ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

Mind Your Language: Market Responses to Central Bank Speeches

Authors	Maximilian Ahrens, Deniz Erdemlioglu, Michael McMahon, Christopher J. Neely, and Xiye Yang
Working Paper Number	2023-013C
Revision Date	September 2024
Citable Link	https://doi.org/10.20955/wp.2023.013
Suggested Citation	Ahrens, M., Erdemlioglu, D., McMahon, M., Neely, C.J., Yang, X., 2024; Mind Your Language: Market Responses to Central Bank Speeches, Federal Reserve Bank of St. Louis Working Paper 2023-013. URL https://doi.org/10.20955/wp.2023.013

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

Mind Your Language: Market Responses to Central Bank Speeches*

Maximilian Ahrens
MAIHEM

Deniz Erdemlioglu
Department of Finance, IESEG School of Management and CNRS

Michael McMahon
Department of Economics, University of Oxford and CEPR

Christopher J. Neely
Research Division, Federal Reserve Bank of St. Louis

Xiye Yang
Department of Economics, Rutgers University

September 28, 2024

Abstract

Post-meeting central bank communication often moves markets, but researchers have paid less attention to the more frequent central bankers' speeches. We create a novel dataset of U.S. Federal Reserve speeches and develop supervised multimodal natural language processing methods to identify how monetary policy news affect bond and stock market volatility and tail risk through implied changes in forecasts of GDP, inflation, and unemployment. We find that forecast revisions derived from FOMC-member speech can help explain volatility and tail risk in both equity and bond markets. Speeches from Chairs tend to produce larger forecast revisions and unconditionally raise volatility and tail risk. There is some evidence that a speaker's monetary policy views, i.e, hawkishness vs. dovishness, may affect the impact of implied forecast revisions after conditioning on GDP growth. We show that central bank communication may *calm* markets, depending on the signals conveyed.

Keywords: Central Bank Communication, Multimodal Machine Learning, Natural Language Processing, Speech Analysis, High-Frequency Data, Volatility, Tail Risk.

JEL: E52, C45, C53, G12, G14

---

*We thank Andrea Ajello, Ilias Chronopoulos, Michael Ehrmann, Klodiana Istrefi, Galina Potjagailo, and participants at the ECONDAT 2023 conference for valuable comments and discussions. We are particularly grateful to Galina Potjagailo for excellent suggestions and discussion on the earlier version of this paper. Michael McMahon gratefully acknowledges financial support from the European Research Council (Consolidator Grant Agreement 819131). Maximilian Ahrens is thankful for support from the Oxford-Man Institute for Quantitative Finance. Our dataset of central bank speeches, time-stamped on the minute of release, is available from the the authors. The views expressed are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

1 Introduction

A large branch of monetary policy research seeks to explain how central bank communication (CBC) influences market dynamics and expectations (Blinder, 2018). Theory suggests that if central bank announcements and speeches convey information on economic conditions, market participants will update their beliefs and therefore their portfolio choices. Central bank communication can thus contribute to revaluing assets and stabilizing market conditions by reducing uncertainty (Bernanke et al., 2005). Empirical research largely corroborates this theoretical prediction and establishes a consensus that central bank communication influences asset prices through its effects on market participants' expectations about economic outlook and policy decisions (Bernanke and Kuttner, 2005; Ramey, 2016). Monetary policy communication also appears to influence the risk premium (Hanson and Stein, 2015; Cieslak and Schrimpf, 2019; Swanson, 2021; Cieslak and McMahon, 2023).

Despite these findings, there are still at least two major challenges: (i) how to identify the news in central bank communication, and (ii) how to identify the effects of such news on market volatility and tail risk. In this paper, we develop a novel multimodal natural language processing (NLP) method to identify macroeconomic signals in central bank speeches and we assess their impact on market volatility and tail risk.

Signals about the economic situation can affect asset prices in different ways. While speeches provide information, we don't think the channel is necessarily the same as the classic "Fed Information Effect" as emphasised in, for example, Romer and Romer (2000) and Nakamura and Steinsson (2018). The "Fed Information Effect" implies that the central bank, either explicitly or implicitly through its policy decision, releases superior information about the economy that is then incorporated into private sector forecasts. Instead, the central bank's alternative economic assessment (as in Byrne et al., 2023) could heighten concerns about a possible monetary policy mistake and thereby create volatility (Caballero and Simsek, 2022; Cieslak and McMahon, 2023). Or, the central bank's views may influence private views about uncertainty (Hansen et al., 2019; Aruoba and Drechsel, 2022). Finally, a cacophony of economic assessments, even if just reflecting different views on the economic outlook, might raise uncertainty (Ahrens and McMahon, 2021).

Furthermore, we introduce novel empirical measures that capture macroeconomic news from the central bank and show that releasing these signals via speeches gives rise to effects on market volatility and tail risk. However, our novel measures cannot distinguish between these possible channels.

The focus on speeches is important because, while central bank policy announcements occur infrequently, i.e., typically every 4-8 weeks, policy makers' speeches influence market expectations much more frequently (Neuhierl and Weber, 2019). Although recent developments in natural language processing (NLP) have allowed economists to analyse text with machine learning methods (see e.g., Bholat et al., 2015; Hansen et al., 2018; Ahrens and McMahon, 2021), researchers have paid only limited attention to speeches so far,[1] partly because their content is difficult to quantify and the field still lacks easily accessible

[1] Recently, Neuhierl and Weber (2019) have investigated the tone of speeches by central bank chairs and vice-chairs while Petropoulos and Siakoulis (2021) use a mixture of machine learning and dictionary methods to calculate sentiment indices from central bank speeches. The latter authors argue that this sentiment predicts financial turmoil. Swanson (2023) highlights

datasets of central bank speeches.

Our methodological framework has two parts. First, we use machine learning methods from the field of multimodal natural language processing to infer implied macroeconomic forecasts from Fed officials' public speeches. Our training dataset consists of Greenbook texts and their respective forecasts, which allows us to learn a mapping from central bank language to central bank forecasts (see Ahrens and McMahon, 2021). In our test dataset, we then apply the learned mapping to central bank speeches to infer how news signals in speeches can predict revisions of public macroeconomic forecasts. Second, we investigate the high-frequency (intradaily) responses of market volatility and tail risk to speech-implied revisions in CPI, GDP, and unemployment outlooks.[2]

Our paper contributes to the literature in several ways. Most importantly, we show that central bankers' speeches significantly influence volatility and tail risk in financial markets. Our novel, multimodal framework identifies the news component of policymakers' signals about GDP growth, CPI, and unemployment outlooks. We compare and contrast the performance of an extensive array of modern machine learning methods for multimodal NLP on our datasets. Our speech-implied forecast revisions predict changes in Survey of Professional Forecasters (SPF) forecasts substantially better than models that use purely tabular data and ignore the textual content of the speeches. These speech-implied forecast revisions explain a sizeable part of realized volatility and tail risk in financial markets, particularly if the speech comes from the Chair.

The remainder of the paper is organized as follows. In the next section, we review the related literature. Section 3 describes the data and section 4 introduces our methodological framework. In sections 5 and 6, we discuss the results of language mapping model performance and nature of speeches, respectively. Section 7 presents our empirical results pertaining to the analyses of speech-implied news and high-frequency market responses. Section 8 concludes the paper.

2 Related Literature

A large literature studies the effects of central bank communication around announcements, and a growing part looks at the effect of speeches. To the best of our knowledge, we are the first to study how bank communication about the economic outlook affects market volatility and tail risk. We focus on financial market uncertainty rather than uncertainty about monetary policy (as in Bauer et al., 2022; Husted et al., 2020; Ozdagli and Velikov, 2020; Tillmann, 2020) or the uncertainty that monetary policymakers face (Cieslak et al., 2023).

2.1 Central Bank Communication Effects on Market Volatility and Tail Risk

Our paper is most closely related to studies of the high-frequency effects of CBC on market uncertainty and volatility. Earlier research has focused on how central bank communication and decisions affect asset

the importance of Fed Chair speeches using an event-study revision decomposition, and Cieslak and McMahon (2023) focus on the communication of Fed stance and its effects on the risk premium.

[2]High-frequency market analysis is common in monetary research; see, for example, Gurkaynak et al. (2005); Gertler and Karadi (2015); Nakamura and Steinsson (2018); Jarociński and Karadi (2020) and Miranda-Agrippino and Ricco (2021).

prices or volatility in financial markets.[3]

To our knowledge, only Hattori et al. (2016) has studied stock and bond market tail risk in response to unconventional policy announcements. Unconventional monetary policy (UMP) increases (decreases) the realized volatility of stocks (bonds), but lowers the tail risk in both markets. Forward guidance (and hence communication) appears to have stronger "dampening effects," compared to other UMP events.[4]

We extend this line of research in three ways. First, we focus on the intraday market responses to policymaker speeches rather than responses to FOMC announcements. Second, we measure the *realized* tail risk instead of the *implied* tail risk from derivatives. The use of realized measures, rather than those from derivatives, allows us to characterize the dynamics of volatility and tails at much higher frequency. In contrast with Hattori et al. (2016), we find that Fed Chair speeches interacted with forecast revision surprises *decrease* realized volatility and tail risk. CBC *may* reduce uncertainty and calm financial markets, depending on the message in the speech, position of the Fed member (Chair versus other FOMC members) and type of macro news, i.e., CPI, unemployment or GDP. Third, we employ a more flexible model that allows for *time-varying* tails and better captures news-induced persistence in intradaily volatility and tail risk. This model allows us to separate extreme volatility responses from the tail responses and, more importantly, to identify the speeches that create *tail cascades*. Unlike the previous studies treating jumps as independent events, e.g., Bauer et al. (2022), we accommodate the stochastic jump intensity that potentially results from heterogeneous interpretation of news by market participants.

2.2 Text Analysis for Monetary Policy

We also contribute to a burgeoning literature that uses natural language processing to analyse monetary policy. Various text analysis methods have been tested in this field. For example, researchers have used topic models (Hansen et al., 2018), combined dictionary methods with classic machine learning models such as XGBoost (Petropoulos and Siakoulis, 2021), and have deployed deep neural network models such as transformers (Cai et al., 2021). In our work, we take a model-agnostic, data-driven approach to reduce modeler bias. That is, we train a variety of NLP models and choose the algorithm that works best in our validation set.

Similarly, researchers have employed various frameworks and datasets to identify monetary policy news. In particular, researchers have often studied the market effects of central bank policy announcements. For instance, Lucca and Trebbi (2009) and Hansen and McMahon (2016) both leverage approaches from

[3]Cieslak and Schrimpf (2019) study the high-frequency effects of the non-monetary news component of communication on risk premiums. Leombroni et al. (2021) explore how CBC influences credit risk premia through high-frequency changes in the yield curve. Ehrmann and Talmi (2020) measure textual differences between central bank announcements and find that higher levels of textual similarity to the previous announcement statement are usually associated with lower market volatility after the announcement date. Bekaert et al. (2013) find evidence that looser policy reduces risk aversion and uncertainty. Gómez-Cram and Grotteria (2022) explore the price discovery process for several asset classes on FOMC announcement days.

[4]In the context of forward guidance, Ehrmann et al. (2019) put forward a model where forward guidance can amplify the reaction of expectations to macroeconomic news. Empirically, they show that the type and horizon of forward guidance—time-contingent, state-contingent, open-ended, short or long horizon—influences the sensitivity of bond yields to news and degree of disagreement among forecasters. For example, while long-horizon forward guidance reduces interest rate sensitivity to macroeconomic news, short-horizon guidance amplifies it. Similarly, state-contingent forward guidance limits bond price responses to macro news but open-ended forward guidance essentially has no statistically significant effect on the response.

computational linguistics within a VAR framework to examine the effect of the content in FOMC statements on macroeconomic variables.[5] Using a deep neural network architecture to identify text-based shocks in FOMC announcements, Handlan (2020) assesses the impact on Fed funds futures and finds that forward guidance in FOMC statements account for four times the variation in Fed funds future prices than do federal funds rate target changes. Gardner et al. (2022) introduce a new FOMC sentiment index using textual analysis, and show that announcement effects on equities depend on good versus bad times. Gómez-Cram and Grotteria (2022) apply a video analysis on words mentioned during central bank press conference videos. Nesbit (2020) proposes a word count based instrumental variable framework to identify monetary policy shocks in FOMC transcripts. Aruoba and Drechsel (2022) use NLP techniques to analyse FOMC meetings in order to measure the information set of the FOMC at the time of policy decisions. Gáti and Handlan (2022) use regularized regressions to map the wording of FOMC statements to Greenbook forecasts of output growth, unemployment and the federal funds rate, arguing that the statement wording implies FOMC expectations fairly well, with the exception of short-run inflation expectations. The authors note that these patterns have changed over time with Fed Chairs.

Rather than focus on FOMC announcements, we join a few recent papers in studying central bankers' speeches that happen with much higher frequency, making central bank communication more nearly continuous (Neuhierl and Weber, 2019).[6] We use a two-step macroeconomic news identification framework, in which we first learn a mapping from central bank language to central bank forecasts with Greenbook data, and then use these learned mappings to infer how FOMC member speeches imply revisions to SPF forecasts of GDP, inflation, and unemployment — an approach which is motivated by Ahrens and McMahon (2021).

To identify the news content of a speech, we control for market expectations with the latest forecasts from the Survey of Professional Forecasters (SPF) conducted by the Federal Reserve Bank of Philadelphia. SPF forecasts directly measure expected GDP growth, inflation, and unemployment. We define a speech-induced news shock as the difference between a speech-implied forecast revision and the latest SPF forecast for that variable at speech time. [7]

3 Federal Reserve and Markets Data

We use several types of data in our paper: Greenbook forecasts, SPF forecasts, intraday volatility and tail risk measures for US stock and bond markets, as well as unstructured text data from the Greenbook and

[5] Lucca and Trebbi (2009) find CBC to be a more important factor than contemporaneous policy rate decisions. Hansen and McMahon (2016) conclude that shocks to forward guidance have stronger effects on markets than communication of current economic conditions.

[6] Neuhierl and Weber (2019) find that the tone of US Fed chair and vice-chair speeches, measured via word count methods, can explain stock market price dynamics. Using a mixture of machine learning and dictionary methods, Petropoulos and Siakoulis (2021) derive sentiment indices from central bank speeches and find that the sentiment predicts financial turmoil. Swanson and Jayawickrema (2023) compare high-frequency changes in interest rates after Fed Chair versus Fed Vice Chair speeches and find that Chair speeches have a much higher impact.

[7] Other researchers have used news media coverage to control for market expectations. Ellen et al. (2022), for example, construct a monetary news series from the difference in narrative between central bank statements and news media coverage. Similarly, Cai et al. (2021) analyse FOMC announcements using BERT (Devlin et al., 2019) and identify monetary policy and information shocks, controlling for market expectations by analysing relevant New York Times articles with NLP methods.

FOMC speeches.[8] These data are split into a training and a test set. We describe these datasets below.

3.1 Federal Reserve Text and Forecast Data

To impute macroeconomic news signals from central bank speeches, we must learn a mapping from Fed words to forecasts. For this, we map Greenbook text sections on forecasts to the respective *internal* Greenbook forecasts. We then apply this learned mapping to FOMC members' *external* speeches and assess how speech-implied forecast revisions affect volatility and tail risk in financial markets.

Our key underlying assumption in this approach is that central bankers use similar vocabulary in Greenbooks and speeches when talking about revisions to the outlook. We believe this to be the case despite the fact that Board of Governors staff economists write the Greenbook text but FOMC members or regional reserve bank economists write the speeches, and despite the fact the two texts are often directed at different audiences for somewhat different purposes. We believe this to be reasonable because the language used to describe economic concepts and data is similar, the staff analysis provides inputs into the speeches, and our use of a supervised topic model to map word co-occurrences to forecasts will capture when different words are semantically similar. Whether this assumption is reasonable is ultimately an empirical question. The out-of-sample R^2 values from our test data confirm that the mappings must be indeed similar.[9] Gáti and Handlan (2022) also find that FOMC statements, another form of externally focused communication, is a good textual description of Greenbook forecasts.

Training set: In the training phase, we learn the mapping of the Fed's Greenbook texts associated with the descriptions of GDP growth, CPI, and unemployment outlooks to the change in the Greenbook forecasts of those variables from the previous forecast. That is, we target the difference in a current period's one-quarter-ahead Greenbook forecast to the previous quarter's forecast, such that for any of our macroeconomic key figures of interest, y, we define $\Delta y_m = y_m - y_{m-1}$, where m indicates the date of the Greenbook forecast.

We focus on near-term forecasts to capture policymakers' assessments of the current state of the economy while avoiding the endogeneity of longer-term forecasts to policy and the tendency of forecasts to revert to long-run values. This is consistent with the approach used in Coibion and Gorodnichenko (2012). Experiments with a one-year-ahead forecast horizon produced a slightly less precise but similar mapping.

The training sample spans 145 Greenbook documents, from January 1, 1995 to December 31, 2013. We only consider the 8,155 Greenbook sections that directly relate to GDP growth, CPI, and unemployment (see Appendix D for a detailed list of section allocations). The average Greenbook section in our dataset has about $3,000$ words; the longest section consists of $31,000$ words and the shortest section contains around 140 words. At any date, we concatenate all Greenbook sections that relate to the same forecasting variable.

Test set: Training the NLP models consists of estimating complex mappings from Greenbook text on each date, for each variable, to the associated revisions from the previous Greenbook to the one-quarter-ahead

[8] The Greenbook forecast information is, in recent years, presented together with Bluebook information in the Tealbook. We continue to refer to it as Greenbook information to make clear we are using the Fed's economic forecasts.

[9] In section 6.2, we explore the individual-specific nature of speeches.

Greenbook forecasts on each date, for each variable. Once the models are trained, we apply the learned mappings to each element of a test set consisting of FOMC members' speeches made from August 1, 2008 to December 31, 2020. The applied mappings imply one-quarter-ahead forecast revisions for GDP growth, CPI, and unemployment.

Table 1 reports the summary statistics of the speeches used in our data. Specifically, it reports the speakers and the characteristics of his/her body of speeches. The table indicates that the time spans in which the speakers gave speeches (i.e., the time between first and last speech in our sample) vary considerably across speakers due to their different terms on the FOMC.

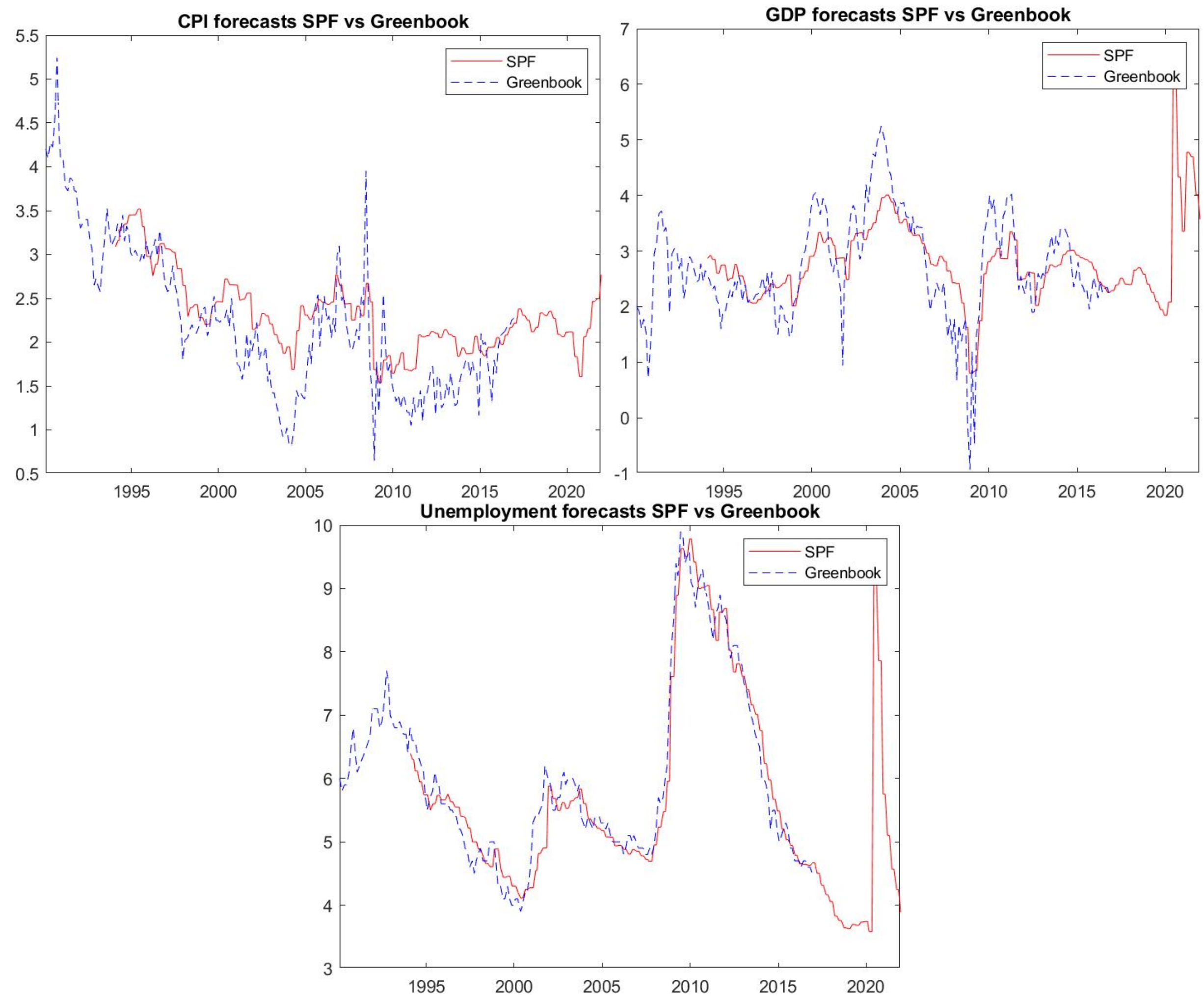
Table 1: Summary statistics of the speeches by FOMC members

FOMC Member	Role	First Speech	Last Speech	Speeches	Time Stamped
Fisher	DAL	18 Aug 2008	9 Mar 2015	106	0
Bernanke	Board	22 Aug 2008	16 Dec 2013	151	144
Lockhart	ATL	27 Aug 2008	14 Feb 2017	114	0
Kroszner	Board	1 Sep 2008	8 Dec 2008	6	0
Plosser	PHL	8 Oct 2008	17 Feb 2015	91	0
Duke	Board	23 Oct 2008	9 May 2013	51	51
Warsh	Board	6 Nov 2008	8 Nov 2010	8	8
Pianalto	DAL	14 Nov 2008	27 Mar 2014	51	0
Kohn	Board	8 Dec 2008	13 May 2010	21	17
Tarullo	Board	19 Mar 2009	4 Apr 2017	77	77
Raskin	Board	15 Jul 2010	17 Jul 2013	19	18
Yellen	Board	15 Jul 2010	29 Nov 2017	82	79
Stein	Board	11 Oct 2012	6 May 2014	16	15
Powell	Board	22 Feb 2013	6 Oct 2020	91	91
Fischer	Board	10 Jul 2014	28 Sep 2017	45	45
Brainard	Board	2 Dec 2014	17 Dec 2020	81	79
Harker	PHL	2 Oct 2015	2 Dec 2020	80	0
Kaplan	DAL	18 Nov 2015	29 Sep 2020	27	0
Quarles	Board	30 Nov 2017	11 Dec 2020	48	48
Clarida	Board	25 Oct 2018	16 Nov 2020	30	30
Bowman	Board	11 Feb 2019	4 Dec 2020	19	19

Notes: The table reports the summary statistics of the statements and speeches by the Fed officials FOMC members in our text dataset. The table presents the names of the speakers, their role, their first and last speech dates in our sample, and the number of speeches overall and for which we have time stamps.

In order to analyse the market effects of the speech signals on the market measures of volatility, we need to have carefully time-stamped speech data. This is not possible for all the speeches in our sample. In fact, it is limited to members of the Board of Governors. Even though they are not time stamped, we nonetheless found it useful to be able to make use of the speeches of some regional Fed presidents in order to control for time and member fixed effects. We describe this adjustment in Section 6.

Figure 1: Comparison of Greenbook and SPF forecasts



Notes: The figure displays the Greenbook and SPF forecasts over time for CPI (top left panel), GDP (top right panel) and unemployment (bottom panel). SPF forecasts are the mean across SPF participants. The two forecasts match quite closely for the majority of the inspected time-series.

The target variables in the test set are the respective changes in one-quarter-ahead SPF forecasts of GDP growth, CPI, and unemployment. The SPF is a publicly available and widely referenced source for economic forecasts. We use the mean SPF forecasts across SPF participants as our proxy for expectations, rather than Greenbook forecasts, because the latter are released to the public with a 5-year delay. We expect that central bank speeches should have similar predictive power for Greenbook and SPF forecast revisions. Figure 1 illustrates that the SPF forecasts are highly correlated with the Greenbook forecasts during 1993 to 2016. We assume that this pattern also holds post 2016, for which there was no public Greenbook data available when the data for this paper were collected.

3.2 High-Frequency Market Data

We use high-frequency transaction prices for 22 Dow Jones Industrial Average (DJIA) stocks, together with 2-year, 5-year, and 10-year U.S. Treasury note and bond futures traded on the Chicago Board of Trade (CBOT). Appendix E lists the individual stocks and bonds. Wharton Research Data Services (WRDS) and Tick Data LLC provide data for individual stocks and bond futures, respectively. As is standard in the literature, we exclude U.S. holidays, Christmas periods, and weekends from our sample. We only consider trading hours from 9:30 EST−16:00 EST and 7:30 CT−14:00 CT, for stock and bond markets, respectively. To reduce the potential impact of market microstructure noise, we filter out *bouncebacks* and irregular quotes that typically occur in ultra high-frequency data. Using our adjusted data, we create equally-spaced 15-second observations, which is an appropriate frequency to implement our response measures. Our sample runs from January 1, 2014 through December 31, 2021.

4 Methodological Framework

Our methodological framework can be broken down into two parts. Section 4.1 explains our multimodal NLP framework used to estimate the mapping from central bank language to forecasts. We test and compare our estimation framework with a variety of machine learning algorithms. Section 4.2 then describes the measurements of the asset price dynamics and their relationship with the speech signals.

4.1 Multimodal NLP Framework

We estimate how central bank speeches influence financial markets. To do so, we map central bank language to macroeconomic forecasts, controlling for the macroeconomic conditions at the time by using macro data as inputs to the mapping function. Conditioning on the macroeconomic situation may be important because the effect of a given forecast revision on financial markets may depend on economic conditions. This economic context requires the multimodal modelling approach. For example, a speech that raised forecast inflation would be a positive signal of improving conditions if inflation was below its desired level. However, the same speech would convey a negative signal if inflation was substantially above target. We employ multimodal machine learning approaches that allow us to use both text and tabular data when mapping central bank language to central bank forecasts and then predicting output, inflation, and unemployment outlook revisions.

4.1.1 Learning Mapping from Central Bank Language to Forecasts

We learn the mapping from the Fed's Greenbook text to the respective Greenbook forecasts. The Greenbooks contain dedicated sections on the Fed's forecasts of GDP growth, CPI, and unemployment, as well as the rationales for the forecasts. These sections allow us to map the Greenbook text - ergo central bank language - to central bank forecasts.

In the training phase, we estimate a separate mapping for each of the three variables, i.e., the one-quarter-ahead forecast change in CPI, GDP growth, or unemployment. We measure the change from the previous $(m-1)$ Greenbook to the current (m) in the one-quarter-ahead forecasts (q_1). We denote CPI by π, GDP growth by g, and unemployment by u. Hence, our three target variables are: $\Delta\pi_{q_1,m}$, $\Delta g_{q_1,m}$, and $\Delta u_{q_1,m}$. For ease of notation in the following equations of our modelling framework, let y serve as a placeholder variable for any of the CPI, GDP growth, and unemployment variables. Our placeholder target variable is $\Delta y_{q_1,m}$.

To capture the economic context, we control for both change and level of the CPI, GDP, and unemployment of the previous Greenbook report, denoted as X_{m-1}. We fit a function, f, that learns parameters, Ω, to map the Greenbook text and tabular inputs to the target output. The equations for CPI, GDP growth, and unemployment have the same explanatory variables, except for the text input, which is specific to the respective Greenbook forecast section. That is, θ_π represents the text features for the CPI corpus, θ_g represents GDP-related text, and θ_u unemployment-related text. We use θ_y as a placeholder for any of the three text inputs, while $\theta_{y,k}$ represents the k^{th} text feature for the respective target variable y. We can now write out our regression equation as

$$\Delta y_{q_1,m} = f\left(X_{m-1}, \theta_{y_m}; \Omega\right). \tag{1}$$

If we assume linearity in function f, the regression equation can be written as follows:

$$\begin{aligned}\Delta y_{q_1,m} = {} & \omega_\pi \pi_{q_1,m-1} + \omega_g g_{q_1,m-1} + \omega_u u_{q_1,m-1} \\ & + \omega_{\Delta u}\Delta u_{q_1,m-1} + \omega_{\Delta\pi}\Delta\pi_{q_1,m-1} + \omega_{\Delta g}\Delta g_{q_1,m-1} \\ & + \sum_{k=1}^{K}\omega_k\theta_{y,k,m} + \epsilon_m.\end{aligned} \tag{2}$$

Here, the ωs represent the regression parameters and ϵ is the measurement error. We use the first 80% of the Greenbook dataset for training and the remaining 20% for validation. The data are demeaned and standardized based on training set values. A respect for the time-series characteristics in the data, i.e., the potential for information leakage, deterred us from randomly splitting the training and validation sets.

We then train the machine learning models to map central bank texts and control variables to the respective target variables. We treat this as a regression problem and use a least squares error loss function, as is typical in econometric applications to monetary policy.

4.1.2 Identifying Information Signals in Central Bank Speeches

In the test phase, we apply the trained models for each of the macroeconomic variables, i.e., CPI, GDP growth, unemployment, to the FOMC member speeches to infer macroeconomic forecast revisions. The coincident tabular data inputs on current economic conditions are the most recent SPF forecast levels and changes on GDP growth, CPI, and unemployment.[10] This procedure maps each central bank speech into an implied revision of the forecasts for CPI, GDP growth, and unemployment. Importantly, speeches that are *not* about the economic situation will not use many forecast-related words and so will be estimated

[10] As shown in Figure 1, the SPF forecasts track the Greenbook forecasts quite closely.

to have very little signal value.[11] Asset prices should only react to relevant news that has not yet been incorporated into asset prices. To isolate the new, i.e., surprising, components of revision signals we thus adjust the speech implied forecasts, $\Delta\hat{y}_{\text{speech},s}$, in two ways. We outline these two adjustments in Section 6.

4.1.3 Machine Learning Methods

We do not know, a priori, which statistical learning model would best approximate the function, f, in equation (1). We have relatively few data points compared to many machine learning projects (e.g. hundreds or thousands rather than millions or billions of data points). Each data point itself is rich in information, however, consisting of a high dimensional feature set. That is, each set of text can be several thousand words long, which presents a problem for many modern language models such as transformer family models (e.g. BERT-based models), which can usually only handle up to around 100-1,000 tokens per data point (Das et al., 2021). Some extensions based on sparse transformers have been proposed such as Child et al. (2019); Zaheer et al. (2020), which can handle sequences of a couple of thousand tokens. However, document lengths of $20,000+$ words would still pose a challenge. Lacking reason to favour a specific class of models, we search broadly for the best model to reduce the a priori modeler bias. That is, we deploy an extensive array of multimodal machine learning algorithms to approximate function f and to learn parameters Ω. We use the multimodal machine learning benchmark suite, AutoGluon (AutoGL) (Erickson et al., 2020), and we add to it the class of multimodal supervised topic models (Card et al., 2018; Ahrens et al., 2021).

4.1.4 AutoGluon

AutoGL is an automated machine learning (AutoML) framework that has been developed to fuse multimodal features such as text, images, and tabular data. We chose this AutoML framework because it outperformed competing frameworks in multimodal benchmark tasks (see Erickson et al., 2020).

Base models: AutoGL fits machine learning *base models* and then combines them through ensembling and stacking to boost performance. AutoGL allows us to apply hyperparameter optimization over all models. The *base models* in AutoGL span the following broad machine learning algorithm classes (See Appendix A for a full description):

1. **K-nearest neighbours** (Dudani, 1976): AutoGL uses two variations of k-nearest neighbours (KNN) that differ in their weighting approaches. One allocates uniform weights to all points while the other weights points according to the inverse of their respective distances.

2. **Random forests** (Breiman, 2001): AutoGL again deploys two variations of this algorithm class. One option uses the information gain of nodes for the assessment of the split quality. The other option uses Gini impurity instead.

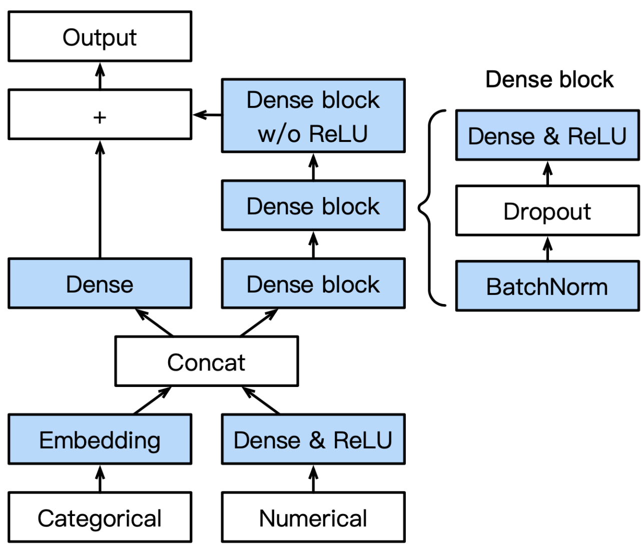
[11] We discuss this in more detail in Remark 3 in Appendix C.

3. **Extremely randomized trees** (Geurts et al., 2006): For the random tree class, AutoGL deploys both an implementation resorting to information gain and another option that uses Gini impurity for the assessment of split quality.

4. **Boosted decision trees**: AutoGL runs (where applicable to the task) Extreme Gradient Boosting (Chen and Guestrin, 2016), Light Gradient Boosting (Ke et al., 2017), Categorical Boosting (Prokhorenkova et al., 2018).

5. **Neural networks**: Figure 2 schematically outlines AutoGL's neural network architecture, which Erickson et al. (2020) details. The architecture has been specifically designed for the multimodal use of categorical (text, images) and numerical data. It uses variable-specific embeddings for each of the categorical features. These are then concatenated with the numerical features into one overall input vector. This vector is in turn fed through a 3-layer feed-forward network as well as through a linear skip-connection (for details see Erickson et al., 2020). Model ensembling and stacking can be applied and are optimally chosen in the validation process.

Figure 2: AutoGL schematic neural network architecture



Notes: The figure displays the AutoGluon schematic neural network architecture, based on the design by Erickson et al. (2020), p. 3. Layers with learnable parameters coloured in blue.

Text representation options: We must also choose how to represent the text in machine-readable format. We define the following approaches:

1. **AutoTab**: Only tabular features are used. Text is excluded. AutoTab is our tabular data baseline next to an OLS regression that only uses tabular data.[12]

[12] AutoGL's *TabularPredictor* approach.

2. **AutoTab + tf-idf**: Use tf-idf weighted word counts of the text as features. Standard text cleaning procedures of removing stopwords and punctuation have been applied.

3. **AutoTab + topics**: Use topic shares from supervised topic models as features (using rSCHOLAR without tabular data for the topic estimation).

4. **AutoMM transformer**: Use the AutoGL's multimodal modelling infrastructure that is based on a large language model (we use Roberta-base (Liu et al., 2019)) for multimodal fine-tuning. Tabular data can be fused into this process as well.[13]

5. **AutoTab + embed**: Use AutoMM transformer as well as AutoTab models that featurize text data as n-grams and ensemble over this zoo of models.[14]

We now proceed to characterize asset price dynamics to assess how forecast-revisions affect financial market volatility and tail risk.

4.2 Asset Price Dynamics

In this section, we first present the continuous-time model for asset prices. Section 4.2.2 then outlines how we measure volatility and realized tail risk from the high-frequency data.

4.2.1 Underlying Continuous-Time Model

We model the intraday behaviour of asset prices with the following continuous-time model: The log-price X of each asset (stock or bond) follows an Itô semimartingale defined on a filtered space (Ω, $\mathcal{F}_t$, $(\mathcal{F}_t)_{t\in[0,T]}$, $\mathbb{P}$) over an interval $[0,T]$. The Grigelionis decomposition (see e.g., Erdemlioglu and Yang, 2022; Boswijk et al., 2018; Dungey et al., 2018) implies that X_t has the following specification:

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \delta * (\mu_t - \psi_t) + (\delta - h(\delta)) * \mu_t, \tag{3}$$

where b_s is the drift term, σ_s is the stochastic volatility component, W is a standard Brownian motion, δ is a predictable function, h is a truncation function (e.g., $h(x) = x1_{\{\|X\|\leq 1\}}$), μ is the jump measure of X, and ψ is its jump compensator, which adopts the decomposition,

$$\psi_t(dt, dx) = [f_t(x)\lambda_t dx]dt,$$

where the function, $f_t(x)$, controls the jump size distribution and λ_t denotes the jump intensity as in Erdemlioglu and Yang (2022) and Boswijk et al. (2018). We focus on the *tail* component of this jump

[13] AutoGL's *MultimodalPredictor* approach.
[14] AutoGL's *TabularPredictor* approach with the *hyperparameter* option being set to *multimodal*.

compensator or λ_t, which captures the jump intensity dynamics.[15] We can specify λ_t as

$$\lambda_t = \lambda_0 + \int_0^t b'_s ds + \int_0^t \sigma'_s dW_s + \int_0^t \sigma''_s dB_s + \delta' * \mu_t + \delta'' * \mu_t^{\perp}, \tag{4}$$

where B is a standard Brownian motion, independent of W, $\mu_t^{\perp}$ is orthogonal to μ_t, and δ', δ'' are predictable. This model, given by equations (3) and (4), satisfies no-arbitrage conditions and leaves the volatility and jump components unrestricted. We now present this model's volatility and tail risk measures.

4.2.2 High-Frequency Measurement of Volatility and Tail Risk

Given the price dynamics in equations (3) and (4), let us define the ith intradaily return on a trading day as $r_{i,t} = X_{i,t}$ - $X_{i-1,t}$. The daily realized volatility (RV) is the square root of realized variance, which is the sum of the squared intraday returns $(1, \ldots, M)$. That is,

$$RV = \sqrt{\sum_{i=1}^{M} r_i^2}. \tag{5}$$

It is well-known that realized variance converges to quadratic variation (see e.g., Andersen et al., 2003, 2001 and Barndorff-Nielsen and Shephard, 2002 for in-depth discussion). We use the scaled version of realized variance and compute RV in the post-signal windows.

Turning to the estimation of $\lambda_{i,t}$ in equation (4), we define the realized intensity (RI) measure as

$$RI = \frac{\Delta_n^{\varpi \widehat{\beta}_i}}{k_n \Delta} \sum_{j=1}^{k_n} g\left(\frac{|r_i|}{\alpha \Delta^{\varpi}}\right) \frac{\alpha^{\widehat{\beta}}}{C_{\widehat{\beta}_i}(k_n)}, \tag{6}$$

where Δ is incremental change between observations, $\alpha\Delta^{\varpi}$ is threshold to retain only large jumps, $g(\cdot)$ admits a specific functional form, k_n is a constant which admits $(1/K \leq k_n \Delta^{\rho} \leq K)$ for $(0 < \rho < 1)$ and $(0 < K < \infty)$, and β_i is the estimator of jump activity index that controls the vibrancy of sharp fluctuations. In equation (6), $g(\cdot)$ is an auxiliary function that separates jump-type movements from the diffusive volatility, based on an α deviation (e.g., α = 2, 3, 6) from the continuous component of the model.[16] We use RI as a proxy for time-varying (high-frequency) *tail risk* (TR), which is very accurate at high frequency, similar to the measures adapted in Bollerslev et al. (2015).[17]

Our tail risk measure RI (equation (6)) has several advantages. RI is relatively easy to implement but does not require strong assumptions about the underlying asset pricing process (see Appendix B.1). Because it simultaneously accounts for time-varying volatility, clustering in extreme price changes (jump clustering) and accommodates tail (jump) activity of the price variation around speeches, it allows us to

[15]See Andersen et al. (2020), who exploit jump intensity process to measure tail risk and assess its equity premium implications.

[16]See e.g., Erdemlioglu and Yang (2022), Boswijk et al. (2018) and Dungey et al. (2018) for implementation details, particularly on the selection of the functional form for $C_{\widehat{\beta}_i}(k_n)$ in (6).

[17]Our tail risk indicator RI is also quite similar to the estimator of Hill (1975). See also Aït-Sahalia and Jacod (2009) for a related discussion on the role of β_i in (6).

accurately measure tail responses to FOMC speeches and to disentangle jumps from shifts in volatility, thereby solving an econometric identification problem. Large (small) values of RI indicate that the returns generate heavy (light) tails.[18]

In summary, communication can create sudden changes in realized volatility as well as asset price jumps and persistently elevated jump intensity. Our approach allows us to first detect the speech-implied jumps, and then assess the 'intensity' of the jump responses.[19]

4.3 Identifying Association Between News and Market Reactions

To measure how realized volatility and tail risk in equity and bond markets react to central bankers' speeches, we regress the market reactions on the forecast revisions implied by the corresponding speech. As the forecast revision itself is a linear combination of the central bank signal and the latest public forecast, we already control for the partial correlation between the SPF forecasts and the market reactions.[20] We do not include additional low-frequency macroeconomic control variables because market prices should already incorporate such publicly available information.

5 Results: Language Mapping and SPF Prediction

In this section, we discuss the language mapping procedure and present the results of model performance. We first learn the mapping from central bank language to central bank forecasts by training our model on the first 80% of the Greenbook sample, holding out the last 20% for validation. In our validation set, we assess the out-of-sample performance of mappings from Greenbook language to Greenbook forecasts.

For each machine-learning class, we select the best performing model from the validation set and then assess how well each maps post-2013 speeches to SPF forecasts. Tables 2, 3, and 4 report the R^2s associated with test set predictions of revisions to CPI, GDP and unemployment SPF forecasts, respectively. Two key takeaways are: (i) the high R^2 values confirm that the Greenbook text and speeches have significant commonality regarding revisions to the economic outlook; (ii) the FOMC member speeches clearly predict changes in public macroeconomic forecasts. For example, the third row of Table 2 indicates that the multimodal neural topic model (MM NTM non-linear) has an R^2 of 0.670 in predicting CPI forecast revisions in the Greenbook training set, 0.830 in the Greenbook validation set, and 0.735 in the test set (speeches).[21]

For each of the three macroeconomic target variables, the best multimodal NLP models markedly outperform models that only use tabular data. Specifically, the multimodal neural topic model (MM NTM) class performs best both in the validation and in the test set. For CPI, Table 2 shows that the MM

[18]The term *intensity* in RI refers to the stochastic intensity of the jump process. While RV in equation (5) estimates the stochastic volatility, RI estimates the stochastic intensity.

[19]As Bollerslev et al. (2018) document, heterogeneous investors often release private information as they trade in the wake of such jumps, creating large price moves, which amplify high-frequency TR. It is also worth mentioning that we aggregate the information in measures by equally weighting the stocks in the panel. We apply the measures to all stocks, obtain the estimates of response measures, equally weight and use the cross-sectional average for a given speech.

[20]See e.g., Frisch and Waugh (1933) and Lovell (1963) for Frisch-Waugh-Lovel theorem.

[21]Appendix F presents all tested machine learning approaches.

NTM (non-linear) model has an R^2 of 0.735 in the test set, which is 15% better than MM NTM (linear) and 44% better than the R^2 of the next best method. Likewise, Table 3 shows that MM NTM (non-linear) has an R^2 of 0.797 in the test set, which is right behind MM NTM (linear)'s R^2 of 0.825. Finally, Table 4 shows that MM NTM (non-linear) performs best again for unemployment, with an R^2 of 0.208, which is markedly better than the second best R^2 of 0.131, achieved by AutoTab.

Interestingly, while both are worse than the best-performing MM approaches, AutoGL's models underperform an OLS regression for CPI inflation and GDP growth. There might be several explanations for this underperformance. First, these datasets contain relatively few data points — a common challenge in macroeconomics and macro-finance, especially for 'data hungry' machine learning methods. AutoGL's machine learning models might therefore struggle to converge or might easily overfit on the limited training data. Second, a linear model might do a good job approximating macroeconomic forecasts (or the revisions).

Table 2: Central bank language to forecast mapping - CPI Q1

Metric: R^2	**train (GB)**	**val (GB)**	**test (speeches)**
OLS	0.288		0.510
MM NTM (linear)	0.600	0.650	0.640
MM NTM (non-linear)	0.670	**0.830**	**0.735**
AutoTab	0.565	0.302	0.475
AutoTab + tfidf	0.953	0.305	0.299
AutoTab + topics	0.370	0.284	0.358
AutoTab + embed	0.573	0.139	0.132
AutoMM transformer	-0.155	-†	-0.292

Notes: The table reports R^2 for training, validation, and test sets for each of the models. Best performing model in validation and test set in bold. †: Model only reports MSE for validation set.

Table 3: Central bank language to forecast mapping - GDP Q1

Metric: R^2	**train (GB)**	**val (GB)**	**test (speeches)**
OLS	0.301		0.785
MM NTM (linear)	0.372	**0.426**	**0.825**
MM NTM (non-linear)	0.483	0.371	0.797
AutoTab	0.497	0.304	0.380
AutoTab + tfidf	0.752	0.240	0.268
AutoTab + topics	0.730	0.253	0.285
AutoTab + embed	0.587	0.220	0.142
AutoMM transformer	0.013	-†	-0.044

Notes: The table reports R^2 for training, validation, and test sets for each of the models. Best performing model in validation and test set in bold. †: Model only reports MSE for validation set.

Table 4: Central bank language to forecast mapping - unemployment Q1

Metric: $\boldsymbol{R}^2$	**train (GB)**	**val (GB)**	**test (speeches)**
OLS	0.231		-0.377
MM NTM (linear)	0.197	0.109	0.066
MM NTM (non-linear)	0.285	**0.457**	**0.208**
AutoTab	0.191	0.058	0.131
AutoTab + tfidf	0.577	0.113	-0.045
AutoTab + topics	0.278	0.053	-0.010
AutoTab + embed	0.415	0.145	-0.044
AutoMM transformer	-0.737	-†	-1.177

Notes: The table reports R^2 for training, validation, and test sets for each of the models. Best performing model in validation and test set in bold. †: Model only reports MSE for validation set.

6 Forecast Revisions from Speeches

We use the model that performed best in the validation set (Greenbook data) to estimate the speech-implied GDP, CPI, and unemployment forecast revisions, $\Delta\hat{y}_{\text{speech},s}$, in the test set (speech data). However, markets should only react to relevant news that has not yet been incorporated into asset prices. That is, a central bank speech must change market expectations to move prices. In this section we first explain how to create the speech signals (6.1) and then characterize the speech data.

6.1 Forecast Revision News and Implied Speech Signals

We seek to learn how forecast revisions implies by FOMC speeches influence market volatility and tail risk. Asset prices should only react to new/surprising information, so we must isolate the surprise components of the forecast revisions. To do so, we calculate deviations from market expectations and then adjust for existing information at the time-of-speech, $\Delta\hat{y}_{\text{speech},s}$.

First, we proxy market expectations with the latest public SPF forecast for each target variable. We then calculate the difference between the most recent SPF forecast change ($\Delta y_{SPF,s}$) available at the time of each speech and the implied forecast change in each speech ($\Delta\hat{y}_{\text{speech},s}$). This difference is our raw signal (capturing the implied forecast revision) given the language in speech s which we label $\psi_{y,s}$ for target variable, y:

$$\psi_{y,s} = \Delta y_{\text{SPF},s} - \Delta\hat{y}_{\text{speech},s}. \tag{7}$$

Second, we adjust for views common to FOMC members at the time of speech s, as well as positions of specific FOMC members that may be predictable. To do this, we regress their macro-variable-specific signal on month and member fixed effects. This leaves us with an adjusted forecast revision, $\nu_{y,s}$, for each speech, s, given by member, m, in month, τ, as:

$$\nu_{y,s} = \psi_{y,s} - \overline{\tau_s} - \overline{\alpha_m}, \tag{8}$$

where $\overline{\tau_s}$ is the average forecast revision signal in the month in which speech s is delivered, and $\overline{\alpha_m}$ is the average forecast revision of all speeches by member m. We denote the CPI revision surprise as $\nu_{\pi,s}$, the GDP revision surprise as $\nu_{g,s}$, and the unemployment revision as $\nu_{u,s}$ for each speech, s.

Figure 3 shows the raw speech signals, $\psi_{y,s}$, for each of the macro variables in the left column. The figure shows that FOMC language often moves together in such a way as to lead to quite large variation.

However, once we adjust for likely predictable signals in the forecast, the implied surprises, $\nu_{y,s}$, in the right column show much less variation, although they display still time-variation, extreme peaks and heterogeneity across the three macro news factors (CPI, GDP, unemployment). We use these $\nu_{y,s}$ surprises in our market analysis.

6.2 Individual Characteristics of the Speech Data

Table 5 reports the summary statistics of the speech signals. The strength of signals varies by type of macro forecast (i.e., CPI, GDP, unemployment) and depends on the speaker. This enables us to examine the heterogeneity that officials convey through their speeches. CPI and GDP forecast revisions tend to be larger than the unemployment signals. This is not surprising as the unemployment is a state variable and, thus, tends to be slower moving.

The length of service of the speaker obviously influences the minimum and maximum surprises. Lael Brainard, Richard W. Fisher, and Richard Clarida have the largest CPI revisions. Lael Brainard, Randal Quarles and Patrick Harker have the smallest. The specific sample has some effect too. Downside GDP revisions tend to be larger for those serving at the start of our sample, which includes the Global Financial Crisis. Though we don't report the speech lengths for brevity, Chairs and Vice Chairs of the Board of Governors, i.e., Jerome H. Powell, Janet L. Yellen and Richard H. Clarida, tend to deliver longer speeches.

7 Effects of Speeches on Intraday Market Volatility and Tail Risk

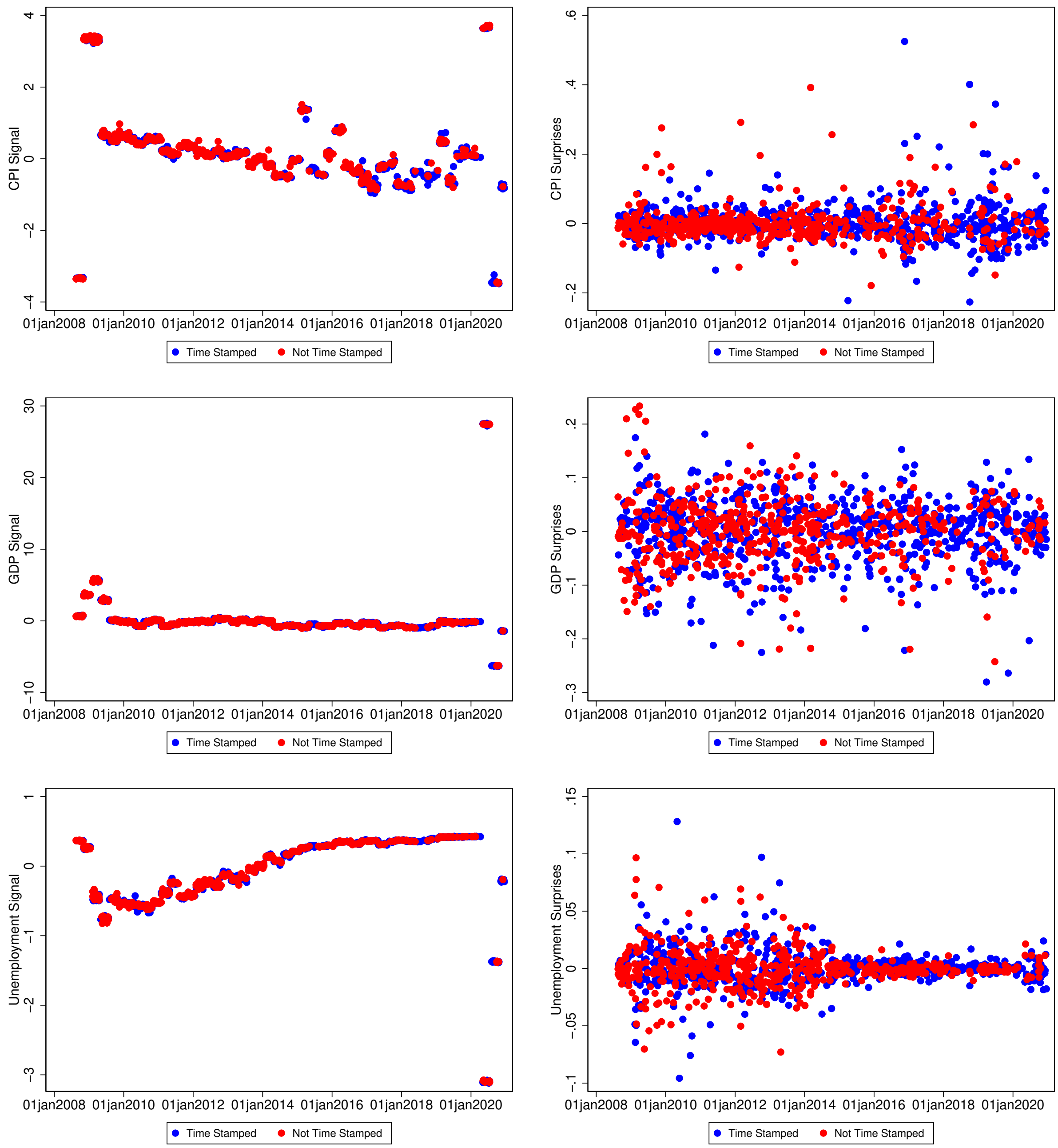
In this section, we investigate the impact of speeches on post-event market volatility and tail risk by regressing the latter variables on speech-implied forecast revisions. We first show our unconditional estimation results in Section 7.1.[22] In Sections 7.2 and 7.3, we explore the effect of the speeches by the Fed Chair and the sign of forecast revisions, respectively. Section 7.4 presents the characteristics of speech signals from hawkish versus dovish FOMC members. In Section 7.5, we examine how speech-implied forecast revisions influence volatility and tail risk when macro variables are low, normal, and high.

7.1 Average News Effects

In Section 4.2.2, we define our volatility and tail risk measures: RV and TR. We regress realized volatility (RV_s) and tail risk (TR_s) in the 30-minute window after a speech on all absolute speech-implied forecast

[22] By "unconditional" we mean that the results are aggregated across all GDP and CPI regimes.

Figure 3: Speech revision signals and implied revision surprises across time by macro variable



Notes: The figure shows raw speech signals across time (left) and the implied surprises (right) for three macro news factors (CPI, GDP and unemployment). The filled blue and red circles indicate the signals that are time stamped and not time stamped, respectively. Equations (7) and (8) show the construction of raw speech signals and implied surprises, respectively. The patterns are generated based on the full sample.

Table 5: Summary statistics of the speech surprises $\nu_{y,s}$

		CPI			GDP			U		
FOMC Member	Role	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Bernanke	Board	-.07	0	.15	-.21	0	.14	-.1	0	.07
Bowman	Board	-.06	0	.05	-.28	0	.09	-.01	0	.01
Brainard	Board	-.22	0	.52	-.22	0	.1	-.02	0	.02
Clarida	Board	-.14	0	.34	-.26	0	.13	-.02	0	.01
Duke	Board	-.09	0	.1	-.23	0	.17	-.06	0	.13
Fischer	Board	-.11	0	.1	-.1	0	.15	-.03	0	.02
Fisher	DAL	-.06	0	.39	-.22	0	.15	-.05	0	.1
Harker	PHL	-.18	0	.28	-.16	0	.08	-.01	0	.02
Kaplan	DAL	-.15	0	.17	-.24	0	.06	-.01	0	.01
Kohn	Board	-.03	0	.03	-.11	0	.12	-.05	0	.06
Kroszner	Board	-.03	0	.05	-.15	0	.15	-.02	0	.01
Lockhart	ATL	-.1	0	.26	-.13	0	.16	-.07	0	.07
Pianalto	DAL	-.13	0	.2	-.18	0	.23	-.05	0	.07
Plosser	PHL	-.06	0	.28	-.13	0	.11	-.07	0	.04
Powell	Board	-.12	0	.4	-.11	0	.12	-.03	0	.02
Quarles	Board	-.23	0	.16	-.2	0	.13	-.02	0	.01
Raskin	Board	-.03	0	.14	-.13	0	.18	-.04	0	.03
Stein	Board	-.05	0	.04	-.09	0	.12	-.02	0	.03
Tarullo	Board	-.09	0	.13	-.18	0	.1	-.08	0	.03
Warsh	Board	-.06	0	.06	-.15	0	.07	-.03	0	.05
Yellen	Board	-.17	0	.25	-.22	0	.12	-.04	0	.05

Notes: The table reports the summary statistics of the speech signals. The table presents the names of the speakers, their role and the descriptive statistics (min, mean, max) of the implied speech signals for each macro revision factor (CPI, GDP, U).

revisions across all regimes to get an average effect, that is,

$$Y_s = \beta_0 + \beta_1 Y_{s,pre} + \beta_2|\nu_{\pi,s}| + \beta_3|\nu_{g,s}| + \beta_4|\nu_{u,s}| + \epsilon_Y. \tag{9}$$

The use of the absolute value means that the $\beta_2 - \beta_4$ coefficients capture whether *the magnitude* of forecast news shocks tends to add or reduce short-term market volatility or tail risk after a news-revealing speech. For example, $\beta_2 > 0$ means that a larger inflation forecast surprise, regardless of whether it is for higher or lower inflation, gives rise to greater volatility. A priori, larger (absolute) forecast revision news might be expected to raise volatility and tail risk. However, if the market was confused about the central bank's thinking on economic conditions, a larger signal could reassure them that the central bank is not making an error and reduce volatility and/or tail risk as in Cieslak and McMahon (2023).

Our first specification considers whether the absolute values of revisions to implied forecasts of CPI, GDP and U predict RV and TR. Preliminary analysis indicated that lagged RV and TR predicted later RV and TR, and that speeches by the Chair were associated with unusually high volatility. Therefore, we always include the pre-speech value of the dependent variable (RV or TR) and a dummy for speeches by the Chair, $D(Chair)$. These variables are more-or-less always statistically significant. To account for the fact that the 4 equations for RV/TR might be related to each other, we estimated a seemingly unrelated regressions (SUR) specification for each 4-equation system for added efficiency.

Table 6 shows the results of each of the two SUR systems. Within each system, most coefficients on CPI and U revisions were negative, of similar magnitude, but usually not statistically significant. The point estimates were consistent with larger revisions reducing volatility and tail risk. Coefficients on GDP revisions were not statistically significant and of varying sign and magnitude. All of the lagged dependent variables and Chair indicators were statistically significant and positive. That is, volatility and tail risk tend to rise during Chair speeches.

Because most of the coefficients on CPI and U revisions were of the same sign and similar magnitude, we considered a pooled version of the model that would restrict the coefficients on each of the three types of forecast revision in each of the four markets to be the same for RV or TR. That is, instead of estimating 20 coefficients plus 4 constants to predict RV in four markets, we will estimate 5 coefficients plus 4 constants to predict RV in four markets.

Whether the coefficients should be pooled—especially across equity and bond RV/TR—is an empirical question. There is no obvious reason that equity RV or TR should behave differently than those of bonds and, in practice, these variables do not seem to behave very differently across markets. Pooling coefficients over both bonds and equities seems to fit the data reasonably well.

Table 7 presents the results of estimating these constrained relations in a SUR framework by maximum likelihood (ML). The BIC narrowly prefers the smaller, pooled models (Table 7) to the unpooled results in Table 6. That is, BIC for the 4-equation RV model in Table 6 was 2566.2, while the BIC for the analogous pooled model in Table 7 was 2557.2. Similarly, the BIC for the unpooled and pooled models for tail risk were 3275.6 and 3275.3.

In contrast to the unpooled results in Table 6, CPI revisions in Table 7 carry negative and significant

Table 6: The impact of forecast revisions on volatility and tail risk

	RV_s				TR_s			
	(1) 2Y	(2) 5Y	(3) 10Y	(4) Eq	(5) 2Y	(6) 5Y	(7) 10Y	(8) Eq
$DepVar_{s,pre}$	0.435	0.417	0.383	0.846	0.409	0.091	0.090	0.216
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\lvert\nu_{\pi,s}\rvert$	-0.048	-0.018	-0.024	-0.015	-0.084	-0.050	-0.026	-0.011
	(0.16)	(0.31)	(0.28)	(0.21)	(0.03)	(0.13)	(0.24)	(0.33)
$\lvert\nu_{g,s}\rvert$	0.034	-0.022	0.009	-0.011	0.021	0.018	0.006	-0.025
	(0.22)	(0.33)	(0.45)	(0.35)	(0.37)	(0.35)	(0.45)	(0.31)
$\lvert\nu_{u,s}\rvert$	-0.080	-0.002	0.001	-0.023	-0.044	-0.062	-0.011	-0.056
	(0.06)	(0.45)	(0.55)	(0.16)	(0.16)	(0.10)	(0.37)	(0.13)
$D(Chair)$	0.161	0.222	0.219	0.086	0.163	0.297	0.275	0.173
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.290	0.239	0.458	0.093	0.347	0.360	0.280	0.135
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.10)
R^2	0.31	0.40	0.29	0.80	0.22	0.13	0.08	0.11
Obs	348	348	348	348	344	344	344	344
BIC				2566.2				3275.6

Notes: The table presents the results of 2 SUR systems—one for RV, one for TR—in which 4 measures of RV and 4 measures of TR were regressed on the lagged dependent variable, an indicator for a speech by the Chair and the absolute revisions to CPI, GDP and unemployment forecasts. The regressors were normalized by dividing by their standard deviations. P-values, in parentheses, were calculated with bootstrapping. The figures for the BIC pertain to the 4-equation RV and 4-equation TR systems, respectively. The sample is the whole sample.

coefficients for predicting TR, and revisions to U carry negative and significant coefficients for both RV and TR. Larger CPI and U revisions reduce tail risk to a statistically significant degree. The coefficient on the Chair indicator is again large, positive and significant, meaning that speeches by the Chair are associated with higher RV and TR.

7.2 Delving Deeper: To Chair, or Not to Chair?

As noted earlier, FOMC Chairs tend give long speeches that produce relatively large forecast revisions. Moreover, the Chair may strongly influence the overall thinking of the FOMC members, potentially by choosing and influencing the Federal Reserve staff who develop the Tealbook forecasts. This would suggest that market reactions place more weight on the Chair's speeches.

On the other hand, the Chair's speeches might simply produce larger speech-implied forecast revisions. Figure 4 plots the kernel density of the speech revision surprises but splits out Chair FOMC speeches from

Table 7: The impact of forecast revisions on volatility and tail risk using a pooled specification

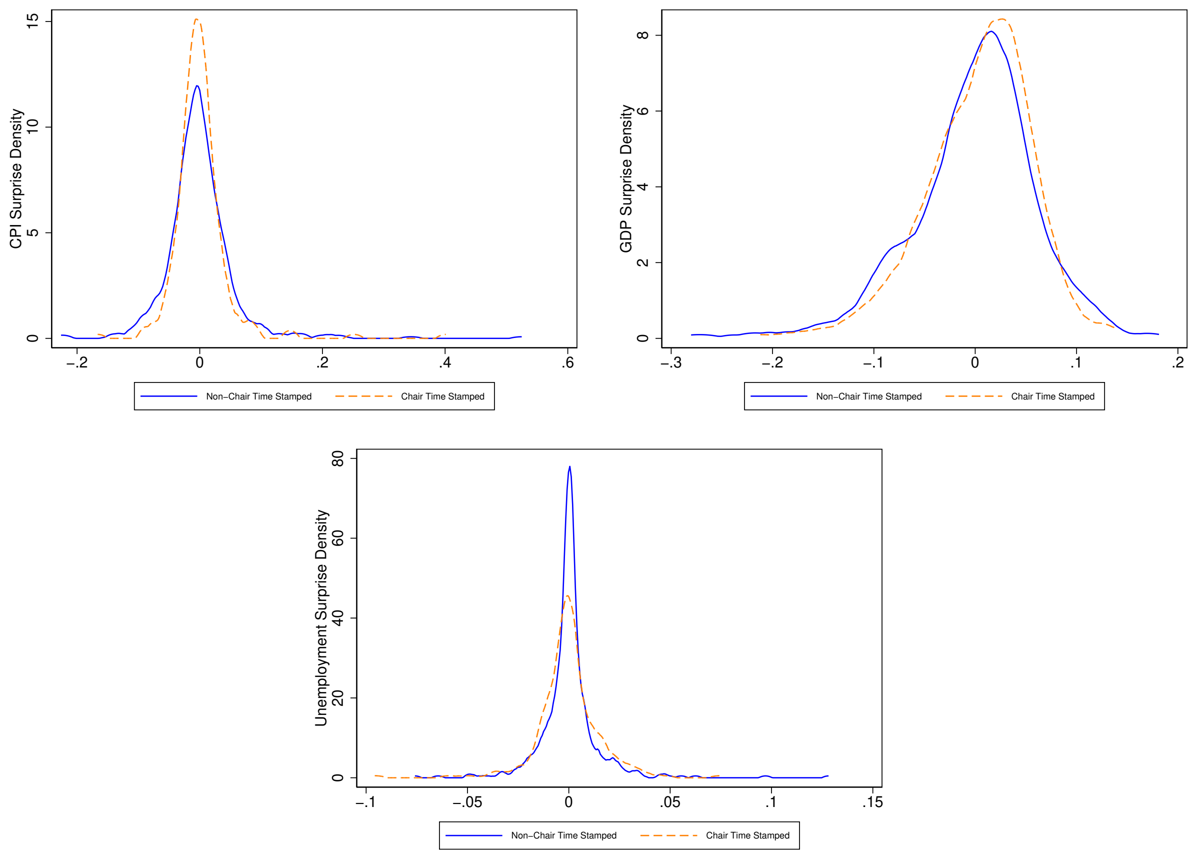
	RV_s (pooled)	TR_s (pooled)
$DepVar_{s,pre}$	0.647	0.126
	(0.00)	(0.00)
$\lvert\nu_{\pi,s}\rvert$	-0.018	-0.044
	(0.14)	(0.04)
$\lvert\nu_{g,s}\rvert$	-0.006	0.012
	(0.39)	(0.33)
$\lvert\nu_{u,s}\rvert$	-0.030	-0.049
	(0.08)	(0.07)
$D(Chair)$	0.12	0.22
	(0.00)	(0.00)
Constant (2YR)	0.166	0.675
	(0.00)	(0.00)
Constant (5YR)	0.317	0.411
	(0.00)	(0.00)
Constant (10YR)	0.465	0.331
	(0.00)	(0.00)
Constant (eq)	0.101	0.191
	(0.01)	(0.00)
R^2 (2YR)	0.30	0.14
R^2 (5YR)	0.40	0.13
R^2 (10YR)	0.28	0.06
R^2 (eq)	0.79	0.06
Obs	348	344
BIC	2557.2	3275.3

Notes: The table presents the results of 2 SUR systems—one for RV, one for TR—in which 4 measures of RV and 4 measures of TR were regressed on the lagged dependent variable, an indicator for a speech by the Chair and the absolute revisions to CPI, GDP and unemployment forecasts. The regressors were normalized by dividing by their standard deviations. The estimates were pooled within each system, requiring each coefficient on the dynamic regressors to be the same across the 4 equations. P-values, in parentheses, were calculated with bootstrapping. The figures for the BIC pertain to the 4-equation RV and 4-equation TR systems, respectively. The sample is the whole sample.

non-Chair speeches (limiting the analysis to the speeches that have time-stamps for our analysis). The distributions of CPI and GDP forecast revisions from non-Chair speeches are more variable than those for Chair speeches and appear to have thicker tails. In contrast, the distribution of unemployment forecast revisions for Chair speeches are more variable than those of non-Chair speeches, as well as possibly slightly thicker tails.

Then, how does the impact of forecast revisions differ for Fed Chair speeches? As previously noted, markets consider speeches by the Chair to be especially important in driving Fed policy even though Chair

Figure 4: Chair vs Non-Chair Speech revision surprises by macro variable



Notes: The figure shows the kernel density of the implied surprises for three macro news factors (CPI, GDP and unemployment). The blue solid lines show the density of the non-Chair members while the orange dashed lines indicate the surprises of Chairs. All densities only include speeches that are time stamped.

has only one vote on the Federal Open Market Committee. The consistent importance of the Chair-speech indicator in Tables 6 and 7 confirm that Chair speeches tend to be associated with higher RV and TR than non-Chair speeches.

To investigate the marginal effect of the effect of forecast revisions created by Fed-Chair speeches, we divide CPI, GDP and unemployment forecast revisions into those created by a Chair speech or a Not-Chair speech. We again estimate the pooled RV and TR SUR systems by maximum likelihood.

Table 8 reports the results of this estimation. Chair interactions with forecast revisions produce negative and significant coefficients for all three types of revisions and the point estimates for "Chair" coefficients are always more negative than the "NotChair" coefficients. That is, while chair speeches themselves raise RV and TR, any forecast revisions implied by those speeches tend to reduce RV and TR. Of the "NotChair" coefficients, only that on the GDP revision was statistically significant. This GDP revision coefficient was positive, indicating that larger GDP revisions by non-Chair speakers was associated with higher TR.

Overall, our analysis can be viewed in parallel with the study of Swanson and Jayawickrema (2023) that also shows that markets often respond differently to Chair speeches. Because Chair speeches seem

Table 8: Impacts of forecast revisions interacted with the Fed Chair indicator on volatility and tail risk using a pooled specification

	RV_s (pooled)	TR_s (pooled)
$DepVar_{s,pre}$	0.645	0.121
	(0.00)	(0.00)
$\lvert\nu_{\pi,s}\rvert \times NotChair$	-0.004	-0.021
	(0.40)	(0.23)
$\lvert\nu_{\pi,s}\rvert \times Chair$	-0.048	-0.091
	(0.09)	(0.04)
$\lvert\nu_{g,s}\rvert \times NotChair$	0.013	0.044
	(0.30)	(0.08)
$\lvert\nu_{g,s}\rvert \times Chair$	-0.087	-0.137
	(0.05)	(0.02)
$\lvert\nu_{u,s}\rvert \times NotChair$	-0.021	-0.039
	(0.19)	(0.15)
$\lvert\nu_{u,s}\rvert \times Chair$	-0.052	-0.079
	(0.10)	(0.06)
$D(Chair)$	0.19	0.34
	(0.00)	(0.00)
Constant (2YR)	0.127	0.614
	(0.01)	(0.00)
Constant (5YR)	0.278	0.348
	(0.00)	(0.00)
Constant (10YR)	0.426	0.268
	(0.00)	(0.00)
Constant (eq)	0.063	0.129
	(0.06)	(0.01)
R^2 (2YR)	0.31	0.14
R^2 (5YR)	0.40	0.13
R^2 (10YR)	0.29	0.07
R^2 (eq)	0.80	0.07
Obs	348	344
BIC	2568.4	3283.4

Notes: The table presents the results of 2 SUR systems—one for RV, one for TR—in which 4 measures of RV and 4 measures of TR were regressed on the lagged dependent variable, an indicator for a speech by the Chair and the absolute revisions to CPI, GDP and unemployment forecasts, as well as the interaction to those revisions with an indicator for Chair speeches. The regressors were normalized by dividing by the standard deviation of the underlying series. For example, the series for CPI revisions interacted with Chair speeches was divided by the standard deviation of all CPI revisions. The estimates were pooled within each system, requiring each coefficient on the dynamic regressors to be the same across the 4 equations. P-values, in parentheses, were calculated with bootstrapping. The figures for the BIC pertain to the 4-equation RV and 4-equation TR systems, respectively.

to be particularly important, we will break down our further results into 2 sets: results based on the full sample and results based only on Chair speeches. Regressions involving only the Chair speeches will not include a Chair dummy because it and the constant would not be separately identified.

7.3 Asymmetric Response: Sign of the Shock

How does asymmetry in the sign of implied forecast revisions predict RV and TR? It is possible that positive and negative absolute forecast revisions might have different effects on market volatility and TR. Therefore, we allow the positive and negative forecast revisions to enter the system with different signs and again estimate system with maximum likelihood. Specifically, we estimate a version of equation (9) but include an interaction between the forecast revisions and indicators for positive-negative revisions.

Table 9 presents those asymmetric results for the full sample on the left and the Chair speeches on the right. Most point estimates and all the significant coefficients are consistent with the hypothesis that larger revisions tend to reduce volatility and tail risk. Negative revisions to CPI and unemployment forecasts are larger and more significant than positive surprises to those variables. Coefficients associated with Chair speeches (right panel) tend to be larger and more statistically significant than those based on the full sample (left panel). As in the previous specifications, speeches by the Fed Chair are associated with higher volatility and tail risk (left panel).

7.4 Hawks versus Doves

We next examine the effects of monetary policy views on the impact of forecast revisions. For brevity, we will henceforth refer to monetary policy views as "hawkishness," although our treatment of hawkish and dovish views is symmetric.

We classify the monetary policy views of each speaker on a five-point scale running from dove (-1), to dove/centrist (-0.5), centrist (0), hawk/centrist (0.5), to hawk (1). Descriptions from multiple sources in the financial press inform our classifications, including those from Reuters, Financial Times, Business Insider, Deutsche Bank, Marketplace and Mitsubishi UFJ Financial Group, Inc. (MUFG). Our approach is similar to that of Bordo and Istrefi (2023). According to these sources, the groups are:

- **Dovish** Stanley Fischer, Ben S. Bernanke, Daniel K. Tarullo, Janet L. Yellen and Lael Brainard.
- **Dovish/Centrist** Dennis Lockhart, Jeremy C. Stein, Jerome H. Powell
- **Centrist** Michelle W. Bowman, Patrick T. Harker, Randal K. Quarles, Sandra Pianalto
- **Hawkish/Centrist** Richard H. Clarida
- **Hawkish** Richard W. Fisher, Robert S. Kaplan, Charles I. Plosser

The construction of the $\nu_{y,s}$ surprises adjusts for member-fixed effects, which controls for differences in mean forecast signals between the groups.[23]

We interacted our measure of hawkishness with each of the three types of forecast revisions and again estimated pooled SUR systems by maximum likelihood.

[23] For instance, the raw GDP signals are, on average, more positive as our hawkish index index increases. But there is no difference in the $\nu_{y,s}$ revisions.

Table 9: Asymmetric impacts of positive/negative forecast revisions on volatility and tail risk using a pooled specification

	Whole Sample		Sample of Chair Speeches	
	RV_s (pooled)	TR_s (pooled)	RV_s (pooled)	TR_s (pooled)
$DepVar_{s,pre}$	0.647	0.127	0.599	0.115
	(0.00)	(0.00)	(0.00)	(0.02)
$\|\nu_{\pi,s}\| \times I(cpi \leq 0)$	-0.059	-0.094	-0.389	-0.559
	(0.08)	(0.05)	(0.02)	(0.01)
$\|\nu_{\pi,s}\| \times I(cpi > 0)$	-0.014	-0.042	-0.047	-0.084
	(0.20)	(0.09)	(0.16)	(0.10)
$\|\nu_{g,s}\| \times I(gdp \leq 0)$	-0.006	0.010	-0.027	-0.058
	(0.38)	(0.40)	(0.37)	(0.24)
$\|\nu_{g,s}\| \times I(gdp > 0)$	-0.007	0.006	-0.002	-0.005
	(0.39)	(0.50)	(0.50)	(0.48)
$\|\nu_{u,s}\| \times I(u \leq 0)$	-0.057	-0.094	-0.126	-0.219
	(0.04)	(0.03)	(0.09)	(0.05)
$\|\nu_{u,s}\| \times I(u > 0)$	-0.010	-0.020	-0.013	0.003
	(0.31)	(0.31)	(0.36)	(0.51)
$D(Chair)$	0.119	0.217		
	(0.00)	(0.00)		
Constant (2YR)	0.183	0.704	0.639	1.425
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (5YR)	0.334	0.439	0.904	1.240
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (10YR)	0.482	0.359	1.059	1.125
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (eq)	0.119	0.220	0.512	0.864
	(0.00)	(0.00)	(0.00)	(0.00)
R^2 (2YR)	0.31	0.15	0.37	0.06
R^2 (5YR)	0.40	0.13	0.26	0.06
R^2 (10YR)	0.29	0.06	0.12	0.01
R^2 (eq)	0.79	0.06	0.65	0.04
Obs	348	344	128	127

Notes: Each panel of the table presents the results of 2 SUR systems—one for RV, one for TR—in which 4 measures of RV and 4 measures of TR were regressed on the lagged dependent variable, an indicator for a speech by the Chair (left panel) and the absolute revisions to CPI, GDP and unemployment forecasts. The latter were split into their positive and negative components. The regressors were normalized by dividing by the standard deviation of the underlying series. For example, the positive CPI revisions were divided by the standard deviation of all CPI revisions. The estimates were pooled within each system, requiring each coefficient on the dynamic regressors to be the same across the 4 equations. P-values, in parentheses, were calculated with bootstrapping. The left-hand panel shows results for the full sample while the right-hand panel shows results for speeches by the Chair.

Table 10: Impacts of forecast revisions interacted with the hawkishness indicator on volatility and tail risk using a pooled specification

	Whole Sample		Sample of Chair Speeches	
	RV_s (pooled)	TR_s (pooled)	RV_s (pooled)	TR_s (pooled)
$DepVar_{s,pre}$	0.645	0.125	0.599	0.116
	(0.00)	(0.00)	(0.00)	(0.02)
$\lvert\nu_{\pi,s}\rvert$	-0.031	-0.048	0.105	-0.014
	(0.15)	(0.15)	(0.25)	(0.44)
$\lvert\nu_{\pi,s}\rvert \times Hawk$	-0.022	-0.010	0.208	0.106
	(0.26)	(0.39)	(0.13)	(0.32)
$\lvert\nu_{g,s}\rvert$	0.017	0.035	-0.415	-0.691
	(0.33)	(0.24)	(0.22)	(0.20)
$\lvert\nu_{g,s}\rvert \times Hawk$	0.038	0.042	-0.382	-0.636
	(0.20)	(0.27)	(0.23)	(0.22)
$\lvert\nu_{u,s}\rvert$	-0.023	-0.037	0.574	4.588
	(0.22)	(0.20)	(0.46)	(0.12)
$\lvert\nu_{u,s}\rvert \times Hawk$	0.017	0.024	0.633	4.670
	(0.40)	(0.37)	(0.46)	(0.12)
$D(Chair)$	0.13	0.23		
	(0.00)	(0.00)		
Constant (2YR)	0.165	0.674	0.59	1.31
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (5YR)	0.316	0.410	0.856	1.128
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (10YR)	0.465	0.330	1.010	1.013
	(0.00)	(0.00)	(0.00)	(0.00)
Constant (eq)	0.101	0.192	0.458	0.754
	(0.00)	(0.00)	(0.00)	(0.00)
R^2 (2YR)	0.30	0.14	0.35	0.09
R^2 (5YR)	0.40	0.13	0.24	0.02
R^2 (10YR)	0.28	0.06	0.11	0.00
R^2 (eq)	0.80	0.06	0.65	0.03
Obs	348	344	128	127

Notes: Each panel of the table presents the results of 2 SUR systems—one for RV, one for TR—in which 4 measures of RV and 4 measures of TR were regressed on the lagged dependent variable, an indicator for a speech by the Chair (left panel only) and the absolute revisions to CPI, GDP and unemployment forecasts, as well as the interaction to those revisions with a measure of “hawkishness" of their monetary policy views. The regressors were normalized by dividing by the standard deviation of the underlying series. For example, the series for CPI revisions interacted with hawkishness was divided by the standard deviation of all CPI revisions. The estimates were pooled within each system, requiring each coefficient on the dynamic regressors to be the same across the 4 equations. P-values, in parentheses, were calculated with bootstrapping. The left-hand panel shows results for the full sample while the right-hand panel shows results for speeches by the Chair.

Table 10 shows that hawkishness interactions are not significant with the pooled model. None of the forecast-revision coefficients are statistically significant. The hawkishness of a speaker's views does not seem to systematically influence the impact of forecast revisions on RV or TR.

Model selection. It is worth highlighting that the BIC for the simple, pooled RV and TR models in Table 7—the most parsimonious models—are generally smaller than the BIC for the larger models presented in Tables 8 through Table 10. This is not surprising as it is well known that the BIC prefers the smaller model for describing the data. We continue to present the results in Tables 8 through Table 10, however, because we think that they do convey useful information. For example, Table 9 shows that negative forecast revisions to CPI and U forecasts have larger and more statistically significant effects on both volatility and tail risk. Similarly, Table 8 shows that the revisions from the Chair are consistently large and statistically significant while those from other FOMC members are consistently not significant. Finally, Table 10 shows that hawkishness doesn't matter.

In sum, pooled systems produce a great deal of evidence that several factors consistently predict post-speech RV and TR. These factors include the lag of the dependent variable, whether the speech is by the Fed Chair, and several characteristics of forecast revisions, including the sign of the forecast revision and whether a Fed Chair speech produced the forecast revision. Speeches by the Fed Chair are associated with higher RV and TR in any specification, while larger (in absolute value) forecast revisions induced by those speeches consistently reduce RV and TR. Negative CPI and U forecast revisions tend to influence RV and TR more than do positive forecast revisions. In contrast, there is no evidence that a speaker's hawkiness/dovishness influenced the impact of the forecast revision on RV and TR.

7.5 Regime-Specific Effects

To complete our empirical analysis, we ask whether the state of the economy influences the financial market effects of speech-implied forecast revisions.[24] That is, we investigate whether a speech-implied forecast revision of inflation, GDP or unemployment has a different effect if the underlying macroeconomic variable is unusually high or low. For this assessment, we categorise each observation on CPI, GDP and the unemployment rate as a *high*, *normal (medium)*, or *low* regime observation. Table 11 outlines the values that determine each classification, while Figure 5 shows the time-series of the regime indicators.

To explore the regime specificity of the results, we regress post-speech RV and TR on the values of the dependent (pre-speech) variable and absolute CPI, GDP and unemployment forecast revisions with the samples split by the regime indicators. In other words, we estimate separate SUR systems for observations

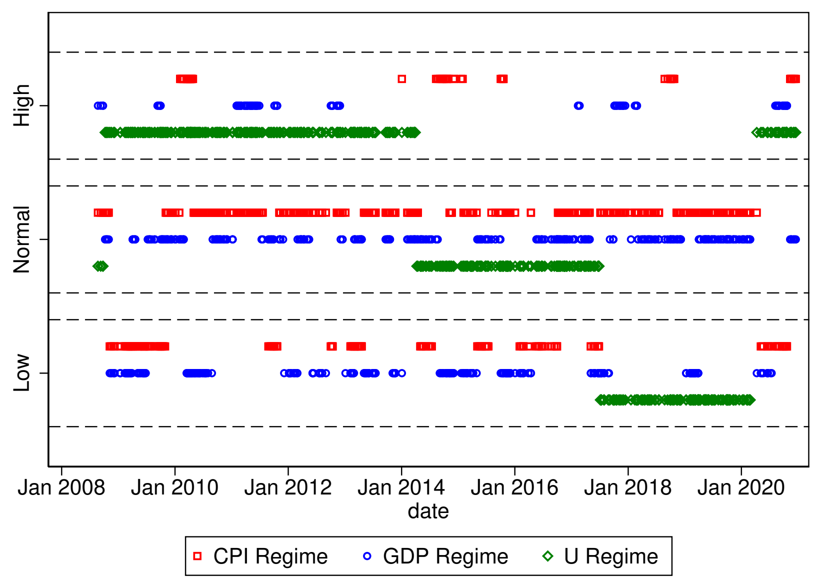
[24] Both theory and empirical evidence suggests that monetary policy is regime dependent. For instance, the empirical results of Tenreyro and Thwaites (2016) suggest that medium- to long-run monetary policy shock effects on the real economy strongly depend on the state of the business cycle. GDP growth is the most consistent factor determining monetary policy effectiveness, and shocks seem to have a more pronounced effect during economic upswings than during downswings. This suggests that monetary policy effects might be subdued during recessions. In related work, Mandler (2012) uses a threshold vector autoregression (VAR) framework to analyse the effectiveness of classical monetary policy shocks, depending on the respective inflationary regime in the US between 1965-2007. The findings of Mandler (2012) indicate that monetary policy shocks have markedly different effects in low and high inflation regimes. To the best of our knowledge, we are the first to investigate regime dependence — with regards to both inflation and GDP growth — of the impact of central bank communication of economic signals.

Table 11: Categories of economic regimes

	CPI	ΔGDP	Unemployment Rate
High	$\pi > 3.9\%$	$g > 3.6\%$	$u > 6.4\%$
Normal	$1.2\% < \pi < 3.9\%$	$1.4\% < g < 3.6\%$	$4.3\% < u < 6.4\%$
Low	$\pi < 1.2\%$	$g < 1.4\%$	$u < 4.3\%$

Notes: The table presents the classification of different economic regimes (high, normal, low) for CPI, GDP and unemployment. The High (Low) regimes are defined by real-time measures of the economic variables being in the upper (lower) quartile of the data over the period August 2008 to December 2020.

Figure 5: Time-series of regimes



Notes: The figure displays the evolution of different economic regimes over time. CPI (red squares), GDP (blue circles) and unemployment (green diamonds).

when CPI inflation is low, medium or high. We do the similar sample splits for low, medium and high GDP growth and unemployment.

The results from these regime-dependent estimations mostly confirmed the patterns noted in the unconditional data. For example, lagged dependent variables and coefficients on Chair indicators were generally significant in all regimes. That is, the level of volatility and tail risk tends to be higher during Chair speeches. We omit the full results for brevity. Intriguingly, however, we noted that the variables interacted with "hawkish" measures often became statistically significant in high and low GDP periods, particularly for Chair speeches, despite the smaller samples.

Table 12 shows the results from the estimation of the pooled "hawkish" model during the sample of Chair speeches. The upper (lower) panel shows the results from the RV (TR) SUR system. In contrast to the unconditional results in Table 10, the hawkish interactions are often statistically significant, although their signs vary depending on the type of forecast revision, i.e., CPI, GDP, or U. Most of the coefficients on the hawkishness interactions are positive, indicating that hawkish forecast revisions are more likely to be associated with increases in RV (upper panel) and TR (lower panel).

In summary, the role of policy view may be regime-specific: forecast revisions interacted with the hawkishness indicator have some statistically significant impact on volatility and tail risk, particularly in periods of low and high GDP regimes. We readily admit, however, that this result is speculative. We leave explorations of reasons for this finding to future research. We report these results and robustness checks in our Supplementary Online Appendix.

8 Conclusion

We introduce a supervised multimodal natural language processing method to map central bank language to forecasts of macroeconomic variables. We benchmark an extensive array of machine learning methods on this task and compare the performance of our proposed model on a dataset of time-stamped speeches from Federal Reserve FOMC members in order to create a novel series of monetary policymakers' implied forecast revisions. These revisions are the differences between central bank speech-implied forecasts and the latest corresponding forecasts from the Survey of Professional Forecasters. To further purge predictable signals, we also control for time and member fixed effects in creating our forecast revision series.

Our multimodal NTM (non-linear) language mappings fit Greenbook forecasts very well in the test period with very high out-of-sample predictive performance. Forecast revisions derived from FOMC-member speeches explain volatility and tail risk in both equity and bond markets.

Our results also indicate that speeches from Fed Chairs, i.e., Powell and Yellen in our sample, tend to produce greater forecast revisions. While Chair speeches are "special" in affecting market volatility and tail risk, we find no strong evidence that specific monetary policy views influence the impact of speech-implied forecast revisions on volatility and tail risk. That is, markets don't react differently to speeches by hawks and doves.

Our analysis also sheds light on the circumstances in which central bank communication calms financial markets and reduces uncertainty. Larger forecast revisions (in absolute value) produced by Chair speeches

Table 12: Regime-specific impacts of forecast revisions interacted with the hawkishness indicator on volatility and tail risk using a pooled specification

Pooled estimates of the RV system

	Low GDP	Medium GDP	High GDP	All observations
$DepVar_{s,pre}$	0.687	0.518	0.561	0.599
	(0.00)	(0.00)	(0.00)	(0.00)
$\lvert\nu_{\pi,s}\rvert$	-0.098	0.098	-0.273	0.105
	(0.22)	(0.38)	(0.02)	(0.26)
$\lvert\nu_{\pi,s}\rvert \times Hawk$	0.298	0.220	-0.541	0.208
	(0.01)	(0.22)	(0.02)	(0.14)
$\lvert\nu_{g,s}\rvert$	0.057	-0.492	-0.004	-0.415
	(0.19)	(0.20)	(0.47)	(0.26)
$\lvert\nu_{g,s}\rvert \times Hawk$	0.143	-0.521	-0.001	-0.382
	(0.01)	(0.19)	(0.54)	(0.26)
$\lvert\nu_{u,s}\rvert$	0.085	1.071	0.072	0.574
	(0.01)	(0.43)	(0.02)	(0.45)
$\lvert\nu_{u,s}\rvert \times Hawk$	0.115	1.184	0.114	0.633
	(0.01)	(0.41)	(0.00)	(0.44)
R^2 (2YR)	0.06	0.56	0.00	0.35
R^2 (5YR)	0.08	0.35	0.07	0.24
R^2 (10YR)	0.07	0.15	0.30	0.11
R^2 (eq)	0.72	0.60	0.80	0.65

Pooled estimates of the TR system

	Low GDP	Medium GDP	High GDP	All observations
$DepVar_{s,pre}$	0.443	0.082	1.727	0.116
	(0.04)	(0.01)	(0.00)	(0.01)
$\lvert\nu_{\pi,s}\rvert$	0.111	-0.067	-0.309	-0.014
	(0.27)	(0.39)	(0.03)	(0.43)
$\lvert\nu_{\pi,s}\rvert \times Hawk$	0.089	0.087	-0.615	0.106
	(0.27)	(0.44)	(0.04)	(0.32)
$\lvert\nu_{g,s}\rvert$	-0.088	-0.973	-0.040	-0.691
	(0.20)	(0.11)	(0.28)	(0.20)
$\lvert\nu_{g,s}\rvert \times Hawk$	0.288	-0.874	0.013	-0.636
	(0.01)	(0.13)	(0.41)	(0.21)
$\lvert\nu_{u,s}\rvert$	0.054	2.557	0.197	4.588
	(0.11)	(0.27)	(0.00)	(0.10)
$\lvert\nu_{u,s}\rvert \times Hawk$	0.146	2.616	-0.013	4.670
	(0.01)	(0.26)	(0.35)	(0.09)
R^2 (2YR)	0.01	0.13	0.15	0.09
R^2 (5YR)	0.01	0.03	0.57	0.02
R^2 (10YR)	0.00	0.01	0.48	0.00
R^2 (eq)	0.12	0.02	0.64	0.03
Obs	37	77	13	127

Notes: The table presents the results of 8 SUR systems—4 for RV and 4 for TR—on samples broken down by low, medium, and high GDP, as well as all observations. The top panel shows results using RV as the dependent variable, while the lower panel shows the results for TR as the dependent variable. Each dependent variable was regressed on the lagged dependent variable, an indicator for a speech by the Chair and the absolute revisions to CPI, GDP and unemployment forecasts in different regimes, those revisions interacted with a variable characterizing the hawkishness of the speaker, and a constant. The regressors were normalized by dividing by their standard deviations. P-values (in parenthesis) were calculated with bootstrapping. The table omits the estimated constants and their p-values for brevity.

tend to reduce volatility and tail risk more than those of other members. We infer that central bank communication may *calm* markets, depending on the message conveyed to markets and if the speech comes from the Chair. The results of Chair effects remain significant, regardless of the economic regimes.

Our findings underpin the importance of analysing the *continuous flow* of central bank communication. In particular, the influence of Chair speeches on tail risk and volatility is consistent with Swanson and Jayawickrema (2023), who document that such speeches have a high market impact. It would be interesting to explore how Fed Chair views impact market uncertainty in both the short/long term and under different market conditions, a direction we leave for future research.

References

Ahrens, M., Ashwin, J., Calliess, J.P., Nguyen, V., 2021. Bayesian topic regression for causal inference, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8162–8188.

Ahrens, M., McMahon, M., 2021. Extracting economic signals from central bank speeches, in: Proceedings of the Third Workshop on Economics and Natural Language Processing, pp. 93–114.

Aït-Sahalia, Y., Jacod, J., 2009. Estimating the degree of activity of jumps in high-frequency data. The Annals of Statistics 37, 2202–2244.

Andersen, T., Fusari, N., Todorov, V., 2020. The pricing of tail risk and the equity premium: Evidence from international option markets. Journal of Business and Economic Statistics 38, 662–678.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H., 2001. The distribution of realized stock return volatility. Journal of Financial Economics 61, 43–76.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71, 579–625.

Aruoba, B., Drechsel, T., 2022. Identifying monetary policy shocks: A natural language approach. CEPR Discussion Papers 17133. C.E.P.R. Discussion Papers.

Barndorff-Nielsen, O.E., Shephard, N., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 253–280.

Bauer, M.D., Lakdawala, A., Mueller, P., 2022. Market-based monetary policy uncertainty. The Economic Journal 132, 1290–1308.

Bekaert, G., Hoerova, M., Duca, M.L., 2013. Risk, uncertainty and monetary policy. Journal of Monetary Economics 60, 771–788.

Bernanke, B., Boivin, J., Eliasz, P.S., 2005. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. The Quarterly Journal of Economics 120, 387–422.

Bernanke, B.S., Kuttner, K.N., 2005. What explains the stock market's reaction to Federal Reserve Policy? Journal of Finance 60, 1221–1257.

Bholat, D., Hansen, S., Santos, P., Schonhardt-Bailey, C., 2015. Text mining for central banks. Available at SSRN 2624811.

Blinder, A.S., 2018. Through a crystal ball darkly: The future of monetary policy communication. AEA Papers and Proceedings 108, 567–71.

Bollerslev, T., Li, J., Xue, Y., 2018. Volume, volatility, and public announcements. The Review of Economic Studies 85, 2005–2041.

Bollerslev, T., Todorov, V., Xu, L., 2015. Tail risk premia and return predictability. Journal of Financial Economics 118, 113–134.

Bordo, M., Istrefi, K., 2023. Perceived FOMC: The making of hawks, doves and swingers. Journal of Monetary Economics 136, 125–143.

Boswijk, H.P., Laeven, R.J.A., Yang, X., 2018. Testing for self-excitation in jumps. Journal of Econometrics 203, 256–266.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Byrne, D., Goodhead, R., McMahon, M., Parle, C., 2023. The central bank crystal ball: Temporal information in monetary policy communication. Discussion Papers 17930. Centre for Economic Policy Research.

Caballero, R.J., Simsek, A., 2022. Monetary policy with opinionated markets. American Economic Review 112, 2353–92.

Cai, Y., Camara, S., Capel, N., 2021. It's not always about the money, sometimes it's about sending a message: Evidence of Informational Content in Monetary Policy Announcements. Working Paper.

Card, D., Tan, C., Smith, N.A., 2018. Neural models for documents with metadata, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics. pp. 2031–2040.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Child, R., Gray, S., Radford, A., Sutskever, I., 2019. Generating long sequences with sparse transformers. Technical Report.

Cieslak, A., Hansen, S., McMahon, M., Xiao, S., 2023. Policymakers' uncertainty. Mimeograph.

Cieslak, A., McMahon, M., 2023. Tough talk: The Fed and the risk premia. Mimeograph.

Cieslak, A., Schrimpf, A., 2019. Non-monetary news in central bank communication. Journal of International Economics 118, 293–315.

Coibion, O., Gorodnichenko, Y., 2012. Why are target interest rate changes so persistent? American Economic Journal: Macroeconomics 4, 126–162.

Das, S., Giggins, C., He, J., Karypis, G., Krishnamurthy, S., Mahajan, M., Prabhala, N., Slack, D., van Dusen, R., Yue, S., Zha, S., Zheng, S., 2021. Context, language modeling, and multimodal data in finance. The Journal of Financial Data Science Summer.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL HLT, 4171–4186.

Dudani, S.A., 1976. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics SMC-6, 325–327.

Dungey, M., Erdemlioglu, D., Matei, M., Yang, X., 2018. Testing for mutually exciting jumps and financial flights in high frequency data. Journal of Econometrics 202, 18–44.

Ehrmann, M., Gaballo, G., Hoffmann, P., Strasser, G., 2019. Can more public information raise uncertainty? The international evidence on forward guidance. Journal of Monetary Economics 108, 93–112.

Ehrmann, M., Talmi, J., 2020. Starting from a blank page? Semantic similarity in central bank communication and market volatility. Journal of Monetary Economics 111, 48–62.

Ellen, S.T., Larsen, V.H., Thorsrud, L.A., 2022. Narrative monetary policy surprises and the media. Journal of Money, Credit and Banking 54, 1525–1549.

Erdemlioglu, D., Yang, X., 2022. News arrival, time-varying jump intensity, and realized volatility: Conditional testing approach. Journal of Financial Econometrics (nbac015), 1–38.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A., 2020. Autogluon-tabular: Robust and accurate automl for structured data.

Frisch, R., Waugh, F.V., 1933. Partial time regressions as compared with individual trends. Econometrica 4, 387–401.

Gardner, B., Scotti, C., Vega, C., 2022. Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. Journal of Econometrics 231, 387–409.

Gertler, M., Karadi, P., 2015. Monetary policy surprises, credit costs, and economic activity. American Economic Journal: Macroeconomics 7, 44–76.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning 63, 3–42.

Gómez-Cram, R., Grotteria, M., 2022. Real-time price discovery via verbal communication: Method and application to Fedspeak. Journal of Financial Economics 143, 993–1025.

Gurkaynak, R.S., Sack, B.P., Swanson, E.T., 2005. Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. International Journal of Central Banking 1.

Gáti, L., Handlan, A., 2022. Monetary communication rules. Working Paper Series 2759, European Central Bank.

Handlan, A., 2020. Text shocks and monetary surprises: Text analysis of FOMC statements with machine learning. Working Paper.

Hansen, S., McMahon, M., 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. Journal of International Economics 99, 114–133.

Hansen, S., McMahon, M., Prat, A., 2018. Transparency and deliberation within the FOMC: A computational linguistics approach. The Quarterly Journal of Economics 133, 801–870.

Hansen, S., McMahon, M., Tong, M., 2019. The long-run information effect of central bank communication. Journal of Monetary Economics 108, 185–202.

Hanson, S.G., Stein, J.C., 2015. Monetary policy and long-term real rates. Journal of Financial Economics 115, 429–448.

Hattori, M., Schrimpf, A., Sushko, V., 2016. The response of tail risk perceptions to unconventional monetary policy. American Economic Journal: Macroeconomics 8, 111–136.

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. The Annals of Statistics 3, 1163–1174.

Husted, L., Rogers, J., Sun, B., 2020. Monetary policy uncertainty. Journal of Monetary Economics 115, 20–36.

Jarociński, M., Karadi, P., 2020. Deconstructing monetary policy surprises—The role of information shocks. American Economic Journal: Macroeconomics 12, 1–43.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30.

Leombroni, M., Vedolin, A., Venter, G., Whelan, P., 2021. Central bank communication and the yield curve. Journal of Financial Economics 141, 860–880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., Allen, P.G., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 .

Lovell, M., 1963. Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association 58, 993–1010.

Lucca, D.O., Trebbi, F., 2009. Measuring central bank communication: An automated approach with application to FOMC statements. Technical Report 15367. National Bureau of Economic Research.

Mandler, M., 2012. Inflation-regime dependent effects of monetary policy shocks. Evidence from threshold vector autoregressions. Economics Letters 116, 422–425.

Miranda-Agrippino, S., Ricco, G., 2021. The transmission of monetary policy shocks. American Economic Journal: Macroeconomics 13, 74–107.

Nakamura, E., Steinsson, J., 2018. High-frequency identification of monetary non-neutrality: The information effect. Quarterly Journal of Economics 133, 1283–1330.

Nesbit, J., 2020. Text as instruments. Working Paper.

Neuhierl, A., Weber, M., 2019. Monetary policy communication, policy slope, and the stock market. Journal of Monetary Economics 108, 140–155.

Ozdagli, A., Velikov, M., 2020. Show me the money: The monetary policy risk premium. Journal of Financial Economics 135, 320–339.

Petropoulos, A., Siakoulis, V., 2021. Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique. Central Bank Review 21, 141–153.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31.

Ramey, V.A., 2016. Macroeconomic shocks and their propagation, in: Handbook of Macroeconomics. volume 2, pp. 71–162.

Romer, C.D., Romer, D.H., 2000. Federal Reserve information and the behavior of interest rates. American Economic Review 90, 429–457.

Swanson, E.T., 2021. Measuring the effects of federal reserve forward guidance and asset purchases on financial markets. Journal of Monetary Economics 118, 32–53.

Swanson, E.T., 2023. The importance of Fed Chair speeches as a monetary policy tool. AEA Papers and Proceedings 113, 394–400.

Swanson, E.T., Jayawickrema, V., 2023. Speeches by the Fed Chair are more important than FOMC announcements: An improved high-frequency measure of US monetary policy shocks. Working Paper.

Tenreyro, S., Thwaites, G., 2016. Pushing on a string: US monetary policy is less powerful in recessions. American Economic Journal: Macroeconomics 8, 43–74.

Tillmann, P., 2020. Monetary policy uncertainty and the response of the yield curve to policy shocks. Journal of Money, Credit and Banking 54, 803–833.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A., 2020. Big bird: Transformers for longer sequences, in: Advances in Neural Information Processing Systems.

Appendix

A Description of the Machine Learning Models

In this section, we provide a detailed description of the machine learning models spanned by our base AutoGL framework. These classes are K-nearest neighbours (KNN), Random Forest, Extremely Randomized Trees, Boosted Decision Trees and Neural Networks.

K-nearest neighbours (KNN)

The K-nearest neighbours (KNN) class that we consider is a widely-used machine learning algorithm, belonging to the family of instance-based, non-parametric learning. It operates on the simple principle of feature similarity, assuming that similar data points can be found near each other in feature space. In both classification and regression, KNN works by finding the k closest training samples to a new data point and then predicts the output based on these neighbours. For classification, the algorithm typically assigns the class most common among its k nearest neighbours, while in regression, it usually takes the average of their values. In fact, KNN is easy to implement and understand, but its performance can significantly decline with high-dimensional data (the curse of dimensionality) and large datasets (due to computational cost).

Random Forest

The other machine learning algorithm that we implemented for performance comparison is the technique called Random forest. This machine learning method is versatile and powerful that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This ensemble learning technique, particularly effective for large datasets, enhances predictive accuracy and controls over-fitting by averaging or *voting* across various trees. Each tree in the forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of them. This randomness, along with the ensemble approach, ensures the model's robustness against overfitting, making Random Forest an appealing choice for many applications in diverse domains ranging from finance to healthcare. We utilize the Random Forest algorithm under the AutoML framework.

Extremely Randomized Trees

Extremely randomized trees (ERT), also known as extra trees, is an ensemble learning technique that constructs a multitude of decision trees at training time. Similar to Random Forests, it operates by averaging predictions for regression tasks or using a majority vote in classification. However, it introduces additional randomness in the way splits are computed: instead of searching for the most discriminating thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This randomness leads to more diversified trees and typically faster training than Random Forest, often with comparable performance.

Boosted Decision Trees

Boosted decision trees involve an ensemble learning technique that combines multiple weak decision tree learners to form a strong predictive model. Unlike methods like Random Forests which build trees in parallel, boosting builds them sequentially. Each tree is trained on the dataset with an emphasis on correctly predicting instances that were misclassified by previous trees. This is achieved through iterative updates to the weights of data points. The final prediction is made based on a weighted vote (in classification) or sum (in regression) of the predictions from individual trees. This method often results in high accuracy, especially for complex datasets, but requires careful tuning to avoid overfitting.

Neural Networks

Neural networks, as our base machine learning model that we put forward in our study, are a foundational model in machine learning, inspired by the structure and function of the human brain. At their core, neural networks consist of layers of interconnected nodes, or *neurons*, each performing simple computations. The network typically includes an input layer to receive the data, one or more hidden layers that process the data, and an output layer that produces the prediction. Each neuron in a hidden layer transforms the values from the previous layer with a weighted linear summation followed by a *non-linear* activation function. These weights are learned during training through a process called backpropagation, which iteratively adjusts the weights to minimize the difference between the network's prediction and the actual data outcomes. Deep neural networks, with many hidden layers, can model complex patterns and relationships in data. They are highly versatile, being applied in fields such as image and speech recognition and natural language processing, as we adopt and extend in our study via multimodal setting.

B Procedures for the Response Measures

In this section, we present the specifics of our procedures with respect to our high-frequency market response measures. To proceed, we first outline the estimation steps of the realized intensity as a high-frequency tail risk measure. We then present a method to assess the accuracy of parameters estimates and stability for both realized volatility and realized intensity. Finally, we present the estimated responses.

B.1 Estimation Steps of the Realized Intensity

We proceed with the details on the estimation of our RI measure (equation (7)) as follows.

Step 1: *Start by defining the jump activity index β:*

$$\beta =: \inf\{r \geq 0; \sum_{0 \leq s \leq t} |\Delta_s X|^r < \infty\}, \tag{10}$$

where $\Delta_s X = X_s - X_{s-}$ is the jump size at time s, and r is the power variation parameter.

Step 2: *Compute the jump activity index β in equation (7):*

$$\widehat{\beta}(t, \varpi, \theta, \theta') := \log \frac{V(\varpi, \theta, g)_t^n}{V(\varpi, \theta', g)_t^n} / \log(\frac{\theta'}{\theta}), \tag{11}$$

for which select $0 < \theta < \theta'$, $0 < \varpi < 1/2$ and

$$V(\varpi, \theta, g)_t^n := \sum_{i=1}^{[t/\Delta_n]} g\left(\frac{|\Delta_i^n X|}{\alpha \Delta_n^{\varpi}}\right), \tag{12}$$

where $g(t)$ is the weight function, choose a form that needs to satisfy the condition $g(x) = |x|^p$ if $|x| \leq a$ for some constant $a > 0$ and even integer $p > 2$.

Step 3: *Choose values for the tuning parameters ϖ, k_n and α in equation (7).*

Step 4: *Compute the g function in equation (7) to disentangle volatility component from the jump component.*

Step 5: *Identify the release times (minutes and seconds) of speeches.*

Step 6: *For each speech, select a window length (e.g., one hour) and estimate RI in equation (7) by using high-frequency returns in this window.*

B.2 Accuracy Assessment

To evaluate the accuracy of the estimated parameters of the response measures, we proceed with the realized intensity first. Let us use TR for $\widehat{\lambda}(k_n)_{t_p}$, instead of λ and continue from this stage. We have

$$\sqrt{\frac{k_n \Delta_n}{\Delta_n^{\varpi\beta}}} \left(\widehat{TR} - TR\right) \xrightarrow{L_{st}} N\left(0, TR \frac{\alpha^{\beta} C_{\beta}(2)}{\left(C_{\beta}(1)\right)^2}\right),$$

where

$$C_{\beta}(k) = \int_0^{\infty} \left(g(x)\right)^k / x^{1+\beta} dx.$$

Therefore, the 95% confidence interval for $\widehat{\lambda}(k_n)_{t_r}$ is given by

$$\widehat{TR} \pm \text{c.v.} \times \sqrt{\frac{\widehat{TR}(\alpha \Delta_n^{\varpi})^{\beta} C_{\beta}(2)}{\left(C_{\beta}(1)\right)^2 k_n \Delta_n}},$$

for which we can use critical value such as c.v. $= 1.96$. The average of the lower and upper bound gives us the estimated intensity.

For spot realized volatility, we have

$$\sqrt{k_n^\sigma}\Big(\hat{c}_{t_r} - c_{t_r}\Big) \xrightarrow{L_{st}} N(0, 2c_{t_r}^2),$$

and the 95% confidence interval is

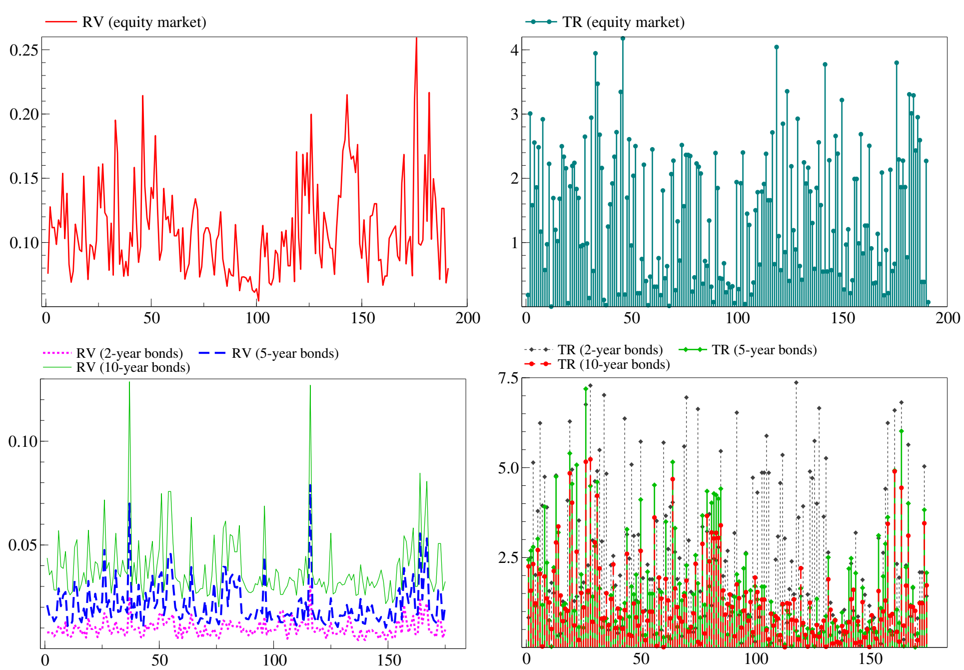
$$c_{t_r} \pm \text{c.v.} \times \sqrt{\frac{2}{k_n^\sigma}} c_{t_r}.$$

In light of these constructed confidence intervals, we assess the fit of the estimates, considering the lower and upper bounds.

B.3 Estimated Response Measures: Realized Volatility and Tail Risk

As we describe in the main text, we use high-frequency data and identify market responses in the forms of realized volatility and tail risk (computed based on realized intensity). Figure 6 displays the estimates of these quantities for each speech in our full sample for both equity and bond markets (upper and lower panels, respectively).

Figure 6: Estimated market response measures for central bank speeches



Notes: The figure shows the estimated response measures for each central bank speech (X-axis) in our dataset. Given the speech release, we compute realized volatility and tail risk—based on the realized intensity (labels RV and TR in the figure). The figure displays the quantities for the equity market (upper panels) and bond market (lower panels). For the equity market, RV and TR estimates are the cross-sectional averages of the individual stocks. For the bond market, the figure shows the estimated RV and TR separately for the 2-year, 5-year and 10-year bond futures.

The figure exhibits a number of features. First, both realized volatility and tail risk vary across central bank speeches. Second, looking at the response patterns of the bond market, we see noticeable differences between the reactions of short- and long-term bonds. That is, while the realized volatility of 2-year bond futures is clearly lower than the realized volatility of 5-year and 10-year bond futures (lower left panel), the realized tail risk identified through 2-year bonds is the highest across all maturities (lower right panel). Finally, central bank speeches tend to create distinct effects on bond and equity markets, which potentially reflects the importance of information signals embedded in the speeches.

C Further Considerations

Remark 1. It is worth emphasizing that the speeches have a much wider content beyond those key macro indicators (CPI, GDP, unemployment) that we rely on in our study. Nevertheless, we do not observe differences in terms of financial market effects mainly because we train our multimodal NLP model, test its out-of-sample performance, and construct the implied speech signals, entirely based on these three macro factors. Our proposed model processes the topics under this setting by utilizing both tabular (macro) data and text (speech) data. Therefore, our framework helps select the most important topics and those that do not carry significant explanatory information are directly excluded. This approach brings an advantage, rather than a setback, as it prevents us from incorrect measurement of market response to other generally important yet irrelevant speeches. Of course, it is possible to extend our model and feed the model by focusing also on other variables beyond macro factors.

Remark 2. When we identify the implied speech signals through our multimodal NLP model, we rely on a time frame for which we evaluate the information content in the entire period. During this process, we "synchronize" the time stamps of the speech and the SPF releases so that when we create the signal, the signal utilizes the information up to the *same* calendar time. Regardless of the time difference between the SPF release time and the speech release time, the time stamp of the signal is the time stamp of the speech and it remains the same as long as both SPF release and speech release fall in the same time frame. In fact, proceeding this way ensures that the process is a *martingale*. That is, the "speech release time" is the time that conditional expectations will be formed, based on all available information (including SPF news) up to speech time. This holds regardless of the past values and the time distance between SPF release and speech release.

Our high-frequency approach allows us to examine the impact of speech immediately after the public release by quantifying the changes in market volatility and market tail risk withing seconds and minutes. When a central bank speech is released *a few weeks* after an SPF release, investors still tend to use the most updated information available to them, perhaps related to market efficiency, so they wait for the release of the central bank speech. As soon as the speech is released and it becomes publicly available, we quantify the market response through our measures. So, the response already incorporates the information content in the SPF news, as investors wait for the new SPF release. As another situation, even if a speech is released, for example, *two days* after an SPF release, the reaction time that we rely on remains the same and hence it is still the speech release time. In this situation, while it is true that investors have a relatively short period of time to *digest* the content of the SPF release, the period is sufficient for those monitoring markets at intradaily levels.

Remark 3: Irrelevant Speeches? One argument would be that only relevant speeches matter and hence irrelevant speeches should not convey important signals. To test this conjecture, we conduct a simple, yet insightful, robustness check. We first rank the speeches in our database in order of their implied signal levels. We then identify the speeches that have the highest and lowest signal estimates (i.e., top ten and bottom ten). We observe that the highest implied signals often derive from the statements about topics on monetary policy, financial stability, economic conditions, and economic outlook. In contrast, the signals with the lowest values are often associated with statements that are indirectly related to the macro environment, financial markets, or monetary policy. For example, these low signal speeches are about the situation of middle-income families (unemployment factor), consumer behavior in credit and payment markets, and small business (GDP factor). Of course, these statements are not necessarily redundant, as they are made by the Fed members and the Chair. However, they are not as directly relevant and hence their signal levels that we measured using our model turn out to be low.

In light of this assessment, we also find that the name of the speaker (e.g., Chair or not) does not play an important role, as we see that Chair speeches can be associated with both lowest and highest signals. This regularity holds for all three news factors (CPI, GDP, unemployment) and for all other Fed members. Therefore, it is our understanding that, by looking at the name and whether the speaker is Chair, it is hard to draw a direct conclusion about which signals should matter. This is largely in line with our additional analyses on speech characteristics. Statements that look similar in terms of the speaker name, time, and title of the talk have different levels of implied forecast revision signals.

D List of Relevant Greenbook Sections

Table 13: Considered Greenbook sections per economic indicator

GDP	**CPI**	**Unemployment**
Ec.GDP	Ec.Prices	Ec.Labor
For.Ec.Overview	For.CostPrice	For.Labor
For.Ec.Summary	Ec.Wages	
For.Outlook		
For.HH		
For.G		
For.Inven		
For.BusInvest		
For.Trade		

Notes: In the table, EC = Economic Conditions Section, For = Forecasts Section.

E Lists of Stocks and Bonds

Table 14: Stock tickers and names

AAPL	Apple	AXP	American	BA	Boeing	CAT	Caterpillar
CSCO	Cisco	CVX	Chevron	DIS	Disney	HD	Home
IBM	IBM	INTC	Intel	JNJ	Johnson	KO	Coca-Cola
MCD	McDonald's	MMM	3M	MRK	Merck	MSFT	MSFT
NKE	Nike	PFE	Pfizer	UNH	UnitedHealth	VZ	Verizon
WMT	Wal-Mart	XOM	Exxon				

Notes: The table lists the tickers and descriptions of the individual stocks used in our empirical analysis.

Table 15: Bond names and maturities

US Treasury Note Futures:	2-Year	5-Year	10-Year

Notes: The table lists the tickers and descriptions of the U.S. Treasury bond futures used in our empirical analysis.

F Additional Results: Language to Forecast Mapping

Table 16: CPI mapping and fit performance

Model \Predictive R^2	**score test**	**score val**	**score train**	**data source**
MM Neural Topic Model (non-lin)	**0.735**	**0.830**	0.670	joint MM tabular + topics
MM Neural Topic Model (linear)	**0.640**	**0.650**	0.600	joint MM tabular + topics
ExtraTreesMSE_BAG_L1	0.588	0.084	0.880	tabular
RandomForestMSE_BAG_L1	0.584	0.052	0.622	tabular + topics
ExtraTreesMSE_BAG_L1	0.584	0.089	0.595	tabular + topics
RandomForestMSE_BAG_L1	0.568	0.047	0.876	tabular
KNeighborsUnif_BAG_L1	0.559	0.141	0.460	tabular + topics
KNeighborsDist_BAG_L1	0.549	0.128	0.798	tabular + topics
KNeighborsUnif_BAG_L1	0.520	0.152	0.439	tabular + tfidf
KNeighborsDist_BAG_L1	0.519	0.146	1.000	tabular + tfidf
KNeighborsUnif_BAG_L1	0.516	0.142	0.442	tabular
NeuralNetFastAI_BAG_L1	0.515	0.233	0.251	tabular + topics
KNeighborsDist_BAG_L1	0.513	0.121	1.000	tabular
OLS	0.512		0.288	tabular
NeuralNetFastAI_BAG_L1	0.494	0.272	0.594	tabular
RandomForestMSE_BAG_L1	0.482	0.103	0.883	tabular + tfidf
WeightedEnsemble_L2	0.475	0.302	0.565	tabular
CatBoost_BAG_L1	0.386	0.200	0.698	tabular
CatBoost_BAG_L1	0.384	0.170	0.905	tabular + tfidf
XGBoost_BAG_L1	0.377	0.169	0.595	tabular + topics
XGBoost_BAG_L1	0.374	0.155	0.937	tabular + tfidf
LightGBMXT_BAG_L1	0.373	0.126	0.295	tabular
XGBoost_BAG_L1	0.368	0.152	0.770	tabular
WeightedEnsemble_L2	0.358	0.284	0.370	tabular + topics
LightGBMLarge_BAG_L1	0.357	0.080	0.646	tabular + tfidf
LightGBM_BAG_L1	0.327	0.136	0.294	tabular
WeightedEnsemble_L2	0.299	0.305	0.953	tabular + tfidf
LightGBM_BAG_L1	0.289	0.138	0.245	tabular + topics
NeuralNetTorch_BAG_L1	0.269	0.210	0.128	tabular + topics
NeuralNetTorch_BAG_L1	0.262	0.247	0.401	tabular
XGBoost_BAG_L1	0.260	0.056	0.783	tabular + embeddings
LightGBMXT_BAG_L1	0.252	0.092	0.348	tabular + tfidf
LightGBM_BAG_L1	0.252	0.131	0.368	tabular + tfidf
LightGBMLarge_BAG_L1	0.251	0.139	0.302	tabular
LightGBMLarge_BAG_L1	0.202	0.156	0.323	tabular + topics
ExtraTreesMSE_BAG_L1	0.193	0.143	0.889	tabular + tfidf
LightGBMLarge_BAG_L1	0.191	0.074	0.440	tabular + embeddings
CatBoost_BAG_L1	0.177	0.250	0.525	tabular + topics
LightGBMXT_BAG_L1	0.162	0.140	0.192	tabular + topics
NeuralNetFastAI_BAG_L1	0.148	0.280	0.912	tabular + tfidf
WeightedEnsemble_L2	0.132	0.139	0.573	tabular + embeddings
CatBoost_BAG_L1	0.126	0.116	0.633	tabular + embeddings
LightGBMXT_BAG_L1	0.116	0.001	0.520	tabular + embeddings
LightGBM_BAG_L1	0.112	-0.018	0.338	tabular + embeddings
NeuralNetTorch_BAG_L1	0.095	0.153	0.500	tabular + tfidf
NeuralNetTorch_BAG_L1	-0.030	0.076	0.161	tabular + embeddings
AutoGluon Multimodal Transformer	-0.292		-0.155	multimodal embeddings

Notes: The table reports the performance (predictive R^2) of different models for the language mapping analysis of the CPI.

Table 17: GDP mapping and fit performance

Model \Predictive R^2	**score test**	**score val**	**score train**	**data source**
MM Neural Topic Model (lin)	**0.825**	**0.426**	0.372	joint MM tabular + topics
MM Neural Topic Model (non-lin)	**0.797**	**0.371**	0.483	joint MM tabular + topics
WeightedEnsemble_L2	0.380	0.304	0.497	tabular
OLS	0.785	0.301		tabular
NeuralNetFastAI_BAG_L1	0.480	0.270	0.443	tabular
WeightedEnsemble_L2	0.285	0.253	0.730	tabular + topics
WeightedEnsemble_L2	0.268	0.240	0.752	tabular + tfidf
WeightedEnsemble_L2	0.142	0.220	0.587	tabular + embeddings
CatBoost_BAG_L1	0.249	0.211	0.552	tabular
RandomForestMSE_BAG_L1	0.302	0.204	0.892	tabular + tfidf
RandomForestMSE_BAG_L1	0.348	0.202	0.892	tabular + topics
ExtraTreesMSE_BAG_L1	0.408	0.193	0.891	tabular
ExtraTreesMSE_BAG_L1	0.381	0.192	0.890	tabular + topics
ExtraTreesMSE_BAG_L1	0.111	0.188	0.891	tabular + tfidf
CatBoost_BAG_L1	0.207	0.187	0.671	tabular + tfidf
LightGBMXT_BAG_L1	0.203	0.178	0.322	tabular
LightGBM_BAG_L1	0.154	0.172	0.367	tabular
XGBoost_BAG_L1	0.141	0.171	0.580	tabular + topics
CatBoost_BAG_L1	0.006	0.169	0.531	tabular + topics
CatBoost_BAG_L1	0.101	0.169	0.552	tabular + embeddings
LightGBM_BAG_L1	0.099	0.162	0.704	tabular + embeddings
NeuralNetTorch_BAG_L1	0.461	0.160	0.341	tabular
LightGBM_BAG_L1	0.101	0.159	0.734	tabular + tfidf
KNeighborsUnif_BAG_L1	0.253	0.158	0.402	tabular + tfidf
LightGBMLarge_BAG_L1	0.245	0.155	0.598	tabular
KNeighborsDist_BAG_L1	0.256	0.151	1.000	tabular + tfidf
NeuralNetTorch_BAG_L1	0.049	0.150	0.553	tabular + tfidf
LightGBMXT_BAG_L1	0.120	0.150	0.348	tabular + tfidf
RandomForestMSE_BAG_L1	0.394	0.150	0.885	tabular
LightGBMLarge_BAG_L1	0.111	0.149	0.536	tabular + topics
LightGBMLarge_BAG_L1	0.181	0.149	0.665	tabular + embeddings
XGBoost_BAG_L1	0.119	0.142	0.567	tabular
NeuralNetFastAI_BAG_L1	0.060	0.136	0.797	tabular + tfidf
KNeighborsDist_BAG_L1	0.255	0.132	1.000	tabular
KNeighborsUnif_BAG_L1	0.248	0.130	0.407	tabular
LightGBM_BAG_L1	0.111	0.126	0.496	tabular + topics
LightGBMXT_BAG_L1	0.105	0.125	0.505	tabular + embeddings
NeuralNetTorch_BAG_L1	-0.071	0.123	0.275	tabular + embeddings
NeuralNetTorch_BAG_L1	0.151	0.108	0.497	tabular + topics
XGBoost_BAG_L1	-0.015	0.107	0.663	tabular + embeddings
LightGBMLarge_BAG_L1	0.108	0.095	0.581	tabular + tfidf
XGBoost_BAG_L1	0.041	0.083	0.564	tabular + tfidf
KNeighborsUnif_BAG_L1	0.286	0.081	0.400	tabular + topics
KNeighborsDist_BAG_L1	0.274	0.074	1.000	tabular + topics
LightGBMXT_BAG_L1	0.097	0.049	0.318	tabular + topics
TextPredictor_BAG_L1	-0.077	-0.123	-0.103	tabular + embeddings
NeuralNetFastAI_BAG_L1	0.407	-0.126	0.438	tabular + topics
AutoGluon Multimodal Transformer	-0.044		0.013	multimodal transformer

Notes: The table reports the performance (predictive R^2) of different models for the language mapping analysis of the GDP.

Table 18: Unemployment mapping and fit performance

Model \Predictive R^2	**score_test**	**score_val**	**score_train**	**data source**
MM Neural Topic Model (non-lin)	**0.208**	**0.457**	0.285	joint MM tabular + topics
WeightedEnsemble_L2	-0.044	0.145	0.415	tabular + embeddings
NeuralNetTorch_BAG_L1	-0.152	0.122	0.313	tabular + embeddings
WeightedEnsemble_L2	-0.045	0.113	0.577	tabular + tfidf
MM Neural Topic Model (linear)	0.066	0.109	0.197	joint MM tabular + topics
CatBoost_BAG_L1	-0.055	0.104	0.690	tabular + tfidf
LightGBMXT_BAG_L1	-0.068	0.074	0.336	tabular + tfidf
NeuralNetTorch_BAG_L1	-0.029	0.070	0.394	tabular + tfidf
WeightedEnsemble_L2	0.131	0.058	0.191	tabular
WeightedEnsemble_L2	-0.010	0.053	0.278	tabular + topics
NeuralNetFastAI_BAG_L1	0.124	0.047	0.237	tabular
CatBoost_BAG_L1	0.021	0.041	0.411	tabular + embeddings
NeuralNetTorch_BAG_L1	0.106	0.033	0.098	tabular
LightGBM_BAG_L1	0.006	0.027	0.349	tabular + embeddings
LightGBM_BAG_L1	-0.035	0.025	0.316	tabular + tfidf
CatBoost_BAG_L1	-0.003	0.021	0.260	tabular + topics
CatBoost_BAG_L1	0.019	0.010	0.095	tabular
RandomForestMSE_BAG_L1	-0.072	0.008	0.868	tabular + tfidf
NeuralNetTorch_BAG_L1	-0.004	0.006	0.022	tabular + topics
XGBoost_BAG_L1	-0.112	0.006	0.883	tabular + tfidf
LightGBMLarge_BAG_L1	-0.001	0.001	0.594	tabular + embeddings
LightGBMLarge_BAG_L1	0.002	-0.003	0.109	tabular + topics
ExtraTreesMSE_BAG_L1	-0.045	-0.003	0.868	tabular + tfidf
LightGBMXT_BAG_L1	-0.001	-0.005	0.084	tabular
LightGBMXT_BAG_L1	0.000	-0.006	0.009	tabular + topics
LightGBM_BAG_L1	0.000	-0.007	0.015	tabular + topics
LightGBMXT_BAG_L1	-0.005	-0.024	0.292	tabular + embeddings
XGBoost_BAG_L1	-0.043	-0.027	0.495	tabular + topics
LightGBM_BAG_L1	-0.002	-0.028	0.170	tabular
LightGBMLarge_BAG_L1	0.013	-0.034	0.094	tabular
NeuralNetFastAI_BAG_L1	0.002	-0.036	0.565	tabular + tfidf
XGBoost_BAG_L1	-0.061	-0.041	0.624	tabular + embeddings
LightGBMLarge_BAG_L1	-0.045	-0.044	0.519	tabular + tfidf
NeuralNetFastAI_BAG_L1	-0.016	-0.058	0.025	tabular + topics
RandomForestMSE_BAG_L1	-0.005	-0.101	0.855	tabular + topics
XGBoost_BAG_L1	-0.048	-0.126	0.277	tabular
ExtraTreesMSE_BAG_L1	0.008	-0.144	0.849	tabular
ExtraTreesMSE_BAG_L1	0.049	-0.163	0.848	tabular + topics
KNeighborsUnif_BAG_L1	-0.013	-0.185	0.188	tabular + tfidf
KNeighborsUnif_BAG_L1	-0.004	-0.187	0.186	tabular
KNeighborsUnif_BAG_L1	-0.048	-0.187	0.195	tabular + topics
TextPredictor_BAG_L1	-0.067	-0.190	-0.070	tabular + embeddings
KNeighborsDist_BAG_L1	-0.003	-0.191	1.000	tabular + tfidf
RandomForestMSE_BAG_L1	-0.034	-0.192	0.842	tabular
KNeighborsDist_BAG_L1	-0.030	-0.210	1.000	tabular + topics
KNeighborsDist_BAG_L1	0.003	-0.215	1.000	tabular
OLS	-0.377		0.231	tabular
AutoGluon Multimodal Transformer	-1.177		-0.737	multimodal transformer

Notes: The table reports the performance (predictive R^2) of different models for the language mapping analysis of the unemployment.