ISSN 1936-5330

kcFED Research Working Papers

Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements

Taeyoung Doh, Dongho Song and Shu-Kuei Yang
June 2022; updated November 2023
RWP 20-14
http://doi.org/10.18651/RWP2020-14

Federal Reserve Bank of Kansas City

Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements*

Taeyoung Doh† Dongho Song‡ Shu-Kuei Yang§

November 14, 2023

Abstract

We present a text-based metric for monetary policy stance using official and alternative Federal Open Market Committee statements. Our advanced natural language processing, with numeric property detection, jointly evaluates quantitative decisions like interest rates and qualitative explanations for these choices from texts. Monetary policy stance is decomposed into expected stance and surprise components by leveraging high-frequency bond futures data around FOMC announcements. We examine responses of stock returns to counterfactual (more dovish or hawkish) policy surprises through alternative language. This investigation yields valuable insights into monetary policy transmission.

JEL Classification: E30, E40, E50, G12.

Keywords: Alternative FOMC statements, counterfactual policy evaluation, monetary policy stance, text analysis, natural language processing.

---

*First version: March 2020. We thank Kyungmin Kim and Suyong Song as well as seminar participants at the Federal Reserve Bank of Kansas City, the Interfed Brown Bag, the KAEA webinar, and National University of Singapore, the University of North Carolina at Chapel Hill who provided helpful comments. All the remaining errors are ours. The opinions expressed herein are those of the authors and do not reflect the views of the Federal Reserve Bank of Kansas City or Federal Reserve System.

†Federal Reserve Bank of Kansas City; taeyoung.doh@kc.frb.org

‡Carey Business School, Johns Hopkins University; dongho.song@jhu.edu

§Federal Reserve Bank of Kansas City; shu-kuei.x.yang@kc.frb.org

1 Introduction

Central banks increasingly turn to public communications as a tool for shaping forthcoming policy actions. This practice, referred to as forward guidance, has gained prominence, particularly in situations where interest rates are restricted by the effective lower bound (as evidenced in Woodford (2005) and Blinder et al. (2008), elaborated upon by Bernanke (2010)). In this context, both the concrete decisions (like setting interest rates or deciding on asset purchases) that central banks make and their explanations about why they make these decisions are crucial pieces of information for understanding how monetary policy functions.

While the literature has been moving towards treating policy statements by central bank officials as analyzable data, significant challenges persist in parsing the textual content of these statements due to their distinct policy-related characteristics. Firstly, policy statements inherently blend a wealth of qualitative information, such as descriptions of economic conditions, with quantitative information like decisions on key interest rates or asset purchases. Secondly, qualitative communications carry intricate nuances vital to policy but challenging to measure. For example, the market focuses on nuanced elements in statements, which reveal economic conditions and policy choices. These attributes require a sophisticated natural language processing (NLP) algorithm that surpasses simple word frequency counting as a metric for textual similarity assessment. A large-scale language model, fine-tuned to detect numeric properties crucial for comprehending policy actions, is better equipped to decipher policy communications.

In this paper, we extract comprehensive information from the Federal Open Market Committee's (FOMC) post-meeting statements using the pre-trained NLP algorithm known as the Universal Sentence Encoder (USE), as introduced by Cer et al. (2018). The USE excels at capturing intricate word relationships via an artificial neural network architecture trained on extensive text data. Crucially, we refine this pre-trained USE algorithm using synthetic data that mirrors numerical information in FOMC statements. For example, it can accurately distinguish that a 4% interest rate is more distant from 3.5% than from 3.75%, and that a 0.50% rate hike is more hawkish than a 0.25% rate hike. This refinement enables a comprehensive assessment of both qualitative and quantitative information, utilizing the algorithm's built-in capacity to transform text statements into numeric vector representations referred to as "embeddings". Our refined embeddings offer a unified framework for capturing information about both policy actions and the underlying rationale from FOMC statements.

Our aim is to construct a text-based metric for monetary policy stance along with its surprise component, uncovered by using intraday bond futures data. To achieve this, our study integrates alternative FOMC statements, available for each meeting since March 2004, which present more dovish (referred to as alternative A) or more hawkish (known as alternatives C or D when available) views compared to the benchmark statement (alternative B). These pre-structured statements allow us to assess the tone of a post-meeting statement by quantifying semantic differences between the post-meeting statement and alternative statements. For example, if the USE representation of the post-meeting statement aligns more closely with alternative A than C, it can be categorized as dovish. The diverse economic outlooks and policy views within these alternatives provide valuable guidance for interpreting the tone of subsequent policy statements.

The Federal Reserve's monetary policy stance is contingent upon economic and financial conditions, and it may vary over time. Utilizing alternative statements enables us to discern the policy stance within this evolving environment. For instance, when the statement downgrades economic outlook significantly but does not change the language on the "gradual rate hike", it could be interpreted as hawkish while the same phrase could be interpreted as dovish if the statement upgrades economic outlook significantly—a subtlety that might go unnoticed when assessing tone through pre-established meanings; the phrase "gradual rate hike" may have different meanings depending on the context which highlights "gradual" or "rate hike". A comparable concept is examined in Laver et al. (2003), streamlining tone identification.

We define a text-based metric for monetary policy stance through the integration of two distinct measures—semantic difference in tone and novelty—across texts. Novelty quantifies semantic differences over time between consecutive FOMC statements while tone captures cross-sectional variation. We then isolate the surprise policy component via post-FOMC intraday bond price movements. Unexpected bond return shifts are interpreted as indicative of monetary policy news, aligning with established literature assumptions, see Gürkaynak et al. (2005). Surprising policy tightening corresponds to decreased returns. Subtracting policy surprises from policy stances leads to expected policy stances. Our approach of obtaining policy surprises is consistent with prior studies in this field, such as Kuttner (2001), Cochrane and Piazzesi (2002), Faust et al. (2004), and others.

We validate our measure through multiple dimensions. Firstly, we confirm that our surprise measure is strongly correlated (around 70 to 80%) with measures in existing studies such as Swanson (2017), Nakamura and Steinsson (2018), Bauer and Swanson (2023). This finding is unsurprising, given our reliance on Eurodollar futures contracts—a well-established instrument in previous studies—for identifying the surprise component. Secondly, we observe that a surprising policy tightening results in a negative stock price reaction, in line with Bernanke and Kuttner (2005), and drives short-term Treasury yields upward as intended. Thirdly, we demonstrate that impulse responses following a contractionary monetary policy lead to declines in real activity and inflation while increasing the credit risk premium (excess bond premium in Gilchrist and

Zakrajšek (2012)). Finally, we verify that assumptions we make to link our tone measure with bond market responses to FOMC announcements are plausible based on the human reading of FOMC transcripts.[1] These analyses underscore the robustness and external validation of our findings.

Our measure of monetary policy stance is derived from text analysis, while its expected (or surprise) component is obtained through the application of intraday bond futures data. This unique feature empowers us to conduct a language counterfactual experiment, enabling the assessment of financial market impacts under alternate communication scenarios. By subtracting the 'unchanged' expected stance from the 'counterfactual' monetary policy stance, recalculated through text analysis, our approach facilitates policymakers' examination of alternative scenarios and their potential effects on the stock market within policy statements. The concept of language counterfactual, which we introduce, represents a novel addition to the existing literature.

We conduct counterfactual analyses using the August 2011 and December 2016 FOMC statements to explore both dovish and hawkish scenarios. To begin, we investigate how the market would have reacted if the August 2011 FOMC announcement had included changes to the composition of the Federal Reserve's balance sheet. Our counterfactual experiment yields a noteworthy revelation: the introduction of a maturity extension policy would likely have sparked an exceptionally positive surprise among stock traders, potentially surpassing the magnitude of the actual market response. Second, we contemplate a more hawkish December 2016 statement, replacing 'only gradual increases' with 'additional gradual increases' in the federal funds rate. In this scenario, the market would more clearly interpret the hawkish signal from the forward guidance concerning rate hikes. These exercises underscore the profound impact of FOMC communication as a pivotal policy tool.

Related Literature. Our work is related to the increasingly popular literature that applies text analysis to the broad field of social science, e.g., Lucca and Trebbi (2009), Schonhardt-Bailey (2013), Meade and Acosta (2015), Hansen and McMahon (2016), Hansen et al. (2017), Jegadeesh and Wu (2017), Shiller (2017), Gentzkow et al. (2019), Ke et al. (2019), Shapiro and Wilson (2019), Shiller (2020), Giavazzi et al. (2020), Husted et al. (2020), Handlan (2022), Shapiro and Wilson (2021), Cieslak et al. (2022), Caldara and Iacoviello (2022), Aruoba and Drechsel (2023), Gorodnichenko et al. (2023), Schmeling and Wagner (2022), Hansen and Kazinnik (2023), Curti and Kazinnik (2023), Shah et al. (2023).

Among them, our research aligns more closely with papers that employ text analysis to study central bank communication. In this specialized field, sentiment and topic analysis emerge as

[1] Romer and Romer (2023) stress the importance of externally validating language model text analysis through careful human reading to assess its plausibility.

prominent methodologies, affording researchers intricate control over individual words and their interpretations. For instance, terms like "hawkish" or "contractionary" are tied to expressions such as interest rate increase or hike. This level of control contributes significantly to the favorability of these approaches, which rely on word frequency counting methods to effectively capture and quantify these specific textual attributes. Several recent papers on this area include Lucca and Trebbi (2009), Hansen and McMahon (2016), Hansen et al. (2017), Husted et al. (2020), Shapiro and Wilson (2021), Acosta (2022), Cieslak et al. (2022), Aruoba and Drechsel (2023), Schmeling and Wagner (2022), among others. While effective for simpler tasks, these methods may fall short in capturing the nuanced intricacies inherent in central bank communications.

We belong to a novel cohort of researchers embracing advanced deep-learning techniques, facilitating a more comprehensive exploration of the intricate complexities embedded in policy communications. Recent instances include XLNet (Generalized Autoregressive Pretraining Method) in Handlan (2022), GPT (Generative Pre-trained Transformer) in Hansen and Kazinnik (2023), BERT (Bidirectional Encoder Representations from Transformers) in Gorodnichenko et al. (2023), Curti and Kazinnik (2023), and Shah et al. (2023). These methods have the potential to reveal deeper layers of meaning hidden in the way central banks communicate. Handlan (2022), contemporaneously with our paper, closely aligns with our research by incorporating alternate FOMC statements and employing advanced natural language processing to derive text-based monetary policy surprises.

We emphasize that our paper is distinct from existing work along two important dimensions. First, from a technical perspective, we fine-tune pre-trained NLP models using an artificial dataset that mimics FOMC statements containing numerical details. This enables us to assess quantitative information like key interest rate decisions and asset purchase amounts. We emphasize that understanding numeric information in a text correctly is important for analyzing central bank communications. To the best of our knowledge, existing studies employing advanced NLP models have overlooked this aspect. Secondly, our methodological advancement allows us to perform language counterfactual experiments, marking our next significant contribution. Our approach to considering pre-labeled tones in alternative statements simplifies the implementation of language counterfactuals. This approach naturally leads to our unique role in the literature, addressing the pivotal question: How might the market have reacted if alternative scenarios (either more dovish or hawkish) were communicated? Our language counterfactual has the potential to scrutinize the causal link from scenario shifts to financial market reactions. This question holds relevance for both policymakers and market participants, and our paper offers an insight.

The paper's structure is outlined as follows: In Section 2, we introduce our NLP technique and elaborate on fine-tuning the algorithm to align with our goals. Section 3 provides a comprehensive explanation of the characteristics of policy alternatives and how they are crafted. Section 4 presents the creation of our text-based metric for monetary policy stance and outlines the identification scheme for monetary policy surprises. Section 5 delves into empirical results and their policy implications. Lastly, Section 6 provides the conclusion.

2 Universal Sentence Encoder for Text Analysis

In this section, we introduce our primary text-analysis method, the Universal Sentence Encoder (USE). USE, a deep learning-based approach, excels in capturing semantic meaning and context within sentences, making it valuable for various NLP tasks. Additionally, we explore FinBERT (Financial Bidirectional Encoder Representations from Transformers), another powerful deep learning-based model tailored for financial text analysis. Furthermore, we introduce two widely used word counting methods: TF-IDF (Term Frequency-Inverse Document Frequency) and Latent Semantic Analysis (LSA).

While deep learning models like USE and FinBERT are adept at learning intricate textual patterns, word counting approaches like TF-IDF and LSA offer simple techniques for text analysis. Our comparative analysis aims to underscore the inherent strengths and advantages of deep learning-based models, especially in their exceptional aptitude for analyzing dense texts, such as monetary policy statements.

To ensure accessibility to general readers, we focus on highlighting the key features of various text analysis methods without delving into intricate technical details.

2.1 Deep learning-based text similarity

Universal Sentence Encoder. Natural language processing tools employ a process called embedding, which converts words or texts into numeric vectors. Alternative methods of embedding differentiate diverse NLP algorithms. The USE presents two versions: 1) Deep averaging of word embeddings and 2) a Transformer-based approach utilizing the self-attention mechanism that determines the embeddings of individual words based on the semantic distance between them in a text. In the transformer-based approach, the embeddings of individual words are jointly determined based on the semantic distance between them in a text.[2] This channel captures

[2]The online appendix describes the transformer-based neural network architecture in detail.

the contextual meaning of each word within the text. These embeddings are then passed through multiple neural networks. In contrast, the deep averaging method applies multiple neural networks directly to the average of word embeddings. Despite their differences, both algorithms produce the embedding of the entire input text as their final output. For our task of calculating text similarity, we utilize the transformer-based version, thanks to its self-attention mechanism that captures context-dependent sentence meanings. This capability allows it to score the similarity between texts more sensibly.

For instance, consider two texts with different word lengths. The USE transforms these texts into numerical representations, obtaining two 512-dimensional vectors (U_1 and U_2) through a deep neural network architecture. After transforming texts into embedding representations, we calculate text similarity using the cosine similarity metric between vector representations of texts, denoted as U_1 and U_2:

$$\mathrm{Sim}(\mathrm{Text}_1, \mathrm{Text}_2) = \mathrm{cosine}(U_1, U_2) = \frac{U_1' U_2}{\sqrt{U_1' U_1}\sqrt{U_2' U_2}}. \tag{1}$$

For further details on the approach, please refer to Appendix A and Cer et al. (2018).

Financial Bidirectional Encoder Representations from Transformers. FinBERT, short for "Financial Bidirectional Encoder Representations from Transformers," is a specialized deep learning-based model tailored for financial text analysis. Similar to USE, it belongs to the BERT family of models, leveraging the powerful transformer architecture and produces a 768-dimensional vector embedding representation of a text. Both FinBERT and USE share the common transformer architecture, which incorporates multiple self-attention channels for efficient information processing for capturing the rich pattern in the contextual meaning. By computing the cosine similarity of FinBERT embedding representations of texts, as depicted in (1), we can evaluate text similarity effectively.

Since FinBERT is trained on financial domain-specific text data, it can comprehend financial jargon, market-specific language, and economic terminologies. The key difference between FinBERT and USE lies in their training data and expertise: FinBERT is specifically trained on labeled financial texts, giving it an advantage in understanding the intricacies of the financial industry, while USE is trained on a more diverse labeled dataset covering various topics. This unique training makes FinBERT a suitable benchmark for analyzing monetary policy statements when we evaluate the performance of USE. For a more detailed explanation of the FinBERT approach, the readers can refer to Araci (2019).

2.2 Word counting-based text similarity

In this section, we provide a brief description of the two methods of text embedding based on word frequency counting. Gentzkow et al. (2019) provide more detailed explanations on these methods. As with the transformer-based approach, we can calculate text similarity using the cosine similarity metric of embedding representations.

Term Frequency-Inverse Document Frequency. One of the most widely used measures for text analysis is the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF calculates how important a word is in a document compared to a collection of documents. It works by looking at the frequency of words in a specific document (term frequency) and scaling it by how rarely the word appears in all the documents (inverse document frequency). This helps to prioritize words that are important in a specific document but not common across all documents.

However, TF-IDF has a significant limitation. While it is effective at identifying important words in a document, it fails to differentiate between contextual meanings created by the varying word orderings, potentially overlooking subtle nuances in the contextual meaning of words. These nuances could be crucial in interpreting monetary policy statements when a change in the word ordering may signal a shift in the priority. For example, a statement describing deteriorating labor market conditions first and then discussing inflation risk can be viewed differently from the one mentioning inflation risk first and then describing deteriorating labor market conditions.

Latent Semantic Analysis. A more sophisticated word counting method known as Latent Semantic Analysis (LSA) considers the co-frequency of words to calculate similarity scores between texts. LSA extracts a low-dimensional representation of the term-frequency/document matrix, with words and documents represented in rows and columns, respectively. This approach effectively captures common themes and underlying structures from large corpora. By rotating term frequency vectors to maximize the co-frequency of words across multiple documents, LSA extracts representations that emphasize co-occurrence patterns of words used in different texts. This makes LSA particularly effective in identifying a few key topics from a large number of texts.

Similar to the TF-IDF approach, LSA also cannot account for complex dependencies between different words beyond co-frequency, which is crucial for comprehending semantic similarity. As a result, it may not fully capture the nuanced semantic relationships between words in monetary policy statements.

Table 1: Similarity scores

	Deep learning-based		Word counting-based	
	USE	FinBERT	TF-IDF	LSA
Sim(How old are you, What is your age)	0.90	0.78	0.00	0.00
Sim(How old are you, How are you)	0.37	0.77	0.78	0.87

Notes: We calculate the cosine similarity scores using both deep learning-based approaches, USE and FinBERT, and word frequency counting-based approaches, TF-IDF and LSA. We apply USE version 5 for these examples.

2.3 Performance checks with simple examples

We demonstrate the superiority of the deep learning-based approaches (USE and FinBERT) in capturing contextual meaning by comparing the similarities between the following sentences:

(S_1) How old are you?

(S_2) What is your age?

(S_3) How are you?

We repeat the same exercise using word counting-based approaches (TF-IDF and LSA) for comparison.

In these examples, sentences S_1 and S_2 convey the same question, while S_3 expresses a different one. An ideal classifier should recognize that S_1 is more similar to S_2 than to S_3. Surprisingly, the similarity scores obtained using TF-IDF and LSA provide the opposite ranking, while the USE and FinBERT models produce more sensible similarity scores (see Table 1).

Word frequency counting methods, such as TF-IDF and LSA, treat words as isolated items in the dictionary and lack the ability to encode contextual linkages between words in sentences. For example, in sentences S_1 and S_3, the word "How" can have different contextual meanings, which these word counting methods fail to capture. Additionally, TF-IDF and LSA represent individual words using one-hot vectors, where only one element is non-zero, limiting their ability to comprehend semantic relationships within the text.

In contrast, deep learning models, like USE, leverage self-attention mechanisms in their neural network architectures, enabling them to understand contextual relationships between words in sentences. USE transforms word embeddings in the current layer of a neural network by computing a weighted average of all word embeddings in the text, in which weights are determined by the degree of closeness (also known as "attention") between embeddings in the previous layer. This allows USE to consider contextual linkages effectively and outperform word counting methods in tasks such as text classification and semantic similarity assessment. For instance, within the transformer architecture, the embedding representation of a word like "How" is transformed by considering the closeness between its previous layer's embedding and the embedding of the word that follows "How" in the given sentence.

2.4 Performance checks with policy-relevant examples

The comparison of similarity scores for the example sentences in Table 1 highlights that deep learning-based approaches perform better in capturing local context over word frequency counting-based methods. Now, we proceed to evaluate their performance on more complex sentences, especially those involving numbers, which are common in FOMC statements. Although pre-trained language models like USE and FinBERT excel at capturing semantic relationships between words and sentences, accurately representing numeric properties associated with numbers may still pose challenges for them.

To address this limitation, we undertake a fine-tuning process on the pre-trained language model using texts that contain numbers, following the approach outlined in Sundararaman et al. (2020). We choose to fine tune USE whose modest representation size can be better aligned with statements which typically consist of a few paragraphs.[3] Specifically, we create a specialized loss function that preserves the numeric distance between numbers in the sentences. In addition, we incorporate a separate loss function for datasets without numbers, ensuring that the semantic distance between sentences, as captured by the original USE representation, is maintained. This fine-tuning process enhances the model's capability to effectively handle numerical information while retaining its proficiency in understanding textual semantics.

For illustration purposes, let's consider two sentences, S_1 and S_2, which involve numbers x_1 and x_2, respectively, and two other sentences, S_3 and S_4, which do not contain numbers. The original USE representations of these sentences are denoted as U_i and U_j, where i and j represent the sentence index. In the fine-tuning step, an additional fully connected feedforward network layer is added on top of the USE representation to encode numeric information.[4]

Suppose that $f(U_i)$ and $f(U_j)$ represent the "fine-tuned" USE representation of the two sentences S_i and S_j, respectively. Training the fine-tuning layer $f(\cdot)$ involves two types of loss functions based on different pairs of sentences.

[3] For texts much longer than statements such as FOMC transcripts, a larger size of FinBERT can be a more suitable choice because it has a higher embedding dimension. However, our fine-turning procedure would work for both architectures.

[4] A fully connected network links each element in the input layer to every element in the output layer, and the feedforward network denotes one-way connections from the input to the output layer. The online appendix describe the details of the fine-tuning procedure.

The first loss function, denoted as $\mathcal{L}_{\text{num}}(U_1, U_2)$, is constructed based on the discrepancy between the numeric distance between x_1 and x_2, and the cosine distance between $f(U_1)$ and $f(U_2)$, i.e., $d\big[f(U_1), f(U_2)\big] = 1 - \text{cosine}(f(U_1), f(U_2))$:

$$\mathcal{L}_{\text{num}}(U_1, U_2) = \left(2\frac{|x_1 - x_2|}{|x_1| + |x_2|} - d\big[f(U_1), f(U_2)\big]\right)^2. \tag{2}$$

However, training the fine-tuning layer $f(\cdot)$ solely with sentences involving numbers may distort the USE representation of sentences without numbers. To address this concern, we augment the loss function $\mathcal{L}_{\text{non-num}}(U_3, U_4)$:

$$\mathcal{L}_{\text{non-num}}(U_3, U_4) = \left(d\big[U_3, U_4\big] - d\big[f(U_3), f(U_4)\big]\right)^2, \tag{3}$$

which aims to preserve the original USE representations as much as possible.

For the actual training, we create a dataset consisting of 252 sentences with numbers and 688 sentences without numbers, based on post-meeting FOMC statements. Each training data point consists of a pair of sentences, resulting in 536,848 pairs of sentences used as the training set. Among them, 63,504 pairs involve numbers, and 473,344 pairs do not. We use $\mathcal{L}_{\text{num}}(U_i, U_{-i})$ for pairs involving numbers and $\mathcal{L}_{\text{non-num}}(U_i, U_{-i})$ for pairs without numbers.

Below, we provide a few selected examples to compare the semantic distance between the original USE representations and the fine-tuned representations.

1. Policy-relevant examples without numbers
   - (P_1) Household spending has been increasing at a solid rate, on net, and business investment has been expanding;
   - (P_2) Household spending is rising moderately and business fixed investment is advancing;
   - (P_3) Household and business spending has been subdued.
2. Policy-relevant examples involving numbers
   - (N_1) FOMC decided to keep the target interest rate at 3.75 percent;
   - (N_2) FOMC decided to raise the target interest rate by 25 basis points to 4.00 percent;
   - (N_3) FOMC decided to lower the target interest rate by 25 basis points to 3.50 percent;
   - (N_4) FOMC decided to raise the target interest rate by 50 basis points to 4.25 percent.

In Panel (A) of Table 2, we present similarity scores for various pairs of (P_1, P_2, P_3) sentences. The statements P_1 and P_3 express differing perspectives on household spending, with P_1 pointing

Table 2: Similarity scores with policy-relevant examples (normalized)

(A) Policy-relevant examples without numbers

Reference	Comparison	Similarity scores		
		Fine-tuned USE	Original USE	FinBert
(P1) has been increasing	(P2) is rising moderately	1.000	1.000	1.000
	(P3) has been subdued	0.763	0.790	0.943
(P2) is rising moderately	(P1) has been increasing	1.000	1.000	1.000
	(P3) has been subdued	0.879	0.851	0.972

(B) Policy-relevant examples involving numbers

Reference	Comparison	Similarity scores		
		Fine-tuned USE	Original USE	FinBert
(N1) keep at 3.75%	(N4) raise by 50 bps to 4.25%	0.999	1.008	1.008
	(N2) raise by 25 bps to 4.00%	1.000	1.000	1.000
	(N3) lower by 25 bps to 3.50%	1.000	1.016	1.008
(N2) raise by 25 bps to 4.00%	(N4) raise by 50 bps to 4.25%	1.000	1.000	1.000
	(N1) keep at 3.75%	0.999	0.917	0.952
	(N3) lower by 25 bps to 3.50%	0.998	0.961	0.999
(N4) raise by 50 bps to 4.25%	(N2) raise by 25 bps to 4.00%	1.000	1.000	1.000
	(N1) keep at 3.75%	0.999	0.924	0.960
	(N3) lower by 25 bps to 3.50%	0.997	0.955	0.997

Notes: We compute the similarity score between the reference sentence and each of the comparison sentences. For both the original version and the fine-tuned version, USE version 5 is applied. In Panel (A), we normalize the similarity score between (P1, P2), and (P2, P1) to 1, respectively. In Panel (B), we normalize the similarity scores between (N1, N2), (N2, N4), and (N4, N2) to 1, respectively.

to solid growth and P_3 describing a subdued pace. On the other hand, P_2 adopts a neutral view, acknowledging a moderate increase in spending. Accordingly, the natural order of similarity scores indicates that $\text{Sim}(P_1, P_3)$ is lower than either $\text{Sim}(P_1, P_2)$ or $\text{Sim}(P_2, P_3)$. To facilitate the interpretation of relative similarity scores, we normalize the similarity score between (P1, P2) to 1, respectively.

Both the original USE representation and our fine-tuned version meet these criteria, indicating that our fine-tuning process retains the USE's proficiency in capturing semantic distance for sentences without numbers. We also observe that FinBERT satisfies these criteria, though with less distinct differentiation compared to the USE.

In Panel (B) of Table 2, we evaluate the performance of the fine-tuned USE in capturing numeric properties. Here, we provide similarity scores for different pairs of sentences (N_1, N_2, N_3, N_4), where the level of the federal funds rate varies. The similarity score should decrease as the difference in the federal funds rate levels increases. Consequently, for any pair of (N_i, N_{-i}), the lowest similarity scores should be found for (N_1, N_4), (N_2, N_3), and (N_3, N_4), respectively. To aid in interpreting relative similarity scores, we normalize the similarity scores between (N1, N2), (N2, N4), and (N4, N2) to 1 for each of the three subpanels in Panel (B) of Table 2. This ranking is preserved in the fine-tuned USE representation, setting it apart from the original USE and FinBERT.

The fine-tuned USE shows improved accuracy in evaluating the similarity between different statement pairs. For example, it recognizes that the phrase "keeps at 3.75%" is approximately equidistant to both "raise by 25bps to 4.00%" and "lower by 25bps to 3.50%," while the original USE considers the latter phrase much more similar due to the presence of the number 3, similar to the original phrase. FinBERT's highest similarity scores being assigned to the pair "keeps at 3.75%" and "raise by 50bps to 4.25%" contradict important criteria for similarity assessment. This discrepancy suggests possible limitations of FinBERT in accurately capturing the nuanced semantic relationships between different statement pairs without fine-tuning to recognize numeracy when numbers are involved.

2.5 Discussion

The objective of these comparisons is to emphasize that even sophisticated pre-trained NLP models may have limitations in accurately analyzing central bank communications, particularly when it comes to capturing numeric properties associated with numbers. Our contribution to the literature lies in demonstrating that fine-tuning the pre-trained NLP model with an artificial dataset mimicking FOMC statements involving numbers enables a more precise assessment of crucial quantitative information, such as decisions on key interest rates, which are integral to monetary policy.

Although the variations in similarity scores for examples involving numbers are relatively small compared to the original USE algorithm, this concern does not apply when analyzing FOMC statements. The rich mixture of qualitative and quantitative descriptions in FOMC statements allows the fine-tuned USE to effectively capture semantic differences and relationships, making it a powerful tool for comprehensively analyzing and understanding monetary policy discussions.

3 Monetary Policy Alternatives

The FOMC issues official statements on interest rates, economic outlook, and policy actions. Additionally, alternative FOMC statements are created by the Federal Reserve Board staff to present various policy options. These alternatives offer a range of policy scenarios beyond the baseline path, helping assess economic risks. Policymakers can evaluate different monetary policy actions and their potential impacts, enhancing transparency in communicating future actions to markets and the public.

In this section, we provide a comprehensive explanation of the characteristics of these policy alternatives and how they are crafted as we intend to utilize information from both official and alternative policy statements. To achieve this, we draw upon the FOMC meeting presentation material dated August 9, 2011.

3.1 Crafting of policy alternatives

These alternative policy statements are intentionally designed to deviate from market expectations in different directions. Alternative A takes a more dovish stance, which typically involves measures such as lower interest rates or other actions aimed at stimulating economic growth and employment. On the other hand, alternative C (or D when available) adopts a more hawkish approach which often involves actions like raising interest rates or reducing monetary stimulus to combat inflation.

To gain insights into crafting alternative statements, let's explore an illustrative case. The Federal Reserve Board staff crafts these alternative statements by drawing insights from the information obtained through the Federal Reserve Bank of New York's survey of primary dealers. This survey methodically compiles data concerning market expectations for alterations in the language of the forthcoming meeting statement. As an example, here are excerpts from the August 2011 Survey of Primary Dealers in response to the question: "Do you expect any changes in the FOMC statement and, if so, what changes?"

> "Many dealers expected that the August statement would contain a downgrade to the characterization of economic conditions, and a few expected the statement to contain reference to the benchmark revisions to GDP and its impact on the outlook for economic growth. A couple of dealers expected that the statement would reference the moderation in headline inflation. The announcement of additional policy action to lengthen the duration of the SOMA portfolio was expected by a couple of dealers,

> as was some indication of the Committee's willingness to ease policy. Some dealers did not expect any substantial changes to the statement."
>
> (FRB of New York — Responses to Survey of Primary Dealers, August 2011)

William English, a former director of the division of monetary affairs at the Federal Reserve Board during the August 2011 meeting, cited the result of the primary dealer survey to gauge the potential market impact of the statement language:

> "A statement along the lines of alternative B would be about in line with the expectations captured by the Desk's survey of primary dealers last week. However, as Brian noted in his briefing, investors have become more concerned about the economic outlook in recent days and reportedly have marked up the odds associated with policy action at this meeting. Thus, the release of a statement like alternative B, with a relatively downbeat assessment of the economy and no policy action, could disappoint some market participants. Bond yields could increase and the foreign exchange value of the dollar rise. Equity prices could decline somewhat."
>
> (William English — FOMC Meeting Transcripts, August 2011)

The primary dealer survey results provide valuable insights, going beyond the market's consensus to include minority perspectives. For instance, in the August 2011 survey, a couple of dealers anticipated additional policy actions, such as extending the duration of the SOMA portfolio. It's noteworthy that this description of extending the SOMA portfolio's duration is also found in alternative A (though absent in alternatives B or C), as detailed under the 'SOMA portfolio' section in Table 3. Conversely, the same primary dealer survey indicated that some dealers didn't anticipate significant changes to the statement. In the previous June FOMC statement, the forward guidance mentioned "exceptionally low levels...for an extended period," often interpreted as implying liftoff from the effective lower bound within a year. Notably, alternative A, as described in the 'Forward guidance' section of Table 3, modified this part of the forward guidance language to provide a more explicit time-based forward guidance. However, this change was not reflected in alternatives B or C.

The existence of various policy alternatives plays a pivotal role in stimulating discussions among FOMC members, which, in turn, promotes well-informed decision-making. For instance, there were vigorous debates involving the dovish camp, led by individuals like Chicago Fed President Charles L. Evans, advocating for the language in alternative A, and the hawkish camp, represented by figures such as Dallas Fed President Richard W. Fisher, favoring minimal

Table 3: Overview of alternatives for the August 9, 2011 FOMC statements

	August Statement	August alternatives A	B	C
Economic activity	considerably slower than the Committee had expected	considerably slower than the Committee had expected	has been slower than the Committee had expected	had been modest of late
Target FFR	0 to $\frac{1}{4}$%	0 to $\frac{1}{4}$%	0 to $\frac{1}{4}$%	0 to $\frac{1}{4}$%
Forward guidance	exceptionally low levels at least through mid-2013	exceptionally low levels at least through mid-2013	exceptionally low levels for an extended period	exceptionally low levels for an extended period
SOMA portfolio	regularly review the size and composition of its securities holdings and is prepared to adjust those holdings as appropriate	purchase $400 billion of Treasuries with maturities 7-30 years sell an equal amount with maturities of 3 years or less... lengthening the average duration...	regularly review the size and composition of its securities holdings and is prepared to adjust those holdings as appropriate	regularly review the size and composition of its securities holdings and is prepared to adjust those holdings as appropriate
	maintain reinvestment policy	maintain reinvestment policy going forward, will use the proceeds to purchase Treasuries with remaining maturities of 7-30 years	maintain reinvestment policy	for the time being, maintain reinvestment policy

Source: Authors' construction based on **FOMC meeting presentation material, August 2011**.

language changes akin to alternative C. Despite three hawkish members dissenting from the decision, it is evident from the released official statement in Table 3 that the FOMC ultimately incorporated elements from alternative A's language, particularly by introducing a more explicit time-based forward guidance, while refraining from adopting the language of the SOMA portfolio.

The August 2011 FOMC meeting serves as a compelling illustration of how the influence of FOMC members with dovish perspectives interacts with the nuanced semantic differences between the publicly released official statement and alternative statements.

3.2 Policy alternatives as policy position benchmarks

Based on our illustration, it is reasonable to assume a potential alignment of market participants' views on the implications of alternative language in FOMC statements with those of the Federal Reserve Board staff.

In our paper, we contend that policy alternatives serve as valuable indicators of the market's expectations for a hawkish or dovish stance on policy possibilities, thereby facilitating the interpretation of the official statement's tone. This idea aligns closely with prevailing practices in text analysis, where the use of pre-labeled words or texts, such as policy alternatives in our context, is widely recognized for classifying the tone or sentiment of unlabeled text. For example, in a study by Grimmer et al. (2022), researchers explored multiple methods to determine a politician's stance by comparing their speeches with various party platforms. We assert that the diverse perspectives on economic outlook and the associated policy recommendations found in alternative statements provide valuable reference points for interpreting the tone of the official statement. The following section formalizes our concept.

4 Text-based Monetary Policy Stance

4.1 Novelty and tone of released FOMC statements

We use the notation F_t to represent the FOMC statement released at time t. To quantify the novelty of each statement, we define novelty as follows:

$$\text{Novelty}_t = 1 - sim(F_t, F_{t-1}), \tag{4}$$

where $sim(F_t, F_{t-1})$ is the cosine similarity based on embeddings that capture the semantic distance between the current statement (F_t) and the previous statement (F_{t-1}) released after respective FOMC meetings. This measure allows us to assess how different or novel each FOMC statement is compared to its immediate predecessor in terms of semantic content.

Furthermore, we consider alternative FOMC statements denoted as F_t^a and F_t^c, corresponding to alternative A and C, respectively. These alternative statements are predefined with specific tones, either dovish (alternative A) or hawkish (alternative C), which enables us to evaluate the tone of the post-meeting statement. We calculate the tone using the formula:

$$\text{Tone}_t = \frac{sim(F_t^c, F_t) - sim(F_t^a, F_t)}{1 - sim(F_t^a, F_t^c)}. \tag{5}$$

Note that the tone value, as defined by (5), ranges from -1 to 1, given that the semantic distance between the dovish and hawkish alternative statements exceeds the difference in semantic distance between the released statement and the respective alternative statements. Our measure aligns with the convention that represents a hawkish tone as a positive number concerning the implied interest rate.

4.2 Monetary policy stance

Our monetary policy stance combines two measures: the novelty measure (4) and the tone measure (5). Building on the concept proposed by Ke et al. (2019), we define the monetary policy stance as:

$$\text{Stance}_t = \underbrace{(1 - sim(F_t, F_{t-1}))}_{\text{Novelty}_t} \underbrace{\left(\frac{sim(F_t^c, F_t) - sim(F_t^a, F_t)}{1 - sim(F_t^a, F_t^c)} \right)}_{\text{Tone}_t}. \tag{6}$$

An interpretation of Stance_t can be framed as follows: the semantic distance captured by our novelty term can be understood as the greatest disparity between the current meeting statement and the preceding one, while our tone measure can be seen as mitigating this disparity in a specific (toward more dovish or hawkish) direction.[5]

Analogous to equation (6), we derive two alternative monetary policy stances, representing the dovish and hawkish perspectives, respectively:

$$\text{Stance}_t^{dove} = -(1 - sim(F_t^a, F_{t-1})), \quad \text{Stance}_t^{hawk} = (1 - sim(F_t^c, F_{t-1})). \tag{7}$$

4.3 Dovish weight in the Federal Reserve's policy communication

Dovish weight in the Federal Reserve's policy communication is defined as

$$\begin{aligned} \text{Stance}_t &= w_t \text{Stance}_t^{dove} + (1 - w_t)\text{Stance}_t^{hawk}, \\ &= 1 - sim(F_t^c, F_{t-1}) - w_t\big(2 - sim(F_t^a, F_{t-1}) - sim(F_t^c, F_{t-1})\big). \end{aligned} \tag{8}$$

[5] To be precise, the distance metric between text embeddings captures not only the semantic difference but also the syntactic difference. Since the structure of FOMC statements rarely changes, the semantic dimension is likely to be dominant. But by comparing the released statement with a similar structure but different tones, we can further isolate the semantic dimension. For more details, see the online appendix.

By re-arranging (6) and (8), we obtain the expression for the dovish weight as follows:

$$w_t = \frac{1 - sim(F_t^c, F_{t-1}) - (1 - sim(F_t, F_{t-1}))\left(\frac{sim(F_t^c, F_t) - sim(F_t^a, F_t)}{1 - sim(F_t^a, F_t^c)}\right)}{2 - sim(F_t^a, F_{t-1}) - sim(F_t^c, F_{t-1})}. \tag{9}$$

When w_t increases (decreases), it brings the monetary policy stance of the released statement into greater alignment with the dovish (hawkish) alternative stance. This observation validates our interpretation of w_t as a dovish weight. We can readily deduce from (9) that $w_t = 1$ (0) when $F_t = F_t^a$ ($F_t = F_t^c$). Given that dovish FOMC members favor a policy stance closer to the dovish alternative, and actively contribute to FOMC deliberations to incorporate language resembling the dovish alternative statement, our text-based measure of dovish weight naturally reflects the relative influence wielded by dovish members within the FOMC.

4.4 Information exchange between the market and the Fed

As detailed in Section 3, the Board staff formulates policy alternatives by leveraging their insights into market perceptions. Therefore, we contend that these policy alternatives serve as informative summaries of the dovish and hawkish possibilities from the market's standpoint.

Financial market participants recognize that the FOMC's policy stance is composed of various alternative policy positions. While real-time access to the precise wording of these policy alternatives is restricted by a five-year publication delay for alternative statements, a wealth of valuable insights can still be extracted by scrutinizing meeting minutes released three weeks after the event or by analyzing intermeeting speeches. These resources offer a diverse array of perspectives on the spectrum of policy views held by FOMC members, facilitating a reasonable estimation of alternative policy boundaries.

For the sake of clarity, we use the term 'mutual information exchange' to describe the dynamic and interactive sharing of insights and information between the Federal Reserve and the market. This exchange fosters a shared understanding of economic conditions, policy intentions, and their potential consequences. In essence, we posit that it is reasonable to assume close alignment between market participants and the Federal Reserve Board staff regarding the impact of alternative language compared to the officially released statement.

4.5 Expected monetary policy stance and its surprise component

Building upon our perspective regarding the mutual information exchange between the market and the Fed, we proceed to formalize the process by which market participants shape their expectations through the introduction of notations.

We designate the anticipated market boundaries as $\text{Stance}_{t-\Delta}^{dove}$ and $\text{Stance}_{t-\Delta}^{hawk}$, which are established at time $t-\Delta$, immediately preceding the monetary policy announcement. It is important to clarify that, while closely related with each other, we do not equate $\text{Stance}_{t-\Delta}^{dove}$ with Stance_{t}^{dove} and $\text{Stance}_{t-\Delta}^{hawk}$ with Stance_{t}^{hawk} in (7), as doing so would assume that the market has perfect access to the precise wording of these policy alternatives.

We introduce $p_{t-\Delta}$ as the market's prevailing dovish probability, which consolidates the market's anticipated positions, encompassing both $\text{Stance}_{t-\Delta}^{dove}$ and $\text{Stance}_{t-\Delta}^{hawk}$. The market's expectation of monetary policy stance is given by

$$E_{t-\Delta}[\text{Stance}_t] = p_{t-\Delta}\text{Stance}_{t-\Delta}^{dove} + (1-p_{t-\Delta})\text{Stance}_{t-\Delta}^{hawk}. \tag{10}$$

The market's response is driven by the unexpected element of the monetary policy stance, defined as

$$MPS_t \equiv \text{Stance}_t - E_{t-\Delta}[\text{Stance}_t]. \tag{11}$$

It's worth noting that simplifications in (11) are possible when we make assumptions about $\text{Stance}_{t-\Delta}^{dove}$ and $\text{Stance}_{t-\Delta}^{hawk}$.

For our baseline analysis, we posit that

$$\text{Stance}_{t-\Delta}^{dove} = -\text{Novelty}_t, \quad \text{Stance}_{t-\Delta}^{hawk} = +\text{Novelty}_t. \tag{12}$$

We establish assumption (12) for two key reasons. From a conceptual perspective, it ensures the validity of the condition expressed in (13),

$$\text{Stance}_{t-\Delta}^{dove} \leq \text{Stance}_t \leq \text{Stance}_{t-\Delta}^{hawk}, \tag{13}$$

which can be derived from (6) due to the fact that Tone_t spans a range from -1 to 1. From an empirical standpoint, as demonstrated in Section 5, we further establish the robustness of our baseline approach by relaxing the assumption in (12) and instead proxying the anticipated market boundaries, $\text{Stance}_{t-\Delta}^{dove}$ and $\text{Stance}_{t-\Delta}^{hawk}$, using pre-meeting information available to the market.

Thanks to (12), we can streamline (11) to

$$\begin{aligned} MPS_t &= \text{Stance}_t - E_{t-\Delta}[\text{Stance}_t], \\ &= \text{Novelty}_t \cdot \text{Tone}_t - \Big(- p_{t-\Delta} \cdot \text{Novelty}_t + (1 - p_{t-\Delta}) \cdot \text{Novelty}_t \Big), \\ &= \text{Novelty}_t \cdot \big(\text{Tone}_t - 1 + 2p_{t-\Delta}\big). \end{aligned} \tag{14}$$

We highlight that the unexpected aspect of monetary policy offers an insightful perspective: the market's surprise doesn't necessarily reflect the maximum allowable deviation from the previous official statement. Instead, it reflects a reduced deviation influenced by the extent to which the market is taken aback by the announcement's tone.

4.6 Inferring monetary policy surprises

In the literature on identifying monetary policy shocks, intraday bond returns are frequently used as instrumental variables for monetary policy shocks, as shown in Gürkaynak et al. (2005). This approach is based on the assumption that changes in bond prices during a short time interval around policy announcements capture the immediate reaction solely to new policy information.

The relationship between intraday bond returns and monetary policy surprise is modeled as

$$r_t = -\beta MPS(p_{t-\Delta}) + \epsilon_t \tag{15}$$

where r_t is demeaned intraday bond returns; ϵ_t is white noise; $MPS(p_{t-\Delta})$, as defined in (14), is expressed (albeit with an abuse of notation) to emphasize its reliance on the probability of a dovish stance, while abstracting away from any evidence dependent on the monetary policy stance.

We search for β and $\{p_{t-\Delta}\}_{t=1}^{T}$ that maximize the rank correlation between high-frequency bond returns $\{r_t\}_{t=1}^{T}$ and the surprise component of monetary policy stance $\{MPS(p_{t-\Delta})\}_{t=1}^{T}$. In doing so, we ensure that $p_{t-\Delta}$ remains within the interval $[0, 1]$ for all t.[6] Specifically, we sort the time series of bond returns $\{r_t\}_{t=1}^{T}$ from most negative to most positive, scaling them with respect to an initial guess β. Let the ordering of the sorted-returns be indicated with new time subscripts $\{\tau_1, ..., \tau_T\}$:

$$\tilde{r}_{\tau_1} = \min\{r_t/\beta\}_{t=1}^{T}, \quad \tilde{r}_{\tau_T} = \max\{r_t/\beta\}_{t=1}^{T}. \tag{16}$$

[6]When $p_{t-\Delta}$ is not time-varying, ($p_{t-\Delta} = \overline{p}$) our estimate is identical to the maximum rank correlation estimator, see Han (1987) and Sherman (1993).

We maximize the following rank correlation function with respect to

$$\{p_{\tau_i-\Delta}\}_{i=1}^{T} = \text{argmax} \sum_{t \neq t'} \mathbf{1}(\tilde{r}_{\tau_t} > \tilde{r}_{\tau_{t'}}; \beta)\mathbf{1}\big(MPS(p_{\tau_t-\Delta}) < MPS(p_{\tau_{t'}-\Delta}); \beta\big). \tag{17}$$

To account for potentially multiple realizations of $\{p_{\tau_i-\Delta}\}_{i=1}^{T}$, we select the one that yields the most negative correlation by iteratively exploring β values, effectively minimizing the influence of ϵ_t in (15). Once we select $\{p_{\tau_i-\Delta}\}_{i=1}^{T}$, we can sort them back to match the original time subscript $\{p_{t-\Delta}\}_{t=1}^{T}$ and construct the corresponding $MPS(p_{t-\Delta})$ for each $t \in \{1, ..., T\}$.

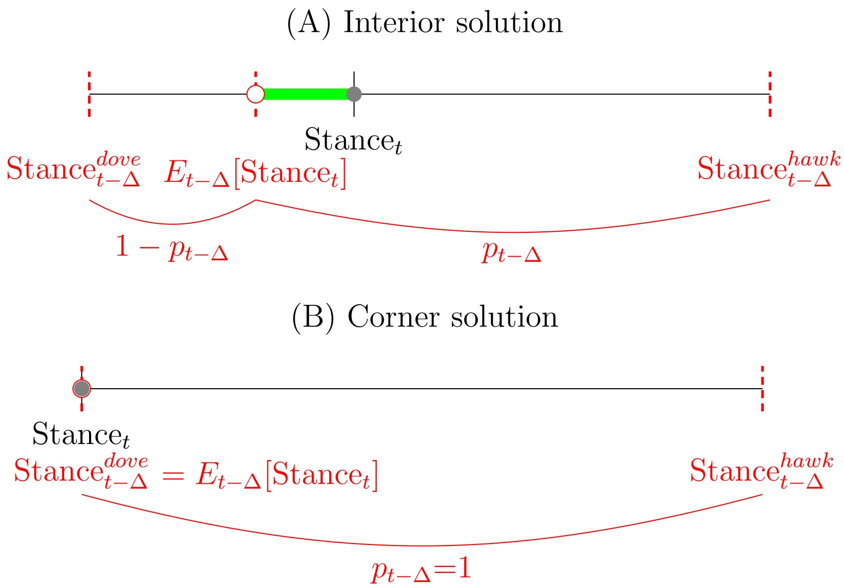
We provide an illustrative example of our maximum rank correlation approach below. For the sake of clarity, we set $\beta = 1$ in this illustration. It becomes apparent that by disregarding ϵ_t in (15), we can invert $MPS(p_{t-\Delta}) = -r_t$ and subsequently solve for $p_{t-\Delta}$ based on (14). Given the intraday bond return response, we can consistently reconstruct the monetary policy surprise, or equivalently, the expected policy stance that satisfies $0 < p_{t-\Delta} < 1$, as long as the boundary conditions are strictly upheld without equality. However, a corner solution for $p_{t-\Delta}$, wherein it is either truncated to 0 or 1, becomes inevitable when one of the boundary conditions holds with equality. Under this case, ϵ_t is bound to explain movements in r_t as it cannot be reconciled by $MPS(p_{t-\Delta})$.

Figure 1 visualizes a scenario where the market interprets a hawkish signal from a monetary policy announcement, implying that the market's expected policy stance is dovish, while intraday bond returns are negative. In Panel (A), we can see that the expected stance derived from the boundary condition $\text{Stance}_{t-\Delta}^{dove} < \text{Stance}_t < \text{Stance}_{t-\Delta}^{hawk}$ aligns with the narrative illustration. What is highlighted in green represents the constructed $MPS(p_{t-\Delta})$, which is positive matching negative intraday bond returns $r_t < 0$, from our maximum rank correlation approach. However, in Panel (B), we present a case where we fail to capture the market's expected stance when $\text{Stance}_{t-\Delta}^{dove} = \text{Stance}_t < \text{Stance}_{t-\Delta}^{hawk}$ and $r_t < 0$. Under this case, our approach sets $p_{t-\Delta} = 1$ and $MPS = 0$ as $p_{t-\Delta} > 1$ is infeasible. Hence, we cannot reconcile our MPS measure with the realized market response using the feasible dovish probability.

As shown in Gürkaynak et al. (2005) since intraday bond returns are frequently used as instrumental variables for monetary policy shocks, we conjecture that ϵ_t would contribute to (nearly) zero in (15).[7] Hence, the frequency of cases where $p_{t-\Delta}$ approaches the boundaries of $[0, 1]$, as illustrated in Panel (B) of Figure 1 is crucial for assessing the empirical validity of our

[7]When noise levels increase significantly, we observe that returns cannot be effectively reconciled by MPS. Through a simulation with noise ϵ_t drawn from a normal distribution with varying variances, we find a consistent decrease in computed correlations as noise magnitudes rise. This suggests that the variable $p_{t-\Delta}$ frequently approaches the boundaries of $[0, 1]$, leading to frequent occurrences of 0 or 1 in the time series, thereby constraining the achievable rank correlation to be well below 1.

Figure 1: Expected monetary policy stance and the role of boundaries



Note: Consider a scenario where the market interprets a hawkish signal from a monetary policy announcement, implying that the market's expected policy stance is dovish. In Panel (A), we can see that the expected stance derived from the boundary condition Stance$_{t-\Delta}^{dove}$ < Stance$_t$ < Stance$_{t-\Delta}^{hawk}$ aligns with the narrative illustration. However, in Panel (B), we present a case where we fail to capture the market's expected stance when Stance$_{t-\Delta}^{dove}$ = Stance$_t$ < Stance$_{t-\Delta}^{hawk}$ when $r_t < 0$.

measurement approach and the associated assumptions, such as those in (12). We discuss this in Section 5.

4.7 Counterfactual monetary policy stance and surprise

By utilizing intraday bond returns, we can determine $\{p_{t-\Delta}\}_{t=1}^{T}$, allowing us to construct the market's expectation of monetary policy stance $E_{t-\Delta}[\text{Stance}_t]$. Keeping $E_{t-\Delta}[\text{Stance}_t]$ constant, we can explore the counterfactual monetary policy surprise using the following expression:

$$MPS_t^{CF}(p_{t-\Delta}; B_{t-\Delta}) = \text{Stance}_t^{CF} - E_{t-\Delta}[\text{Stance}_t], \tag{18}$$
$$\text{Stance}_t^{CF} = (1 - sim(F_t^{CF}, F_{t-1}))\left(\frac{sim(F_t^c, F_t^{CF}) - sim(F_t^a, F_t^{CF})}{1 - sim(F_t^a, F_t^c)}\right),$$

where Stance$_t^{CF}$ represents the stance measure when the released statement is instead F_t^{CF}. This counterfactual approach enables us to investigate the impact of different statements on monetary policy expectations while holding the market's anticipation constant.

It is crucial to note that our monetary policy stance is solely derived from text analysis, while its surprise component is inferred from intraday bond returns, following the standard approach in the literature (e.g., Gürkaynak et al. (2005)). This unique structure allows us to conduct language counterfactual experiments as outlined in (18). By adopting this approach, we enable policymakers to explore alternative descriptions of the economy and policy prescriptions and their potential impact on financial markets when crafting policy statements.

5 Empirical Results

5.1 Text data

The Federal Reserve Board staff began preparing alternative FOMC statements from the March 2004 FOMC meeting. They are released to public with a five-year lag. Our dataset consists of 99 FOMC statements issued between March 2004 and December 2016. We have excluded two inter-meeting announcements (August 2007 and January 2008) and four meetings due to the unavailability of alternative statements (September 2005, December 2005, August 2008, and April 2009). In cases where multiple versions of hawkish alternative statements exist (e.g., C and D), we utilize the most extreme one (D) to identify the tone of the released statement.

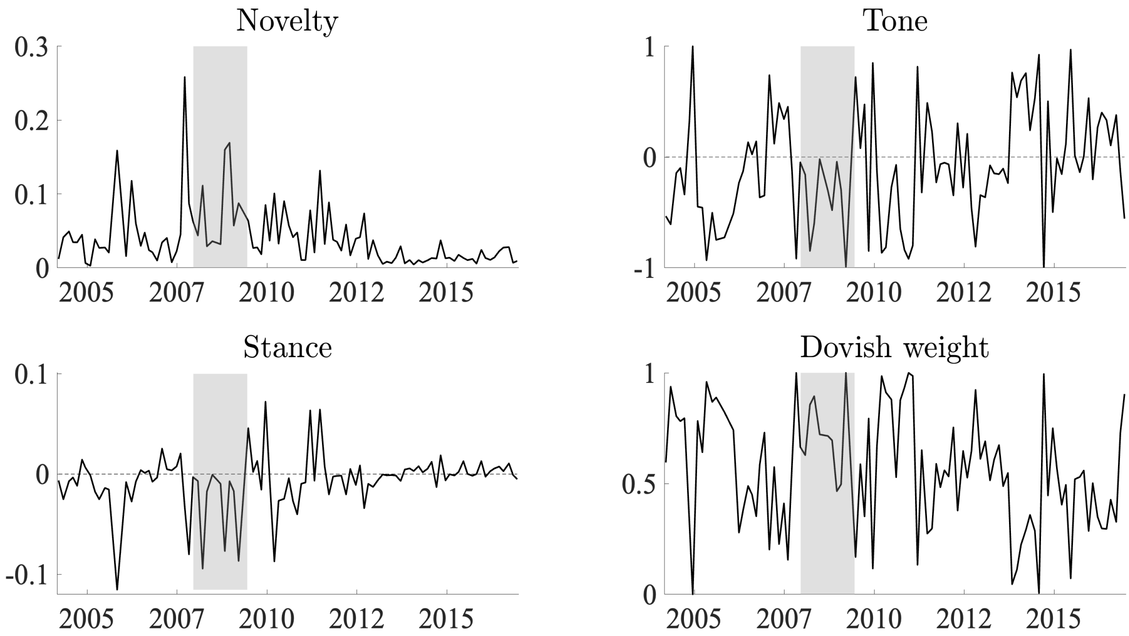
During our analysis period, Lawrence H. Meyer, a former Federal Reserve Board governor (1996-2002), distributed 'monetary policy insights' newsletters to paid clients. These newsletters, issued on the last Friday before each FOMC meeting, analyzed inter-meeting macroeconomic data and FOMC member speeches to predict changes in upcoming FOMC meeting statements. Starting from the September 2008 FOMC meeting, Mr. Meyer's statement drafts closely resembled the structure of actual FOMC statements. We use Mr. Meyer's statement drafts as empirical proxies for the expected FOMC statement drafts by market participants to validate our MPS measure construction assumption.

5.2 Text-based monetary policy measures

Figure 2 presents the time-series of novelty, tone, and stance, as derived from the USE using the definitions provided in (4), (5), and (6), respectively. Additionally, it shows the dovish weight defined in (9).

Our novelty measure demonstrates significant variability over time, allowing us to gauge the differences between the released statement and its immediate predecessor. Additionally, our tone measure spans from -1 to 1, signifying that the semantic distance between the dovish

Figure 2: Monetary policy stance and its components



Notes: We present the time-series of novelty, tone, and stance, as derived from text analysis using the definitions provided in (4), (5), and (6), respectively. Additionally, we show the dovish weight defined in (9).

and hawkish alternative statements surpasses the disparity between the released statement and the respective alternative statements.[8] Our stance measure indicates the direction in which the released statement differs from its immediate predecessor, either in a dovish or hawkish manner. For instance, during the Great Recession, our measure shows a significant shift in a dovish direction for the released statement. The dovish weight in the Federal Reserve's policy communication reveals that the released statement is not consistently the average of the two alternatives, as the weights can vary significantly from low to high values over time. For example, the weight for the August 2011 FOMC meeting is approximately 0.3, which reflects intense debates between the dovish and hawkish camps, as discussed in Section 3. The dovish camp successfully introduced explicit time-based forward guidance in the official statement. However, much of the SOMA portfolio policy descriptions did not come from the dovish alternative statements. This discrepancy is indicated by the dovish weight being 0.3. The dovish weight nearly doubled during the September 2011 FOMC meeting when the FOMC adopted a portfolio

[8]The sole exception occurs in September 2014, where the released statement closely resembles Alt A but is slightly more distant from Alt C than Alt A. Alt C replaces 'considerable time after the asset purchase program ends' with 'some time after the asset purchase program ends' for forward guidance on the interest rate path, while Alt A offers threshold-based guidance based on future inflation rate, making it challenging to classify their distance. Due to this uncertainty, we assigned the most dovish tone score to the released statement for that episode.

Table 4: Dovish weight and macroeconomic predictors observed before the FOMC announcement

	Correlation	p-value
Nonfarm payrolls surprise	-0.16	0.12
Employment growth	-0.13	0.20
S&P 500 change	-0.11	0.28
Yield curve slope	0.06	0.53
Commodity prices	0.09	0.39
Treasury skewness	0.05	0.61

Notes: We consider the six predictors examined in Bauer and Swanson (2023) and present their methodology for constructing these six series here. Nonfarm payrolls surprise: the surprise component of the most recent nonfarm payrolls release prior to the FOMC announcement, measured as the difference between the released value of the statistic minus the median expectation for that release from the Money Market Services survey. Employment growth: the log change in nonfarm payroll employment from one year earlier to the most recent release before the FOMC announcement, as used in Cieslak (2018). S&P 500 change: the log change in the S&P500 stock market index from three months (65 trading days) before the FOMC announcement to the day before the FOMC announcement. Yield curve slope: the change in the slope of the yield curve from three months before the FOMC announcement to the day before the FOMC announcement, measured as the second principal component of one- to ten-year zero-coupon Treasury yields from Gürkaynak et al. (2007). Commodity prices: the log change in the Bloomberg Commodity Spot Price index (BCOMSP) from three months before the FOMC announcement to the day before the FOMC announcement. Treasury skewness: the implied skewness of the ten-year Treasury yield, measured using options on ten-year Treasury note futures with expirations in 1–3 months, averaged over the preceding month, from Bauer and Chernov (2022).

policy similar to what is described in alternative A, a feature that was missing in alternative C.

Table 4 calculates the correlation between the dovish weight and key macroeconomic predictors considered by Bauer and Swanson (2023) which are acknowledged to exhibit an intuitive connection with the monetary policy of the Federal Reserve. Given the range of the corresponding p-values, we narrow our focus to the three cases characterized by comparatively lower p-values. As anticipated, we discover that the dovish weight shows an inverse correlation with positive surprises in labor market indicators and the stock market return.

5.3 Monetary policy surprises inferred from EuroDollar futures

We utilize intraday EuroDollar futures returns with a 12-month maturity computed over a 5-minute window around the time of announcements to infer monetary policy surprises, following the approach of maximizing rank correlation as explained in Section 4.6.

As a reminder, our baseline approach hinges on the assumption detailed in (12) as a foundation for constructing the expected monetary policy stance, which is equivalent to inferring monetary policy surprises. Within this framework, the market has the potential to anticipate the tone of monetary policy as either highly dovish (-1) or significantly hawkish (+1). We maintain this assumption because we believe that accurately predicting the extent of divergence between the official statement and its preceding counterpart can be reasonably accomplished. Our confidence in this assertion is supported by the findings in Section 3, where we present survey results from primary dealers who were asked about anticipated changes in the FOMC statement within the Survey of Primary Dealers.

In this section, we compile empirical proxies for the statement draft, readily accessible to market participants just before the meeting. This not only demonstrates the empirical validity of our assumption (12) but also serves as a robustness check for our baseline approach. We denote Mr. Meyer's statement draft before the FOMC meeting at time t as $F^{LM}_{t-\Delta}$ and directly proxy the anticipated market boundaries as follows:

$$\text{Stance}^{dove}_{t-\Delta} = -\text{Novelty}^{LM}_{t-\Delta} \equiv -(1 - sim(F^{LM}_{t-\Delta}, F_{t-1})), \quad \text{Stance}^{hawk}_{t-\Delta} = +\text{Novelty}^{LM}_{t-\Delta}. \tag{19}$$

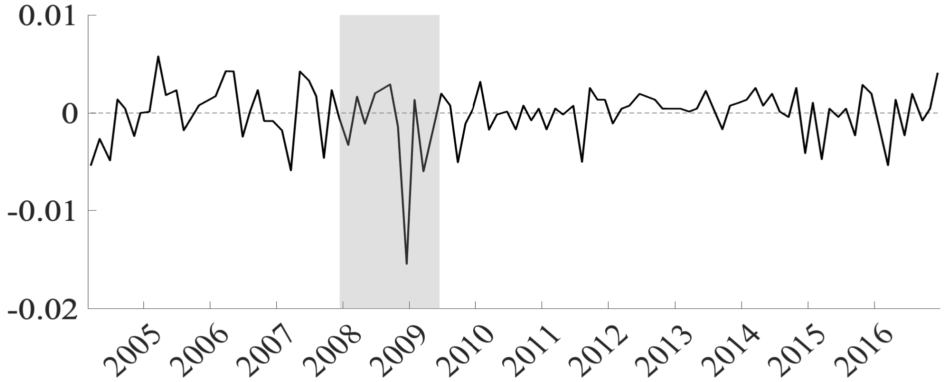
When examining data from September 2008 to December 2016, we observe a substantial sample correlation of approximately 0.75 between Novelty_t and $\text{Novelty}^{LM}_{t-\Delta}$. This noteworthy correlation, derived solely from Mr. Meyer's statement drafts, underscores the plausibility of assuming that market participants possess a reasonably accurate anticipation of the potential deviations or new elements that might be introduced in the forthcoming meeting statement. It's conceivable that incorporating additional information could further enhance this dimension, reinforcing our argument that predicting the extent of divergence between the official statement and its preceding counterpart is indeed feasible.

We have established two measures of monetary policy surprises: one utilizing the baseline approach, and the other drawing on Meyer's speeches. Remarkably, these two surprise measures exhibit a notably high correlation coefficient of 0.87, thereby reinforcing the robustness of our baseline approach.

As detailed in Section 4.6, we investigate instances where $p_{t-\Delta}$ approaches the boundaries of $[0, 1]$. We identified five cases: (i) on December 14, 2004 and June 17, 2015, where it approached 0; and (ii) on May 3, 2005, September 17, 2014, and December 14, 2016, where it approached 1. As we illustrated in Panel (B) of Figure 1, only these five instances suggest that the boundary conditions we impose may not always lead to feasible dovish probabilities. Thus, we argue that our approach reasonably captures the market's expected stance, which is most often within the boundary set by alternative statements.

Throughout the rest of the paper, we explore the diverse implications of monetary policy surprises derived from the baseline approach. The associated time-series is depicted in Figure 3.

Figure 3: Monetary policy surprise



Notes: We provide the time-series of monetary policy surprises (= stance − expected stance), which are defined in (14). These surprises are inferred from intraday EuroDollar futures returns with a 12-month maturity, computed over a 5-minute window around the time of announcements, following the approach of maximizing rank correlation as explained in Section 4.6.

Table 5: Comparison with other measures of monetary policy surprises

	MP surprise	MP stance
Bauer and Swanson (2023)	0.85	-0.01
- Orthogonalized measure	0.74	-0.10
Bu et al. (2020)	0.56	0.14
Nakamura and Steinsson (2018)	0.81	0.00
Swanson (2017) (FFR+FG+LSAP)	0.73	0.08
- Federal funds rate (FFR) factor	0.36	-0.01
- Forward guidance (FG) factor	0.69	-0.06
- Large-scale asset purchase (LSAP) factor	-0.20	-0.22

Notes: The correlation with other existing measures of monetary policy surprises is computed using the available samples. Bauer and Swanson (2023)'s orthogonalized measure is the residual from regressing the original Bauer and Swanson (2023)'s monetary policy surprise measure onto six predictors in Table 4 observed before the FOMC announcement. The factors in Swanson (2017) are largely distinguished by their different loadings on the maturity spectrum of the underlying interest rate data.

5.4 Comparison with other monetary policy surprise measures

Our surprise measure exhibits a high correlation with other measures of monetary policy shocks. Table 5 presents a sample correlation of our surprise measure with those measures.

Each of these approaches provides unique methods to identify monetary policy shocks, thereby contributing valuable insights into understanding the effects of monetary policy decisions on different financial assets and the overall economy. Bauer and Swanson (2023) identify monetary policy shocks using a structural vector autoregression (VAR) model with external instruments. They use the first principal component of high-frequency changes in four Eurodollar futures rates as the instrument. Bu et al. (2020) construct monetary policy shocks based on the idea that the variance of daily bond returns is higher on FOMC days compared to non-FOMC days due to the monetary policy announcement. They use information from the entire yield curve, including near-term and long-term maturities (up to thirty years). Nakamura and Steinsson (2018) identify monetary policy shocks based on the assumption that unexpected changes in a 30-minute window during FOMC announcements arise from news about monetary policy. Swanson (2017) identifies multiple dimensions of monetary policy shocks using eight different asset prices, including three Treasury bond yields (with maturities of 2, 5, 10 years) and five interest rate futures used in Nakamura and Steinsson (2018).

Indeed, the high correlations observed between our surprise measure and those of Bauer and Swanson (2023), Nakamura and Steinsson (2018), and Swanson (2017) indicate that our measure effectively captures the surprise component of monetary policy decisions surrounding FOMC announcements. Notably, sharing the highest correlation with Bauer and Swanson (2023) is not unexpected, as both approaches rely on intraday EuroDollar futures returns, which can capture information about both the current federal funds rate target and the future policy path. This similarity in data sources likely contributes to the strong alignment between the two measures.

5.5 Asset return prediction with monetary policy surprises

To examine the effects of monetary policy surprises on Treasury bonds and stock prices, we conduct regressions using changes in log prices (or yields) as the dependent variable and "standardized" monetary policy surprises as the independent variable. The standardization process imposes that the daily change in the one-year maturity Treasury yield, corresponding to the standardized monetary policy surprise, is fixed at 0.25 or 25 basis points.

Panel (A) of Table 6 presents the estimation results using intraday Treasury futures of maturities 5- and 10-years, respectively, computed over various time intervals for computing returns. The coefficient loading on policy surprise is consistently found to be strongly and negatively significant across different permutations, indicating a notable impact of monetary policy surprises on these Treasury futures.

Panel (B) of Table 6 displays the estimation results from the regression involving daily one-year Treasury yield changes, considering both zero-coupon and instantaneous forward rates. A one-unit increase in monetary policy surprise is observed to significantly increase the one-year

Table 6: Return predictability regression

(A) Intraday Treasury bond return prediction

Return interval	α	β	t-stat (α)	t-stat (β)	R^2
5-year maturity futures					
10 minutes	0.02	-1.40	1.57	-5.90	0.57
30 minutes	0.03	-1.67	2.08	-8.93	0.60
60 minutes	0.04	-1.88	1.84	-7.12	0.49
90 minutes	0.06	-2.00	2.43	-7.26	0.51
10-year maturity futures					
10 minutes	0.03	-1.85	1.30	-4.12	0.39
30 minutes	0.05	-2.15	1.62	-5.07	0.43
60 minutes	0.07	-2.47	1.74	-5.12	0.39
90 minutes	0.08	-2.66	2.21	-5.58	0.42

(B) Daily Treasury yield change prediction

	α	β	t-stat (α)	t-stat (β)	R^2
1-year maturity zero-coupon rate					
	-0.00	0.25	-0.47	5.27	0.35
1-year maturity instantaneous forward rate					
	-0.00	0.36	-0.00	4.89	0.38

(C) Intraday stock return prediction

Return interval	α	β	t-stat (α)	t-stat (β)	R^2
10 minutes	0.05	-2.14	1.19	-4.52	0.27
30 minutes	0.09	-1.81	1.58	-2.55	0.14
60 minutes	0.20	-3.38	3.28	-4.84	0.31
90 minutes	0.18	-3.22	2.41	-5.61	0.23

(D) Daily stock return prediction

α	β	t-stat (α)	t-stat (β)	R^2
0.39	-3.99	3.32	-3.09	0.15

Notes: Panel (A) examines intraday Treasury futures returns for maturities of 5- and 10-years. Panel (B) utilizes daily Treasury yield data from Gürkaynak et al. (2007). Panel (C) examines intraday E-min S&P 500 futures returns. Panel (D) examines daily CRSP value-weighted stock returns. For panels (A) and (C), intraday returns are computed using 10-, 30-, 60-, and 90-minute time intervals. To address the generated regressor's influence, t statistics are computed with bootstrap standard errors.

Treasury zero-coupon and forward rates by 25 and 37 basis points, respectively. The significant and consistent results from the regression analyses indicate that monetary policy surprises, as identified by the market reaction during short windows around FOMC announcements, have lasting impacts on short-term Treasury yields beyond narrow windows, moving in the intended direction.

Given that the Federal Reserve intervenes in Treasury markets to influence interest rates, it is not surprising that bonds react to monetary policy surprises. However, the impact of monetary policy extends beyond bond returns and influences the broader financial market conditions, including the stock market. Understanding the connections between monetary policy and asset prices, beyond bond returns, is crucial for evaluating monetary policy transmission channels, as emphasized by Bernanke and Kuttner (2005). The estimation results, summarized in Panel (C) and (D) of Table 6, indicate that monetary policy surprise significantly predicts stock returns at various window intervals. On average, an increase in surprise that leads to a 25 basis point rise in one-year Treasury yields results in a 2-3 percentage point drop in stock prices. The R^2 values are approximately 20-30% across different return window intervals. Our findings support Bernanke and Kuttner (2005), showing a significant decrease in stock returns after monetary tightening. The difference in magnitude (approximately 1-2 percentage points) is not large given the sampling and estimation uncertainty.[9]

5.6 Macroeconomic impacts of monetary policy surprises

In this section, we employ a similar approach to Bauer and Swanson (2023) to assess the effects of a monetary policy shock on various macroeconomic variables. Specifically, we investigate the effects of a monetary policy shock on log industrial production, log consumer price index, the excess bond premium by Gilchrist and Zakrajšek (2012), and the two-year Treasury bond yield. We use our orthogonalized monetary policy surprise measure as an instrument z_t for the four-variable monthly VAR, with a lag length of 12 represented by[10]

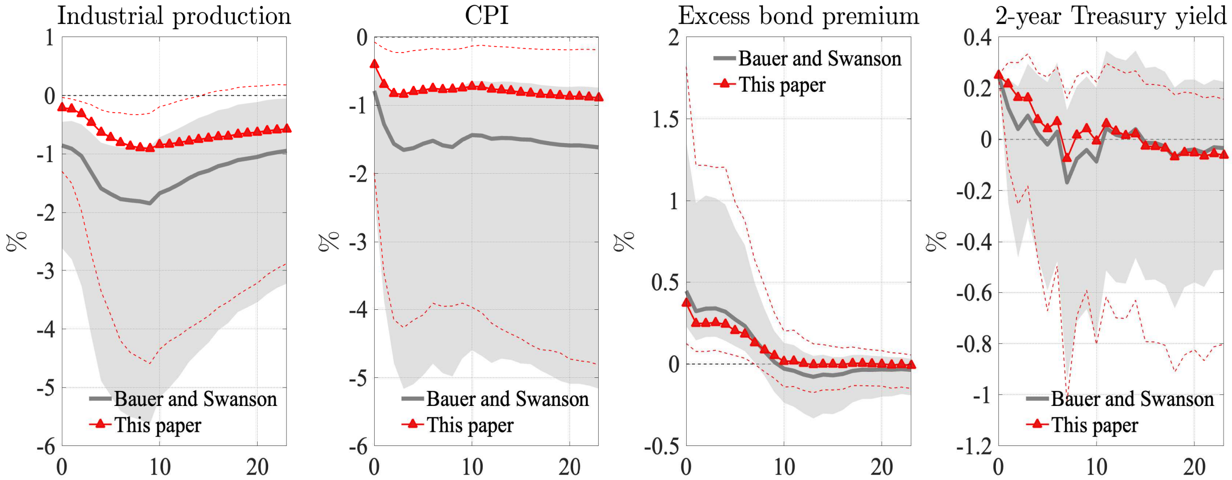
$$y_t = A_0 + \sum_{k=1}^{12} A_k y_{t_k} + u_t, \quad u_t = S\epsilon_t \sim (0, \Omega), \quad \epsilon_t \sim (0, I_4). \tag{20}$$

Here, y_t denotes the vector of macroeconomic variables, and A_0, A_k, and u_t are the coefficient matrices and structural residuals, respectively. The matrix S contains the identifying structural shocks ϵ_t from reduced-form VAR residuals u_t.

[9]Our sampling period of 2004-2016 does not overlap with Bernanke and Kuttner (2005).

[10]The orthogonalization procedure for the monetary policy surprise is detailed in the legend of Table 5.

Figure 4: SVAR with high-frequency monetary policy surprise as an external instrument



Notes: Following Bauer and Swanson (2023), we estimate a SVAR with a 12-month lag using the following variables: log industrial production, log consumer price index, excess bond premium Gilchrist and Zakrajšek (2012), and the two-year Treasury bond yield. To instrument our analysis, we employ the orthogonalized monetary policy surprise measure derived from the residuals of a regression on six predictors explained in Table 4 observed before the FOMC announcement. While we estimate the VAR coefficients using a longer sample from 1973:M1 to 2019:M12, the orthogonalized high-frequency monetary policy surprises are available only from 2004:M3 to 2016:M12. To facilitate comparison, we present the impulse responses obtained by utilizing the orthogonalized high-frequency monetary policy surprise instrument employed in Bauer and Swanson (2023).

We standardize a one-standard deviation positive monetary policy shock to raise the two-year Treasury bond yield by 25 basis points. Subsequently, we estimate the impulse responses of the vector of macroeconomic variables y_t to this monetary policy shock $\epsilon_{mp,t}$. The VAR coefficients are inferred from an extended sample covering the period from 1973:M1 to 2019:M12, utilizing orthogonalized high-frequency monetary policy surprises observed between 2004:M3 and 2016:M12. To facilitate comparison, we re-estimate the SVAR using Bauer and Swanson (2023)'s instrument while matching the same sample availability and overlay the results. The respective impulse response functions of the VAR variables are displayed in Figure 4.

Given the high correlation between our surprise measure and that of Bauer and Swanson (2023) (as shown in Table 5), it is not surprising that the two sets of impulse responses exhibit similarities. However, due to the relatively short estimation sample from 2004 to 2016, the confidence intervals are wide, which may impact the statistical significance of the impulse responses. This observation is reflected in the uncertainty surrounding the estimates, particularly in the short-term.

In summary, our findings suggest that our text-based surprise measure, being highly correlated with Bauer and Swanson (2023)'s measure, demonstrates similar impulse responses of macroeconomic variables to a monetary policy shock, thus further supporting the effectiveness of our approach in capturing the impact of monetary policy decisions.

5.7 Counterfactual policy evaluation

To make things easier to understand how we conduct language counterfactual experiment, we recap Section 4.7. Our method involves analyzing text and using intraday bond futures data to figure out what the expected monetary policy is and how it might surprise the market. This approach lets us conduct an experiment where we change the language used to communicate monetary policy and see how it affects financial markets. We do this by comparing the 'unchanged' expected stance to a 'counterfactual' stance based on our text analysis. This helps policymakers explore different scenarios and how they could impact the stock market, as seen in policy statements. Our study introduces the 'language counterfactual' concept, which is an important addition to existing research.

We delve into both dovish and hawkish alternatives as part of our counterfactual analysis. In practical terms, the words we choose in this analysis affect how the market expects the Federal Reserve to act in the next FOMC meeting. However, it's important to note that our analysis focuses solely on the one-time impact on the market and considers this a secondary concern.

First, we consider the dovish alternative counterfactual experiment. For the August 2011 meeting, we assess the stock market impact of releasing a more dovish statement (alternative A) announcing changes in the composition of the Federal Reserve's balance sheet (known as "System Open Market Account", SOMA in short) shown in Table 3. The released statement does not announce changes in the balance sheet as in alternatives B and C.

Using the values $\widehat{\alpha} = 0.18$ and $\widehat{\beta} = -3.22$ from Panel (C) of Table 6 (last row), we generate predicted stock returns based on the counterfactual monetary policy surprise component, allowing us to assess its impact on the stock returns (defined in the 90-minute interval). It is worth emphasizing that this analysis focuses on the replacement of a single data point, corresponding to the release date of the August 2011 FOMC statement, while maintaining all other factors constant. This exercise involves modifying the dovish weight, originally calculated at 0.30 based on the released statement, to reflect an alternative scenario where the counterfactual released statement aligns with alternative A implying dovish weight of one.

In Panel (A) of Table 7, we present the counterfactual predictions associated with increasing dovish weights as we replace some paragraphs in the released statement by the corresponding paragraphs in alternative A, starting from 0.30. Firstly, we change the first paragraph P_1 describing the economic activity in the statement. Since the released statement is very similar

Table 7: Counterfactual policy communication and its impact on the stock market

(A) August 2011 FOMC meeting

Actual value	Prediction	Counterfactual prediction	Dovish weight in (9)	Adjustment
		0.87	0.30	N/A
		0.87	0.30	Replace P_1
		2.51	0.44	Replace part 2 of P_4
0.63	0.87	6.85	0.84	Replace part 1 of P_4 with \$200 bil
		7.12	0.86	Replace part 1 of P_4 with \$300 bil
		8.21	0.96	Replace P_1 and P_4
		8.75	1.00	Replace all with Alt A

(B) December 2016 FOMC meeting

Actual value	Prediction	Counterfactual prediction	Dovish weight in (9)	Adjustment
		-0.39	0.90	N/A
-0.65	-0.39	-1.14	0.77	Replace P_1
		-3.76	0.34	Replace P_1 and P_2
		-5.82	0.00	Replace all with Alt C

Source: We take the OLS estimates, $\widehat{\alpha} = 0.19$ and $\widehat{\beta} = -3.22$, as presented in Panel (C) of Table 6. We present the observed stock returns, the predicted values generated by these OLS estimates, as well as the counterfactual prediction values across a range of dovish weight scenarios. The description in the parenthesis denotes paragraphs changed into those in alternative statements. All values are expressed in percentage terms.

to alternative A in describing economic activities, the dovish weight changes only minimally with this modification.

Next, we focus on modifying the fourth paragraph (P_4) that outlines the balance sheet policies concerning the SOMA portfolio. This alteration is aligned with the content found in alternative A. We implement these changes in three distinct ways. To simplify the discussion, we divide P_4 into two components: "Part 1," which refers to "purchasing \$400 billion of Treasuries...," and "Part 2," which pertains to "maintaining reinvestment policy...," as indicated in Table 3. Our analysis reveals intriguing results. When we solely modify part 2 of P_4 while leaving part 1 unchanged, it leads to an increase in the dovish weight, reaching 0.44. However, when part 1 is transformed to announce a maturity extension policy, we observe a substantial spike in the dovish weight, soaring to a range between 0.84 and 0.86. What's particularly noteworthy is that, due to our numerical training, the algorithm accurately predicts that \$300 billion is closer

Table 8: Overview of alternatives for the December 8, 2016 FOMC statements

	December Statement	December alternatives A	B	C
Inflation Compensation	moved up considerably but still are low	moved up but remain low	moved up considerably but still are low	moved up considerably
Target FFR	$\frac{1}{2}$ to $\frac{3}{4}$%	$\frac{1}{4}$ to $\frac{1}{2}$%	$\frac{1}{2}$ to $\frac{3}{4}$%	$\frac{1}{2}$ to $\frac{3}{4}$%
Forward guidance	gradual adjustments in the stance of monetary policy	gradual adjustments in the stance of monetary policy	gradual adjustments in the stance of monetary policy	additional gradual adjustments in the stance of monetary policy
	warrant only gradual increases	warrant only gradual increases	warrant only gradual increases	warrant additional gradual increases

Source: Authors' construction based on **December 2016 Tealbook B**.

to \$400 billion than \$200 billion is to \$400 billion.

When the dovish weight reaches one, it signifies that the counterfactual released statement is entirely equivalent to alternative A. It's crucial to exercise caution when interpreting the counterfactual prediction value. In the hypothetical scenario where alternative A had been released, the stock return within the 90-minute interval would have surged by over 8 percentage points. As advocated by Leeper and Zha (2003), we find it prudent to restrict our consideration to a modest counterfactual policy that doesn't significantly disrupt agents' expectations regarding policy actions, especially within the context of our assumption of linear dynamics, where we are essentially multiplying the counterfactual MPS with the OLS estimates. Consequently, a complete shift of the dovish weight from 0.30 to 1.00 would not be considered modest.[11] Nevertheless, our analysis still offers an intriguing insight: had the Federal Reserve adopted the maturity extension policy, it would have triggered an exceptionally positive reaction among stock traders. Our counterfactual prediction would appear more plausible when considering a modest language change, such as modifying only "Part 2" of P_4.

Shifting our attention to the December 2016 FOMC meeting in Table 8, we conduct another counterfactual analysis, this time exploring the scenario of what might have occurred if the more hawkish alternative C had been released instead. As we reduce the dovish weight from 0.90 toward zero by replacing some paragraphs in the released statement to those in alternative C,

[11] Given that intraday stock returns typically fall within the -2% to 2% range, it's fair to assert that a substantial shift from a dovish weight of 0.30 to 1.00 doesn't align with the modest standards established by Leeper and Zha (2003).

the counterfactual statement aligns increasingly with alternative C. On inflation compensation, alternative C mentions that the measure "moved up considerably" but does not add that the level is still low unlike alternatives A and B as well as the released statement. On forward guidance, the description in alternative C implies that additional increases in the federal funds rate would have happened sooner than in alternatives A or B. This "additional" language is used in both the second paragraph and the fourth paragraph. As before, we change the released statement incrementally to make it identical to alternative C. Applying the same coefficients, our findings suggest that the release of alternative C would have led to a stock market return reduction of more than 5-6 percentage points, as depicted in Panel (B) of Table 7. The market's response is evidently influenced by the hawkish signal conveyed through the forward guidance, particularly as discussed in the second paragraph (P_2). Furthermore, emphasizing the upward trend in inflation compensation exclusively within the first paragraph (P_1) without altering the forward guidance would have also resulted in a roughly 1% decrease in stock market returns. This exercise underscores the profound impact of FOMC communication, highlighting its pivotal role as a policy tool.

6 Conclusion

Since March 2004, the FOMC has deliberated on alternative policy statements prepared by the Federal Reserve staff before each FOMC meeting. Two alternative statements capture the hawkish or dovish deviation from the market's central tendency of expectations, offering cross-sectional variations around the released statement. We use a novel natural language processing algorithm based on a deep learning architecture to analyze alternative FOMC statements, with the goal of identifying the novelty and tone of the released statement. This USE algorithm detects the contextual meaning of words within the statement and quantifies the information conveyed by language in alternative statements. Furthermore, we fine-tune the USE algorithm using artificial text datasets to enhance its ability to recognize numeric values, as FOMC statements often involve numeric variables for describing policy actions.

We construct a measure of monetary policy surprises by combining high-frequency bond returns around FOMC announcements with text analysis of policy statements using the USE algorithm. We find that unexpected policy tightening leads to an average decline in stock market returns, while easing results in an increase. This finding supports the notion that FOMC communication has consistently influenced financial market conditions in the expected direction since at least 2004, consistent with recent empirical research. Leveraging monetary policy stance defined by text analysis, we create counterfactual monetary policy surprises by altering languages in the released statement to what is in alternative statements. Two language counterfactual exercises suggest that altering language in FOMC statements could significantly impact financial market conditions, underscoring the importance of FOMC communication as a policy tool.

The literature on large-scale language models is rapidly expanding and improving human language understanding and reasoning (see Manning (2022)). Our paper suggests that incorporating such tools into economic analysis can be valuable for conducting rigorous assessments of economic narratives. While our primary focus is evaluating FOMC communications through post-meeting statements, our method of fine-tuning a pre-trained large-scale language model with task-specific datasets has broader applicability in quantifying economic narratives (see Shiller (2017) and Shiller (2020)).

References

Acosta, Miguel, "A New Measure of Central Bank Transparency and Implications for the Effectiveness of Monetary Policy," 2022. Manuscript.

Araci, Dogu, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," 2019.

Aruoba, Boragan and Thomas Drechsel, "Identifying Monetary Policy Shocks: A Natural Language Approach," 2023.

Bauer, Michael D. and Eric T. Swanson, "A Reassessment of Monetary Policy Surprises and High-Frequency Identification," *NBER Macroeconomics Annual*, 2023, *37* (1), 87–155.

_ **and Mikhail Chernov**, "Interest Rate Skewness and Biased Beliefs," *Journal of Finance*, 2022.

Bernanke, Ben S., "Central Bank Independence, Transparency, and Accountability," in "Speech of Chairman Ben S. Bernanke at the Institute for Monetary and Economic Studies International Conference" 2010.

Bernanke, Ben S and Kenneth N Kuttner, "What explains the stock market's reaction to Federal Reserve policy?," *The Journal of finance*, 2005, *60* (3), 1221–1257.

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen, "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence," *Journal of Economic Literature*, 2008, *46* (4), 910–945.

Bu, Chunya, John Rogers, and Wenbin Wu, "Forward-Looking Monetary Policy and the Transmission of Conventional Monetary Policy Shocks," 2020.

Caldara, Dario and Matteo Iacoviello, "Measuring geopolitical risk," *American Economic Review*, 2022, *112* (4), 1194–1225.

Cer, Daniel, Yinfei Yang, Shengyi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Universal Sentence Encoder," 2018. Manuscript.

Cieslak, Anna, "Short-Rate Expectations and Unexpected Returns in Treasury Bonds," *The Review of Financial Studies*, 04 2018, *31* (9), 3265–3306.

__ , **Stephen Hansen, Michael McMahon, and Song Xiao**, "Policymakers' Uncertainty," *Manuscript*, 2022.

Cochrane, John H. and Monika Piazzesi, "The Fed and Interest Rates - A High-Frequency Identification," *American Economic Review*, May 2002, *92* (2), 90–95.

Curti, Filippo and Sophia Kazinnik, "Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences," *Journal of Monetary Economics*, 2023.

Faust, Jon, Eric T. Swanson, and Jonathan H. Wright, "Identifying VARS based on high frequency futures data," *Journal of Monetary Economics*, 2004, *51* (6), 1107–1131.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Giavazzi, Francesco, Felix Iglhaut, Giacomo Lemoli, and Gaia Rubera, "Terrorist Attacks, Cultural Incidents and the Vote for Radical Parties: Analyzing Text from Twitter," Technical Report, National Bureau of Economic Research 2020.

Gilchrist, Simon and Egon Zakrajšek, "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, June 2012, *102* (4), 1692–1720.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera, "The voice of monetary policy," *American Economic Review*, 2023, *113* (2), 548–584.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart, *Text as data: A new framework for machine learning and the social sciences*, Princeton University Press, 2022.

Gürkaynak, Refet S, Brian Sack, and Eric Swanson, "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements," *International Journal of Central Banking*, 2005, *1*, 55–93.

__ , __ , **and Jonathan Wright**, "The US Treasury yield curve: 1961 to the present," *Journal of Monetary Economics*, 2007, *54* (8), 2291–2304.

Han, Aaron K., "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 1987, *35*, 303–316.

Handlan, Amy, "Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning," 2022. Manuscript.

Hansen, Anne Lundgaard and Sophia Kazinnik, "Can ChatGPT Decipher Fedspeak?," 2023. Manuscript.

Hansen, Stephen and Michael McMahon, "Shocking language: Understanding the macroeconomic effects of central bank communication," *Journal of International Economics*, 2016, *99*, S114–S133. 38th Annual NBER International Seminar on Macroeconomics.

_ , _ , and Andrea Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics*, 2017, *133* (2), 801–870.

Husted, Lucas, John Rogers, and Bo Sun, "Monetary policy uncertainty," *Journal of Monetary Economics*, 2020, *115*, 20–36.

Jegadeesh, Narasimhan and Di Andrew Wu, "Deciphering Fedspeak: The information content of FOMC meetings," *Available at SSRN 2939937*, 2017.

Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu, "Predicting Returns with Text Data," 2019. Manuscript.

Kuttner, Kenneth N, "Monetary policy surprises and interest rates: Evidence from the Fed funds futures market," *Journal of Monetary Economics*, 2001, *47* (3), 523–544.

Laver, Michael, Kenneth Benoit, and John Garry, "Extracting Policy Positions from Political Texts Using Words as Data," *American Political Science Review*, 2003, *97* (2), 311–331.

Leeper, Eric M. and Tao Zha, "Modest policy interventions," *Journal of Monetary Economics*, 2003, *50* (8), 1673–1700.

Lucca, David O and Francesco Trebbi, "Measuring central bank communication: an automated approach with application to FOMC statements," Technical Report, National Bureau of Economic Research 2009.

Manning, Christopher D, "Human language understanding & reasoning," *Daedalus*, 2022, *151* (2), 127–138.

Meade, Ellen E. and Miguel Acosta, "Hanging on Every Word: Semantic Analysis of the FOMC's Postmeeting Statement," FEDS Notes, Board of Governors of the Federal Reserve System (U.S.) 2015.

Nakamura, Emi and Jón Steinsson, "High-frequency identification of monetary non-neutrality: the information effect," *The Quarterly Journal of Economics*, 2018, *133* (3), 1283–1330.

Romer, Christina D. and David H. Romer, "Presidential Address: Does Monetary Policy Matter? The Narrative Approach after 35 Years," *American Economic Review*, June 2023, *113* (6), 1395–1423.

Schmeling, Maik and Christian Wagner, "Does Central Bank Tone Move Asset Prices?," *Available at SSRN 2629978*, 2022.

Schonhardt-Bailey, Cheryl, *Deliberating American Monetary Policy: A Textual Analysis*, MIT Press, 2013.

Shah, Agam, Suvan Paturi, and Sudheer Chava, "Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis," 2023.

Shapiro, Adam and Daniel J Wilson, "The Evolution of the FOMC's Explicit Inflation Target," *Federal Reserve Bank of San Francisco Working Paper 2019-02*, 2019, *2019*, 12.

_ and _ , "Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis," *The Review of Economic Studies*, 12 2021, *89* (5), 2768–2805.

Sherman, Robert P., "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 1993, *61* (1), 1123–137.

Shiller, Robert J, "Narrative economics," *American economic review*, 2017, *107* (4), 967–1004.

_ , "Narrative economics," in "Narrative Economics," Princeton University Press, 2020.

Sundararaman, Dhanasekar, Shijing Si, Vivek Subramanian, Guoyin Wang, Devamanyu Hazarika, and Lawrence Carin, "Methods for Numeracy-Preserving Word Embeddings," in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)" 2020, pp. 4742–4753.

Swanson, Eric T, "Measuring the effects of Federal Reserve forward guidance and asset purchases on financial markets," Technical Report, National Bureau of Economic Research 2017.

Woodford, Michael, "Central Bank Communication and Policy Effectiveness," *Proceedings of the Federal Reserve Bank of Kansas City Economic Symposium at Jackson Hole*, 2005, pp. 399–474.

Supplementary Online Appendix (Not for publication)

A Details about Text Analysis

A.1 Text Analysis using a Large Language Model

We briefly describe the procedure that we use to analyze text data using a large language model.[1] The raw text data is unstructured and can be represented as a string of characters, including letters, numbers, and symbols. We denote the total number of relevant characters as $|C|$. Consequently, each text is a string of characters with a variable length, as illustrated in the following example:

$$\text{Text}_t = [H, o, w, <>, a, r, e, <>, y, o, u, ?] \in |C|^{12}. \tag{A-1}$$

Because this data is unstructured, distinguishing different text data using a distance metric can be challenging. The embedding of text data by a large language model provides a representation of the original string data as a numeric vector, allowing us to define various text data features within a vector space. In the model, this embedding process involves two distinct steps: 1) tokenization, 2) neural network architectures.

A.1.1 Tokenization

Characters do not have linguistic or statistical meaning by themselves. A language model converts the sequence of characters into the sequence of tokens that are more interpretable linguistically or statistically. Language models transform the raw input text into the sequence of tokens by using tokenizers. One of popular tokenizers is PTB tokenizer, which obeys the English grammar and separates the contraction into two units. USE applies the PTB tokenizer to the raw input text. After the tokenization, the input data is transformed from the sequence of characters to the sequence of tokens, in which each token typically corresponds to a distinct

[1] Technical discussion in this section heavily draws on lectures notes on a large language model offered by the computer science department at Stanford University. For more details, see the above link and references therein.

vocabulary.

$$\phi : |C|^l \rightarrow |V|^L \, , \, \text{Text}_{v,t} = \phi(\text{Text}_t), \tag{A-2}$$

$$[H, o, w, <>, a, r, e, <>, y, o, u, ?] \rightarrow [How, <>, are, you, ?]. \tag{A-3}$$

The length of tokens in the sequence (L) might differ from the length of characters in the sequence (l) due to the pairing of characters during tokenization. Each token has its own embedding as a numeric vector. We can denote this token embedding as w. The result of tokenization is a sequence of token embeddings that represent the input text as $[w_1, \ldots, w_L]$. Each token is represented as a 64-dimensional numeric vector.

A.2 Neural Network Architecture

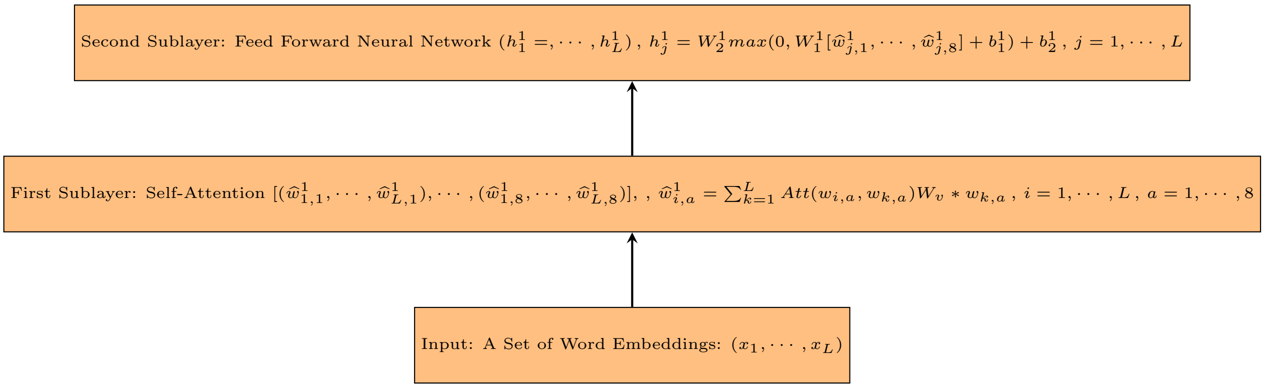
The USE has the transformer architecture consisting of six neutral network layers, each of which has two sublayers with eight self-attention heads. We describe tho original architecture and then explain how to fine tune it to obtain the paragraph level decomposition of similarity scoring across statements.

The first neural network in the USE is built by linking two sublayers as shown in Figure A-1 after taking a group of word embeddings that represents the source sentence as input. For the first layer, the same token embeddings are used for each attention head as input. In other words, $w^1_{i,a} = w_i \, , \forall a = 1, \cdots, 8$. The first layer generates the eight sequence of word embeddings $([\widetilde{w}^1_{1,1}, \cdots, \widetilde{w}^1_{L,1}], \cdots, [\widetilde{w}^1_{1,8}, \cdots, \widetilde{w}^1_{L,8}])$ as output and feeds this as input for the second layer. The actual USE architecture is slightly more complicated than presented below. It involves 1)positional embedding in which the order of any given word is also mapped into the embedding of that word,[2] 2)residual connection in which input bypasses attention and feed-forward neural network channels with a certain probability known as the dropout rate, 3)output from the layer is normalized to have mean zero and standard deviation of one.

The self attention channel can be best understood as looking up the dictionary value of all the words ($W_v * w_j \, , \forall j = 1, \cdots, L$) in the input text to match the query ($W_q * w_i$). The strength of match with the query word for each word in the look-up table is determined by the key ($W_k * w_j$). Here, attention weights for the a-th head are determined by how strong the key is with respect to the query word. In the attention sublayer, every word embedding is linearly transformed to

[2]The position of each word embedding is used to generate since and cosine functions of different frequencies and these values are used the position embedding of the i-th word (P_i) and added to the input embedding w_i.

Figure A-1: First Neural Network Layer



have the key, the query, and the value representation through W_k, W_q, W_v.

$$\widehat{w}_{i,a} = \sum_{k=1}^{L} Att(w_{i,a}, w_{k,a}) W_v * w_{k,a} \ , \ Att(w_{i,a}, w_{k,a}) = \frac{e^{w'_{i,a} W'_k W_q w_{k,a}}}{\sum_{k=1}^{L} e^{w'_{i,a} W'_k W_q w_{k,al}}}. \tag{A-4}$$

Once word embeddings based on eight attention heads are obtained, USE concatenates these eight embeddings of each word into one big embedding and apply the feedforward neural network to this sequence of L embeddings. The final output from the second sublayer is the sequence of L embeddings.[3]

The second layer takes the output of the first layer as input and split the 512-dimensional vector representation (h_i^1) into eight 64-dimensional vector representations $([w_{i,1}^2, \cdots, w_{i,8}^2])$.

The entire USE algorithm works by vertically stacking six neural network layers which take the sentence embedding output in the previous layer as input and generate another sentence embedding as output. Figure A-2 describes the entire process.

To train parameters in the neural network architecture, we need to define the loss function that compares outcomes based on sentence embeddings from the USE with those based on human judgement. For example, if we define the relation between two texts as one of 3 classes (entail,contradict,neutral), we can apply the softmax classifier (f) to the difference between two embeddings. In this case, we can choose parameters in the neural network architecture to minimize the loss function that measures the distance between the machine-classified outcome $(f(U^i, U^j))$ and the one judged by humans $(f^{\text{human}}(\text{Text}_i, \text{Text}_j))$ where U^i is the 512-dimensional

[3]Each embedding representation has 512 dimension because we concatenate eight transformations of 64 dimensional original token embeddings.

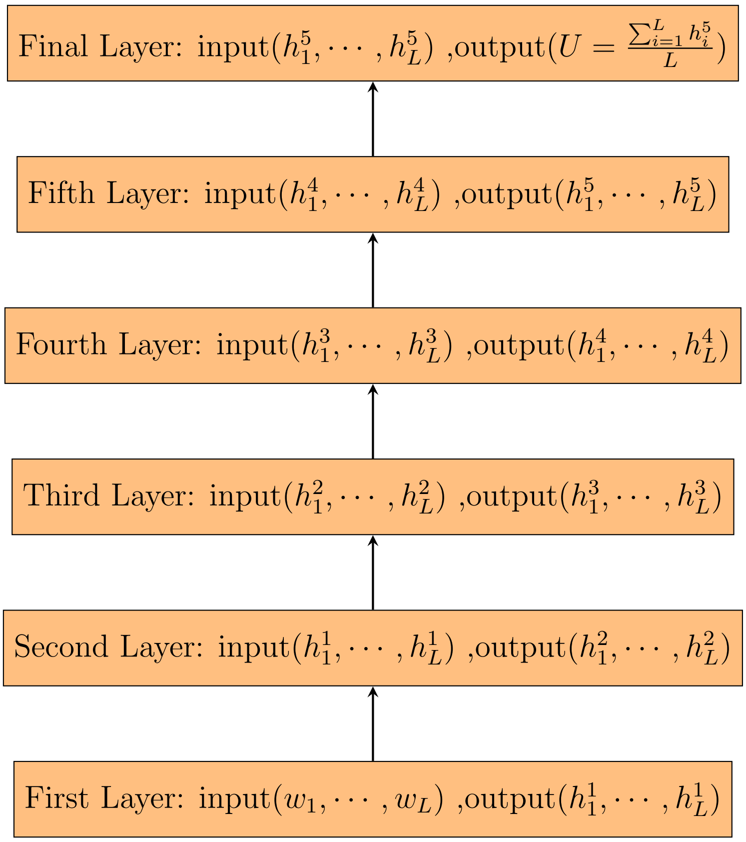


Figure A-2: Neural Network Architecture

USE representation of Text$_i$. In addition, two other natural language processing tasks are run to train the model.

- **Skip-thought task**: conditional on the center sentence, predict neighboring sentences (previous and next). The training dataset is from wikipedia articles.
- **Question-answer prediction**: predict the correct response for a given question among a list of correct answers and other randomly sampled answers. The training dataset is from web question-answer pages and discussion forums.
- **Natural language inference**: given a premise sentence and a hypothesis sentence, extract the relation between them. Let U_p and U_h be the sentence embeddings of the premise and the hypothesis, respectively. A fully-connect layer and and a 3-way softmax classifier are applied for the concatenated input of (U_p, U_h, $|U_p - U_h|$, and $U_p - U_h$). The three-way classifier predicts if the premise entails, contradicts, or is neutral to the

hypothesis. The training dataset is the Stanford Natural Language Inference (SNLI) corpus.

A.3 Paragraph Level Decomposition of the USE Representation

In some cases, paragraph-by-paragraph comparison may provide more interpretable results. For instance, we may be interested in which paragraph drives the similarity score between different statements. For this, we obtain paragraph level USE representations and approximate the statement level USE representation by a weighted average of paragraph level USE representations.

Denote the USE representation of the released FOMC statement at time t by S_t^R. Similarly, S_t^i, $(i = A, B, C, D)$ denotes the USE representation of alternative statements. The USE representation of the j-th paragraph of the FOMC statement at time t is $P_{j,t}^i$. To calculate $P_{j,t}^i$, we run the USE algorithm for each paragraph j. The idea is to construct $\sum_k w_k P_{k,t}^i$ that can mimic S_t^i best in terms of minimizing the squared difference between two representations of the FOMC statement at time t.

- **Step 1: Paragraph Padding** Some statements are longer than others, meaning that the corpus of FOMC statements has an unequal length depending on the statement. An easy way to fix this is to pad a shorter statement with empty paragraph encodings. Suppose that n_{max} is the maximum number of paragraph of any given FOMC statement from the entire corpus of our dataset including both released statements and alternative statements. Then, we can extract the following array of the paragraph USE representation of the FOMC statement.

  $$P_t^R = [P_{1,t}^R, \cdots, P_{n_{max},t}^R]. \tag{A-5}$$

  If the number of paragraphs in the statement at time t $(n_{R,t})$ is smaller than n_{max}, we add $(n_{max} - n_{R,t})$ zero vectors of 512 dimensions. The purpose of this operation is to make the USE representation of any FOMC statement have the same number of the USE representations at the paragraph level.

- **Step 2: Approximate the Statement Level USE Representation by a Weighted Average of Paragraph Level USE Representations**

  The goal is to select weights $(w_j\,, j = (1, \cdots, n_{max})$ that can mimic this statement-level USE representation using paragraph-level USE representations. We consider the following squared loss:

  $$\sum_{i \in R,A,B,C,D} \sum_t (S_t^i - \sum_j w_j P_{j,t}^i)^T (S_t^i - \sum_j w_j P_{j,t}^i). \tag{A-6}$$

We can put the non-negativity and unit-sum constraints on w_j such that $w_j >= 0$, $\sum_j w_j = 1$. Once we find the solution for weights, we can mimic P_t^i by $\sum_j w_j P_{j,t}^i$. But the numerical optimization routine might be non-convex when you put the constraints directly. So we may consider the following transformation of w_j to make the problem an unconstrained minimization problem:

$$w_j = \frac{e^{\alpha_j}}{\sum_{k=1}^{n_{max}} e^{\alpha_k}}, \tag{A-7}$$

where α_j is an unconstrained parameter. Notice that w_j still satisfies the constraints but we are minimizing the loss function with respect to $(\alpha_1, \cdots, \alpha_{n_{max}})$.

- **Step 3: Decomposing the Similarity Scoring**

  For the unit-vector, the cosine similarity is simply the inner product. So we can renormalize the USE representation to have a unit length. In that case, we have the following nice decomposition of the similarity scoring between texts.

$$Sim(P_t^i, P_t^j) \propto Sim(\sum_{k=1}^{n_{max}} w_k P_{k,t}^i, \sum_{k=1}^{n_{max}} w_k P_{k,t}^j) = \sum_k \sum_{k'} w_k w_{k'} Sim(P_{k,t}^i, P_{k',t}^j). \tag{A-8}$$

A.4 Details of Fine-tuning

As explained in the text, we add an additional layer to the USE representation of the text to train the final embedding output to recognize numeric properties better. We consider a fully connected feed-forward network with a rectified linear unit as an activation function. For the original USE representation of a FOMC statement $U_{FOMC} = [U_1, \cdots, U_{512}]$, our additional layer performs the following transformation:

$$f(U_{FOMC}) = [\max(W_1' U_{FOMC} + b_1, 0), \cdots, \max(W_{512}' U_{FOMC} + b_{512}, 0)]. \tag{A-9}$$

Let's stack parameters governing this transformation by $\vartheta = [W_1, \cdots, W_{512}, b]$ where $b = [b_1, \cdots, b_{512}]$. As described in the text, we generate two separate training datasets to optimize ϑ in order to minimize loss functions set out in equation (2) and (3).