Decision Tree: Decision trees partition the feature space through recursive binary splits based on information gain or Gini impurity. Each internal node represents a decision rule on a linguistic feature, while leaf nodes contain class predictions:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{3}$$

where p_i is the proportion of samples belonging to class i in set S. Key hyperparameters include maximum depth, minimum samples per split, and minimum samples per leaf to control overfitting.

Random Forest: Random Forest combines multiple decision trees through bootstrap aggregating (bagging) and random feature selection. The final prediction aggregates votes from all trees:

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_B(\mathbf{x})\} \tag{4}$$

where $h_b(\mathbf{x})$ represents the prediction of the b-th tree and B is the total number of trees. This ensemble approach reduces overfitting while maintaining interpretability through feature importance scores.

XGBoost (Extreme Gradient Boosting): XGBoost iteratively constructs an ensemble of decision trees through gradient boosting, where each subsequent tree corrects the errors of previous trees:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(\mathbf{x}) \tag{5}$$

where f_t represents the t-th weak learner, η is the learning rate, and the objective function combines prediction accuracy with regularization:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{6}$$

Key hyperparameters include learning rate (η), maximum depth, subsample ratio, feature subsampling, and L1/L2 regularization parameters.

Naive Bayes: Naive Bayes applies Bayes' theorem with the assumption of conditional independence between features. For continuous features, we use Gaussian Naive Bayes:

$$P(y|\mathbf{x}) = \frac{P(y) \prod_{i=1}^{d} P(x_i|y)}{P(\mathbf{x})} \tag{7}$$

where $P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$ for Gaussian features. Despite its simplicity, Naive Bayes often performs well with text-derived features and provides probabilistic interpretations.

All models undergo hyperparameter optimization through cross-validation, and SHAP analysis provides interpretability across all algorithmic approaches, enabling comparison of feature importance and model decision-making processes.

C. Tools and Libraries

All experiments will be implemented in Python using the following open-source tools:

- **NLP:** `transformers`, `nltk`, `spacy`, `gensim`, `bertopic`
- **Finance Data:** `yfinance`, `pandas-datareader`
- **Machine Learning:** `scikit-learn`, `xgboost`, `shap`
- **Visualization:** `matplotlib`, `seaborn`, `plotly`
- **Version Control:** GitHub
- **Environment:** Google Colab / AWS SageMaker

V. Results and Metrics

The effectiveness of our approach in linking Federal Reserve speech content to market movements requires a comprehensive evaluation framework that addresses both the predictive accuracy of our models and the economic significance of the relationships we uncover. Given the multi-faceted nature of our analysis—encompassing regression modeling, classification tasks, and event study methodologies—we establish a robust metrics framework that evaluates performance across different analytical dimensions while ensuring statistical rigor and practical relevance.

Our evaluation strategy recognizes that financial market prediction presents unique challenges compared to traditional machine learning applications. Market data exhibits high volatility, non-stationarity, and complex temporal dependencies that require specialized validation approaches. Moreover, the economic interpretation of our results demands that we move beyond simple predictive accuracy to assess whether identified relationships are statistically significant and economically meaningful in real-world trading scenarios.

A. Evaluation Metrics

Our comprehensive evaluation framework employs distinct metric categories tailored to each analytical component of our study. This multi-dimensional approach ensures that we capture both the statistical performance of our models and the economic significance of the linguistic features we extract from Federal Reserve communications.

- **Regression Models:**
  For our continuous return prediction models, we employ metrics that capture both explanatory power and predictive accuracy. The coefficient of determination (R^2) and adjusted R^2 measure the proportion of market return variance explained by our linguistic features, with adjusted R^2 penalizing model complexity to prevent overfitting. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors in percentage terms, providing an interpretable measure of model accuracy that directly relates to potential trading performance. Statistical significance tests using t-statistics and p-values validate the robustness of individual feature contributions, ensuring that observed relationships are not due to random chance.
  - Coefficient of Determination (R^2), Adjusted R^2
  - Mean Absolute Error (MAE)
  - Statistical significance tests (t-statistics and p-values)
- **Classification Models:**
  Our binary classification task of predicting market direction requires metrics that balance sensitivity to both positive and negative market movements. Accuracy provides an overall performance measure, while precision and recall for each class reveal potential biases toward bull or bear market predictions. The F1-score harmonizes precision and recall, particularly important given potential class imbalance in market direction data. ROC-AUC scores evaluate the model's ability to discriminate between market directions across different decision thresholds, with values above 0.6 indicating meaningful predictive power in financial contexts. Confusion matrices provide detailed insight into classification patterns and reveal systematic prediction biases.

– Accuracy, Precision, Recall, and F1-score
– ROC–AUC Score
– Confusion Matrix for classification performance visualization

- **Event Study Metrics:**
Event study methodology requires specialized metrics that capture both immediate and cumulative market responses to Federal Reserve communications. Average Abnormal Return (AAR) measures the mean market response across all speech events, controlling for expected market movements through benchmark models. Cumulative Abnormal Return (CAR) aggregates abnormal returns over specified event windows (e.g., [0,+1], [0,+3]), capturing both immediate and delayed market reactions to policy communications. Statistical significance testing using t-tests validates whether observed abnormal returns exceed random market fluctuations, with significance thresholds adjusted for multiple hypothesis testing across different event windows and market indices.
  – Average Abnormal Return (AAR)
  – Cumulative Abnormal Return (CAR)
  – Significance testing using t-tests across event windows

B. Validation Strategy

Financial time series prediction demands validation approaches that respect temporal dependencies and avoid look-ahead bias that could artificially inflate performance metrics. Our validation framework combines temporal splitting with robustness checks to ensure that our findings generalize to out-of-sample periods and are not artifacts of specific market conditions or data peculiarities.

The temporal nature of financial markets necessitates chronological data splitting rather than random cross-validation, which could introduce future information into training sets. Our approach simulates realistic prediction scenarios where models trained on historical data must predict future market movements without access to contemporaneous information. Additionally, robustness checks help distinguish genuine linguistic signals from spurious correlations that might arise from specific market regimes or external economic factors.

- **Temporal Train-Test Split:** We implement a chronological split using data from 1996–2015 for training and 2016–2020 for testing, simulating realistic prediction scenarios where models must forecast future market behavior based solely on historical patterns. This approach prevents look-ahead bias and evaluates model performance across different market regimes, including the 2008 financial crisis (in training) and post-crisis recovery period (in testing).
- **Cross-Validation:** Within the training period, we employ 5-fold time-series cross-validation with expanding windows to tune hyperparameters and assess model stability across different sub-periods. This approach maintains temporal order while providing multiple validation sets for robust parameter selection and performance estimation.
- **Robustness Checks:** We implement several robustness tests to validate the authenticity of our findings, including placebo tests using randomized event dates to verify that results are not due to calendar effects or coincidental timing. Control variables for prevailing macroeconomic conditions such as federal funds rates, inflation (CPI), and market volatility (VIX) help isolate the independent effect of speech content from broader economic contexts. Additionally, we test model stability across different market regimes (bull/bear markets, high/low volatility periods) to ensure consistent performance.

VI. Detailed Weekly Timeline

Table I outlines the planned weekly milestones and deliverables from proposal submission until the final project deadline on December 5, 2025. Each week, the team will report progress, demonstrate intermediate results, and adjust based on feedback.

VII. Conclusion

This research proposal presents a comprehensive framework for systematically analyzing the relationship between Federal Reserve Governors' speeches and short-term stock market performance through advanced natural language processing and data mining techniques. By addressing the critical gap in understanding how linguistic nuances in central bank communications translate into measurable market reactions, this study contributes to both theoretical knowledge and practical applications in financial markets.

Our methodology integrates state-of-the-art NLP tools, including FinBERT sentiment analysis and domain-specific financial lexicons, with rigorous event study approaches and machine learning models to extract predictive signals from Federal Reserve communications. The comprehensive evaluation framework, incorporating both statistical significance testing and economic interpretation through SHAP analysis, ensures that our findings are both methodologically sound and practically meaningful for market participants.

The expected outcomes of this research extend beyond academic contributions to offer tangible benefits for investors, policymakers, and financial analysts. By developing a systematic approach to quantify the informational content in central bank language, we provide a novel tool for understanding monetary policy transmission mechanisms and predicting market responses to future Federal Reserve communications. This work demonstrates the growing importance of textual analysis in financial forecasting and establishes a foundation for future research in computational finance and central bank communication analysis.

The successful completion of this project will advance our

TABLE I
WEEKLY PROJECT TIMELINE

Week	**Dates**	**Milestone / Planned Accomplishment**
1	Oct 20 – Oct 26	Finalize proposal draft, confirm dataset sources, assign responsibilities for data collection and NLP pipeline development.
2	Oct 27 – Nov 2	Collect and clean all Federal Reserve speech data and market data; align datasets by date and speaker metadata.
3	Nov 3 – Nov 9	Preprocess text data (tokenization, lemmatization, and stopword removal).
4	Nov 10 – Nov 16	Apply FinBERT and Loughran–McDonald lexicon for sentiment and uncertainty scoring.
5	Nov 17 – Nov 23	Conduct event study; calculate abnormal and cumulative abnormal returns.
6	Nov 24 – Nov 30	Train and test ML models (Logistic Regression, Random Forest, XGBoost).
7	Dec 1 – Dec 5	Conduct SHAP analysis; finalize report and presentation.

understanding of how communication serves as a monetary policy instrument while providing practical insights that could enhance investment decision-making and risk management strategies in modern financial markets.

REFERENCES

[1] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with Applications*, vol. 73, pp. 125–144, 2017.

[2] H. Wu, W. Zhang, W. Shen, and J. Wang, "Hybrid Deep Sequential Modeling for Social Text-Driven Stock Prediction," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, Torino, Italy, 2018, pp. 1627–1630.

[3] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," *arXiv preprint arXiv:2003.08137*, 2020.

[4] Y. Soun, J. Yoo, M. Cho, J. Jeon, and U. Kang, "Accurate Stock Movement Prediction with Self-supervised Learning from Sparse Noisy Tweets," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 1691–1700.

[5] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*, Buenos Aires, Argentina, 2015, pp. 2327–2333.

[6] Y. Liu, Q. Zeng, H. Yang, and A. Carrio, "Stock Price Movement Prediction from Financial News with Deep Learning and Knowledge Graph Embedding," in *Pacific Rim Knowledge Acquisition Workshop*, 2018.

[7] K. A. Keith and A. Stent, "Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls," *arXiv preprint arXiv:1906.02868*, 2019.

[8] A. Petropoulos and V. Siakoulis, "Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique," *Central Bank Review*, vol. 21, no. 4, pp. 141–153, 2021.

[9] Z. Zhou, L. Ma, and H. Liu, "Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading," *arXiv preprint arXiv:2105.12825*, 2021.

[10] Araci, D. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint arXiv:1908.10063*, 2019.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017