STAFF
REPORTS

NO. 1081
NOVEMBER 2023

Fed Transparency and Policy Expectation Errors: A Text Analysis Approach

Eric Fischer | Rebecca McCaughrin | Saketh Prazad | Mark Vandergon

FEDERAL RESERVE BANK of NEW YORK

Fed Transparency and Policy Expectation Errors: A Text Analysis Approach
Eric Fischer, Rebecca McCaughrin, Saketh Prazad, and Mark Vandergon
Federal Reserve Bank of New York Staff Reports, no. 1081
November 2023
https://doi.org/10.59576/sr.1081

Abstract

This paper seeks to estimate the extent to which market-implied policy expectations could be improved with further information disclosure from the FOMC. Using text analysis methods based on large language models, we show that if FOMC meeting materials with five-year lagged release dates—like meeting transcripts and Tealbooks—were accessible to the public in real time, market policy expectations could substantially improve forecasting accuracy. Most of this improvement occurs during easing cycles. For instance, at the six-month forecasting horizon, the market could have predicted as much as 125 basis points of additional easing during the 2001 and 2008 recessions, equivalent to a 40-50 percent reduction in mean squared error. This potential forecasting improvement appears to be related to incomplete information about the Fed's reaction function, particularly with respect to financial stability concerns in 2008. In contrast, having enhanced access to meeting materials would not have improved the market's policy rate forecasting during tightening cycles.

JEL classification: E43, E52, E58, C80
Keywords: interest rates, monetary policy, central bank and policy, sentiment analysis

Fischer, McCaughrin: Markets Group, Federal Reserve Bank of New York (email: eric.fischer@ny.frb.org, mccaughrin@ny.frb.org). Prazad: Research Analyst, Research Group, Federal Reserve Bank of New York (email: saketh.prazad@ny.frb.org). Vandergon: Technology Group, Federal Reserve Bank of New York (email: mark.vandergon@ny.frb.org). The authors thank Michael Bauer, Ryan Bush, Richard Crump, Matthew Gentzkow, Seung Lee, Eric Offner, Chiara Scotti, Adam Shapiro, Derek Tang, and seminar participants at the Federal Reserve Bank of New York, the Federal Reserve Board, the Federal Reserve Bank of Atlanta Quantitative Skills Conference, and the Western Economic Association International Conference for helpful comments. They also thank Federal Reserve Bank of New York Markets Group analysts for assistance with sentence annotation and Thierno Diallo for excellent research assistance.

This paper presents preliminary findings and is being distributed to economists and other interested readers solely to stimulate discussion and elicit comments. The views expressed in this paper are those of the author(s) and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. Any errors or omissions are the responsibility of the author(s).

To view the authors' disclosure statements, visit
https://www.newyorkfed.org/research/staff_reports/sr1081.html.

1 Introduction

Over the last few decades, central banks have made substantial changes to their communications and transparency practices in an effort to improve market expectations of future monetary policy. Clear and effective communications help to ensure that policy shifts are properly transmitted to financial conditions and the real economy. Recent work on monetary policy expectations has documented that while market expectations may often be unbiased predictors of the future policy rate, there can be substantial ex-post predictability of market expectation errors, particularly during significant policy easing episodes (Cieslak, 2018), (Schmeling, Schrimpf and Steffensen, 2022), (Bauer and Swanson, 2023). These expectation errors are believed to be ex-ante unpredictable and driven by an underestimation of the central bank's sensitivity to economic downturns. In this paper, we argue that Fed transparency plays some role in these expectation errors, and demonstrate that a portion of the expectation errors are predictable ex-ante, given enhanced access to information from monetary policy meetings.

We define transparency in terms of *information loss* between the public and the Fed with respect to expectations of future monetary policy. In a procedural sense, central banks are not fully transparent. Monetary policy meeting deliberations are confidential. Transcripts and related briefing materials are only released to the public with a significant lag. To be sure, central banks have legitimate reasons to maintain confidentiality or institute lags in providing access to information. Such opacity, which we define as the inverse of transparency, may be deemed necessary, for instance, to maintain independence and encourage objective and vigorous debate among policymakers. If internal discussions and staff economic forecasts were to become immediately available to the public, market participants could perceive such forecasts as a commitment to certain future actions by the central bank, which could in turn reduce the flexibility of the central bank in the future.

In lieu of immediate access to meeting materials, the public receives a wide range of communications from the Fed, including statements, meeting minutes, speeches, interviews, and press conferences. If these communications can be used to accurately predict the Fed's future policy decisions, despite the confidentiality of meetings, then we posit that there is no information loss. In this paper, we do not attempt to weigh the trade-offs of increased transparency against potential improvements in market policy forecasts. Rather, our goal is to highlight potential costs of opacity that have been previously unacknowledged by the literature.

We measure information loss by comparing actual market policy expectations with our predictions of what market expectations would have been in a counterfactual world in which the Fed releases meeting transcripts and Tealbooks immediately instead of with a five year lag.[1] These counterfactual predictions are generated using our "FedSpeak model", a forecasting model that predicts the change in the fed funds rate h months after meeting t, using fed funds futures from the day after meeting t, confidential sentiment and topic content from meeting t, and confidential Tealbook forecasts from meeting t.[2] One day after a meeting, the futures market should have priced in all relevant information released before the meeting, including meeting minutes, speeches, and macroeconomic data releases.

We produce time-varying measures of sentiment and topic content of meeting transcripts and Tealbooks. Topic modeling and sentiment analysis are standard machine learning techniques that allow for a detailed quantitative assessment of how subjective content of text changes over time. We measure sentiment and topic at the sentence-level. A sentence can have a sentiment label of "Positive", "Negative", or "Neutral". The same sentence can also have a topic label of "Economic Growth", "Inflation", "Labor Markets", "Financial Stability", and "Monetary Policy". A sentence may also have no associated topic label. The text analysis is conducted using supervised machine learning methods and large language models, which we train using a dataset of sentences from FOMC speeches, statements, minutes, transcripts, and interviews. The topic and sentiment labels for this training dataset were determined by analysts from the Markets Group at the Federal Reserve Bank of New York. We show that our machine learning-based approaches generate improvements in accuracy on a held-out test set relative to the lexical approaches commonly used in economics.

We show that the out-of-sample predictions of the FedSpeak model significantly outperform fed funds futures at three-, six-, and nine-month forecasting horizons, thus indicating opacity. The outperformance we find for the full 1996 to 2016 period reveals an important asymmetry between tightening and easing cycles. During the 2001 and 2008 easing cycles, the gap between fed funds futures and the FedSpeak model's predictions was as much as 125 basis points at the six month

[1] Before 2010, the two main staff briefing documents were the Greenbook, which summarized economic developments, and the Bluebook, which summarized monetary policy options. In 2010, the two documents were merged into the Tealbook. For the rest of this paper, when we refer to "Tealbook", we refer to the pre-2010 Greenbooks and the post-2010 Tealbooks. We do not consider Bluebooks.

[2] In this counterfactual world, we assume that the Fed does not adjust the contents of its meetings in response to greater transparency.

horizon, equivalent to a 40-50% reduction in mean squared error. We find no such outperformance during tightening cycles and during the zero lower bound. These results imply that Fed communications may be less informative during easing cycles, to the extent that meeting materials contained policy-relevant information not reflected in market pricing.

One potential explanation for our results is that although the market takes into account communications released before meeting t, a large portion of the important information within meeting transcripts is released to the public in the speeches and minutes released after meeting t. If post-meeting communications are highly informative, then the opacity we identify in our main results may be temporary. To account for this possibility, we implement our forecasting exercise with a modified FedSpeak model that pairs market expectations one day after meeting t with meeting materials from meeting $t-1$ (rather than from meeting t). If post-meeting communications eliminate opacity, then information from meeting $t-1$ should no longer be important since Fed communications released between meeting $t-1$ and meeting t would have been priced in by the market. Instead, we find the modified FedSpeak model continues to outperform the market during easing cycles, though by a somewhat lower magnitude. This suggests that opacity persists many weeks after a meeting.

Using an alternative metric of transparency, we verify the implication that the outperformance of the FedSpeak model reflects less informative communications during easing cycles. To do this, we test whether a hypothetical investor could have used meeting minutes in real time to predict the contents of confidential meeting transcripts. The investor estimates the historical relationship between the contents of meetings minutes and the contents of meeting transcripts and then applies that historical relationship to new minutes observations in order to generate predictions of transcript content. We compare these predictions with actual transcript content in order to assess transparency. We continue to find an asymmetry between tightening cycles and easing cycles. In particular, predicted transcript sentiment tends to be more positive than actual transcript sentiment during easing cycles, which we posit may explain why the market underestimated future rate cuts.

We offer two main caveats to our analysis. First, our results cannot identify whether recent innovations in Fed communications, like the Summary of Economic Projections (SEP) and press conferences, have improved Fed transparency. Due to the lagged release of meeting transcripts, we can simultaneously observe the SEP, press conferences, and meeting transcripts only from 2011 through 2017. These years were mostly characterized by the zero lower bound (ZLB) and explicit forward guidance, both of which effectively reduce the scope for market expectation errors. Evaluating the effects of the SEP and press conferences would require more post-ZLB meeting transcripts to be released. Second, our results do not consider the possibility that FOMC members could respond to enhanced transparency by altering the content discussed at monetary policy meetings. These behavioral responses are emphasized in Hansen, McMahon and Prat (2018), who study the response of policymakers to the transparency-enhancing reforms in 1993.

The asymmetry we find between tightening and easing cycles is consistent with the results of Cieslak (2018) and Schmeling, Schrimpf and Steffensen (2022), who find large excess returns in Treasuries and fed funds futures during easing cycles, but not during tightening cycles. They attribute the excess returns to market expectation errors, rather than to changing risk premia. Cieslak (2018) argues that while these forecasting errors may be predictable ex-post, they are difficult to predict in real-time, even for policymakers. Other papers finding asymmetries in monetary policy expectations between tightening and easing cycles include Bauer, Pflueger and Sunderam (2022), who find that professional forecasters perceive policy decisions to be less dependent on macroeconomic conditions during easing cycles and therefore less predictable.

We also contribute to a long-standing literature on Fed transparency. Most work on transparency and central bank communication has studied the optimal level of transparency and the conditions under which signaling the path of future rates is welfare-enhancing.[3] Less work has been done on the extent to which the Fed achieves transparency in practice. We fill this gap by empirically testing the extent to which markets could have historically improved their policy rate forecasting with broader access to information from central bank meetings.

Many prior studies of central bank transparency have focused on relative improvements in transparency. For example, Swanson (2006) finds that the private sector has become better at forecasting monetary policy since the 1980s, likely due to improved Fed transparency. But these studies can only measure relative changes in transparency and cannot tell us how transparent communications are overall. The only studies that attempt to measure central bank transparency in an absolute sense rely on a qualitative lens, such as Eijffinger and Geraats (2006) and Dincer and Eichengreen (2018), who develop qualitative indices of transparency over time for central banks around the

[3] For details, see Woodford (2005), Cukierman (2009), Morris and Shin (2005), among others.

world. While these indices are useful for comparing a very diverse set of institutions, they rely on coarse binary criteria.

Our second set of results identifies the specific pieces of information within meeting transcripts and Tealbooks that explain the large gaps between the FedSpeak model's predictions and market expectations during easing cycles. We generate variable importance measures from the FedSpeak model to determine which pieces of information within meeting deliberations would have been most valuable for policy-sensitive rate markets to have known in real-time. In 2007-2008, the most important variables were financial stability topic frequency, financial stability sentiment, economic growth sentiment, and the sentiment of the FOMC's leadership. In 2000-2003, the most important variables were aggregate sentiment, economic growth sentiment, and the sentiment of leadership. In both periods, the importance of these variables often far surpassed the importance of market expectations.

We interpret the variable importance results through the lens of a simple monetary policy rule, where the future fed funds rate is determined by the committee's economic outlook and a time-varying reaction function. According to Nakamura and Steinsson (2018), Campbell, Evans and Justiniano (2012), and Romer and Romer (2000), the Fed has an "information advantage" about the economy and can therefore make better forecasts than the public about future economic conditions. Based on this view, the FedSpeak model might outperform the market because the meeting materials contain important information about the Fed's economic forecasts.

The view that the Fed has stronger forecasting abilities than market participants has been challenged by Hoesch, Rossi and Sekhposyan (forthcoming) and Bauer and Swanson (2023), among others, who find that Tealbook forecasts have not been more accurate than private sector forecasts in recent years. Consistent with these findings, we show that Tealbook forecasts were relatively unimportant within the FedSpeak model during easing cycles and pointed towards tighter policy rather than looser policy. Since Tealbook forecasts are commonly assumed within the literature to reflect the FOMC's economic outlook (see, for example, Shapiro and Wilson (2019) and Bauer and Swanson (2023)), the low importance of these forecasts suggests that outlook-related information was unlikely to be responsible for reduced transparency during easing cycles.

As an additional test of the importance of outlook-related information, we re-estimate the FedSpeak model under a hypothetical scenario where the Fed has perfect foresight of future economic conditions. We modify the FedSpeak model to include macroeconomic data releases, like nonfarm payrolls and CPI inflation, from the month *after* meeting t. Even with such extreme assumptions about the Fed's knowledge of the economy, the topic and sentiment variables retain at least 75% of their predictive power. This suggests that the text-derived variables are capturing information unrelated to the state of the economy and are thus providing information about policymakers' reactions to incoming economic news.

We find that sentiment and topic frequency variables are often more important than market expectations for forecasting future policy during easing cycles, providing a compelling affirmative case for a reaction function-based explanation of easing cycle opacity. Sentiment is important because it allows us to directly observe the reactions of FOMC members to incoming economic data rather than having to infer their reaction based on historical relationships between macroeconomic variables and monetary policy. Topic frequency variables should also be interpreted as related to the reaction function. Holding topic-specific sentiment constant, if the FOMC discusses a certain topic more often, then the members may implicitly be weighting that topic more heavily in their reaction function.

Our results emphasizing incomplete information about policymakers' reactions are consistent with a growing literature. Bauer and Swanson (2023) propose a "Fed response to news" channel for explaining monetary policy surprises. Schmeling, Schrimpf and Steffensen (2022) find that market expectation errors occur contemporaneously with Taylor Rule deviations. Cieslak (2018) finds that a large portion of unexpected easing comes from unscheduled FOMC meetings, suggesting that FOMC members eased more aggressively than markets expected in response to surprise economic news. Bauer, Pflueger and Sunderam (2022) find that professional forecasters update their beliefs about the Fed's reaction function in response to monetary policy shocks, indicating imperfect information about the reaction function. We contribute to this literature by providing the first quantitative evidence of the magnitude of these information frictions.

A reaction function-based explanation of the FedSpeak model's performance has quite different implications for Fed transparency policy than a forecast or outlook-based explanation. Information about forecasts may be relatively easy to convey to market participants, through instruments like the Summary of Economic Projections. But as Woodford (2005) emphasizes, conveying information about the Fed's reaction function to the public is difficult because of the large number of different contingencies and scenarios that may arise, each demanding a different response from the central bank. The easing cycles that we emphasize may be examples of infrequent contingencies, where conveying information about the Fed's response in advance may be difficult in practice. In a similar vein, Schmeling, Schrimpf and Steffensen (2022) argue that the Fed's response to negative macroeconomic shocks is inherently difficult for markets to learn because of the relatively few observations.

Finally, we contribute to a large literature that uses textual analysis of FOMC documents and communications to study FOMC communications and the transmission of communications to financial markets. Some important papers include Hansen and McMahon (2016), Acosta (2015), Gardner, Scotti and Vega (2022), Schmanski et al. (2023), Chernulich, Li and McGinn (forthcoming), and Hansen and Kazinnik (2023).

The rest of the paper proceeds as follows: in Section 2, we discuss the text analysis methods used to summarize the qualitative content of Fed documents. In Section 3, we describe the construction of the FedSpeak model and show its superior forecasting ability relative to market expectations. In Section 4, we use variable importance measures to argue that market forecasting errors during easing cycles were due to incomplete information about the FOMC's reaction function rather than the economic outlook. Section 5 offers concluding remarks.

2 Text Analysis Methodology

In this section, we describe the text analysis methods we use for topic classification and sentiment analysis. We define topic as the subject of a speaker's sentence, as it relates to the economy or monetary policy. For example, in the phrase "market liquidity is worsening", the topic is clearly market functioning, or financial stability more broadly. In the phrase, "the CPI release shows mixed signals", the topic is clearly inflation.

We define sentiment as the subjective attitude that a speaker conveys through their language. For example, the phrase "market liquidity is worsening" conveys negative sentiment because the speaker is making an opinionated statement with a negative directionality. On the other hand, the phrase "the outlook for the labor market looks bright" conveys positive sentiment. Some sentences may have neutral sentiment, either because they contain both positive and negative attitudes ("the labor market is strong, but inflation is weak") or because they contain explicitly neutral language ("the CPI release shows mixed signals") or because they are purely factual ("the FOMC has a 2 percent inflation target.")

2.1 Human Annotation Exercise

To develop and validate text analysis methods, we took a random sample of sentences from FOMC documents, which were then categorized by topic and sentiment by analysts at the Federal Reserve Bank of New York. The specific details of the annotation exercise are included in the Appendix. The exercise resulted in a dataset of roughly 2,350 sentences, each with a topic and sentiment label. Each sentence was annotated by three different reviewers and a consensus label selected. If a sentence did not receive a majority vote, then we discarded the sentence from the dataset. The human annotators chose among four sentiment labels: Positive, Negative, Neutral, No Sentiment and among six topic labels: Economic Growth, Labor Market, Financial Stability, Inflation, Monetary Policy, and No Topic.

The Labor Market and Inflation topics relate to the Fed's dual mandate. Economic Growth relates to discussion of economic output or economic activity, including GDP growth, conditions related to different sectors, housing, business investment, consumer spending, etc. Financial Stability relates to discussion of financial market risks and vulnerabilities. Monetary Policy relates to any discussion of policy, including policy rate decisions, asset purchases, policy implementation framework, etc. These topics were selected to roughly correspond to the components of a central bank's loss function. Shapiro and Wilson (2019) find that the Fed's loss function has historically included not just unemployment and inflation, but also output growth and financial variables. For sentences unrelated to any of the five main topics,"No Topic" was selected. Crucially, if a sentence contained more than one topic, then the annotator could attach more than one topic label.

2.2 Text Analysis Methodology

In industry, sentiment analysis is usually conducted using supervised machine learning methods and large language models, such as BERT and GPT. While these tools are generally powerful off-the-shelf, they often require domain-specific fine-tuning. We utilize a pre-trained, finance-domain large language model called FinBERT, developed by Huang, Wang and Yang (2023), and further fine-tune it to the central banking domain using our annotated dataset of Fed sentences drawn from policymaker speeches, interviews, FOMC meeting statements, minutes, and transcripts.

In economics, sentiment analysis is usually done using rules-based, lexical methods. This method involves creating a pre-defined list of words and phrases that correspond to categories of interest. An algorithm will assign a sentiment label to a sentence based on the number of matches in the sentence with these pre-defined lexicons. Examples of economics papers that use a lexical approach include Shapiro, Sudhof and Wilson (2020), Shapiro and Wilson (2019), and Hansen and McMahon (2016). The most widely used lexicon for sentiment analysis is from Loughran and McDonald (2011). This lexicon, from here referred to as "LM", was developed for a general economics and finance audience in order to analyze SEC filings and does not include many important pieces of central bank terminology and jargon. Therefore, we construct our own sentiment lexicon specifically tailored to central banking text, adopting methods from Correa et al. (2021) and Picault and Renault (2017).

Analysis of topic content in economics and in industry is often done using unsupervised machine learning methods, such as the Latent Dirichlet Analysis (LDA) method used by Hansen, McMahon and Prat (2018). These methods are often referred to as "topic modelling". LDA observes the co-occurrences between words in sentences to find groupings of words that can be interpreted as topics. While this method is often effective at finding optimal groupings, it has a clear limitation: the researcher cannot pre-specify a list of desired topics, only a desired number of topics. LDA finds the requested number of groupings and outputs the most common words within each grouping, leaving it to the researcher to label these groupings after the fact. This can be very useful in situations where the underlying set of topics is unknown to the reader, making LDA a valuable tool in exploratory analysis of text data. But in the context of FOMC communications, the researcher has very strong priors about the true set of topics. Additionally, certain specific topics in central banking have strong theoretical value, like inflation and labor markets.

For example, applied to FOMC communications, LDA could output a topic that mixes together words related to labor markets and inflation. This might be an optimal way to group the text, especially considering the connections between labor markets and inflation. But this mixed topic is less meaningful and less interpretable than having a separate labor market and inflation topic, especially since our priors indicate that the true set of topics separates labor markets and inflation.

Rather than analyzing topic content as an unsupervised modelling problem, we instead analyze topics as a supervised classification problem. We train supervised machine learning models using our annotated dataset to classify each sentence by topic. Our model is able to output multiple topic labels if a sentence contains more than one topic. Details about our topic classification model, as well as our use of FinBERT for sentiment analysis, are available in the Appendix. We also include results based on a ChatGPT-like generative AI models for sentiment and topic analysis.

2.3 Validation and Machine Learning

We validate our text analysis approaches using our human-annotated dataset. We split the sentences into a training set and test set. We train the machine learning models on the training set and validate both the lexical methods and the machine learning methods on the held-out test set.

For topic classification, the machine learning model outputs all of the topics that it believes are contained within a given sentence. We test accuracy topic-by-topic. When testing for topic T, we determine whether the list of predicted topics contains topic T. Suppose that a sentence contains topics T_1 and T_2 and the model thinks that the sentence contains only topic T_1. When testing for topic T_1, we score the sentence as a successful classification, but when testing for topic T_2, we score the sentence as an unsuccessful classification. If the sentence does not contain T_3 and the model predicts that the sentence does not contain T_3, then we score the sentence as a successful classification. In Table 1, we provide the model's accuracy for each topic. We find that the model is highly proficient at topic classification, with accuracy of roughly 90% or greater and average F1 score to be above 0.7 (out of 1) for all five topics. F1 score is a common classification performance metric defined as the harmonic mean of precision and recall scores, which are important for datasets with high class imbalance.

Table 1: Test Set Performance of Topic Classification for Supervised Machine Learning

Topic	**Accuracy**	**Average F1**
Economic Growth	0.88	0.80
Financial Stability	0.88	0.75
Inflation	0.98	0.91
Labor Market	0.97	0.87
Monetary Policy	0.89	0.83

For sentiment, we test whether the predicted sentiment label is equal to the actual sentiment label for each methodology. In Table 2, we provide accuracy metrics for each methodology. We find that our central banking-specific sentiment lexicon outperforms the Loughran-McDonald lexicon. However, the machine learning-based methods ultimately scored the best, particularly the fine-tuned FinBert model. We thus use the sentiment output of the fine-tuned FinBert model for the main analysis in this paper.

Table 2: Test Set Performance of Sentiment Analysis by Methodology

Methodology	**Accuracy**	**Average F1**
Lexical: FMPV *	0.75	0.69
Lexical: Loughran-McDonald	0.64	0.61
ML: FinBert off-the-shelf	0.73	0.75
ML: FinBert fine-tuned *	0.78	0.78

2.4 Sentiment Time Series

We use our machine learning models to generate a sentiment prediction and topic prediction for every sentence in our corpus of FOMC text. We convert the sentiment predictions to numerical scores by assigning a score of +1 to a prediction of "Positive", a score of 0 to a prediction of "Neutral" or "None", and a score of -1 to a prediction of "Negative. We aggregate these sentence-level sentiment scores to produce document-level sentiment scores. We generate a sentiment score for document *d* at time *t* by averaging over the sentences *s* within the document:

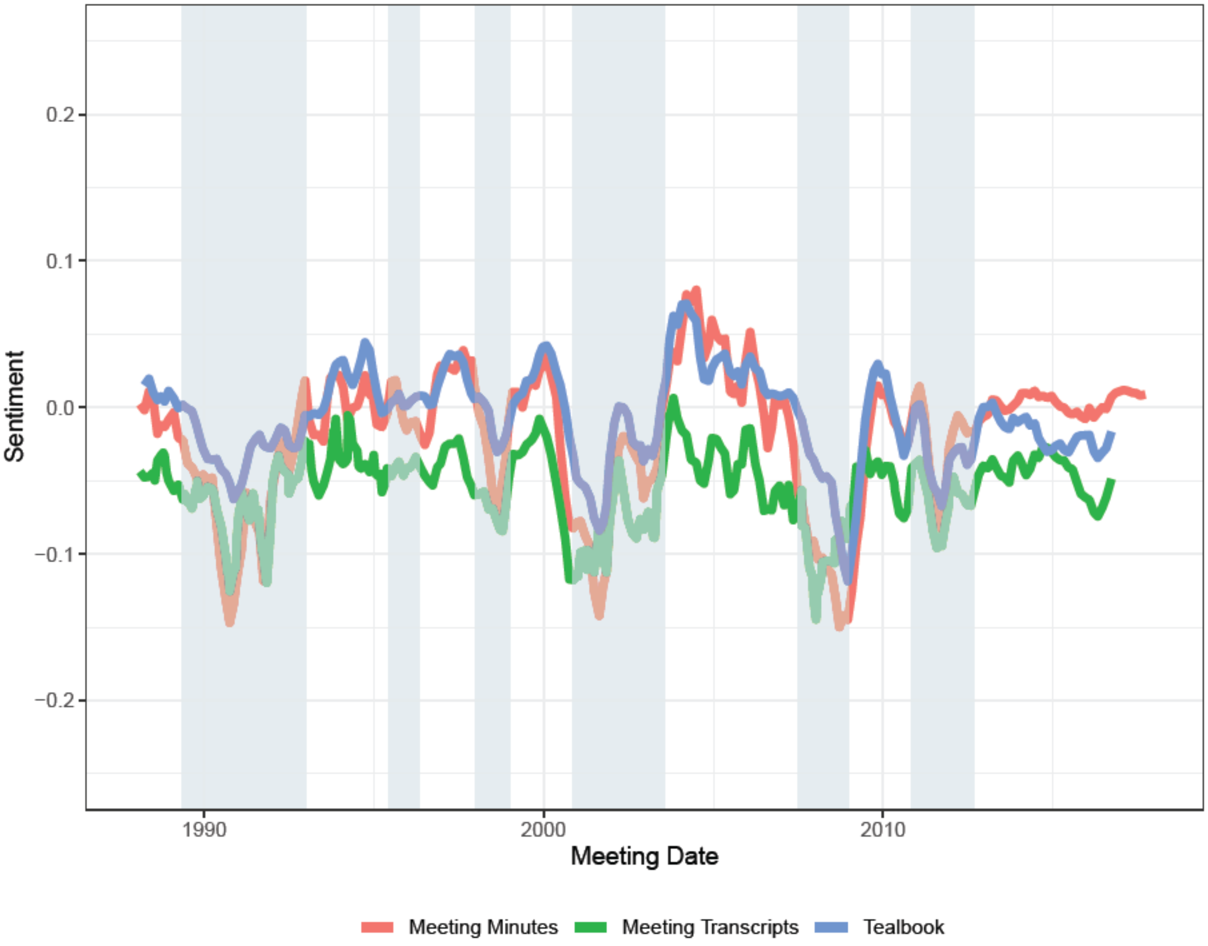
$$Sentiment_{d,t} = \frac{1}{n}[\sum_{s=1}^{n} Sentiment_{s,d,t}] \tag{1}$$

We also generate time series for subsets of the text corresponding to certain topics or speakers. We group the text by communication type, meeting date, topic, and/or speaker and calculate the average sentiment within each group. For the topic-specific sentiment time series, we average all of the sentence-level sentiment scores within a given document, subsetting on sentences that have at least one mention of topic *k*. This means that the same sentence could potentially be considered for multiple topic-specific sentiment series.

In Figure 1, we plot the sentiment time series for the three main document types used in this paper: meeting transcripts, Tealbooks, and meeting minutes, presented as a four-meeting rolling average.[4]

[4] This is roughly equivalent to a six month rolling average.

Figure 1: Aggregate Sentiment Time Series



Notes: In this figure, we plot sentiment time series as explained in Section 2.3 for meeting minutes, meeting transcripts, and Tealbooks. We apply a four meeting rolling average to each series. We highlight periods during which the FOMC cut interest rates. We also highlight the November 2010 through September 2012 period during which the FOMC announced two rounds of quantitative easing and implemented calendar-based forward guidance.

Overall, while the sentiment scores for the different types of documents appear to be closely correlated, the meeting transcripts appear to be consistently less positive than meeting minutes or Tealbooks. For example, the sentiment derived from meeting minutes is on average 0.15 points higher than meeting transcript sentiment. The aggregate sentiment time series also closely tracks changes in policy rates. The shaded areas in Figure 1 highlight periods where the FOMC loosened monetary policy. During nearly all of these easing periods, aggregate sentiment fell sharply for all three document types.

3 Measuring Transparency using Market Policy Expectations

3.1 FedSpeak Model Setup

We define transparency in terms of information loss between the Fed and the market with respect to expectations of future monetary policy. We compare actual, realized market policy expectations versus what those expectations would have been in a counterfactual world in which the Fed immediately released its meeting transcripts and Tealbooks. We estimate the counterfactual expectations using a forecasting model that we refer to as the "FedSpeak model". We first provide the equation of the model and then explain its components.

$$AvgEFFR_{t+h} - EFFR_t = \alpha + \beta(MktExp_{t+h} - EFFR_t) + Text_t\gamma + \varepsilon_t \qquad \text{(FedSpeak Model)}$$

We index each FOMC meeting by t. $AvgEFFR_{t+h}$ is the effective fed funds rate h months after meeting t, averaged over the month. $EFFR_t$ is the average effective fed funds rate over the five days following meeting t.[5]

$MktExp_{t+h}$ is the market-forecasted average EFFR for h months after meeting t, based on fed funds futures pricing from the day after the meeting.[6] We assume that one day after the meeting, the futures market should have priced in any relevant macroeconomic data, any new information revealed by the Fed's policy announcement at meeting t, and all Fed communications (like minutes and speeches) that were released prior to the meeting.

The fact that pre-meeting communications should be priced into policy expectations is very important to our interpretation of the FedSpeak model. In his study of Fed chair speeches, Swanson (2023) notes that policy decisions at FOMC meetings are increasingly signaled ahead of time through speeches by FOMC members, resulting in fewer monetary policy surprises. Swanson (2023) shows that since 1990, Fed chair speeches have had even more influence than FOMC announcements on a range of asset prices, except very short-term interest rate futures. If the FedSpeak model outperforms the market, despite pre-meeting signaling from speeches, then we can conclude that Fed communications were not fully transparent.

[5] The five day average is intended to smooth out volatility in the effective fed funds rate.

[6] Data retrieved from Bloomberg. The payout of a fed funds futures contract is based on the average effective fed funds rate during the expiration month.

While risk premia are included in policy expectations derived from fed funds futures, these risk premia were likely very small in magnitude for the short term horizons (two to nine months) that we use in this analysis. Schmeling, Schrimpf and Steffensen (2022) calculate term premia using Blue Chip survey-based expectations. They find that term premia on three month and six month fed funds futures were very low throughout our sample period, usually less than 25 basis points. The average over the period was slightly negative. In fact, during the 2001 and 2008 easing cycles that we focus on later in our results, term premia were sharply negative. This means that markets underestimated future policy easing even more than implied by fed funds futures. Other papers finding relatively very low risk premia over short horizons include Piazzesi and Swanson (2008) and Crump, Eusepi and Moench (2018).

$Text_t$ is a matrix containing variables derived from meeting materials, such as transcript sentiment and Tealbook forecast variables. We emphasize that this matrix contains information that the markets could not directly have accessed because of the five year release lag of the meeting transcripts and Tealbooks. In the next section, we provide further details about the variables included in $Text_t$.

To test for transparency, we conduct an out-of-sample forecasting exercise. We study regularly scheduled FOMC meetings from 1989 through 2016. At meeting t, we train a predictive model on observations 1 through $t-h-1$. This restriction on the training sample ensures out-of-sample validity: if we had trained on observations 1 through $t-1$ instead, then we would be training on observations of the outcome that had not yet been realized in real-time. We use the trained model to generate a forecast based on observation t. We impose a starting sample size of 60 observation, so our first prediction is for the December 1996 meeting. For every meeting afterwards, we re-estimate the FedSpeak model using newly available data. Thus, at every iteration, the sample used to train the model becomes one observation larger.

We do not impose any variable selection. We instead allow the models to down-weight unimportant variables through regularization. We test two different classes of models: LASSO and random forests. For LASSO, the regularization parameter is selected using ten-fold cross validation within the training set. Due to the rolling nature of the forecasting exercise, a new regularization parameter is selected at every iteration.

To account for the 2009-2015 period, we impose a zero lower bound on the forecasts: if a forecast implies a negative fed funds rate, then we adjust the forecast upwards until it implies a funds rate of zero. We also add two more variables to the model: the smoothed effective fed funds rate after meeting t and the smoothed effective fed funds rate after meeting $t-1$. These two variables can help the model recognize that future policy changes can differ based on the current level of rates. For example, there is less room to cut rates when the fed funds rate is already low.

After generating an out-of-sample prediction for every meeting, we compare the FedSpeak model's predictions to fed funds futures. If the FedSpeak model outperforms the market, then $Text_t$ contains important policy-relevant information that was not accessible to the markets and was thus not priced into the futures market. We interpret this outcome as reflecting opacity.

3.2 Meeting Content Variables

The following variables are included in $Text_t$. We derive the variables from meeting transcripts and Tealbooks, using our text analysis techniques when relevant. We linearly interpolate missing values. For all meeting transcript-related variables, we only consider remarks by FOMC members, thus excluding presentations by staff.

1. Aggregate meeting transcript sentiment

This is the average of all the sentence-level sentiment scores across a given meeting transcript. This is equivalent to the graphs shown in Figure 1 without a rolling average.

2. Topic-specific meeting transcript sentiment.

For topic k, we average all the sentence-level sentiment scores within a given meeting transcript for sentences that have at least one mention of topic k. The topics we consider are: Economic Growth, Labor Market, Inflation, Financial Stability, and Monetary Policy.

3. Dispersion of meeting transcript sentiment across members.

For each member i, we average all sentence-level sentiment scores for sentences spoken by member i, regardless of topic. We then calculate the standard deviation of these member-level sentiment scores.

4. Transcript sentiment for specific members

For each member or group of members *i*, we average all of the sentence-level sentiment scores for sentences spoken by member(s) *i*, regardless of topic. We consider the Chair, the FOMC leadership (Chair, Vice Chair, President of the Federal Reserve Bank of New York), Reserve Bank Presidents, and Board of Governors.

5. Frequency of mention for meeting transcript topics

For topic *k*, we find the number of sentences that mention topic *k* and divide it by the total number of sentences.

6. One quarter ahead Tealbook forecasts

We obtain these variables from the Philadephia Fed's Tealbook dataset. We use one quarter[7] ahead forecasts for unemployment, GDP, and core PCE inflation. We also use the staff's assumption for the fed funds rate over a one quarter horizon.

7. Aggregate Tealbook sentiment

The Tealbooks have large sections where the staff provide their qualitative views on the economy and monetary policy. We find the average of all of the sentence-level sentiment scores, regardless of topic.

8. Topic-specific Tealbook sentiment

For topic *k*, we average all the sentiment-level sentiment scores for sentences that have at least one mention of topic *k*. The topics we consider are: Economic Growth, Labor Market, Inflation, Financial Stability, and Monetary Policy.

3.3 Forecasting Results

In Figure 2, we calculate the mean squared error of market expectations and the FedSpeak model over the full 1996 through 2016 period. We use the Diebold-Mariano test to test the null hypothesis that the mean squared error of market expectations is equal to the mean squared error of the FedSpeak model. We repeat this aggregation for tightening cycles, easing cycles, and the zero lower

[7] We use the one-quarter-ahead forecasts in order to maintain a consistent set of Tealbook forecast variables for all forecasting horizons. Additionally, as emphasized by Shapiro and Wilson (2019), short-run Tealbook forecasts are especially informative because they are unlikely to be contaminated by the staff's assumptions about the path of monetary policy. Due to data availability, we do not include FFR forecasts for model predictions for 2015 and after.

bound. We consider the November 2000 through July 2003 period and the July 2007 through December 2008 period to be easing cycles.[8] Non-easing cycles are considered to be tightening cycles.

When we use random forests, the FedSpeak model outperforms market expectations over the entire 1996-2016 forecasting period. The results for the full period hides a significant asymmetry between easing cycles and tightening cycles. During easing cycles, the FedSpeak model outperforms market expectations at all eight forecasting horizons. The magnitude of the outperformance is large; the mean squared error (MSE) of the FedSpeak model is 40-50% less than that of market expectations. However, during tightening cycles, market expectations perform somewhat better than the FedSpeak model. We interpret these results as suggesting that the FOMC is more transparent during tightening cycles than during easing cycles.

The version of the FedSpeak model estimated using lasso regularization performs slightly better than random forests during easing cycles and slightly worse during tightening cycles. However, lasso regression performs worse during the zero lower bound, which drives its worse performance for the full period. This may occur because random forests are better able to capture the nonlinearities that the zero lower bound introduces.

[8] These dates were selected based on the peak and trough fed funds rate during the two easing cycles.

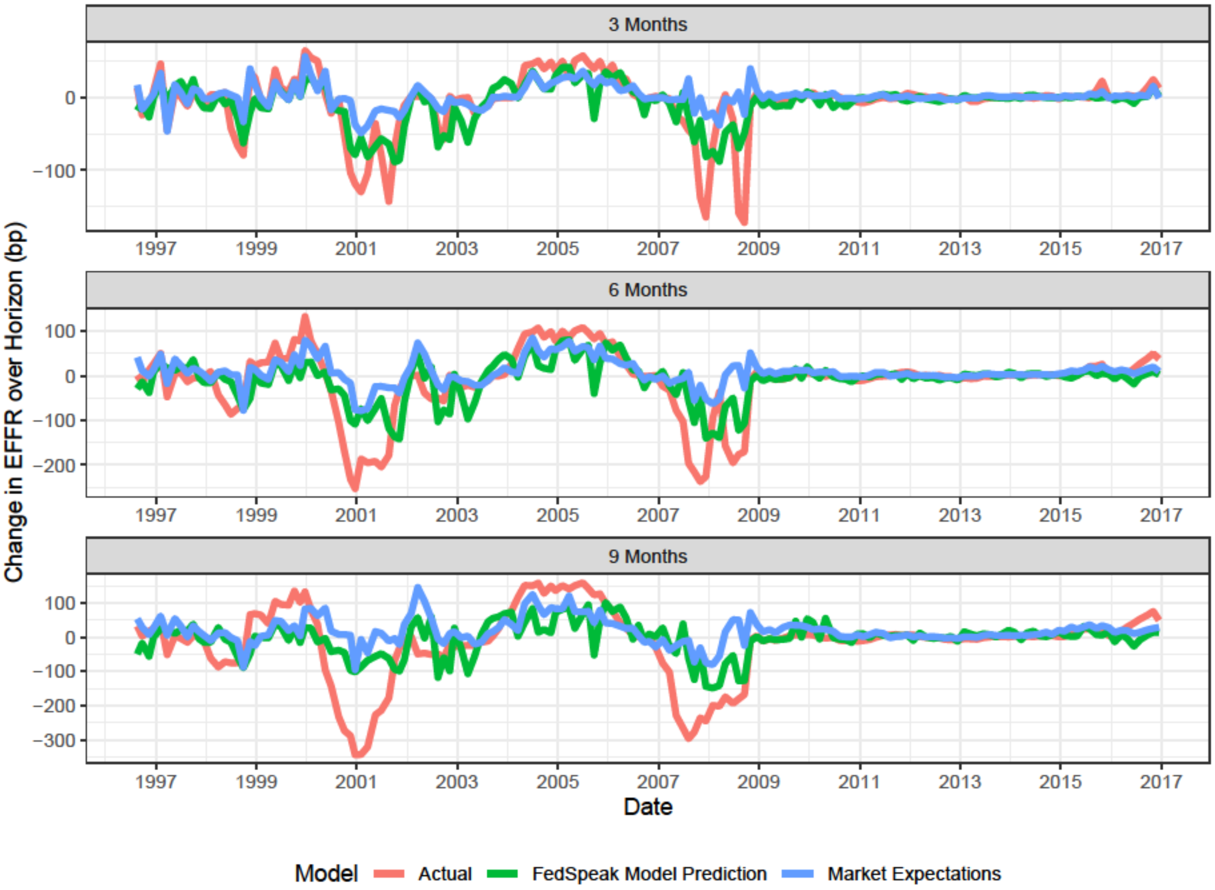
Figure 2: Forecasting Errors of Market Expectations and FedSpeak Model

	Full Period		Tightening Periods		Easing Periods		Zero Lower Bound	
Horizon (Months)	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE
Lasso								
3	0.109	0.078	0.035	0.052*	0.448	0.243**	0.002	0.012***
6	0.379	0.28*	0.172	0.249*	1.446	0.74***	0.004	0.043***
9	0.888	0.736	0.644	0.827*	2.848	1.323***	0.020	0.259***
Random Forest								
3	0.109	0.072*	0.035	0.049	0.448	0.235**	0.002	0.004***
6	0.379	0.271**	0.172	0.229*	1.446	0.8***	0.004	0.006
9	0.888	0.668***	0.644	0.763	2.848	1.52***	0.020	0.025

Notes: This table aggregates fed funds futures market expectation and the out-of-sample predictions of the FedSpeak model (defined in Section 3.1). The first two columns calculate the mean squared error of market expectations and the FedSpeak model over the December 1996 through December 2016 period. The third and fourth columns calculate mean squared errors from December 1996 through December 2016, except from November 2000 to July 2003 and July 2007 to December 2008. The fifth and sixth columns calculate mean squared errors from November 2000 to July 2003 and July 2007 to December 2008. We use the Diebold-Mariano test on the null hypothesis that the MSE of market expectations is equal to the MSE of the FedSpeak model. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

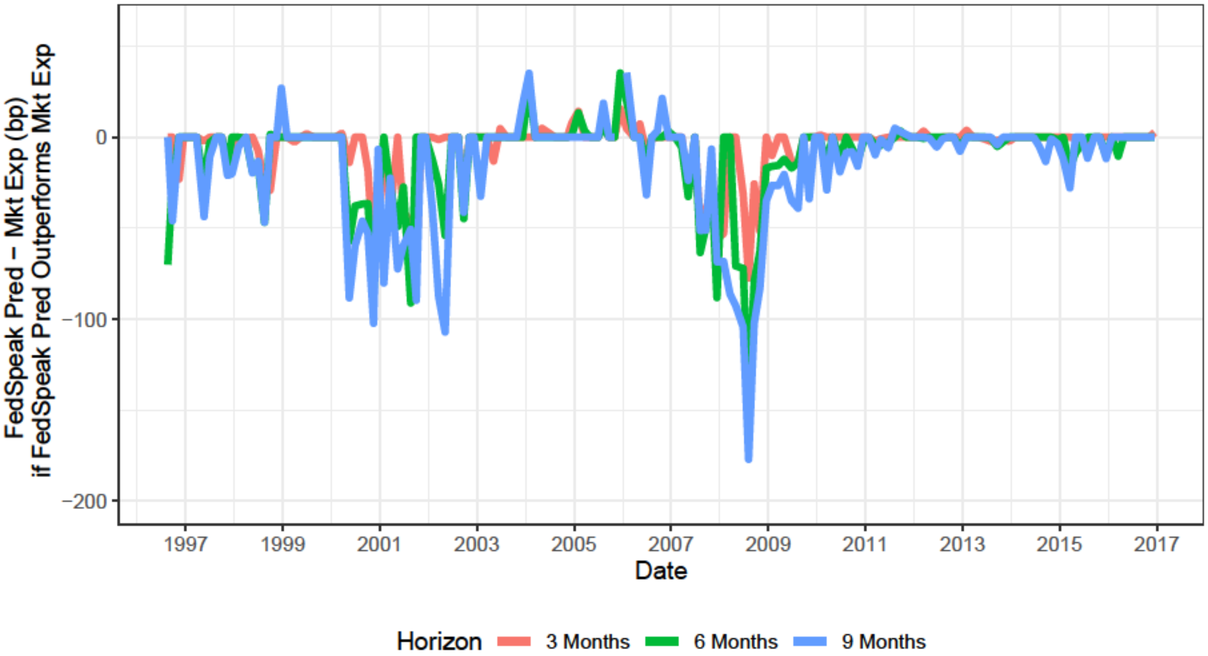
In Figure 3, we plot the out-of-sample forecasts underlying the mean squared error results. In Figure 4, we calculate the gap between market expectations and the FedSpeak model predictions during periods in which the FedSpeak model outperforms market expectations. If market expectations outperform during a given period, then we set the gap to zero. The figure can be interpreted as the improvement in forecasting accuracy that the FedSpeak model achieves relative to market expectations and can therefore be seen as a plot of changes in opacity over time. The figure shows that if markets had real-time access to meeting materials during easing cycles, they could have predicted as much as 125 basis points of additional rate cuts at the six month horizon and as much as 175 basis points of additional rate cuts at the nine month horizon.

Figure 3: Out-of-Sample Predictions



Notes: This figure plots the out-of-sample predictions of the FedSpeak model, fed funds futures-based market policy rate expectations, and actual realizations of the fed funds rate for the three, six, and nine month forecasting horizons, as described in Section 3.1.

Figure 4: Changes in Opacity over Time



Notes: This figure plots the out-of-sample predictions of the FedSpeak model minus fed funds futures-based market expectations for the three, six, and nine month forecasting horizons. If the FedSpeak model does not outperform market expectations for a particular prediction, then we plot a value of zero.

One potential concern about our results is the possibility that speeches and meeting minutes released after meeting t may reveal information about the meeting that were not immediately available to the public one day after the meeting. If this were true, then the opacity we identify during easing cycles may be temporary. Under this hypothesis, once the meeting minutes are released and post-meeting speeches are delivered, the markets would no longer benefit from reading the transcripts and Tealbooks because the communications provide all relevant information from the meeting.

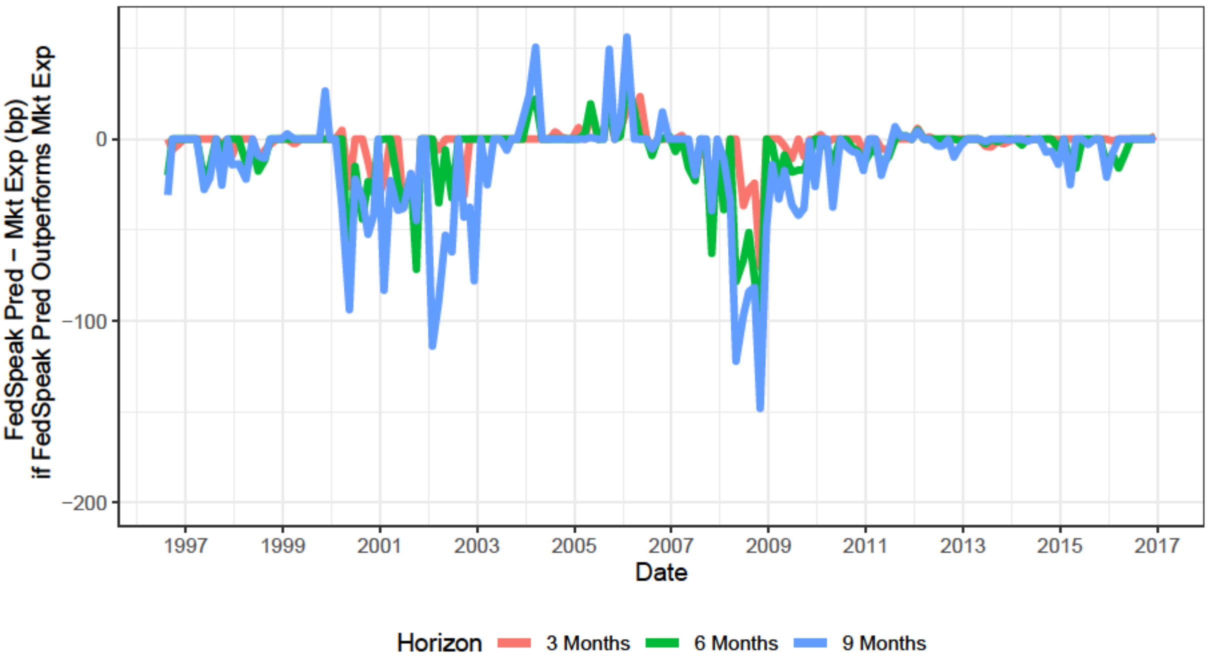
One testable implication of this hypothesis is that one day after meeting t, the market would no longer benefit from having access to the materials from meeting $t-1$. The materials from meeting $t-1$ would no longer be informative because the meeting minutes and speeches released between meeting $t-1$ and meeting t would provide all relevant information from meeting $t-1$. To test this possibility, we modify the FedSpeak model to use transcripts and Tealbooks from meeting $t-1$ rather than meeting t. In other words, we test whether market expectations measured one day after meeting t would benefit from accessing the materials from meeting $t-1$, rather than the materials from meeting t. We find in Figure 5 and Figure 6 that during easing cycles, this modified FedSpeak model still substantially outperforms market expectations, albeit with a lower magnitude. This implies that post-meeting communications like minutes and speeches do not eliminate easing cycle opacity.

Figure 5: Forecasting Errors of Market Expectations and FedSpeak Model with Variables from Meeting $t-1$

	Full Period		Tightening Periods		Easing Periods		Zero Lower Bound	
Horizon (Months)	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE	Market MSE	FedSpeak MSE
Lasso								
3	0.109	0.085	0.035	0.044	0.448	0.292**	0.002	0.013***
6	0.379	0.312*	0.172	0.218	1.446	0.957***	0.004	0.045***
9	0.888	0.808	0.644	0.764	2.848	1.866***	0.020	0.224***
Random Forest								
3	0.109	0.102	0.035	0.059**	0.448	0.36	0.002	0.003**
6	0.379	0.344	0.172	0.252**	1.446	1.102**	0.004	0.005
9	0.888	0.809	0.644	0.806*	2.848	2.111***	0.020	0.023

Notes: This table aggregates fed funds futures market expectation and the out-of-sample predictions of the FedSpeak model (defined in Section 3.1). We modify the FedSpeak model to use meeting materials from meeting $t-1$ rather than from meeting t. The first two columns calculate the mean squared error of market expectations and the FedSpeak model over the December 1996 through December 2016 period. The third and fourth columns calculate mean squared errors from December 1996 through December 2016, except from November 2000 to July 2003 and July 2007 to December 2008. The fifth and sixth columns calculate mean squared errors from November 2000 to July 2003 and July 2007 to December 2008. We use the Diebold-Mariano test on the null hypothesis that the MSE of market expectations is equal to the MSE of the FedSpeak model. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Figure 6: Changes in Opacity over Time with FedSpeak Variables from Meeting $t-1$



Notes: This figure plots the out-of-sample predictions of the FedSpeak model minus fed funds futures-based market expectations for the three, six, and nine month forecasting horizons. We modify the FedSpeak model to use meeting materials from meeting $t-1$ rather than from meeting t. If the FedSpeak model does not outperform market expectations for a particular prediction, then we plot a value of zero.

The asymmetry we observe between tightening and easing cycles is consistent with the literature on short rate expectation errors. Cieslak (2018) and Schmeling, Schrimpf and Steffensen (2022) find large excess returns on Treasuries, fed funds futures, and OIS during easing cycles, but not during tightening cycles. They attribute these excess returns to incorrect expectations from investors rather than time-varying risk premia. Our results highlight that these expectation errors could have been significantly reduced if the FOMC had disclosed more information from its internal meeting deliberations on a timely basis.

Our results cannot draw any conclusions about whether recent innovations in Fed communications, such as the Summary of Economic Projections (SEP) and press conferences, have improved transparency and information transmission. The SEP began in 2007 and the press conferences began in 2011. The time period after these communications were introduced was largely characterized by the zero lower bound (ZLB) and explicit calendar-based forward guidance, both of which severely reduce the scope for market expectation errors. The only observations in our sample period from the post-ZLB period are from 2016 to 2017 because of the lagged release of the meeting transcripts and the Tealbooks. Evaluating the effects of the SEP and press conferences using methods similar to those introduced in this paper will likely be possible only after many more meeting transcripts are released from the post-ZLB period.

Importantly, our analysis does not consider the possibility that if meeting transcripts were immediately available to the public, then FOMC members could respond by altering the content discussed at monetary policy meetings. Hansen, McMahon and Prat (2018) find that after the FOMC adopted a set of transparency-enhancing reforms in 1993, including the lagged release of meeting transcripts to the public, the members altered their behavior. They find a positive "discipline effect", where members became more data-driven and were more likely to be influenced by less experienced members. But they also find a negative "conformity effect", where members became less likely to voice dissenting opinions. Although they argue that the positive discipline effect dominated, other transparency reforms may cause negative behavioral changes that outweigh positive changes. These behavioral changes may push valuable policy-relevant discussion out of formal meetings and into more informal settings. Thus, our results should be viewed as reflective of a hypothetical world in which there were no behavioral responses. Our results should not be viewed as predicting what would have happened in the real world had the FOMC began contemporaneously releasing transcripts at the beginning of our sample period. The possibility that important policy-making discussions occur outside of formal meetings implies that our estimates of opacity may be a lower bound.

3.4 Alternative Measure of Transparency

As a robustness check of our main results, we show that the asymmetry in the informativeness of communications between easing and tightening cycles can be independently found using an alternative method of measuring transparency. We test the extent to which a hypothetical investor could have used meeting minutes in real time to predict the contents of unreleased meeting transcripts. We construct the following out-of-sample forecasting exercise:

1. One day after meeting t, an investor wants to know the sentiment of meeting t (as captured by the sentiment of the associated meeting transcript). She has access to the meeting minutes associated with meeting t. Denote the unknown meeting transcript sentiment as y_t and the

known meeting minutes sentiment as x_t. She wants to use the observation of x_t to predict y_t.

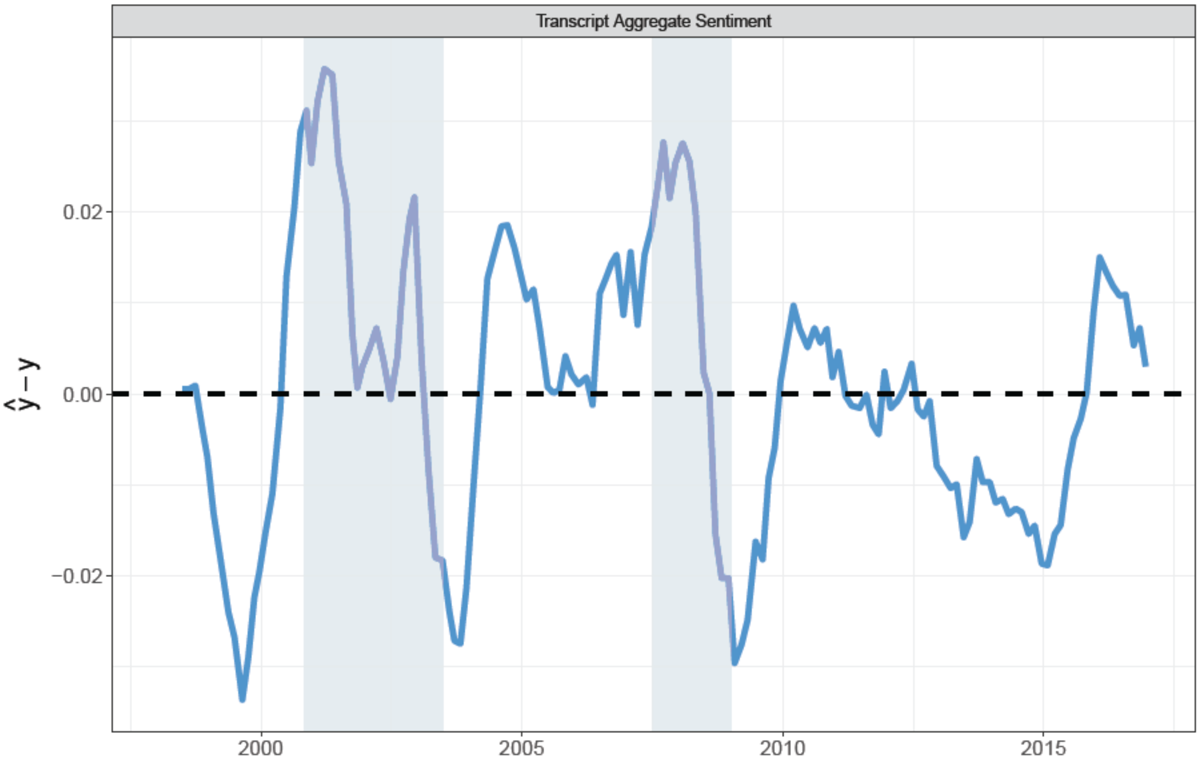
2. Due to the five year lagged release of the meeting transcripts, the investor only has access to meeting transcripts $Y = \{y_1, ..., y_{t-k}\}$, where $t-k$ is the closest meeting from at least five years ago. The corresponding meeting minutes observations are $X = \{x_1, ..., x_{t-k}\}$.

3. The investor regresses Y on X to estimate function g_t, which is then used to predict meeting transcript sentiment for meeting t: $\hat{y}_t = g_t(x_t)$.

We generate these out-of-sample forecasts for regularly scheduled meetings from 1998 to 2016. In the above example, we compare aggregate meeting transcript sentiment to aggregate meeting minutes sentiment. We repeat the exercise for topic-specific transcript sentiment and transcript topic mention frequency. For each variable, we predict the transcript outcome using the corresponding variable derived from the meeting minutes. For example, we predict transcript financial stability topic frequency using minutes financial stability topic frequency.

In Figure 7, we graph the difference between the predicted transcript outcome and the actual transcript outcome $(\hat{y}_t - y_t)$ for aggregate sentiment. We apply an eight meeting rolling average to the plot. We find an asymmetry between easing cycles and tightening cycles. During easing cycles, highlighted in gray, predicted transcript sentiment is farther from the actual transcript sentiment, suggesting that the minutes implicitly convey a more optimistic message during such periods.

Figure 7: Minutes-Implied Transcript Sentiment versus Actual Transcript Sentiment

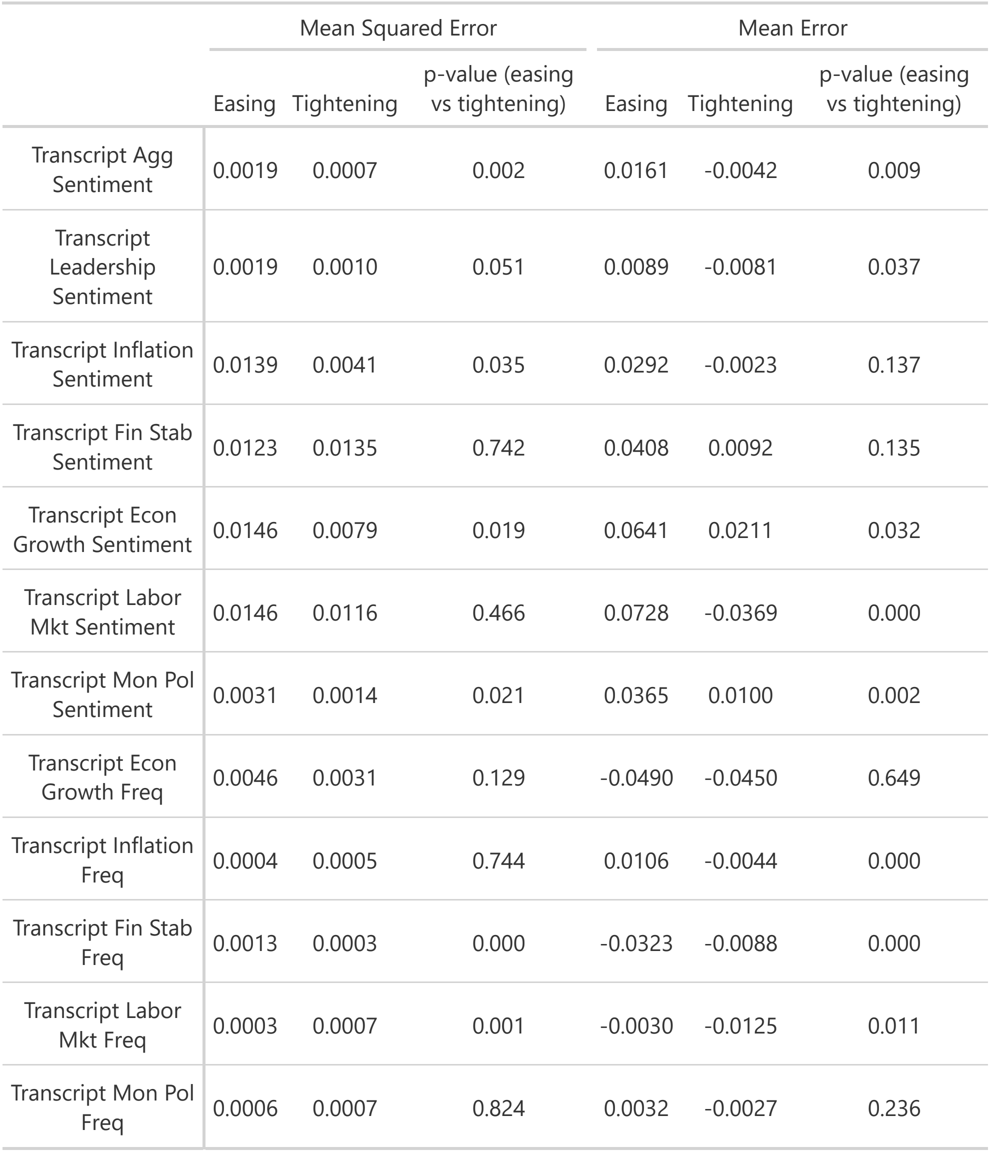


Notes: This figure plots the minutes-implied prediction of aggregate transcript sentiment minus actual transcript sentiment, with an eight meeting rolling average. We highlight two easing periods: November 2000 through July 2003 and July 2007 through December 2008.

In Figure 8, we record the mean squared error and mean error for the predictions of all of the transcript-derived variables we use in the FedSpeak model. For leadership sentiment, we compare transcript leadership sentiment to aggregate minutes sentiment. Differences in mean squared error during easing cycles and tightening cycles tell us whether or not the minutes are less informative in general during easing cycles. Differences in mean error tell us whether minutes tend to be more optimisic or more pessimistic during easing cycles)

The mean squared error and mean error calculations are done within easing periods and tightening periods. We then run hypothesis tests for the null hypothesis that the error is the same during easing periods and tightening periods. For many of the variables, we find that during easing cycles, the forecasting performance of the meeting minutes, as measured by mean squared error, deteriorated by a statistically significant amount. For sentiment variables, we often find that the mean error during easing cycles was significantly higher than during tightening cycles, indicating that the minutes conveyed a more optimistic message during easing periods.

Figure 8: Prediction Aggregation for Important Topics



	Mean Squared Error			Mean Error		
	Easing	Tightening	p-value (easing vs tightening)	Easing	Tightening	p-value (easing vs tightening)
Transcript Agg Sentiment	0.0019	0.0007	0.002	0.0161	-0.0042	0.009
Transcript Leadership Sentiment	0.0019	0.0010	0.051	0.0089	-0.0081	0.037
Transcript Inflation Sentiment	0.0139	0.0041	0.035	0.0292	-0.0023	0.137
Transcript Fin Stab Sentiment	0.0123	0.0135	0.742	0.0408	0.0092	0.135
Transcript Econ Growth Sentiment	0.0146	0.0079	0.019	0.0641	0.0211	0.032
Transcript Labor Mkt Sentiment	0.0146	0.0116	0.466	0.0728	-0.0369	0.000
Transcript Mon Pol Sentiment	0.0031	0.0014	0.021	0.0365	0.0100	0.002
Transcript Econ Growth Freq	0.0046	0.0031	0.129	-0.0490	-0.0450	0.649
Transcript Inflation Freq	0.0004	0.0005	0.744	0.0106	-0.0044	0.000
Transcript Fin Stab Freq	0.0013	0.0003	0.000	-0.0323	-0.0088	0.000
Transcript Labor Mkt Freq	0.0003	0.0007	0.001	-0.0030	-0.0125	0.011
Transcript Mon Pol Freq	0.0006	0.0007	0.824	0.0032	-0.0027	0.236

Notes: This table shows the minutes-implied prediction of various text-based outcomes from meeting transcripts minus the actual meeting transcript outcomes during easing cycles ("Easing") versus all other periods ("Tightening"). We define the following periods as easing cycles: November 2000 to July 2003, and July 2007 to December 2008. The procedure for generating minutes-implied predictions is provided in Section 5. The "Unpaired Diff in Means" column finds the difference between the first column and the second column. The "p-val" column displays the p-value for the hypothesis test with null hypothesis that easing and tightening cycles have the same gap between predicted transcript outcomes and actual transcript outcomes.

4 Characterizing the Information Value of Meeting Materials

In this section, we identify the specific pieces of information contained in meeting transcripts and Tealbooks during easing cycles that was not priced into the futures market. We conduct this analysis through the lens of a simple monetary policy rule, where the future fed funds rate is a function of the FOMC's economic outlook and reaction function. In the following decomposition of the future fed funds rate, i_{t+h} is the policy rate h months after meeting t, X_t is the FOMC's economic outlook, and ε_{t+h} is a monetary policy shock relative to the information available at meeting t. f_t is a time-varying reaction function that maps the committee's economic outlook to a future policy rate.

$$i_{t+h} = f_t(X_t) + \varepsilon_{t+h}$$

The economic outlook X_t captures the committee's assessment of the current state of the economy and the committee's forecasts of future economic developments. The reaction function f_t captures the degree of sensitivity to those developments. The reaction function reflects the fact that two central banks with the same macroeconomic data and the same forecasts might set policy differently based on their preferences and risk sensitivities. These preferences include the committee's inflation target, their view on potential trade-offs between output and inflation, and the weight they place on financial stability concerns.

Suppose that at meeting t, the markets are predicting 50 basis points of rate cuts over the next 3 months. The FedSpeak model predicts 100 basis points of rate cuts. By meeting $t+3$, the economy has shed 500,000 jobs and the fed funds rate has been cut by 100 basis points. What information did the FedSpeak model have that the markets did not have?

One possibility is that the Fed, at meeting t, accurately forecasted the loss of 500,000 jobs within three months. The 100 basis point cut was a mechanical response of monetary policy to economic deterioration. The market only predicted a 50 basis point cut because the market only predicted a loss of 100,000 jobs. The Fed's pessimistic forecast is reflected in their meeting transcripts and Tealbooks. The FedSpeak model incorporated this information and predicted 100 basis points of cuts. A key part of this story is that the market knew the Fed's sensitivity to employment news. Thus, if the market had also predicted 500,000 jobs lost, then the market would have also predicted

100 basis points of rate cuts. The view that the Fed has an "information advantage" over the public about the state of the economy has been argued in the literature, such as Nakamura and Steinsson (2018), Campbell, Evans and Justiniano (2012), and Romer and Romer (2000).

Another possibility is that the Fed, at meeting t, shared the market's forecast of 100,000 jobs lost on the basis of the same publicly available economic data. But the Fed judges inflation to be fairly subdued and thinks that the risks to the employment mandate outweigh the risks to the inflation mandate. The Fed decides to react more aggressively to news of 100,000 jobs lost than they may have in previous cycles. The Fed's newfound sensitivity to employment news is reflected in their meeting transcripts. The FedSpeak model incorporated this information and thus predicted steep rate cuts. Even with the exact same forecast of future job loss as the Fed, the market could not have predicted the extent of future rate cuts without knowing the adjustments to the Fed's reaction function. The view that the public has incomplete knowledge about the Fed's reaction function has been argued in Bauer and Swanson (2023), Bauer, Pflueger and Sunderam (2022), Schmeling, Schrimpf and Steffensen (2022), and others. We present evidence that the FedSpeak model's performance is more consistent with a reaction function-based explanation than an outlook-based explanation.

4.1 Variable Importance

To calculate the relative importance of each piece of information incorporated into the FedSpeak model, we use variable importance metrics to decompose each of the out-of-sample predictions into additive components. Since lasso regression and random forests performed very similarly during easing cycles, according to Figure 2, we focus on the lasso-based results to allow for simpler and more interpretable variable importance metrics. We use the following expression, often referred to as "Linear SHAP" values in the machine learning literature (Lundberg and Lee, 2017)[9]:

$$\phi_{it} = \beta_{it}(x_{it} - E[x_{it}])$$

For horizon h at meeting t, we estimate the FedSpeak model on meetings 1 through $t-h$. β_{it} is the estimated coefficient for variable i at meeting t. x_{it} is the value of variable i at time t. $E[x_{it}]$ is the average of variable i over meetings 1 through $t-h$.

[9] SHAP values are often used as a model-agnostic method to explain a machine learning model's prediction. The expression for SHAP values that we present are specific to linear models.

Intuitively, SHAP values compare the contribution of a variable toward a prediction relative to what the model would have predicted if the variable was at its average historical value. A SHAP value of -50 basis points implies that a variable pushed the final prediction in the direction of policy easing, with a magnitude of 50 basis points.

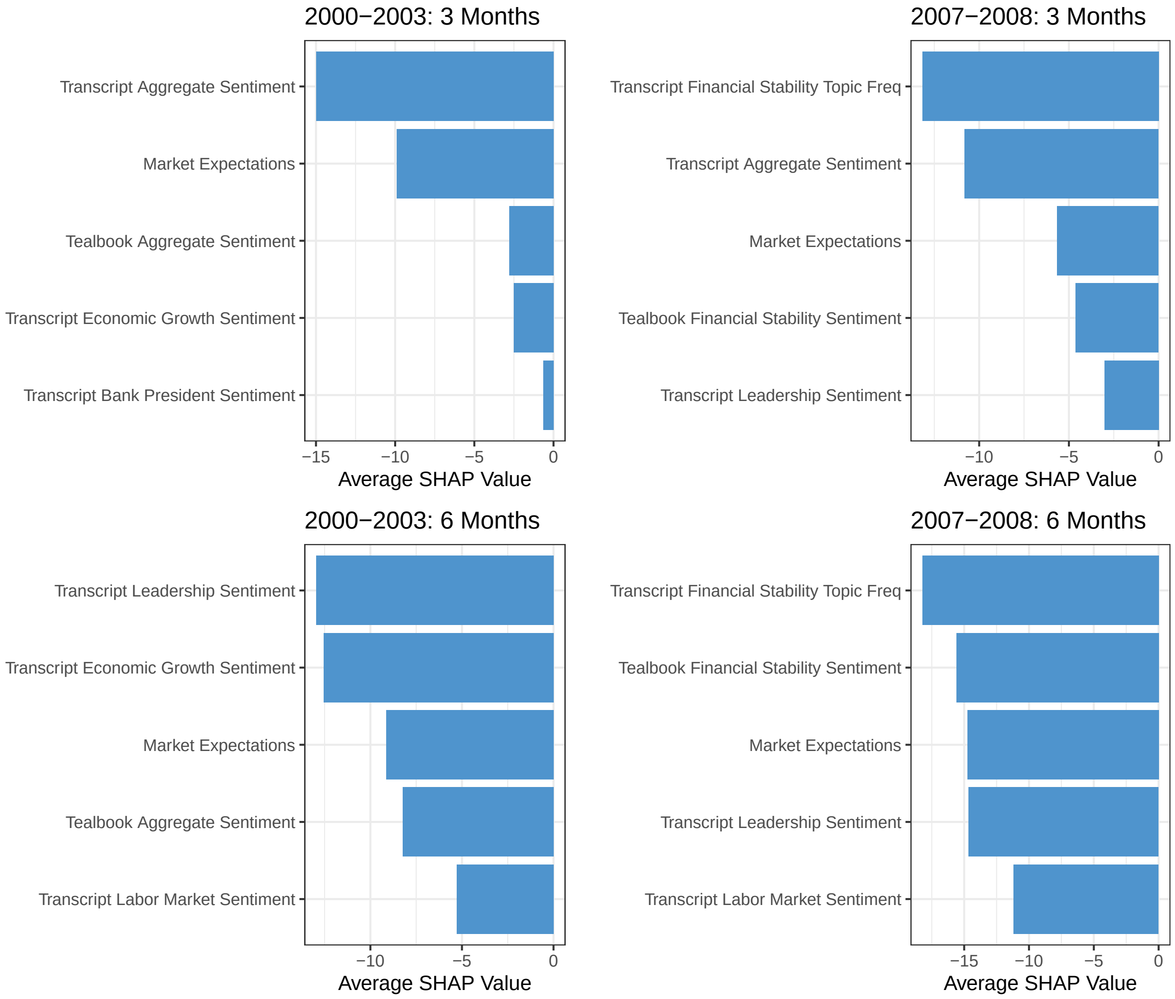
SHAP values are additive in the sense that the sum of the SHAP values for all of the variables equals the overall prediction minus the model's average prediction over the training data. For a fixed t, we can decompose the sum of the SHAP values as follows:

$$\begin{aligned}\sum_i \phi_i &= \sum_i (\beta_i x_i - \mathbb{E}[\beta_i x_i]) \\ &= \sum_i (\beta_i x_i) - \sum_i \mathbb{E}[\beta_i x_i]) \\ &= \beta_0 + \sum_i (\beta_i x_i) - \beta_0 - \sum_i \mathbb{E}[\beta_i x_i]) \\ &= \hat{f}(x) - \mathbb{E}[\hat{f}(x)]\end{aligned}$$

In Figure 9, we average the SHAP values for each variable over each period and forecasting horizon. We then rank the variables by average importance as measured by the absolute value of the average SHAP values and then select the top five. We find that in the 2000-2003 period, the most important variables were aggregate sentiment, economic growth sentiment, and the sentiment of the FOMC's leadership. In the 2007-2008 period, the most important variables were financial stability topic frequency, financial stability sentiment, economic growth sentiment, and FOMC leadership sentiment.

Interestingly, Tealbook forecasts of the future fed funds rate rank relatively low in the variable importance measures. This is consistent with Cieslak (2018), who finds that policymakers and central bank staff appear to display similar errors in their forecasts on short rates as public forecasters. Two potential explanations arise. First, staff members may not have sufficient access to FOMC meeting materials to be able to make the observations of policymaker forecasts and preferences that the FedSpeak model is able to observe. Second, Figure 9 suggests that the FedSpeak model is fairly complex: a wide range of text-derived variables are important for predicting policy easing. Replicating the FedSpeak model's predictions may therefore be difficult for human analysts, particularly without access to the computational text analysis tools used in this paper.

Figure 9: SHAP Values

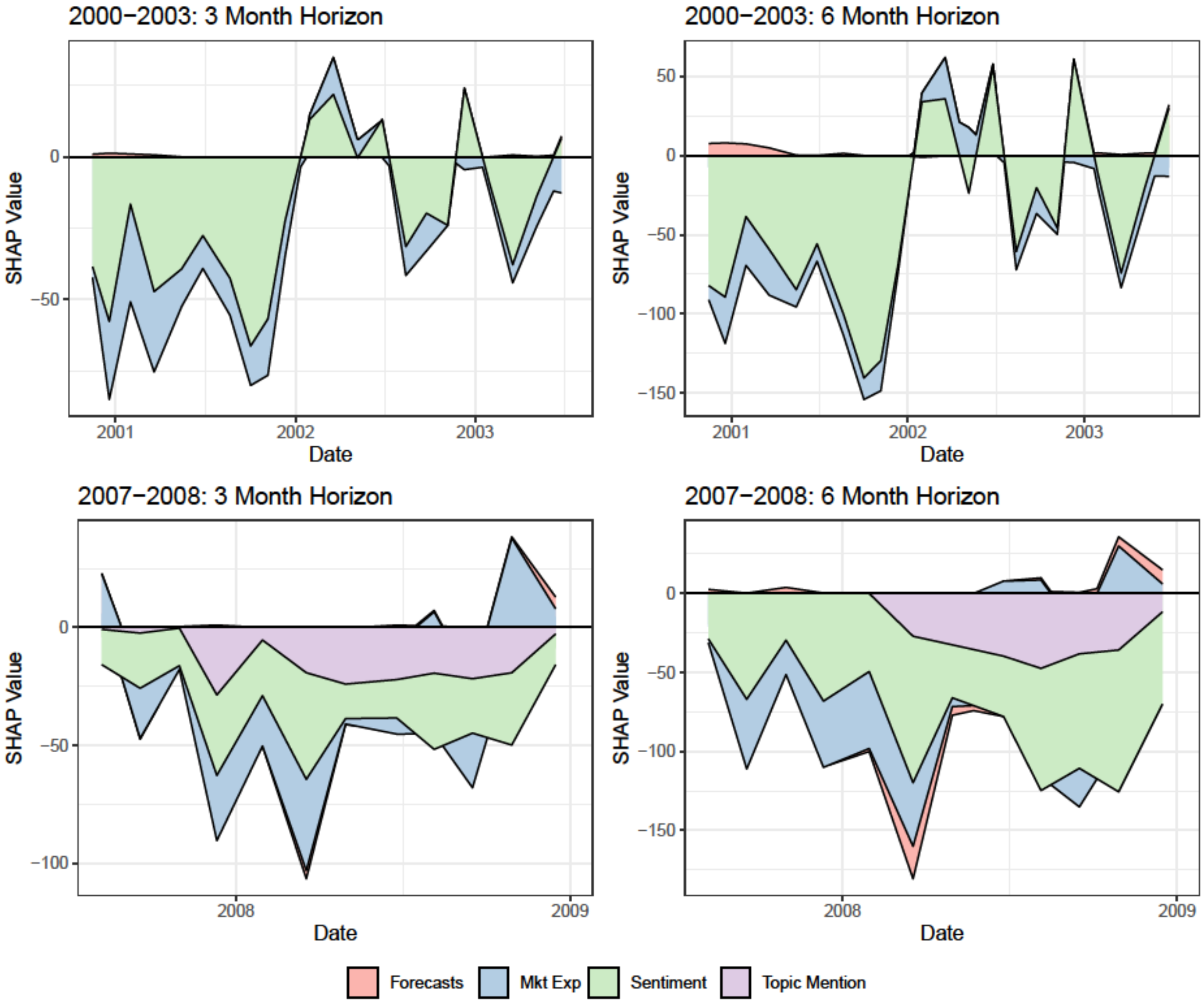


Notes: In this figure, we average SHAP values for each variable within each period and forecasting horizon. We then rank the variables by absolute average SHAP value and select the top five.

4.2 Testing for Outlook-Related Information

If incomplete information about the FOMC's economic outlook is responsible for easing cycle opacity, then we might expect Tealbook forecast errors to rank highly in the variable importance analysis. A large literature has assumed that the Tealbook forecast errors are a good proxy for the forecasts of the FOMC members. In Figure 10, we group the variables by type of information (forecasts, sentiment, topic mention, market expectations) and plot the sum of their SHAP values over time. We find that Tealbook forecasts have an extremely low importance and are thus unable to explain why the market underestimated policy easing.

Figure 10: SHAP Values for 6 Month Horizon Predictions, Grouped by Variable Type



Notes: In this figure, we plot the SHAP values for the FedSpeak model variables over time, as described in Section 4.1. We use the three month and six month forecasting horizons. We group variables by the type of information they provide (Tealbook forecasts, sentiment, topic mention frequency, and market expectations) and sum the SHAP values within each group. In the first panel, we plot SHAP values for the November 2000 through July 2003 period. In the second panel, we plot SHAP values for the July 2007 through December 2008 period.

The observation that Tealbook forecasts are unimportant does not rule out outlook-related explanations of opacity. The Tealbook forecast variables may be unimportant because they are highly correlated with sentiment variables, which are in turn highly correlated with the FOMC's outlook. To test this hypothesis, we re-implement the FedSpeak model under a hypothetical scenario where investors had access to macroeconomic data releases associated with the months *after* meeting t.

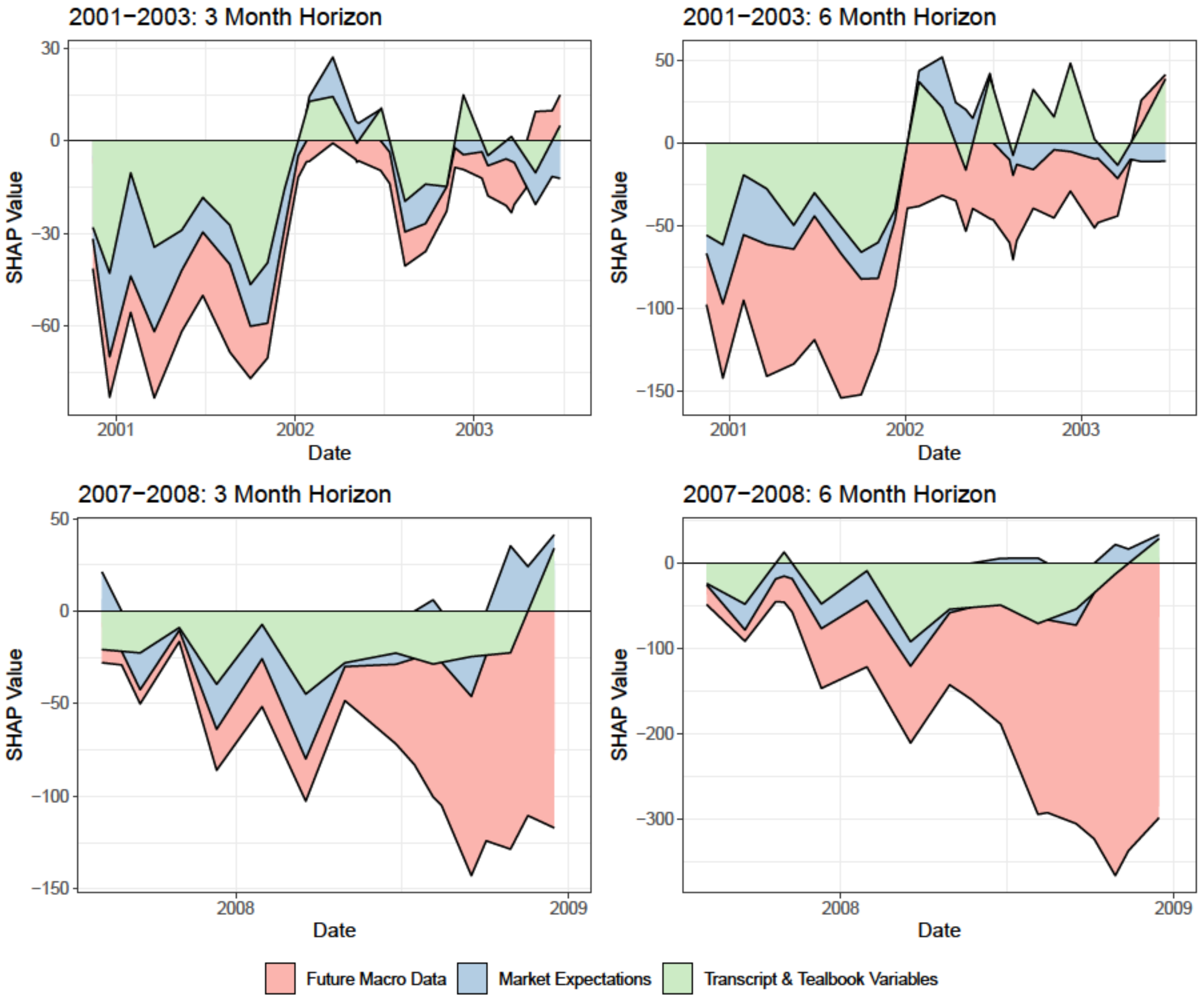
This is equivalent to a scenario where the Fed has perfect foresight of future economic conditions. We modify the FedSpeak model by adding macroeconomic data releases from one month and two months into the future to the set of text-based predictors. We provide the equation for the modified FedSpeak model below. $\zeta_1 Macro_{t+1}$ refers to macro data releases associated with the month after meeting t. $\zeta_2 Macro_{t+2}$ refers to macro data releases two months after meeting t.

In our set of macro data releases, we consider change in nonfarm payrolls and consumer price index. Following Bauer and Swanson (2023), we also use the Brave-Butters-Kelley index from Brave, Butters and Kelley (2019), which aggregates hundreds of macroeconomic data releases to arrive at a single economic activity index.

$$AvgEFFR_{t,h} - FFR_t = \alpha + \beta(MktExp_{t,h} - FFR_t) + Text_t\gamma + \zeta_1 Macro_{t+1} + \zeta_2 Macro_{t+2} + \varepsilon_t \tag{FedSpeak Model with Future Macro Data}$$

Importantly, given the long lags with which monetary policy operates, real macroeconomic outcomes over a short horizon are plausibly exogenous with respect to the FOMC's policy decision at meeting t. Note that due to the lagged release of data, the macro data associated with the month after meeting t would only be released approximately two months after meeting t.

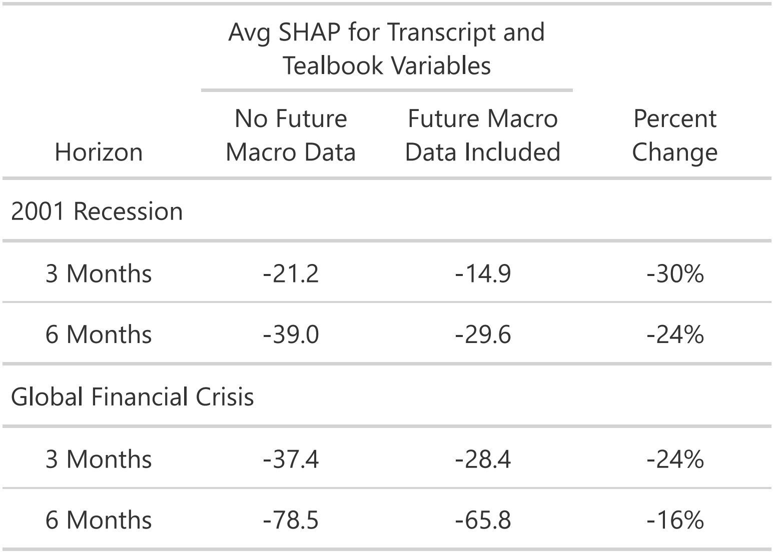
Figure 11: SHAP Values for 6 Month Horizon Predictions, Including Future Macro Data



Notes: In this figure, we plot the SHAP values for the FedSpeak model variables over time, as described in Section 4.1. We use the three month and six month forecasting horizon. We group all of the variables derived from meeting materials into a category called "Transcript & Tealbook Variables". We also plot the SHAP values for market expectations and future macro data. In the first panel, we plot SHAP values for the November 2000 through July 2003 period. In the second panel, we plot SHAP values for the July 2007 through December 2008 period.

Even with these extreme assumptions about the FOMC's knowledge of the state of the economy, the text-based variables retain most of their predictive power. In Figure 11, we group the variables and plot their SHAP values over time. In Figure 12, we provide the SHAP values of text-derived variables before and after controlling for future macro data. We find that the text-based variables lose at most 29% of their predictive power when controlling for future macro data. We interpret the remaining predictive power as reflecting information about the Fed's reaction to economic developments rather than superior knowledge of the economy itself. Given the unrealistic assumptions about the Fed's knowledge of the economy that this exercise uses, it would be reasonable to conclude that more than 71% of the FedSpeak model's outperformance is due to reaction function-related information.

Figure 12: Average SHAP Values of Transcript and Tealbook variables, Controlling for Future Macro Data



	Avg SHAP for Transcript and Tealbook Variables		
Horizon	No Future Macro Data	Future Macro Data Included	Percent Change
2001 Recession			
3 Months	-21.2	-14.9	-30%
6 Months	-39.0	-29.6	-24%
Global Financial Crisis			
3 Months	-37.4	-28.4	-24%
6 Months	-78.5	-65.8	-16%

Notes: In this table, we calculate the average SHAP values of transcript and Tealbook variables within each period and forecasting horizon. In the second column, we report the values when we do not control for future macro data. The third column adds future macro data as controls. The fourth column finds the percent change between the second and third column.

The importance of sentiment and topic frequency variables in Figure 10 also points toward an affirmative case for a reaction function-based interpretation of opacity. Shapiro and Wilson (2019) show that the sentiment expressed by the FOMC within meeting transcripts is a good proxy for the FOMC's loss in the context of a loss function. Using narrative evidence from the meeting transcripts, they extensively validate the connection between sentiment and the committee's loss function.[10]

Suppose we estimate a very simple version of the FedSpeak model:

$$AvgEFFR_{t+h} - EFFR_t = \alpha + \beta(MktExp_{t+h} - EFFR_t) + \gamma Sentiment_t + \delta Forecast_t + \varepsilon_t$$

[10] Shapiro and Wilson (2019) use sentiment to estimate text-implied inflation targets for individual members and find that these estimated inflation targets largely match explicitly stated inflation targets.

Based on Shapiro and Wilson (2019), we assume that

$$Sentiment_t = \zeta_0 + \zeta_1 Forecast_t$$

Theoretically, $\hat{\delta}$ should be a good estimate of the FOMC's reaction function because it directly maps the committee's forecast to changes in policy. But if the reaction function is time-varying, then $\hat{\delta}$ may no longer be accurate and may thus produce inaccurate predictions of future policy. Sentiment is valuable because it allows us to bypass the problem of estimating a time-varying δ using historical data. Instead, we can directly observe members' reactions to economic news by observing $Sentiment_t$. This leaves us only with the task of estimating γ, the function that maps the members' reactions to future policy rates. If γ does not vary a lot over time, then $\hat{\gamma} Sentiment_t$ will be more predictive of future policy than $\hat{\delta} Forecast_t$.

We also argue that the specific timing of opacity within the easing cycles is consistent with a reaction function-based interpretation. In Figure 13, we group variables by the topic they are associated with and then sum the SHAP values within each group. For example, in the Economic Growth category, we group Tealbook GDP forecast, transcript Economic Growth sentiment, Tealbook Economic Growth sentiment, and transcript Economic Growth topic frequency. We also create an "Aggregate" category for aggregate transcript and tealbook sentiment and we create a member-specific category for sentiment related to leadership, the chair, the vice chair, reserve bank presidents, and governors.

In 2007-2008, the text-derived variables became especially important in early 2008, around the collapse of Bear Stearns. This is especially true of the financial stability topic, which had low importance throughout 2007. The financial stability topic became especially important around the Lehman Brothers collapse in September 2008. This is more consistent with a narrative where the Fed adjusted policy faster than the markets expected in response to financial sector turmoil compared to a narrative where the Fed anticipated financial sector turmoil ahead of the public. In 2000-2003, the text-derived variables, particularly aggregate sentiment and leadership sentiment, became especially important around the September 11th attacks. This is more consistent with the Fed reacting more strongly to news of the attacks than the markets expected, rather than the Fed having superior ability to forecast the economic damage that the attacks would create.

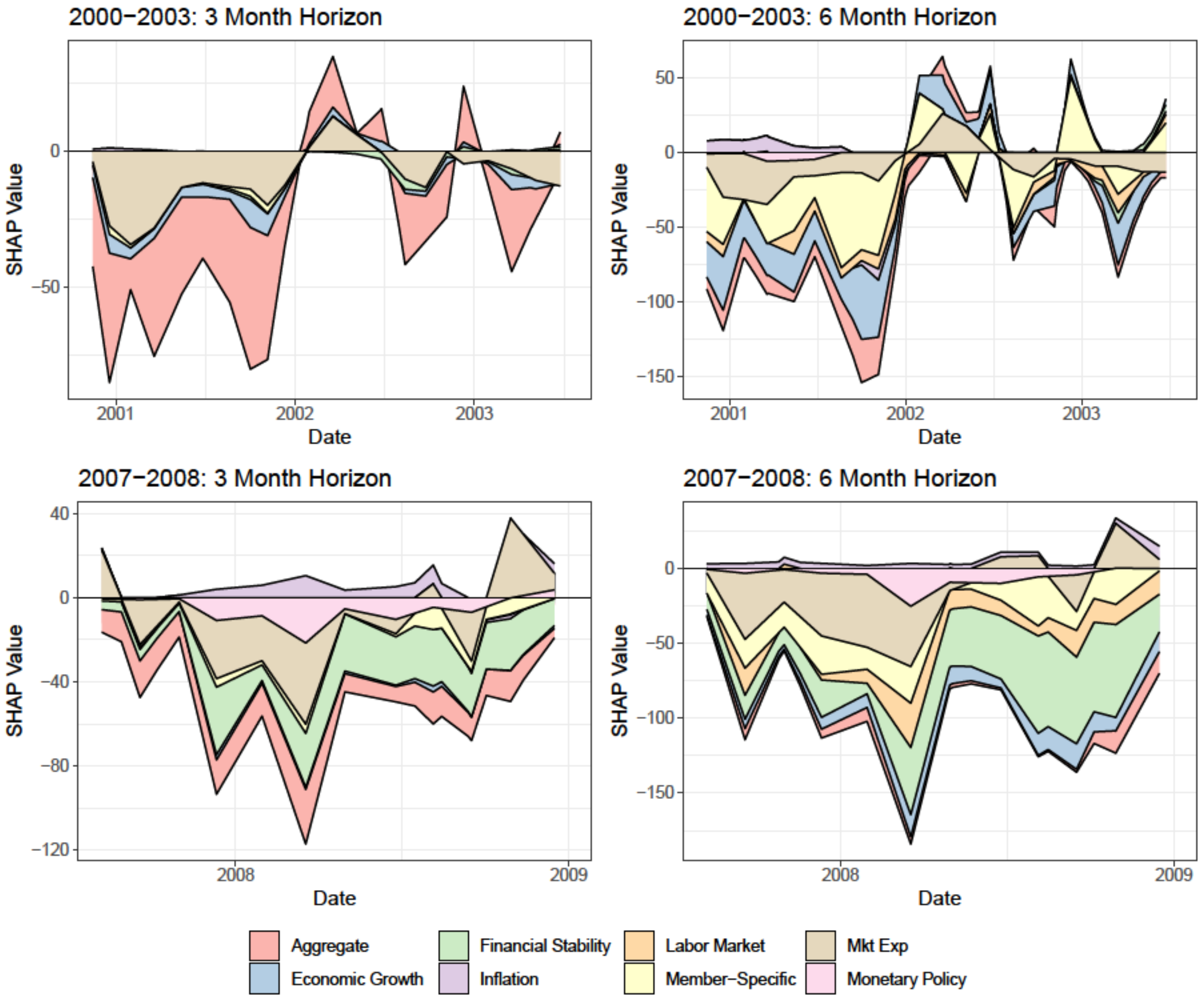
The extent of the committee's rate cuts during these cycles may have been surprising to market participants who had a dual mandate-centric view of the reaction function. For much of 2001, unemployment[11] was below the Congressional Budget Office's estimate of the natural rate of unemployment.[12] Realized Core PCE hovered very close to the assumed 2% target before the 9/11 attacks.[13] In 2007 and 2008, realized unemployment only started to significantly increase in mid-2008 and realized Core PCE was above 2% until after the Lehman collapse. But according to Figure 13, FedSpeak variables related to labor markets and inflation were of very low importance throughout both periods.

[11] Retrieved from the FRED UNRATE series.

[12] Retrieved from the FRED NROU series.

[13] Retrieved from the FRED PCEPILFE series.

Figure 13: SHAP Values, Grouped by Topic



Notes: In this figure, we plot the SHAP values for the FedSpeak model variables over time, as described in Section 4.1. We use the three month and six month forecasting horizons. We group the variables by topic and sum the SHAP values within each group. In the first panel, we plot SHAP values for the November 2000 through July 2003 period. In the second panel, we plot SHAP values for the July 2007 through December 2008 period.

5 Conclusion

In this paper, we use text analysis tools to show that during easing cycles, market participants could have significantly improved their forecasting of future monetary policy if they had real-time access to meeting transcripts and Tealbooks. Markets could have improved their policy forecasting during easing cycles by as much as 150 basis points at the six month horizon and 300 basis points at the nine month horizon. The meeting materials contained important information about policymakers' reactions to incoming economic data, particularly their reactions to financial stability concerns during the 2008 financial crisis. The markets may have misinterpreted policymakers' reactions due to the optimism of meeting minutes during easing cycles relative to tightening cycles.

These results have important implications for both policy and future academic research. First, there has been an ongoing effort across central banks to enhance transparency over the last few decades. Increased transparency may not always be beneficial. There are trade-offs of increased transparency that are worth considering, and central banks need to strike a balance between transparency and legitimate needs for confidentiality. This paper allows for a fuller discussion of those trade-offs by highlighting the conditions under which the Fed has historically been most opaque. Our methodology can also, in principle, be applied to any central bank that releases information to the public in a lagged manner.

Second, we show that during easing cycles, FOMC communication are less informative about policymakers' sensitivity to incoming economic developments. Future work should seek to clarify the incentives and strategic reasons that explain why policymakers are more restrained in their external communications during these periods. Easing cycles are typically associated with growth downturns, which may make policymakers more sensitive to downside risks than upside risks and more sensitive to the possibility of spooking the markets. Easing cycles are also times of great uncertainty about the economic outlook, which may further incentivize policymakers to be restrained in their communications.

References

Acosta, Miguel. 2015. "FOMC Responses to Calls for Transparency." *Finance and Economics Discussion Series* 2015(060):1–44.

Araci, Dogu. 2019. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.". arXiv:1908.10063 [cs].

Bauer, Michael D, Carolin E Pflueger and Adi Sunderam. 2022. "Perceptions about Monetary Policy." *Working Paper* .

Bauer, Michael D. and Eric T. Swanson. 2023. "An Alternative Explanation for the 'Fed Information Effect'." *American Economic Review* 113(3):664–700.

Brave, Scott R., Andrew Butters and David Kelley. 2019. "A New 'Big Data' Index of U.S. Economic Activity." *Federal Reserve Bank of Chicago Economic Perspectives* .

Campbell, Jeffrey R., Fisher Jonas D.M. Evans, Charles L. and Alejandro Justiniano. 2012. "Macroeconomic Effects of Federal Reserve Forward Guidance." *Brookings Papers on Economic Activity* .

Chernulich, Aleksei, Mengheng Li and Eamon McGinn. forthcoming. "Does the Fed say it all? Textual analysis of public communications and private discussions." *Working Paper* .

Cieslak, Anna. 2018. "Short-Rate Expectations and Unexpected Returns in Treasury Bonds." *The Review of Financial Studies* .

Correa, Ricardo, Keshav Garud, Juan M Londono and Nathan Mislang. 2021. "Sentiment in Central Banks' Financial Stability Reports*." *Review of Finance* 25(1):85–120.

Crump, Richard K, Stefano Eusepi and Emanuel Moench. 2018. "The Term Structure of Expectations and Bond Yields." *Federal Reserve Bank of New York Staff Reports* .

Cukierman, Alex. 2009. "The Limits of Transparency." *Review of Banking, Finance and Monetary Economics* .

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.". arXiv:1810.04805 [cs].

Dincer, N. Nergiz and Barry Eichengreen. 2018. "Central Bank Transparency and Independence: Updates and New Measures." *34th issue (March 2014) of the International Journal of Central Banking* .

Eijffinger, Sylvester C. W. and Petra M. Geraats. 2006. "How transparent are central banks?" *European Journal of Political Economy* 22(1):1–21.

Gardner, Ben, Chiara Scotti and Clara Vega. 2022. "Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements." *Journal of Econometrics* 231(2):387–409.

Hansen, Anne Lundgaard and Sophia Kazinnik. 2023. "Can ChatGPT Deciper FedSpeak?" *Working Paper* .

Hansen, Stephen and Michael McMahon. 2016. "Shocking language: Understanding the macroeconomic effects of central bank communication." *Journal of International Economics* 99:S114–S133.

Hansen, Stephen, Michael McMahon and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133(2):801–870.

Hoesch, Lukas, Barbara Rossi and Tatevik Sekhposyan. forthcoming. "Has the Information Channel of Monetary Policy Disappeared? Revisiting the Empirical Evidence." *American Economic Journal: Macroeconomics* .

Huang, Allen H., Hui Wang and Yi Yang. 2023. "FinBERT: A Large Language Model for Extracting Information from Financial Text." *Contemporary Accounting Research* 40(2):806–841.

Loughran, Tim and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66(1):35–65.

Lundberg, Scott and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions.".

Morris, Stephen and Hyun Song Shin. 2005. "Central Bank Transparency and the Signal Value of Prices." *Brookings Papers on Economic Activity* 2005(2):1–66.

Nakamura, Emi and Jón Steinsson. 2018. "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect*." *The Quarterly Journal of Economics* 133(3):1283–1330.

Piazzesi, Monika and Eric T. Swanson. 2008. "Futures prices as risk-adjusted forecasts of monetary policy." *Journal of Monetary Economics* 55(4):677–691.

Picault, Matthieu and Thomas Renault. 2017. "Words are Not All Created Equal: A New Measure of ECB Communication." (ID 2980777).

Romer, Christina D. and David H. Romer. 2000. "Federal Reserve Information and the Behavior of Interest Rates." *American Economic Review* 90(3).

Schmanski, Bennett, Chiara Scotti, Clara Vega and Hedi Benamar. 2023. "Fed Communication, News, Twitter, and Echo Chambers." *Working Paper* .

Schmeling, Maik, Andreas Schrimpf and Sigurd A. M. Steffensen. 2022. "Monetary policy expectation errors." *Journal of Financial Economics* 146(3):841–858.

Shapiro, Adam H. and Daniel J. Wilson. 2019. "Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis." *Federal Reserve Bank of San Francisco, Working Paper Series* pp. 01–74.

Shapiro, Adam Hale, Moritz Sudhof and Daniel J. Wilson. 2020. "Measuring news sentiment." *Journal of Econometrics* .

Swanson, Eric T. 2006. "Have Increases in Federal Reserve Transparency Improved Private Sector Interest Rate Forecasts?" *Journal of Money, Credit and Banking* 38(3).

Swanson, Eric T. 2023. "The Importance of Fed Chair Speeches as a Monetary Policy Tool." *AEA Papers and Proceedings* 113:394–400.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. "Attention Is All You Need.". arXiv:1706.03762 [cs].

Woodford, Michael. 2005. "Central Bank Communication and Policy Effectiveness." *Economic Policy Symposium Proceedings. Jackson Hole: Federal Reserve Bank of Kansas City* .

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg and Gideon Mann. 2023. "BloombergGPT: A Large Language Model for Finance." *arXiv preprint arXiv:2303.17564* .

Yang, Honyang, Xiao-Yang Liu and Christina Dan Wang. 2023. "FinGPT: Open-Source Financial Large Language Models." *arXiv preprint arXiv:2306.0603* .

Zhang, Boyu, Hongyan Yang and Xiao-Yang Liu. 2023. "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models." *arXiv preprint arXiv:2306.12659* .

Appendix

A Human Validation and Data Collection

We randomly generated a dataset of sentences from policymaker speeches, transcripts, minutes, press conferences, and interviews. We obtained a stratified sample of sentences over time and across communication type, aiming to maintain a consistent proportion of sentences from a given year and communication type, notwithstanding large imbalances in document length. The dataset spanned the time period 2000 to 2022 and contained 2,500 sentences in total. Once this dataset was established, we set aside approximately 450 sentence for validation. The remaining 1,600 sentences were used for model development.

We then divided the dataset into five subsets of 500 sentences each and assigned each subset to a different set of staff to audit. Each team consisted of three members who had relevant domain knowledge and experience. The teams were given instructions on the proper interpretation of the topic and sentiment labels prior to the start of the exercise, and were requested to label sentences as follows:

1. Assign labels to each sentence according to our pre-defined topics: {Economic Growth, Inflation, Financial Stability, Labor Market, Monetary Policy, No Topic}.
2. Return to the same sentences to assign sentiment: {Positive, Negative, Neutral, None}.

We aggregated the annotations from each team and chose the consensus label for each sentence. The consensus label was defined as the label that received at least two votes from the three analysts. If no sentence received a majority vote for at least one label, we discarded the label from the dataset. We obtained a higher quality dataset with this additional filtering, as only around 200 sentences were discarded (8%).

B Machine Learning Models

This work leveraged techniques from supervised learning to train and fine-tune models using the training set discussed above. We utilized two specifications of machine learning models for topic and sentiment prediction, opting to use our FinBERT model in our sentiment results and Ensemble in the topic classification results.

B.0.1 Bag-of-Words and Neural Network Ensemble

The first model followed the bag-of-words model approach familiar to the literature, but with an added small neural network component. The implementation for this text classifier utilized the spaCy natural language processing Python library. The purpose of the added neural network (with attention) is to supply information about the whole sentence to the model during inference, in contrast to the linear component. It is not utilized in our results of this work because it offered too little performance gain over a lexical approach for the loss of interpretability. With a larger training set, the results may improve.

This bag-of-word component has similarities to dictionaries: both operate using a lexicon of terms. However, maintaining dictionaries by-hand is time-intensive, motivating a data-dependent approach that potentially reduces over-classification. Work by Picault and Renault (2017)show how corpus statistics can be used in the lexicon creation process in this domain. Informally, the linear component of this method applies Bayes Rule to the input sentence and class probability:

$$P(c|s) = \frac{P(s|c)P(c)}{P(s)}$$

where c is a class instance like "Economic Growth", s is a sentence instance. $P(s|c)$ is thereby decomposed, forming a "bag" of independent random variables representing terms in the given sentence. The chosen class is given by the decision rule:

$$c_{bayes} = argmax_{c \in C} P(c_i) \prod_{t \in T} P(t|c)$$

where T is the set of one or more words present in the sentence, referred to as uni-grams, bi-grams, up to n-grams. N-grams are determined using corpus term-frequency statistics. Order beyond this term boundary is not modelled, granting it the "bag" term.

B.0.2 Transformer Model

The second model leveraged recent advancements in the field of machine learning via the Transformer class of models. A Transformer model is a model architecture characterized by its series of ”self-attention” neural network layers (Vaswani et al. (2017)). The BERT model (Devlin et al. (2019)) popularized the use of this architecture, and while there are now successors to this approach, it is at this time widely used in Natural Language Processing literature and other applications.

Practically speaking, a Transformer-based text classifier is developed with two distinct training sets. First, a large (often web-scale) set of text data is used in what is referred to as the ”pretraining” stage of development. Language representations are learned in the target languages using either a cloze (masked language model) or autogregressive (next token) objective. This computationally intensive stage requires a representative set of text for both natural language and, especially in our case, the domain of interest. Because of the large cost in developing these models, there are various models made available open-source to the research community, which we leveraged in this work.

The second, ”fine-tuning” stage brings the model's representations semantically closer to a specific domain and/or performing a discrete task. For our classification task, this means supervised learning on a set of annotated sentences. Since we specified sentiment in the broader finance and economics domain, we leveraged existing work to train domain-specific BERT models. FinBERT (Araci (2019) and Huang, Wang and Yang (2023)) is a BERT model pre-trained on financial communication text. It is trained on a broad, 4.9 billion token financial communications dataset, including SEC EDGAR filings and analyst reports. Huang, Wang and Yang (2023) develop and publish models from both stages, providing a ”finbert-tone” sentiment classifier built on a Transformer backbone.

We evaluated the authors' sentiment classification model ”off-the-shelf” and with additional fine-tuning using our internally-sourced training and validation set. The fine-tuning stage is considerably cheaper computationally, passing over roughly 1,600 sentences with a cross-entropy loss minimization objective. This provided a 3-point improvement on our validation set using an F1 score, a mean of precision (i.e., minimization of false positives) and recall (i.e., recovery rate of true positives). Since F1 scores are computed on a per-class basis, we calculated the weighted average of the scores based on the class frequency. For topic F1 scores, we chose a (more conservative) simple average due to the high class imbalance on a per-class basis. Implementation was supported by the HuggingFace transformers Python library.

B.0.3 Generative Pre-Trained Transformer Model

Recent developments in the generative or ”next-token prediction” variety of Transformer model have demonstrated state-of-the-art performance on a wide range of natural language processing tasks. These GPT models come in strong commercial and publicly-available implementations. The introduction of ”in-context learning” (i.e., providing information about the task to the model along with the task itself) has useful properties for NLP applications like sentiment analysis. Of particular benefit is the avoidance of costly model parameter updates via this instruction. The research community is still actively discussing the use of in-context learning over parameter fine-tuning.

As a preliminary result, we have evaluated the performance of the ”Vicuna” open-source model (developed by the Large Model Systems Organization, which in turn leveraged Llama2 by Meta) with an input (prompt) that contained examples from our training set and descriptions of the label categories. We tested the smallest model (7 billion parameters) and applied quantization. This approach yielded an accuracy of 0.72 for topics and 0.63 for sentiment classification on a small evaluation set taken from our training data. This represents slightly weaker performance than our findings using Loughran-McDonald. Results based on ”FinGPT” (Yang, Liu and Wang (2023), Zhang, Yang and Liu (2023), Wu et al. (2023) yielded an accuracy rate of 0.71. In both cases, we used a small selection of examples in the model prompt.

In conclusion, there is future work needed to develop and evaluate this newer class of GPT models. For example, we found in some cases that the model particularly struggled to classify Neutral sentences. Given that BERT architectures showed improvements via fine-tuning, we hypothesize the same could be true of the GPT class of models.