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

Taeyoung Doh[†]   Dongho Song[‡]   Shu-Kuei Yang[§]

June 10, 2022

Abstract

We construct a new measure of monetary policy surprise based on a natural language processing algorithm designed to capture contextual nuances in FOMC statements. Specifically, we exploit cross-sectional variations across alternative FOMC statements to identify the statement's tone, and compare current and previous FOMC statements to obtain the novelty. We use high-frequency bond price movements around FOMC announcements to compute the surprise component of the monetary policy announcement. According to our measure, the stock market declines after unexpected policy tightening. Our text-based approach allows us to assess the counterfactual effects of an altered FOMC statement on the stock market.

[†]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 have increasingly relied on public communications to provide guidance about future policy actions, a practice known as forward guidance (see, for example, Woodford (2005) and Blinder et al. (2008)). This practice became more prevalent when monetary policy is constrained by the effective lower bound (see Bernanke (2010)). In this regard, both quantitative decisions (e.g., about interest rates or asset purchases) made by central banks and their qualitative descriptions of the economic factors that lead to these decisions serve as important information variables for understanding monetary policy.

While the profession has moved toward treating policy statements by central bank officials as data to be analyzed, significant challenges remain in parsing of the textual content of these policy statements due to their unique characteristics of policy statements. First, policy statements typically contain a rich mixture of qualitative information (such as descriptions of economic condition) and quantitative information (such as a decision on the key interest rate or amount of asset purchases) communications. Second, qualitative communications often have subtlety and nuances that are an important part of the policy but comparatively difficult to measure. For example, the market's focus will be on any nuance in statements that sheds light on the economic conditions and policy decisions. These characteristics call for an advanced natural language processing algorithm that goes beyond evaluating the textual similarity of documents based on the frequency of overlapping words. Such an algorithm should also recognize numeric properties associated with numbers, such as ordering.

In this paper, we refine information in the Federal Open Market Committee's (FOMC) post-meeting statements using a pre-trained natural language processing algorithm known as the **Universal Sentence Encoder** (USE) introduced by Cer et al. (2018). The USE is able to capture complex dependencies of different words by training an artificial neural network model over large amounts of textual data. We fine-tune the USE algorithm, which already performs well in capturing qualitative communications, with an artificial dataset that mimics FOMC statements involving numbers to assess quantitative information as well.

In addition to methodological contribution, we also bring in alternative FOMC statements, which are available for each FOMC meeting since March 2004, that contain a more dovish (known as Alt A) or a more hawkish (known as Alt C or Alt D) statement than the benchmark

statement (Alt B).[1] Using alternative statements has two advantages. First, these statements increase the corpora, thereby improving the text analysis. Second, and more importantly, these alternative statements have pre-defined tones, allowing us to assess the tone of the post-meeting statement by calculating the semantic distance between the post-meeting statement and alternative statements. For example, if the USE representation of the post-meeting statement is closer to Alt A rather than Alt C or Alt D, then we can classify it as dovish. The different views of the economic outlook and associated policy prescriptions contained in the alternative statements provide important anchoring points for interpreting the tone of the policy statement released after the meeting. By taking the product of the tone and novelty of monetary policy statements, i.e, the distance in terms of the USE representation between the current post-meeting statement and the previous post-meeting statement, we can obtain the monetary policy stance. Note that the identification of the monetary policy stance relies solely on the text analysis of the FOMC statements (the importance of which we emphasize later).

We then tease out the expected versus surprise components of the monetary policy stance. Specifically, we exploit high-frequency bond price movements around FOMC announcements to compute the expected and surprise components of the monetary policy stance. As is common in the literature, we proceed with the assumption that unexpected changes in high-frequency bond returns during FOMC announcements reflect news about the monetary policy stance. By subtracting the expected component from our monetary policy stance, we therefore construct a new measure of the monetary policy surprise. A tightening policy surprise is associated with a decline in bond return (alternatively, an increase in bond yield).

To evaluate the plausibility of our measure, we perform two robustness checks. First, we examine high-frequency stock return data. Our measure shows that a tightening policy surprise generates a negative stock price reaction, consistent with the stock market impact of a conventional monetary policy shock in Bernanke and Kuttner (2005). Specifically, a positive one-standard-deviation surprise leads high-frequency stock prices to drop by about 20 to 40 basis points (bps) on average across various event intervals considered. Second, we verify that our measure of monetary policy shocks is highly correlated (37 to 70%) with those identified in the existing literature, such as Swanson (2017), Nakamura and Steinsson (2018), Bu et al.

---

[1]Like the FOMC transcripts, alternative statements are released with a five year lag. Dovish and hawkish alternative statements generally incorporate information on market expectations regarding the upcoming policy statement because they are written to surprise the market in respective directions. The staff of the Federal Reserve Board prepare alternative statements as a part of discussion materials at each FOMC meeting, See FRB (2004) for a detailed description of statement language.

([2020](#)). Both checks serve as external validation of our measure.

The key advantage of our approach is that it allows us to evaluate how markets would react to counterfactual policy statements. Because our measure of the tone of FOMC statements is derived from texts (independent from financial markets), it has an important advantage over other commonly used approaches (including ones that rely on the inversion of high-frequency bonds across different maturities). Specifically, our measure allows us to assess the impact of a counterfactual policy statement by changing its tone (via replacing sentences described in post-meeting statements) while keeping the market expectations constant. This is possible because our tone measure is not inverted from high-frequency bond data. Thus, our approach enables policymakers to assess alternative scenarios and their impact on the stock market when writing policy statements.

We consider both dovish and hawkish scenarios in the counterfactual analysis by working with FOMC statements released in November 2010 and December 2016, respectively. We first analyze what the market's reaction would have been had the November 2010 FOMC announcement included more explicit time-dependent forward guidance on the future path of the federal funds rate. Specifically, we replace the guidance that the funds rate would remain low "for an extended period," which was included in the post-meeting statement, with "at least through the mid-2012" which was included in the (dovish) Alt A statement. We find that the counterfactual policy statement would have led to stock market returns to increase by about 39 bps instead of the 11 bps decline after the actual post-meeting statement. Next, we examine the counterfactual impact of a more hawkish alternative statement from the December 2016 FOMC meeting. Here, the description about forward guidance in the actual post-meeting statement was that the Committee expected "only gradual increases in the federal funds rate." We replace this guidance with "additional gradual increases in the federal funds rate" from the (hawkish) Alt C statement. We find that the release of a more hawkish statement would have led to stock returns to drop by 110 bps, in sharp contrast to the 6 bps increase after the actual post-meeting statement. Interestingly, the magnitude is comparable to the effect of an unexpected raise of the federal funds rate target by 25 bps (1% decline in the stock market return) found in Bernanke and Kuttner ([2005](#)). These two counterfactual exercises highlight the importance of narrative information in the Fed communication.

Related Literature. Our paper is related to multiple lines of research.

First, our work is closely related to papers that identify monetary policy shocks using high-frequency bond data, e.g., Gürkaynak et al. ([2005](#)), Swanson ([2017](#)), Nakamura and Steinsson

(2018), Bu et al. (2020), Bauer and Swanson (2020), Hoesch et al. (2020), Bauer and Swanson (2022). Since Gürkaynak et al. (2005), most identification approaches impose assumptions on the factor structure of bonds of different maturities to extract central bank communication information. A key feature of our approach is that we can assess which part of the statement is perceived as dovish or hawkish without relying on those maturity assumptions or even without bond data. This approach is appealing given that certain assumptions may not be empirically validated. For example, policy that is intended to move the longer- (shorter-) maturities may have consequences for shorter- (longer-) maturities if asset purchases signal the future path of the short-term interest rate, e.g., Bauer and Rudebusch (2014) or forward guidance affects term premium by reducing interest rate uncertainty, e.g., Bundick et al. (2022).

Second, our work is also related to the increasingly popular literature that applies text analysis to the fields of economics and finance, (e.g., Gentzkow et al. (2019), Ke et al. (2019), Hansen et al. (2017), Schonhardt-Bailey (2013), Shapiro and Wilson (2019), Jegadeesh and Wu (2017), Meade and Acosta (2015), Giavazzi et al. (2020), Lucca and Trebbi (2009), Handlan (2020), Caldara and Iacoviello (2022), Drechsel and Aruoba (2022), Gorodnichenko et al. (2021), Shiller (2017), Shiller (2020)). We adapt the method proposed by Ke et al. (2019) who demonstrate how to predict stock returns by combining new information and compute sentiment scores based on news articles. We extend their method in two important directions. First, we highlight the unique characteristics of policy statements and show how to capture subtlety and nuances in qualitative communications. For this, we introduce a more sophisticated natural language processing technique that produces context-aware representation of the text.[2] Second, we leverage alternative FOMC statements to better identify the tone of the released FOMC statement, which allows us to conduct interesting counterfactual policy experiments.

Two contemporaneous works are closely related to our paper. Handlan (2020) accommodates alternative FOMC statements and uses an advanced natural language processing algorithm (XL-Net) to produce a context-aware, vectorized representation of sentences, which is important for assessing policy communications. However, unlike Handlan (2020), we fine-tune the USE algorithm to accurately capture the basic numeracy of numbers such as comparing the magnitude of the target interest rate in FOMC statements. Drechsel and Aruoba (2022) is another closely related paper. They construct sentiment scores for individual words used in FOMC statements

---

[2]Other than Handlan (2020) and Giavazzi et al. (2020), none of the above-listed papers uses the context-aware representation of the text. Gorodnichenko et al. (2021) add the tone of the voice of the Federal Reserve Chair during press conferences to back out additional information about the overall sentiment of FOMC communications on top of what is expected in the statement.

based on the dictionary in Loughran and McDonald (2011). However, the sentiment classification of individual words may not be able to capture subtlety and nuances in qualitative communications, which we believe are a very important part of policy statements. Neither of the two papers conducts counterfactual policy experiments as we do in this paper.

Outline of the structure of the paper is as follows. Section 2 introduces our natural language processing technique and explains how we fine-tune the algorithm to fit our purpose. Section 3 describes identification scheme of monetary policy surprises using alternative statements. Section 4 discusses empirical results and policy implications. Section 5 concludes.

2 Universal Sentence Encoder for Text Analysis

2.1 Universal sentence encoder

Natural language processing tools convert words or texts into numeric vectors. Such a process is called embedding and how it is done differentiates many natural language processing algorithms. Cer et al. (2018) describe two versions of the USE; 1) deep averaging of word embeddings, 2) transformer-based approach using the self-attention channel. We apply the transformer-based version of the USE to calculate the similarity between texts because the self-attention channel that links each word in the text with all the other words in the text is powerful in capturing the context-dependent meaning of sentences. The USE is able to capture the dependencies between even distant words by training deep neutral networks that can recognize complex dependencies of different words based on large corpora. Hence, it can score the similarity between texts in a more sensible way. For example, imagine that there are two sentences consisting of n_1 and n_2 words respectively:

$$\left\{ S_1 = (w_{1,1}, \cdots, w_{1,n_1}) \quad , \quad S_2 = (w_{2,1}, \cdots, w_{2,n_2}) \right\} \tag{1}$$
$$\downarrow$$
$$\left\{ U_1 = (U_{1,1}, \cdots, U_{1,512}) \quad , \quad U_2 = (U_{2,1}, \cdots, U_{2,512}) \right\}.$$

USE will find out numerical representations of S_1 and S_2 by two 512 dimensional vectors (U_1 and U_2) using a deep neural network architecture. The process of transforming texts into vector representations is called embedding. Natural language processing algorithms differ largely by

how the embedding is done. The embedding representation of the USE is trained to perform a variety of tasks such as text classification similar to humans and predicting some part of the text based on the rest of it. USE is available through **Google Tensor Flow**. We calculate the similarity between the two texts based on the cosine similarity between two embedding vectors:

$$\text{Sim}_{\text{USE}}(\text{Text}_1, \text{Text}_2) = \text{cosine}(U_1, U_2) \quad = \quad \frac{U_1' U_2}{\sqrt{U_1' U_1} \sqrt{U_2' U_2}}. \tag{2}$$

Notice that we are not restricting on pre-fixed features (e.g., the frequency of overlapping words) of the text to calculate sentence embeddings and similarity scores. The training process of embedding representations captures rich features of the text not necessarily confined to the frequency of words. This is the main difference of the USE from methods relying on word-counting that we describe below.

2.2   Word-counting methods

We describe two commonly used word-counting methods for text analysis.

Term Frequency-Inverse Document Frequency. One of the most widely used measures is the Term Frequency-Inverse Document Frequency (TF-IDF) method. Here, similarities between multiple (N) documents are determined by the frequency of words that show up in all of these documents. Specifically,[3]

$$W_{i,j} = \frac{n_{i,j}}{\sqrt{\sum_k n_{k,j}^2}} \ln(\frac{N+1}{df_j + 1} + 1),$$
$$\text{Sim}_{\text{TF-IDF}}(\text{Text}_1, \text{Text}_2) = \text{cosine}(W_{.,1}, W_{.,2}), \tag{3}$$

where $n_{i,j}$ is the count of the j-th word in the i-th document and df_j is the number of documents that contain the j-th word. The term frequency is weighted by the inverse of the frequency of documents that the word appears because common words showing up in many documents can be syntactically important but not semantically. However, the main problem of this method is that it is too coarse to fine-tune the semantic similarity between different words and the algorithm cannot be trained to incorporate the contextual meaning.

---

[3]We follow the default method used in Python.

Latent Semantic Analysis. A more sophisticated word counting method is available, which is known as the Latent Semantic Analysis (LSA). LSA considers the co-frequency of words in calculating the similarity score between texts. LSA uses low-dimensional objects obtained by the singular value decomposition of $W = [W_{.,1}, W_{.,2}]$ to calculate the similarity between texts.[4] Specifically,

$$W = U\Sigma V',$$
$$\mathrm{Sim}_{\mathrm{LSA}}(\mathrm{Text}_1, \mathrm{Text}_2) = \mathrm{cosine}(X_{.,1}, X_{.,2}), \; X = U_{.,1:2}\Sigma_{1:2,1:2}. \tag{4}$$

By rotating term frequency vectors to maximize the co-frequency of words across multiple documents, the LSA extracts representations that highlight the co-frequency of words used in different documents. For this reason, it is widely used in identifying a few key topics from a large number of texts. However, similar to the TF-IDF approach, it does not take into account complex dependencies between different words beyond the co-frequency, which is important for understanding semantic similarity.

2.3 Illustration with simple examples

We illustrate the advantage of the USE in capturing the contextual meaning by comparing the similarity 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 with TF-IDF and LSA for comparison.

It is obvious that S_1 and S_2 are asking the same question, whereas S_3 is not. Hence, the ideal classifier should recognize that S_1 is more similar to S_2 than S_3. However, the similarity score under the TF-IDF or LSA provides an opposite ranking whereas the USE provides a more sensible similarity score. We provide information on the results in Table 1.

The mechanism that makes the USE capture the contextual similarity of different texts is the self-attention channel behind the deep neural network architecture. "How" ($w_{1,1}$) is contextually connected with "old" ($w_{1,2}$) in S_1 while "How" ($w_{3,1}$) is related to "are" in S_3. The contextual

---

[4]The dimension reduction is especially powerful when we try to capture the common theme from large text corpora.

Table 1: Similarity scores

	TF-IDF	LSA	USE
Sim(How old are you, What is your age)	0.00	0.00	0.91
Sim(How old are you, How are you)	0.78	0.78	0.28

Notes: We apply the Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), and Universal Sentence Encoding (USE) approaches to calculate the cosine similarity scores.

representation of "How" should be able to reflect this difference, which is not possible in TF-IDF and LSA because they represent words as an item in the dictionary without encoding contextual linkages in embeddings. For example, if the dictionary contains $|V|$ words, $w_{i,j}$ is represented by a $|V|$-dimensional vector in which the j-th element is one and all the other elements are zeros under these approaches. This representation is known as a one-hot vector encoding. The USE does not restrict word embedding to one-hot vector encoding, allowing multiple non-zero elements.

Arrows in Figure 1 illustrates how attention weights link a particular word in S_1 with all the other words.[5] Since the USE does not use one hot vector encodings and any element in $w_{i,j}$ can be non-zero. The attention-weighted average transforms word embeddings to perform tasks in the training stage such as text classification and word prediction better by making contextually linked words have close embeddings. Notice that unlike one-hot vector encoding, elements in word embeddings are parameters set to minimize the loss function in the training stage.[6]

Figure 1: Illustration of self-attention: S_1
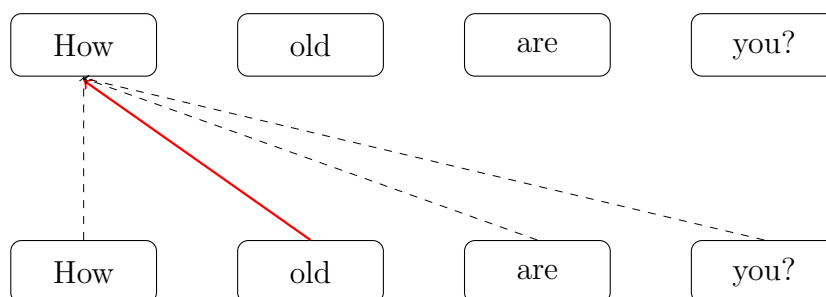
Notes: The red arrow highlights the contextual link between "How" and "old".

---

[5]This way of illustrating the self-attention channel follows Vaswani et al. (2017).

[6]One of the training dataset contains web-based question and answer texts, which facilitate the USE to detect the contextual differences in the meaning of "How" in S_1 and S_3 better.

Furthermore, the USE transforms the given word embedding by the weighted average of embeddings of other words in the text in which weights are determined based on the cosine similarity between two embeddings.[7] This is how the self attention channel works to generate the contextual representation of any word in the text.

2.4 Illustration with policy-relevant examples

Having illustrated the superiority of the USE over word-counting methods in capturing the local context, we now examine how the USE performs on more complicated sentences such as those involving numbers, which are common in FOMC statements. While pre-trained language models like the USE effectively capture semantic relationships between different words or sentences, they are less effective at capturing numeric properties associated with numbers.

Following Sundararaman et al. (2020), we fine-tune the pre-trained language model using texts involving numbers. We set up a specific loss function that preserves the numeric distance between numbers. We also add a separate loss function for datasets not involving numbers because we want to preserve the semantic distance between sentences without numbers as generated by the original USE representation. For the convenience of exposition, let us consider two sentences S_1 and S_2 involving numbers (x_1 and x_2) and other two sentences S_3 and S_4 that do not involve numbers. Recall that the original USE representation of the two sentences S_i and S_j are denoted as U_i and U_j, respectively. The fine-tuning step adds an additional layer of a fully connected feedforward network on top of the USE representation of texts to encode numeric information.[8]

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 $f(\cdot)$ requires two types of loss functions, i.e., one based on S_1 and S_2 and the other with S_3 and S_4.

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

$$\mathcal{L}_{\text{num}}(U_1, U_2) = \left(2\frac{|x_1 - x_2|}{|x_1| + |x_2|} - d\big[f(U_1), f(U_2)\big]\right)^2. \tag{5}$$

---

[7]The appendix describing the neural network architecture behind the USE provides detailed explanation on how these weights are determined.

[8]A fully connected network links each element in the input layer to every element in the output layer while the feedforward network means the direction of the connection is only one way from the input layer to the output layer.

However, training the fine-tuning layer $f(\cdot)$ by minimizing the loss function only with sentences involving numbers may distort the USE representation of the sentence without numbers.

- For this reason, we augment the following loss function

$$\mathcal{L}_{\text{non-num}}(U_3, U_4) = \left(d\big[U_3, U_4\big] - d\big[f(U_3), f(U_4)\big] \right)^2 \tag{6}$$

to make the trained parameters in $f(\cdot)$ preserve the original USE presentation as much as possible.

For the actual training, we create 252 and 688 sentences with and without numbers, respectively based on post-meeting FOMC statements. Since each training data point consists of a pair of sentences, we use 536,848 pairs of sentences as the training set. Out of 536,848 training data points, 63,504 are those involving numbers and 473,344 are those without numbers. Thus, we use $\mathcal{L}_{\text{num}}(U_i, U_{-i})$ for the former and $\mathcal{L}_{\text{non-num}}(U_i, U_{-i})$ for the latter.

We provide a few selected examples below to compare the semantic distance from the original USE representation and with the one from our fine-tuned representation.

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.

Table 2: Similarity scores with policy-relevant examples

		...has been increasing...	...is rising moderately...	...has been subdued
(P1)	Household spending has been increasing at a solid rate, on net, and business investment has been expanding	1	0.8032 (0.8230)	0.6128 (0.6498)
(P2)	Household spending is rising moderately and business fixed investment is advancing		1	0.7060 (0.7001)
(P3)	Household and business spending has been subdued			1

Notes: Similarity scores are provided based on the fine-tuned USE (original USE).

Table 2 shows similarity scores for possible pairs of (P_1, P_2, P_3). By pointing out solid growth (subdued pace) in household spending, P_1 (P_3) takes more optimistic (pessimistic) outlook than P_2 that reflects neutral view as it acknowledges a moderate increase in spending. Hence, the natural ordering of similarity scores would imply that $\text{Sim}(P_1, P_3)$ is lower than either of $\text{Sim}(P_1, P_2)$ or $\text{Sim}(P_2, P_3)$. It turns out that both the original USE representation and the fine-tuned version satisfy this restriction. This illustration highlights that our fine-tuned version retains the original USE representation which already does a good job in capturing the semantic distance for sentences without numbers.

At the same time, the fine-tuned version is trained to capture the numeric properties. Table 3 provides the similarity scores based on the examples N_1, N_2, N_3, N_4. Note that the similarity score monotonically decreases as the difference in the level of the federal funds rate increases. Consequently, for any pair of (N_i, N_{-i}), we find the lowest similarity scores for (N_1, N_4), (N_2, N_3),

Table 3: Similarity scores with policy-relevant examples involving numbers

		Fine-tuned USE	Original USE
(N1) keep at 3.75%	raise by 50 bps to 4.25% (N4)	0.9968	0.8831
	raise by 25 bps to 4.00% (N2)	0.9976	0.8765
	lower by 25 bps to 3.50% (N3)	0.9974	0.8903
(N2) raise by 25 bps to 4.00%	raise by 50 bps to 4.25% (N4)	0.9982	0.9559
	keep at 3.75% (N1)	0.9976	0.8765
	lower by 25 bps to 3.50% (N3)	0.9960	0.9190
(N4) raise by 50 bps to 4.25%	raise by 25 bps to 4.00% (N2)	0.9982	0.9559
	keep at 3.75% (N1)	0.9968	0.8831
	lower by 25 bps to 3.50% (N3)	0.9950	0.9133

Notes: The four texts we are comparing are "FOMC decided to keep the target interest rate at 3.75 percent, FOMC decided to raise the target interest rate by 25 basis points to 4.00 percent, FOMC decided to lower the target interest rate by 25 basis points to 3.50 percent, FOMC decided to raise the target interest rate by 50 basis points to 4.25 percent." We calculate the similarity between the two texts based on the cosine similarity between two embedding vectors.

and (N_3, N_4), respectively. It is important to note that this ranking is not preserved in the original USE representation while it is preserved in the fine-tuned USE representation.[9]

---

[9]The magnitude of the variations across the similarity scores for examples involving numbers is small relative to that based on the original USE algorithm. We emphasize that this concern does not carry over to the exercises involving FOMC statements because of a rich mixture of qualitative and quantitative descriptions therein.

3 Identification of Monetary Policy Stance

3.1 Defining monetary policy stance

The fine-tuned USE representation appears to be successful in capturing the numeric properties of numbers as well as the semantic differences across sentences. This feature of our algorithm allows us to capture subtlety and nuances in FOMC statements, which we believe are a very important part of the policy.

Beyond a technical point of view, we also bring in alternative FOMC statements, which are available for each FOMC meeting since March 2004, that contain a more dovish (Alt A) or a more hawkish (Alt C or Alt D) statement than the benchmark one (Alt B). There are two advantages of using alternative statements. First, it increases corpora size resulting in better performance. Second, and more importantly, these alternative statements have pre-defined tones, and thus, we can assess the tone of the post-meeting statement by calculating the semantic distance between the post-meeting and alternative statements. For example, if the USE representation of the post-meeting statement is closer to Alt A rather than Alt C or Alt D, then we can classify it as dovish. This approach is sensible if the semantic distance between the dovish and hawkish alternative statements is greater than the difference in the semantic distance between the post-meeting statement and respective alternative statements (e.g., the semantic distance between the post-meeting statement and the hawkish alternative statement minus the semantic distance between the post-meeting statement and the dovish alternative statement). For future reference, we refer to this condition as "tone identification condition." We find that the tone identification condition mostly holds in our sample period between March 2004 and December 2016.[10]

We characterize the "monetary policy stance" communicated by each FOMC statement. The "tone" of monetary policy announcements is obtained by computing the similarities between the released statement and the alternative statements. We define the "novelty" of monetary policy announcements by computing the semantic distance between the current statement and the previous statement released after respective FOMC meetings. By taking the product of the

---

[10]The only exception is September 2014 in which the post-meeting statement is nearly identical to Alt A but slightly more distant from Alt C than Alt A. Regarding forward guidance on the future path of the interest rate, Alt C replaces "considerable time after the asset purchase program ends" with "some time after the asset purchase program ends" whereas Alt A provides the threshold-based guidance based on the future inflation rate. Unfortunately, without information on how soon inflation threshold is likely to be reached, our algorithm cannot classify if Alt A is more distant from Alt C than the post-meeting statement. For this episode, we set the most dovish tone score for the post-meeting statement.

tone and novelty of monetary policy announcements, following Ke et al. (2019), we obtain the monetary policy stance.

Specifically, we define the (benchmark) monetary policy stance from the post-meeting statement as

$$\text{MP stance } (t) = \underbrace{(1 - \text{Sim}(FOMC_t, FOMC_{t-1}))}_{\text{Novelty}} \underbrace{\left(\frac{\text{Sim}(FOMC_t, FOMC_{C,t}) - \text{Sim}(FOMC_t, FOMC_{A,t})}{1 - \text{Sim}(FOMC_{A,t}, FOMC_{C,t})} \right)}_{\text{Tone}} \quad (7)$$

and the (alternative) dovish and hawkish monetary policy stance as[11]

$$\begin{aligned} \text{Dovish MP stance } (t) &= -|1 - \text{Sim}(FOMC_t, FOMC_{t-1})|, \quad (8) \\ \text{Hawkish MP stance } (t) &= |1 - \text{Sim}(FOMC_t, FOMC_{t-1})|, \end{aligned}$$

respectively. Novelty in the current benchmark FOMC statement relative to the previous one quantifies the change in the FOMC's intended policy stance. As long as the tone identification condition holds in the data, we observe the following monotonicity in the tone of respective policy stance:

$$-1 = \text{tone(Dovish MP stance)} \leq \text{tone(MP stance)} \leq \text{tone(Hawkish MP stance)} = 1. \quad (9)$$

As conventional, we sign a positive tone as a hawkish stance and a negative tone as a dovish stance. Note that the tone measure is normalized to be between -1 and 1.

The above analysis is based on the USE representation at the statement level. But statements typically consist of multiple paragraphs and we may be interested in isolating the relevance of a particular paragraph. To do this, we compute the USE representation of the j-th paragraph in the i-type FOMC statement where i denotes the different versions of the statement. Let this be $P_{j,t}^i$. By comparing $P_{j,t}^i$ with $P_{j',t'}^{i'}$, we can identify which paragraph contributes most to the change in the similarity score between statements.[12]

---

[11]Here, we assume alternative stances differ from the benchmark stance only in terms of the tone but not novelty. We tried to use variations in alternative statements for both novelty and tone, by defining hawkish monetary policy stance as $1 - Sim(FOMC_{C,t}, FOMC_{t-1})$, for example. But variations in the novelty across alternative statements sometimes offset variations in the tone across them, complicating our tone identification. Since we think cross-sectional variations across alternative statements are most informative about the tone identification, we chose not to use alternative statements for computing the novelty factor.

[12]Note that the statement level USE representation is not an equal average of paragraph-level USE representations. When we approximate the statement level USE representation by a weighted average of paragraph-level USE representations, the first and second paragraphs take most weights. While this is not the exact replication

To identify monetary policy surprises around FOMC announcements, we have to estimate the market expectations for the MP stance right before the FOMC meeting. We do this by using a weighted average of the dovish MP stance and the hawkish MP stance based on the assumption that alternative statements mimic expectations of market participants with more extreme views.[13]

$$E_{t-\Delta}\text{MP stance}(t) = (1 - p_t) \times \text{Hawkish MP stance } (t) + p_t \times \text{Dovish MP stance } (t). \quad (10)$$

Note that the weight, p_t, can vary over time. Next, we discuss how to obtain the estimates of p_t.

3.2 Measuring the market expectations from bond prices

At the time of FOMC announcement, the reaction of the high-frequency asset i's prices can be captured by

$$r^i_{t-\Delta_l, t+\Delta_h} \equiv \ln\left(\frac{P_{i,t+\Delta_h}}{P_{i,t-\Delta_l}}\right) = \alpha_i + \beta_i\text{MPS}(p_t; t - \Delta_l) + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim (0, \sigma_i^2). \quad (11)$$

The surprise component of the FOMC announcement measured at $t - \Delta$ is

$$\text{MPS}(p_t; t - \Delta_l) = \text{MP stance } (t) - E_{t-\Delta_l}\text{MP stance } (t). \quad (12)$$

It is important to understand that $\text{MPS}(p_t; t - \Delta_l) > 0$ corresponds to tightening monetary policy.

The underlying assumption of (10) and hence (12) is that the market is aware of the two alternative stances which serve as bounds when forming expectations. This assumption may sound too strong given that alternative statements are publicly available with a five year lag. One practical justification is from the bluebook in 2004. The staff of the Federal Reserve Board rationalizes alternative statements by intentionally beating market expectations in the hawkish or the dovish direction. Hence, alternative statements reflect the Board staff's best guess for two

of the statement-level USE representation, our findings suggest that the ordering of language matters because it sets the context from which the subsequent words are interpreted.

[13]Lucca and Trebbi (2009) construct a measure of monetary policy stance based on the systematic co-occurrence of words in FOMC statements with pre-labeled words (e.g., hawkish or dovish) in news articles covering FOMC announcements. However, they equate market expectations to ones from the previous meeting, ignoring the market reaction to developments during the inter-meeting period.

extreme market beliefs. Roughly speaking, we are capturing the marginal investor's expectation as a weighted average of these two extreme expectations. As long as the marginal investor's expectation is within the bounds set by the survey of market participants, this assumption is plausible.

3.3 Constructing the surprise component of monetary policy stance

We calibrate the weight $\{p_t\}_{t=1}^T$ that maximizes the rank correlation of the high-frequency bond returns (left-side of (11)) and the surprise component of monetary policy announcements $MPS(p_t; t - \Delta)$, aka, monetary policy shocks. When p_t is not time-varying, $(p_t = \overline{p})$ our estimate is identical to the maximum rank correlation estimator, see Han (1987) and Sherman (1993). Here, we assume that a dovish surprise should lead to a positive bond return because bond prices move inversely with bond yields.[14] Specifically, we maximize the following rank correlation function with respect to p_t:

$$(p_{\tau_i})_{i=1}^T = \arg\max \sum_{t \neq t'} \mathbf{1}(r^b_{\tau_t - \Delta_l, \tau_t - \Delta_h} > r^b_{\tau_{t'} - \Delta_l, \tau_{t'} - \Delta_h}) \mathbf{1}(MPS(p_{\tau_t}) < MPS(p_{\tau_{t'}})). \tag{13}$$

This (negative) rank correlation is maximized by calibrating p_t based on the sorted bond return. Specifically, the time series of bond returns $\{r^b_{t-\Delta_l, t+\Delta_h}\}_{t=1}^T$ are sorted from most negative to most positive. Let the ordering of the sorted-returns be indicated with new time subscripts $\{\tau_1, ..., \tau_T\}$:

$$
\begin{aligned}
r^b_{\tau_1 - \Delta_l, \tau_1 + \Delta_h} &= \min\{r^b_{t-\Delta_l, t+\Delta_h}\}_{t=1}^T \\
r^b_{\tau_T - \Delta_l, \tau_T + \Delta_h} &= \max\{r^b_{t-\Delta_l, t+\Delta_h}\}_{t=1}^T.
\end{aligned}
\tag{14}
$$

For a strictly negative value of β_b in (11), we have that

$$
\begin{aligned}
MPS(p_{\tau_T}) \leq ... \leq \quad MPS(p_{\tau_t}) \quad &\leq ... \leq MPS(p_{\tau_1}) \\
r^b_{\tau_1 - \Delta_l, \tau_1 + \Delta_h} \leq ... \leq \quad r^b_{\tau_t - \Delta_l, \tau_t + \Delta_h} \quad &\leq ... \leq r^b_{\tau_T - \Delta_l, \tau_T + \Delta_h}
\end{aligned}
\tag{15}
$$

where $\tau_t \in \{\tau_1, ...\tau_T\}$. Because it is possible that there are (potentially) multiple realizations of $\{p_{\tau_1}, ..., p_{\tau_T}\}$ that satisfy (15), we pick the one that achieves the largest negative correlation between $\{MPS(p_{\tau_1}), ..., MPS(p_{\tau_T})\}$ and $\{r^b_{\tau_1 - \Delta_l, \tau_1 + \Delta_h}, ..., r^b_{\tau_T - \Delta_l, \tau_T + \Delta_h}\}$. This can be done via

---

[14]The sign of the correlation is negative because p_t corresponds to a dovish probability, which contributes to a negative surprise in monetary policy stance.

grid search (with respect to p_{τ_t}). Once we select $\{p_{\tau_1}, ..., p_{\tau_T}\}$, we can sort them back to match the original time subscript $\{p_1, ..., p_T\}$ and construct the corresponding $MPS(p_t)$ for each p_t, $t \in \{1, ..., T\}$.

3.4 Discussion

A key feature of our approach is that we are able to identify the tone of the post-meeting statement from the pre-labeled (dovish and hawkish) alternative statements. We then use high-frequency bond prices to back out the market expectations. It is important to understand that because the tone of the statement is defined in the space of texts (independent from financial markets), it creates an important advantage over commonly used approaches in the existing literature. Specifically, we can assess the impact of a counterfactual policy statement by changing its tone while keeping the market expectations constant. This is possible because our tone measure is not inverted from high-frequency bond data. Our approach enables policymakers to assess alternative descriptions of the economy and policy prescriptions and their impact on financial market when writing policy statements. In the next section, we conduct two counterfactual policy evaluations to showcase this idea.

In principle, we can also consider obtaining market expectations based on text analysis. For this, we can train a neural network architecture that mimics market expectations of the upcoming post-meeting statement using public communications available before the meeting (previous post-meeting statement and inter-meeting speeches by FOMC participants). Doh (2020) trains a sequential neural network model that predicts the tone of the released FOMC statement based only on public information at that time. While training relies on the pre-labeled alternative statements, once trained, the model can predict the tone of the released statement without relying on alternative statements. However, given the short nature of the availability of alternative statements (from 2004 to 2016), we find that this approach is not empirically appealing.

4 Empirical Results

4.1 Data for alternative FOMC statements

The Federal Reserve Board staff started to prepare alternative FOMC statements from the March 2004 FOMC meeting. The latest available statement is the one prepared for the December 2016

Figure 2: Monetary policy stance
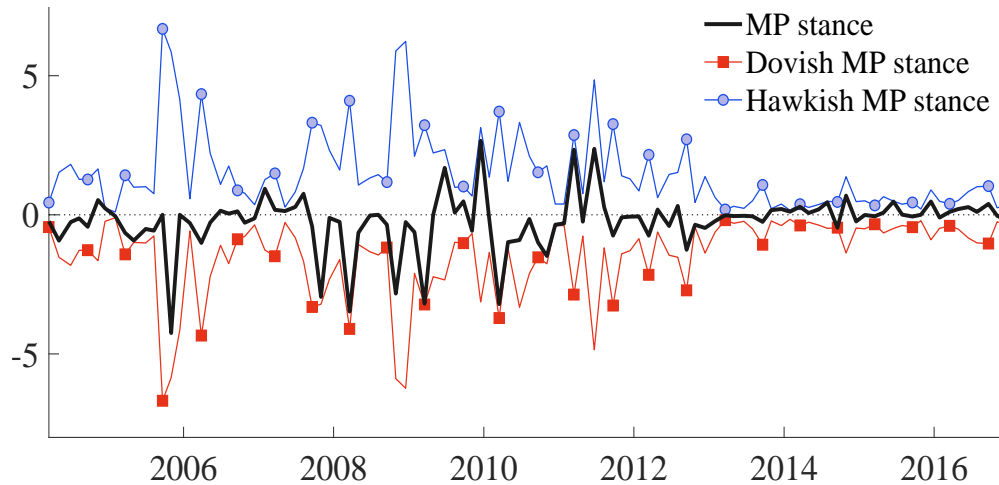
Notes: All measures are normalized to have a unit variance.

FOMC meeting. We have 103 FOMC statements (March 2004 to December 2016) excluding two inter-meeting announcements (Aug 2007, Jan 2008). When multiple versions of hawkish or dovish alternative statements are available (e.g., Alt A1 or Alt D), we use the most extreme one to identify the tone of the released statement.

4.2 Monetary policy stance and surprises

Figure 2 provides the time series of monetary policy stance based on (7) and (8) when the textual similarity is calculated by the fine-tuned USE representation. Our measure captures the change in the policy stance including both the current action (e.g., change in the federal funds rate target) and the expected future action (e.g., forward guidance about the future interest rates).

To extract monetary policy surprises, we construct the market expectation of monetary policy stance which is the weighted average of the dovish and hawkish monetary policy stance. The market-based probability that the dovish and hawkish alternative statement would be released are parameterized by p_t and $1 - p_t$, respectively. To calibrate the market-based probability p_t, we use high-frequency bond market return data around FOMC announcements. One virtue of our maximum rank correlation approach is that it sidesteps the burden of estimating $\alpha_i, \beta_i, \sigma_i^2$ when identifying $p_{1:T}$ in (11). To provide robustness to our claim, we rely on bond futures

returns of various combinations of $\Delta_l, \Delta_h \in \{10, ..., 120\}$min to obtain $p_{1:T}$ and the corresponding MPS($p_{1:T}$). We provide the results in Figure 3. The median values of $\hat{p}_{1:T}$ are highly correlated with each other, e.g., 0.96 or higher. This finding implies that the dovish probabilities extracted from bond returns are fairly robust to different window intervals or instruments. The robustness of our result is different from Bu et al. (2020) who find large differences in monetary policy shock estimates depending on the maturity of the bond data. We interpret this as demonstrating the value of our text-based analysis in isolating core information from FOMC statements.

Having obtained the MPS measure, we quantify its impact on bonds. Specifically, we regress the intraday 5-year Treasury futures returns on our MPS measure and report the results in Panel (A) of Table 4. We find that R^2 is over 50% across many different event windows. Panel (B) of Table 4 contains the estimation results from the regression involving daily one-year Treasury yield changes. A one standard deviation positive shock to our MPS measure is found to increase the one-year Treasury bond yield by 2 basis points.

4.3   Stock market responses to monetary policy surprises

Since the Federal Reserve intervenes in Treasury markets to influence the interest rate, it is not surprising that bond returns react to monetary policy surprises. But the monetary policy transmits to the real economy by affecting broad financial market conditions including stock market as well as Treasury markets. For this, understanding the links between monetary policy and asset prices above and beyond bond returns is important as highlighted by Bernanke and Kuttner (2005) who find that an unanticipated 25 bps cut in the federal funds rate leads to about 1 percent increase in the stock market return. We turn to the stock market reaction to our measure of monetary policy surprise to check if our text-based measure captures similar stock market responses.

For the benchmark case, we select the 5-year Treasury bond futures returns with window intervals $\Delta_l = \Delta_h = 10$min to back out the probability weights and construct MPS($\hat{p}_t$). Conditional on this output, we conduct the regression analysis using stock returns as an external validation check. Specifically, we regress stock returns $r^s_{t-\Delta_l, t+\Delta_h}$ on the bond market-implied MPS($\hat{p}_t$). In essence, we are estimating (11) using an OLS with stock returns. The estimation results summarized in Panel (C) of Table 4 imply that the bond market-implied MPS($\hat{p}_t$) significantly predicts stock returns measured at various window intervals. Because we normalize MPS($\hat{p}_t$) to have a unit variance, we can directly interpret the magnitude of β coefficient in assessing the

Figure 3: Dovish probability comparison

1-year Eurodollar futures

5-year Treasury futures

10-year Treasury futures
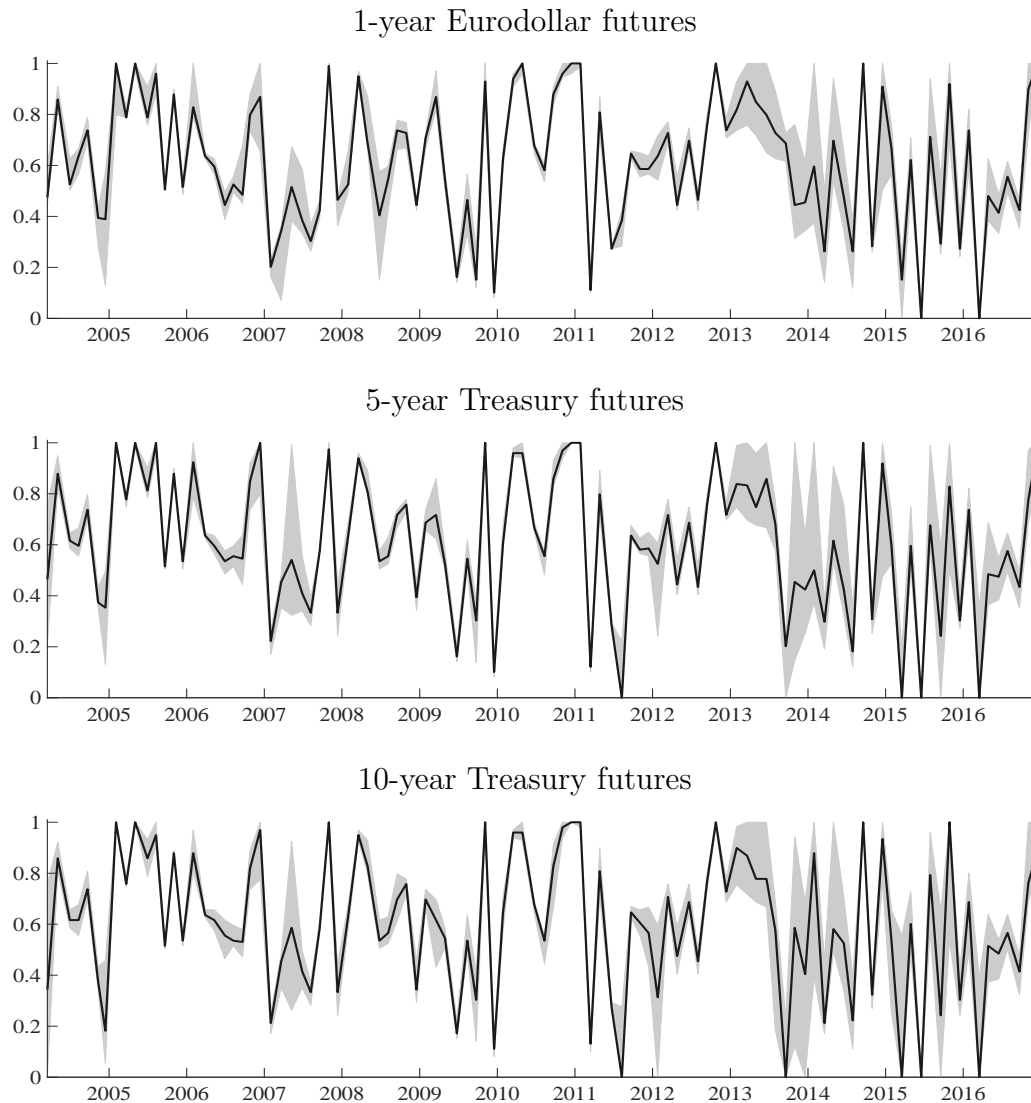
Notes: We rely on the 5-year and 10-year Treasury bond futures returns and the 1-year Eurodollar futures returns. Returns are defined with the following interval $\Delta_l, \Delta_h \in \{10, ..., 120\}$min. The median values are indicated with solid lines.

economic significance of MPS($\hat{p}_t$). On average, we find that a positive one-standard-deviation surprise leads to 20-40 bps drop in stock prices. The R^2 values are around 20-25% across various return window intervals.

Table 4: Return predictability regression

(A) Intraday bond return prediction

$[\Delta_l$	$\Delta_h]$	α	β	t-stat (α)	t-stat (β)	R^2
			Symmetric window			
$[-10$	$10]$	0.06	-0.18	4.23	-10.10	0.60
$[-60$	$60]$	0.10	-0.25	4.39	-9.91	0.55
$[-120$	$120]$	0.12	-0.26	5.15	-11.40	0.54
			Asymmetric window			
$[-1$	$10]$	0.07	-0.18	4.10	-10.25	0.60
$[-1$	$60]$	0.09	-0.24	4.02	-10.26	0.54
$[-1$	$120]$	0.10	-0.24	3.92	-8.73	0.50

(B) Daily bond yield change prediction

		α	β	t-stat (α)	t-stat (β)	R^2
		-0.01	0.02	-1.45	3.82	0.20

(C) Intraday stock return prediction

$[\Delta_l$	$\Delta_h]$	α	β	t-stat (α)	t-stat (β)	R^2
			Symmetric window			
$[-10$	$10]$	0.11	-0.25	2.08	-3.51	0.24
$[-60$	$60]$	0.29	-0.38	3.59	-3.14	0.26
$[-120$	$120]$	0.25	-0.44	2.66	-3.83	0.21
			Asymmetric window			
$[-1$	$10]$	0.07	-0.23	1.37	-2.88	0.21
$[-1$	$60]$	0.27	-0.38	3.30	-3.29	0.25
$[-1$	$120]$	0.21	-0.45	2.05	-3.37	0.20

Notes: In panels (A) and (B), based on the median value of $\hat{p}_{1:T}$, we regress stock and bond returns (defined at various window intervals) on $MPS(\hat{p}_{1:T})$. For this, we rely on the intraday E-mini stock futures and the 5-year Treasury bond futures. For panel (C), we regress daily changes of the (annualized) 1-year Treasury bond yields on $MPS(\hat{p}_{1:T})$. The t statistics are computed based on bootstrap standard errors to account for the fact that our policy surprise measure is a generated regressor.

4.4 Comparison with other monetary policy surprise measures

Our $MPS(\hat{p}_t)$ is highly correlated with other measures of monetary policy shocks based on the high-frequency asset market data around FOMC announcements. We explain a few of them here.

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 consisting of three Treasury bond yields (maturities of 2, 5, 10 years) on top of the five interest rate futures used in Nakamura and Steinsson (2018). He computes the three principal components that account for common variations in these eight different asset prices around FOMC announcements: 1) federal funds rate (FFR) factor that affects the current month federal funds rate futures, 2) forward guidance (FG) factor that is orthogonal to the change in the current month federal funds rate futures, and 3) large-scale asset purchase (LSAP) factor that is also orthogonal to the change in the current month federal funds rate futures and plays a minimum role in explaining the data before the federal funds rate reached the effective lower bound in December 2008.[15] Bu et al. (2020) construct monetary policy shocks using the idea that the variance of the daily bond return is higher on FOMC days relative to non-FOMC days due to the monetary policy announcement. In addition to near-term maturities, they use information from the entire yield curve (up to the maturity of thirty years).

Table 5 provides sample correlation of our MPS measure with those measures. We find that our MPS measure is highly correlated with the monetary policy shock measure of Bu et al. (2020) and the FG and LSAP factors of Swanson (2017).[16] The fact that our MPS measure is highly correlated with the FG and LSAP factors indicates that we are effectively capturing communication channel of monetary policy that impacts longer term interest rates.

---

[15]In practice, these factors are largely distinguished by their different loadings on the maturity spectrum of the underlying interest rate data. The FFR factor has a large non-zero loading on the current month federal funds futures while the other factors have zero loadings. In addition, the loadings of FG factor are concentrated in the one-to-five year maturity spectrum while the LSAP factor has the largest loading on the ten-year Treasury yield.

[16]The negative correlation with the LSAP factor is due to the fact that Swanson (2017) normalized a positive innovation to the LSAP as larger asset purchases than expected, resulting in policy easing.

Table 5: Comparison with other measures

	MP stance: surprise	MP stance: level
Bu et al. (2020)	0.51	0.14
Nakamura and Steinsson (2018)	0.37	0.00
Swanson (2017) (FFR+FG+LSAP)	0.70	0.08
- Federal funds rate (FFR) factor	0.10	0.00
- Forward guidance (FG) factor	0.48	-0.06
- Large-scale asset purchase (LSAP) factor	-0.65	-0.22

Notes: Based on the median value of $\hat{p}_{1:T}$, we construct $MPS(\hat{p}_{1:T})$ and compute correlation with other existing measures of monetary policy factors. The last three factors are from Swanson (2017): 1) federal funds rate (FFR) factor; 2) forward guidance (FG) factor; and 3) large-scale asset purchase (LSAP) factor.

4.5   Discussion on the information channel

Here, we discuss the empirical relevance of the Fed information effect, i.e., an FOMC tightening communicates that the economy is stronger than expected, shown in Nakamura and Steinsson (2018).

In our sample that ranges from 2004 to 2016, we find that a positive one-standard deviation increase of our MPS measure leads to 20-40 bps drop in stock prices as shown in Table 4, which is consistent with implications from standard macro-finance models. Our evidence implies that the empirical support for the information channel may be weak for our sample period. However, we cannot draw the same conclusion for the pre-2004 period that is included in the analysis of Nakamura and Steinsson (2018). We rather suspect that more efforts made by the Federal Reserve staff to fine-tune statement language since 2004 might have increased the effectiveness of monetary policy communications on the asset markets. This hypothesis is consistent with the finding in Lunsford (2020) who shows that the information channel effect was present before August 2003 but disappeared for the later sample as the FOMC provides a more explicit policy inclination in statements. A similar observation was made by Hoesch et al. (2020). Bu et al. (2020) argue that the information channel effect highlighted by Nakamura and Steinsson (2018) is present mainly because Nakamura and Steinsson (2018) consider only the near-term interest rate data in constructing their measure of monetary policy shock. Bu et al. (2020) show that when bonds of longer-term maturities than what is used in Nakamura and Steinsson (2018) are included instead, the information channel effect dissipates.[17] On the other hand, Bauer and

---

[17]The maturity choice does not seem to play an important role in our study as our policy surprise measure is

Swanson (2020) find weak support of the information channel even for the sample including the pre-2004 period and argue for the importance of controlling information from economic indicators released during the inter-meeting period.

4.6 Counterfactual policy evaluation

The beauty of our approach is that we are able to conduct a counterfactual policy evaluation by replacing sentences in the released statement with those from either one of the alternative statements. We consider both dovish and hawkish alternative statements in the counterfactual analysis.

First, we consider the dovish alternative counterfactual experiment. For the November 2010 meeting, we assess the stock market impact of releasing a more dovish statement (Alt A) providing an explicit forward guidance shown in Table 6. The released post-meeting statement does not provide an explicit forward guidance and maintains "for an extended period" as in Alt B (which are close to the released statement).

Conditional on $\hat{\beta} = -0.23$ (the smallest coefficient) in Panel (C) of Table 4, we multiply the counterfactual monetary policy surprise component to assess the impact on the stock returns (defined in the 10-minute interval). For this, we replace the monetary policy stance with the counterfactual one and subtract the previously determined expected monetary policy stance obtained from the high-frequency bond returns. It is important to note that we are only replacing one data point (that corresponds to the November 2010 FOMC statement release date) while keeping all else equal in this exercise. We find that the counterfactual monetary policy stance turns out to be more dovish leading to an 39 bps increase of stock prices relative to the 11 bps decrease with the actual released statement.

Next, we perform another counterfactual exercise for the December 2016 FOMC meeting. As shown in Table 7, forward guidance in Alt C suggests that further increases in the federal funds rate will come sooner than what is described in Alt A or Alt B. Using the same coefficient $\hat{\beta} = -0.23$, we find that the release of Alt C would have decreased stock market return by 110 bps. Since the stock return increased by 5 bps upon the release of the actual post-meeting statement, our exercise suggests that net effect of releasing Alt C would have been a 115 bps

---

robust to various maturities of bond returns, see Figure 3.

Table 6: Alternative language for the November 2010 FOMC announcement

	Alternative A	Alternative B	Alternative C
Federal funds rate target	0 to 0.25%	0 to 0.25%	0 to 0.25%
Forward guidance	0 to 0.25% <u>at least until mid-2012</u>	exceptionally low levels for an extended period	exceptionally low levels for an extended period

Source: Authors' construction based on **FOMC historical materials**.

Table 7: Alternative language for the December 2016 FOMC announcement

	Alternative A	Alternative B	Alternative C
Federal funds rate target	0.25 to 0.5%	0.5 to 0.75%	0.5 to 0.75%
Forward guidance	only gradual increases in the federal funds rate	only gradual increases in the federal funds rate	<u>additional gradual increases</u> in the federal funds rate

Source: Authors' construction based on **FOMC historical materials**.

decline in the stock return.[18] Since we took the most conservative estimate of the stock return predictability regression coefficient in Table 4, our assessment is more likely to serve as a lower bound on the effect of changing language in FOMC statements. Even so, the magnitude of market response is not negligible, consistent with the view that FOMC communication may be an important policy tool. In particular, in the case of a hawkish counterfactual exercise, the stock market response is comparable to what is expected from an unanticipated raise of the federal funds rate target by 25 bps found in Bernanke and Kuttner (2005).

5   Conclusion

The central bank's public communications about current and future policy actions have increasingly received attention as a policy tool. Since March 2004, the FOMC has deliberated on

---

[18]One caveat in our exercise is that when we change the entire statement by the alternative statement not only the tone factor but also the novelty factor might be affected. For the November 2010 episode, the change in the novelty explains only about 1/3 of the change in the monetary policy surprise measure with the rest explained by the change in the tone factor. However, the change in the novelty factor explains about 60 percent of the variation in the monetary policy surprise component for the December 2016 episode.

alternative policy statements prepared by the Federal Reserve staff before each FOMC meeting. Two alternative statements capture the hawkish or dovish deviation from the central tendency of the market expectations right before the meeting, providing cross-sectional variations around the released statement. We apply a novel natural language processing algorithm based on a deep learning architecture to alternative FOMC statements in order to identify the tone of the released statement. This USE algorithm detects the contextual meaning of words in the statement and quantifies the information provided by language in alternative statements. Furthermore, we fine-tune the USE algorithm with artificial text datasets to enhance its ability to detect the numeracy of numbers because FOMC statements often involve numeric values for policy variables.

We construct a new measure of monetary policy surprises by combining the high-frequency bond returns around FOMC announcements with a text analysis of alternative statements by the USE. Our text-based monetary policy surprise measure is robust to the choice of the maturity of bond returns used. We find that an unexpected policy tightening (easing) leads to a decline (an increase) in the stock market return on average. This finding reinforces that the FOMC's communication has affected financial market conditions in an expected direction at least since 2004, which is consistent with the recent empirical findings. Two counterfactual exercises suggest that changing language in FOMC statements may alter financial market conditions in a significant way, highlighting the importance of FOMC communication as a policy tool.

A literature on large-scale language model is rapidly growing and improving human language understanding and reasoning (see Manning (2022)). Our paper suggests that bringing in such a tool into economic analysis can be useful for performing a rigorous analysis of economic narratives. Although we focus on the assessment of FOMC communications via post-meeting statements, our method of fine-tuning a pre-trained large-scale language model with task-specific datasets can be more generally applied in quantifying economic narratives (see Shiller (2017) and Shiller (2020)).

References

Bauer, Michael and Glenn D Rudebusch, "The signaling channel for Federal Reserve bond purchases," *International Journal of Central Banking*, 2014.

Bauer, Michael D and Eric T Swanson, "The Fed's response to economic news explains the" Fed information effect"," Technical Report, National Bureau of Economic Research 2020.

_ **and** _ , "A reassessment of monetary policy surprises and high-frequency identification," Technical Report, National Bureau of Economic Research 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.

Bundick, Brent, Trenton Herriford, and A Lee Smith, "The Term Structure of Monetary Policy Uncertainty," 2022.

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.

Doh, Taeyoung, "Identifying the Tone of FOMC Statements Using a Natural Language Processing Tool," 2020. Manuscript.

Drechsel, Thomas and Boragan Aruoba, "Identifying Monetary Policy Shocks: A Natural Language Approach," 2022.

FRB, "Some General Issues Concerning Statement Language," in "Bluebook, 2004 March FOMC Meeting" 2004.

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.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera, "The voice of monetary policy," Technical Report, National Bureau of Economic Research 2021.

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.

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," 2020. Manuscript.

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

Hoesch, Lukas, Barbara Rossi, and Tatevik Sekhposyan, "Has the information channel of monetary policy disappeared? Revisiting the empirical evidence," *CEPR Discussion Paper No. DP14456*, 2020.

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.

Loughran, Tim and Bill McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of finance*, 2011, *66* (1), 35–65.

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.

Lunsford, Kurt G, "Policy Language and Information Effects in the Early Days of Federal Reserve Forward Guidance," *American Economic Review*, 2020.

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.

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

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.

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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in "Advances in neural information processing systems" 2017, pp. 5998–6008.

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.
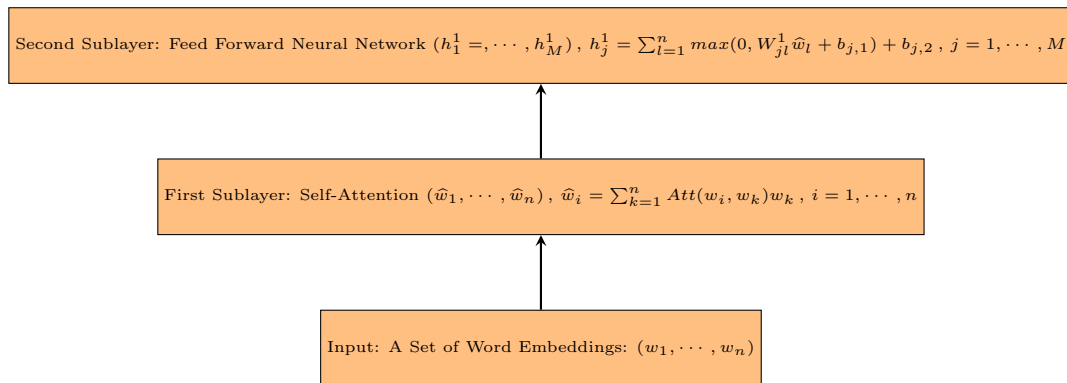
Appendix

A   Technical Appendix

A.1   Details of the Neural Network Architecture in the USE

The USE architecture in this paper is based on six neutral network layers, each of which has two sublayers with a self-attention channel. We first describe tho original architecture and then explain how to fine tune it to obtain the paragraph level decomposition of similarity scoring across statements.

A.2   Deep Neural Network Layers in the USE

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. The first layer generates the sentence embedding vector $(h_1^1, \cdots, h_M^1)$ as output and feeds this as input for the second layer.[19]

Figure A-1: First Neural Network Layer

Second Sublayer: Feed Forward Neural Network $(h_1^1 =, \cdots, h_M^1)$, $h_j^1 = \sum_{l=1}^n max(0, W_{jl}^1 \widehat{w}_l + b_{j,1}) + b_{j,2}$, $j = 1, \cdots, M$

First Sublayer: Self-Attention $(\widehat{w}_1, \cdots, \widehat{w}_n)$, $\widehat{w}_i = \sum_{k=1}^n Att(w_i, w_k) w_k$, $i = 1, \cdots, n$

Input: A Set of Word Embeddings: $(w_1, \cdots, w_n)$

---

[19]The actual USE architecture is slightly more complicated than presented below. It involves 1)sub-word (character) level embedding, 2)positional embedding in which the order of any given word is also mapped into the embedding of that word, 3)residual connection in which input bypasses attention and feed-forward neural network channels with a certain probability known as the dropout rate, 4)output from the layer is normalized to have mean zero and standard deviation of one, and 5)8 multihead attention channels are applied in the attention sublayer.
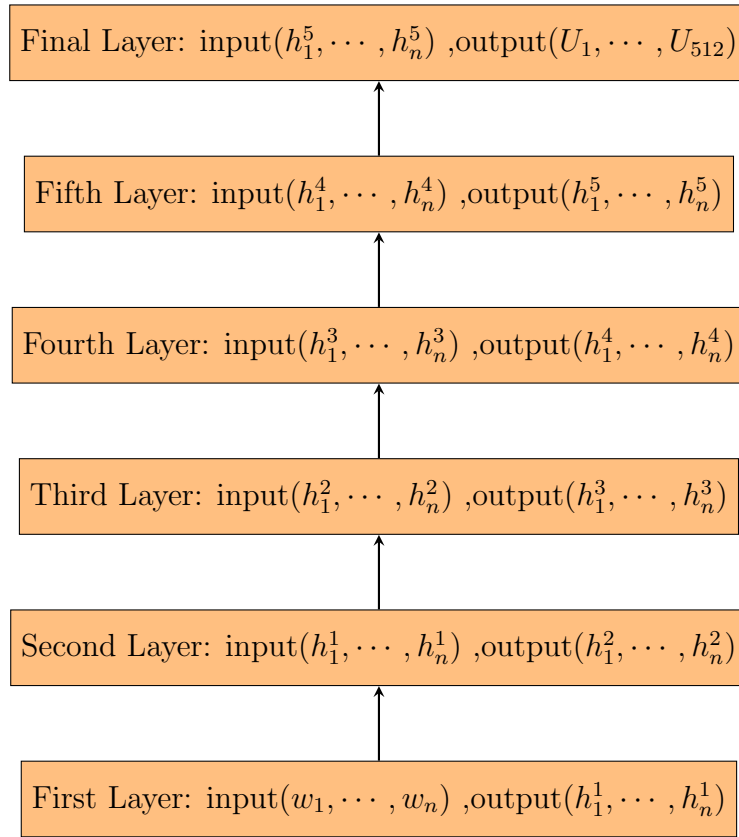
Figure A-2: Neural Network Architecture

Here, attention weights are determined by the distance between different word embeddings as follows:

$$\widehat{w}_{i,j} = \sum_{k=1}^{n_i} Att(w_{i,j}, w_{i,k}) w_{i,k} \,,\; Att(w_{i,j}, w_{i,k}) = \frac{e^{w'_{i,j} w_{i,k}}}{\sum_{l=1}^{n_i} e^{w'_{i,j} w_{i,l}}}. \tag{A-1}$$

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 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 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.

$$P_t^R = [P_{1,t}^R, \cdots, P_{n_{max},t}^R]. \tag{A-2}$$

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:

  $$\sum_{i \in R,A,B,C,D} \sum_t (S_t^i - \sum_j w_j P_{j,t}^i)^T (S_t^i - \sum_j w_j P_{j,t}^i). \tag{A-3}$$

  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:

  $$w_j = \frac{e^{\alpha_j}}{\sum_{k=1}^{n_{max}} e^{\alpha_k}}, \tag{A-4}$$

  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.

$$Sim(P_t^i, P_t^j) \propto Sim(\sum_{k=1}^{n_{max}} w_k P_{k,t}^i, \sum_{k=1}^{n_{max}} w_k P_{k,t}^j) = \sum_k \sum_{k'} w_k w_{k'} Sim(P_{k,t}^i, P_{k',t}^j). \tag{A-5}$$

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:

$$f(U_{FOMC}) = [\max(W_1' U_{FOMC} + b_1, 0), \cdots, \max(W_{512}' U_{FOMC} + b_{512}, 0)]. \tag{A-6}$$

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 (5) and (6).