Word2Prices: Embedding central bank communications for inflation prediction*

Douglas Araujo (BIS) Nikola Bokan (ECB)
Fabio Alberto Comazzi (ESM) Michele Lenza (ECB and CEPR)

September 6, 2024

Abstract

The embeddings of the European Central Bank's introductory statements at monetary policy press conferences significantly improve out-of-sample core inflation forecasts multiple quarters ahead, even when the embeddings are learned from relatively simple Word2Vec models. Other common textual analysis techniques, such as dictionary-based metrics or sentiment metrics do not obtain the same results. These results hold for periods of significant inflation volatility such as after the Covid pandemic. We show that both principal component analysis and spike-and-slab regressions effectively use the information in the embeddings, suggesting rich signal content. The embeddings Granger-cause inflation in out-of-sample exercises even when controlling for central bank inflation forecasts. Embeddings learned from more sophisticated, pre-trained large language models improve the forecasting performance, although in these cases the exercise can only be considered to be purely out-of-sample after each models' training data cutoff date. This work contributes to the literature by documenting a novel and simple approach to leverage central bank texts for policy-relevant forecasting.

JEL classification: E31, E37, E58
Keywords: Inflation, forecasting, central bank texts, embeddings

*The authors thank (Sujit Kapadia, Michael, Marek, Hanno, Joan) and several seminar participants at the European Central Bank [and Bank for International Settlements (BIS)] for comments and Johannes Damp for estimating the first proof-of-concept at request. The opinions in this paper are those of the authors and do not necessarily reflect the views of the European Central Bank, the Eurosystem, the CEPR, the BIS or the ESM. Contact author: Michele Lenza (michele.lenza@ecb.europa.eu).

1 Introduction

Recent years have seen growing interest in understanding economically-relevant signals from textual data. The wider availability of text datasets combined with an advanced analytical toolbox (Gentzkow et al. (2019), Dell (2024)) offers opportunities to understand the macroeconomic effects of texts that are rich in economic content, such as those produced by central banks. Well-recognised as an integral part of the policy toolkit (Blinder et al. (2008)),[1] these texts convey to different degrees the central bank's assessments, intentions, forecasts or scenarios, and policy directions. Such content is aimed at influencing and driving both public and financial market expectations (Jansen and De Haan (2007), Hansen et al. (2019), Ahrens and McMahon (2021), Ahrens et al. (2023)).

Building on these results, the central question we attempt to address in this paper is whether central bank texts can improve inflation forecasts. More precisely, can we extract systematically useful information about future inflation from Central Bank texts: do Central Bank texts Granger-cause inflation? To explore this question, we conduct a real-time extraction of simple, continuous vector space numerical representation (embeddings) of the relevant "concepts" in the transcripts of the ECB press conference introductory statements. We then compare the inflation forecasting mean squared error (MSE) of these embeddings to (vector-)autoregressive ((V)AR) benchmark at up to four quarters in the future. Important to point out, we conduct the analyses within a full-fledged out-of-sample framework to ensure the robustness of our results[2].

Our key finding is that central bank text embeddings improve inflation forecasting. In other words, the information extracted from Central Bank texts helps to predict inflation. This result is invariant to whether we summarize information contained in the embeddings via principal components or we estimate Bayesian regression with hierarchical prior distribution by exploiting the full recursive space of available embeddings. These pure out-of-sample results are valid for the time period between 2008 to 2023, which encompasses the Global Financial Crisis (GFC), the Euro sovereign crisis, and the period around and after the Covid-19 pandemic.This last period was remarkable due to the historically high inflation variability and forecast errors (Chahad et al. (2022)). Relative to other non-embedding forms of extracting information from texts, such as placebo word counts or sentiments on inflation, our measure based on word embeddings additionally improves inflation forecasts. We further demonstrate that the embeddings series continue to improve inflation forecasts even when the Eurosystem's inflation forecasts are also considered. Our findings suggest that relatively simple and sufficiently flexible techniques (as

[1] For example, the European Central Bank recognised as such in its strategy review of monetary policy communication (available at https://www.ecb.europa.eu/home/search/review/html/monpol-communication.en.html).

[2] The embeddings model for any given press conference statement only trains on information up until the date of corresponding analysed text.

our NLP based embeddings are) of text information extraction indeed contain novel signals that consistently contribute to predicting inflation.

Recently, a fast-growing literature started exploring the nuances of central bank communication and its consequences. As already mentioned, several above referenced papers highlight how central bank communications can influence market outcomes and inflation expectations, underscoring their importance as a tool for managing economic expectations. Focusing on the Euro Area, Rosa and Verga (2007), Tobback et al. (2017), Cour-Thimann and Jung (2021), Ehrmann and Talmi (2020), Ehrmann and Wabitsch (2022), Fraccaroli et al. (2022), Gáti and Handlan (2022) and Pavelkova (2022)) investigate the informational content of ECB communications. . Hansen et al. (2019) and Mumtaz et al. (2023) study the effects of the Bank of England's Inflation Report on asset prices, with a focus on long-run interest rates.

Another voluminous body of work deals with inflation forecasting with textual data and the information content of central bank texts. [3] Barbaglia et al. (2023) and Kalamara et al. (2022) forecast inflation and other macroeconomic variables such as GDP and industrial production using sentiment extracted from news, while de Bandt et al. (2023) report that news are correlated with inflation and inflation expectations. Carriero et al. (2024) report that time series foundational models, which are based on the transformer architecture (Vaswani et al. (2017)) uderpinning large language models, do not perform generally better than econometric modes at a range of macroeconomic series that includes inflation.

In another stream of work, the signal contained in central bank textual communication has been studied, among others, by Cieslak and Schrimpf (2019), Hansen et al. (2017), Jarociński and Karadi (2020), Ahrens and McMahon (2021), Ahrens et al. (2023). Baranowski et al. (2021) study similar texts as in our setting, the ECB introductory statements, and find that an orthogonal "tone shock" is correlated with future policy decisions.

In relation to the existing literature on the use of text as a data and the inflation forecasting literature, our paper contributes by proposing a relatively simple, low-cost novel method for processing in real-time central bank written communications and evaluating its out-of-sample informational content in a policy-relevant context. It sets the stage for a more focused investigation into whether central bank texts, focusing on ECB, can systematically provide useful information for predicting future inflation. While central banks have more generally found a number of uses (such as gauges of inflation expectation) for third-party texts (i.e. from social media) to inform monetary policy (Araujo et al. (2023)), this work might drive a reflection in policy circles about the effects of own texts on the inflation path. This paper also contributes to the broader group

[3] See Faust and Wright (2013) for a survey of the inflation forecasting literature

of manuscripts that use text as data in econometric analyses (Gentzkow et al. (2019); Hirschbühl et al. (2021); Ash and Hansen (2023)).

The subsequent sections of the paper detail the methodology employed, present the empirical results, and discuss the implications and limitations of our findings. More specifically, section 2 describes the features of our out-of-sample forecasting exercise, including the specific proposal to transform text in data. Section 3 presents the original dataset we use, and Section 4 discusses the transformation of the textual data into dense vectors through "embedding" techniques. Section 5 reports and discusses the empirical results and Section 6 concludes, also highlighting some remaining challenges.

2 Forecasting exercise

Our aim is to establish whether a certain text-based variable (m_t) has explanatory power for euro area inflation (π_t); the construction of m_t is detailed later. If m_t Granger-causes π_t, we conclude that it has explanatory power. In other words, our measure of explanatory power is based on predictive content. Moreover, we require the evidence of Granger-causality to be strong enough to survive also out-of-sample over a meaningfully long period of time.

The benchmark against which the predictive power of m_t will be tested is an autoregressive (AR) model of quarter-on-quarter inflation (π_t). Such simpler, naive benchmark forecasts are difficult to beat in practice in out-of-sample forecast, as shown by Atkeson and Ohanian (2001). The AR model is:

$$\pi_t = a + b(L)\pi_{t-1} + \varepsilon_t, \tag{1}$$

with $a \in \mathbb{R}$ a constant, $b(L)$ a row vector of coefficients with the lag operator (ie, the coefficient is a scalar if the lag order is one), and ε_t a zero-mean, i.i.d residual.

If m_t has predictive power, then the following vector autorregressive (VAR) model for quarter-on-quarter inflation (π_t) and text-based measure (m_t) or alternative predictors would perform better than the AR in Eq. 1:

$$y_t = \begin{bmatrix} \pi_t \\ m_t \end{bmatrix} = A + B(L)y_{t-1} + \eta_t, \tag{2}$$

where $A \in \mathbb{R}^2$ is a vector of constants, $B(L)$ are the coefficients and the lag operators, and η_t is a zero-mean, i.i.d residual.

We compare the out-of-sample inflation forecasting performance of the two models at different horizons. The measure of forecasting accuracy is the ratio of the out-of-sample mean squared error (MSE) of the VAR to that of of the AR model for each horizon, as in Eq. 3, where h is the period ahead. Values smaller than 1 indicate that the VAR model is more accurate than the AR. The MSE for a specific forecasting horizon h is computed by looking at the cumulative inflation rate over the forecast horizon. In other words, by assuming that p_t is the logarithm of consumer prices, we compute the forecast error for the MSE at time t by using $p_{t+h} - p_t$ as our target variable:[4]

$$\text{MSE}_{\text{VAR},h}/\text{MSE}_{\text{AR},h} = \sum_{t=2008Q1}^{2023Q4} \frac{(\hat{p}_{t+h}|(B(L)[\pi_{t-1}, m_{t-1}]) - p_{t+h})^2}{(\hat{p}_{t+h}|(b(L)\pi_{t-1}) - p_{t+h})^2}. \tag{3}$$

Note that the MSE metrics for each forecasting horizon are calculated with out-of-sample forecasts from 2008 Q1 onwards in a "rolling out-of-sample" way. We start with AR and VAR models estimated exclusively with data from 2002 Q1 to 2007 Q4 as a first estimation sample. Using these first estimates, we compute our forecasts for the horizons from 1 to 4 quarters ahead. Then, we increase the estimate sample by one quarter and we repeat the whole exercise, ie estimate the AR and VAR with data from 2002 Q1 to 2008 Q1 and forecast the 1 to 4 quarters ahead. This is done until the last estimation sample in 2023 Q2. This results in 63 forecasts for each of the four horizons. For the 1 and 2 quarters ahead, we can evaluate all of them. For the horizons of three and four quarters ahead, one and two forecasts, respectively, fall outside of the available sample for inflation and we cannot evaluate them. The parameters of both models are estimated with Bayesian techniques, using Minnesota priors calibrated for stationary data in a hierarchical model as in Giannone et al. (2015). The models are estimated with four lags.

The results below refer to the point forecasts, obtained for both models by setting the VAR parameters to the posterior mode.

[4] In other words, $p_{t+h} = p_t \prod_{s=1}^{h}(1 + \pi_{t+s})$.

3 Data

Our dataset comprises mainly the inflation data and the ECB policy texts, namely the introductory statements at the beginning of each policy press conference. Both datasets start at the first quarter of 2022; the texts end at the second quarter of 2023 while our inflation series ends on the fourth quarter of that year.

The inflation series π_t is the quarter-on-quarter growth rate of the core, harmonised index of consumer prices (HICPex), p_t. This series is presented in Fig. X.

The introductory statements are prepared texts read out by the President of the ECB at the beginning of each policy press conference.[5] Their content follows a regular structure where important conjectural topics are brought up first, followed by commentary and the ECB's outlook on economic activity, inflation, risks to economic activity and financial and monetary conditions. This exact content was studied also by Baranowski et al. (2021) and Anderes et al. (2021), amongst others.

Central bank communication modalities evolve over time, but they usually range from scheduled press conferences to explain policy decisions to periodic reports and ad hoc official speeches.[6]

4 Embedding central bank communication

A crucial step to extract meaningful information from the introductory statements is to "embed" these texts in a vector space. The result, commonly referred to as an embedding, is a dense vector that encodes the information content of each particular text in a way that makes it quantifiably different from the other texts. Consider an embedding function M that maps words[7] of a text A into a dense ϕ-dimensional real vector space, $M : |w_A| \rightarrow \mathbb{R}^{\phi}$. Each of these vectors is an embedding of the word in question. The vector dimension ϕ is a choice of the analyst; higher dimensional representations allow for more nuance in differentiating words but intuitively require bigger and more diverse textual datasets and more sophisticated architectures to learn these concepts satisfactorily.

[5] Available at https://www.ecb.europa.eu/press/press_conference/monetary-policy-statement/html/index.en.html.

[6] Increasingly, central banks engage with social media as well, eg Masciandaro et al. (2024).

[7] Or any other subdivision of texts such as parts of words (commonly referred to as tokens), or sentences, paragraphs, etc.

These embeddings can be interpreted as a latent representation of the texts in a given vector space. The mapping is learned through optimisation to minimise a loss function with respect to the collocation of words in massive textual corpora. The practical incarnation of M is almost always through a neural network, although the way they are set up (their so-called "architectures") can be considerably varied. Over time, a few architectures - all based on neural networks - have proven to efficiently encode different latent meanings in a way that makes it useful across applications and domains.

An early such architecture to gain prominence is Word2Vec (Mikolov et al. (2013)), while other models appeared around that time such as GloVe (Pennington et al. (2014)) and later, ELMo (Peters (2018)). The transformer model (Vaswani et al. (2017)), unveiled in 2017 and then successfully deployed in a flexible end-to-end large language model in 2018, the BERT model (Devlin et al. (2018)), spawned a newer generation of embedding models that take the broader context of each text into account. More recently, the GPT-class of models further pushed the frontier of embeddings. Nowadays, various other alternatives push frontiers of the state of the art. The variety of modern embedding models is best illustrated in open leaderboards such as the Massive Text Embedding Benchmark (MTEB; Muennighoff et al. (2023)).[8]

Other techniques allow numerical analyses of texts with varying levels of sophistication, but these often present considerable disadvantages compared to embeddings. Dictionary-based methods are only sensitive to the words defined by the analyst, and may fail to capture semantically close words not in the exact list (false-negatives) or wrongfully count mentions to words in the wrong context (false-positives). A related technique that consists of representing each word by a dummy vector, so-called "one-hot encoding", has the added disadvantages that regressions become unnecessarily high-dimensional and that each concept is implicitly assumed to be orthogonal to each another. Topic modelling, as done for example with latent Dirichlet allocation (Blei et al. (2003)) usually assign sets of words to single topics, and does not necessarily offer continuous metrics that reflect the meanings of each word or topic.

In contrast, embeddings offer a flexible representation of texts that remains amenable for quantitative analyses, at the expense of the lack of direct interpretability. Measured vector proximity should in principle capture semantic and syntactic similarity between the words. Words representing similar concepts will be closer neighbours in the embedding space, and directions roughly reflect human ideas of overall word relatedness, including "surprising" relatedness along various salient semantic dimensions (Russell and Norvig (2022)). But each of the $1, \ldots, \phi$ individual items in the m_t vector usually does not have a direct interpretation: it would take significant analysis to understand what exactly the different values of nth item represent along the real line.[9] Therefore

[8] Available at https://huggingface.co/spaces/mteb/leaderboard.

[9] An added interpretability challenge is that the location in the vector is not guaranteed to be exactly the same

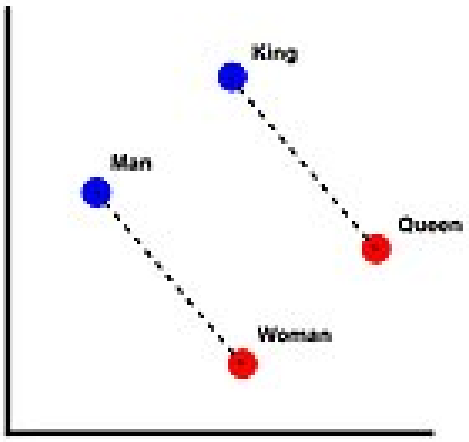


Figure 1: Source: https://en.wikipedia.org/wiki/Word2vec

interpreting embeddings require more sophisticated techniques.[10]

A typical illustration of the algebraic manipulation of text concepts enabled by embeddings is to reconstruct the concept of "queen" by adding the values for "woman" and "king", minus the embedding for "man" (see Fig. 1). But where do embeddings come from? The function M is learned from reams of textual data that relate words to one another in semantically and logically-coherent texts. In practice, the functional form of M is highly non-linear and may depend only on the word itself, or in more sophisticated applications also on the larger context in which the word appears. For example, for a sufficiently diverse text dataset, a model that learns from scratch to map out each word to a vector will end up with the embedding for "apple" consisting of a weighted average of the meaning for a fruit and for the IT company. But embeddings from models that also include contextual information would be able to differentiate these two cases depending on the text they appear in. Specialised references as for example Ash and Hansen (2023) discuss the intuition in more detail.

for two embedding models learned with the same neural network architecture trained on the same data, if the random initialisations are different.

[10] One example is semantic projection (Grand et al. (2022)), a technique that projects related embeddings onto the line connecting extremes of a concept. For example, if the embeddings of different assets are projected onto the line connecting the embeddings for "safe" and "risky", the vector for "T-bills" would be closer to the former and "equities" to the latter.

The different ways mentioned above to learn embeddings have practical implications when studying the informational content of texts. More sophisticated models might result in more precise quantification of the meaning of a text by taking advantage of broader context windows and leveraging higher-level representations of the concepts; in fact, many such models *need* to be used from a pre-trained or foundation version because it is impractical to train them from scratch. But while using such pre-trained models carry a number of practical advantages (Araujo (2023a)), one important disadvantage for evaluation of forecasting power is that the embeddings are truly out-of-sample only for dates after their pre-training cutoff date. Before that date, all embeddings incorporate *some* information about the future (up to the cut-off) because they are jointly estimated. This makes running complete out-of-sample forecasting exercises with such pre-trained models difficult. Conversely, researchers can set up simpler applications such as Word2Vec models to be trained from scratch with a more explicit control over the sample date cut-offs. As discussed later, while these simpler models are not state of the art, their performance is good enough and researchers maintain a crucial control over the information set used to train them.[11]

4.1 Baseline embedding method

In our application, we follow the Word2Vec (Bengio et al. (2000) and Mikolov et al. (2013)) method, using $\phi = 100$. The model is trained from scratch using only the vocabulary μ_T in the documents contained in each sample period. That is, at the cutoff date $T = 2007Q4$, the vocabulary comprises all the unique words appearing in the texts up to that date: $mu_{2007Q4} = \{w : w \in \cup_{t=2002Q1}^{2007Q4} M_t\}$. Then, for the next period, the cutoff date T moves forward by one quarter and the exercise is repeated. We do not use text pre-cleaning given the well-structured nature of our dataset. In our implementation, the context window is from -5 to +5 positions and $\phi_{\text{Word2Vec}} = 100$.[12]

Word2Vec parameterizes the probability of a certain *word* w being v given its *context* C(w) as:

$$P(w = v|C(w)) = f(\alpha * \rho_w, ...),$$

with $\rho_v =$ embedding vector for word v; $\alpha =$ context vector (average of embeddings of neighbouring words). Note that α and ρ_v estimated by maximizing the predictive accuracy of the model in the

[11] Other papers using text embeddings include Kogan et al. (2021), who measure of workers' exposure to technological innovation from the embeddings of job descriptions, and Hansen et al. (2021)'s summaries of the skills required of executives in job descriptions.

[12] For brevity, subsequent mentions of the vector dimensionality ϕ will not indicate in the subscript the technique used unless it is unclear from the context.

whole corpus of text fed to the model. Eg, if $P(w = v|C(w))$ is high, ρ_v tends to have "large" entries where context vector has large entries (similar to the context).

As a result, words that appear in similar corpus contexts tend to have similar embeddings, as seen in Fig. 2. This graph ploits the concepts in the press conference data are related to one another, transformed by a dimensionality-reduction technique called t-SNE (Van der Maaten and Hinton (2008)) to represent in two dimensions the ϕ-dimensional embeddings for each word. In Fig. 2, we can see that words related to monetary policy ("refinancing", "purchase", "TLRO") are close to one another, as are real-economy factors that commonly affect inflation ("wages", "bottlenecks", "production", with "supply-side" also somewhat close). Another group of words include "pandemic" and "coronavirus", while the words "inflation", "2" and "remain" are also close to one another as a group.

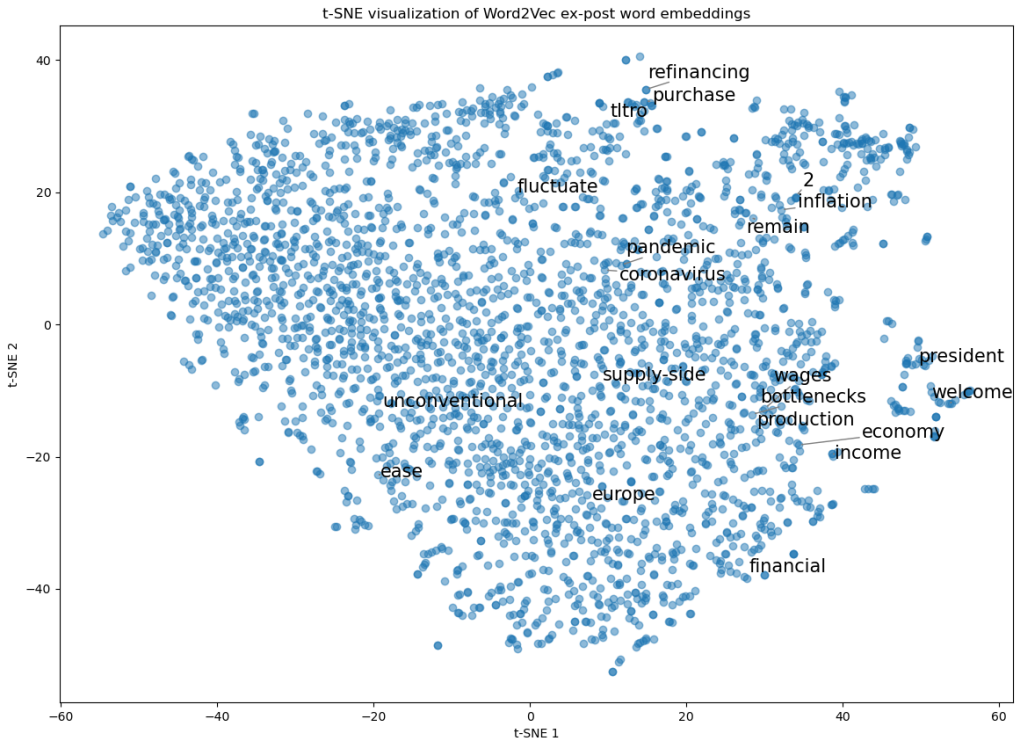


Figure 2: t-SNE representation of the embedding concepts

Note that our embedding step does not incorporate any information about inflation beyond what is organically contained in the texts themselves. Instead, the embeddings are learned from the straightforward exercise of predicting the word given neighbouring words. Put differently, the

embedding model is not learning to optimize latent representations that explicitly target inflation forecasting.

One important aspect when evaluating Granger-causality of text signal is to careful consider the time dimension of the embedding function M to compare true out-of-sample forecasting, because the meaning of words and expressions change over time. Consider M_t as an embedding function whose parameters are set using only information up to time t. M_T instead includes information of the complete time series, even beyond t, to learn how to embed the words. For example, Let $t = 2012, T = 2015$ and A_{2012} = "Whatever it takes". Then it might be that $M_{2012}(A_{2012}) \neq M_T(A_{2012})$. Hence, due attention must be paid to the use of LLMs to embed texts when evaluating forecasting power. Particular care is warranted to avoid assigning meanings to texts of which we might not be aware, in real-time. Of course, a downside of this is that out-of-sample embedding might not benefit from higher-quality representation of language nuances that tends to be found in newer models.

The process to go from the word embeddings to the quarterly series $\{m_t\}$ involves averaging the vectors of words in each speech, then those in each quarter. In our baseline regressions, each quarter t's press conferences at day $d \in t$ A_d are embedded separately and their embeddings are averaged, resulting in $m_t = 1/|A_d| \sum_{d \in t} M(A_d)$ (note that for notational simplicity, $M(A_d)$ is the average of the word embeddings of a text A). All the elements in m_t are encoded through Word2Vec-type of models to jointly represent some meaningful information from the ECB texts, as learned by a shallow, two-layer neural networks trained to reconstruct the linguistic contexts of words in our central bank corpus. While not directly interpretable, Fig. 3 shows the time series of each of the 100 elements in the embedding vector to illustrate how each component changes over time.

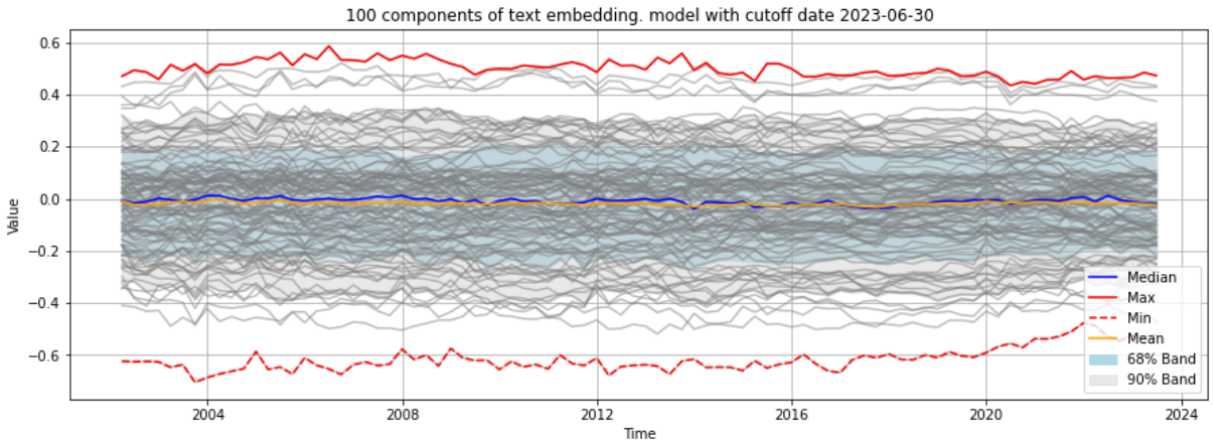


Figure 3: Time series of the embedding components (cutoff on 30 June 2023).

One can also use pre-trained word embeddings. In particular, embeddings can be seen as a by-product of the embedding layer of most common language models. We test these with OpenAI and BERT, for which we use the pre-trained embeddings - mostly because it is too costly or impossible to train the model ourselves.

4.2 Principal components of embedding vectors

Figure 4: In-sample estimate of PC - Word2Vec

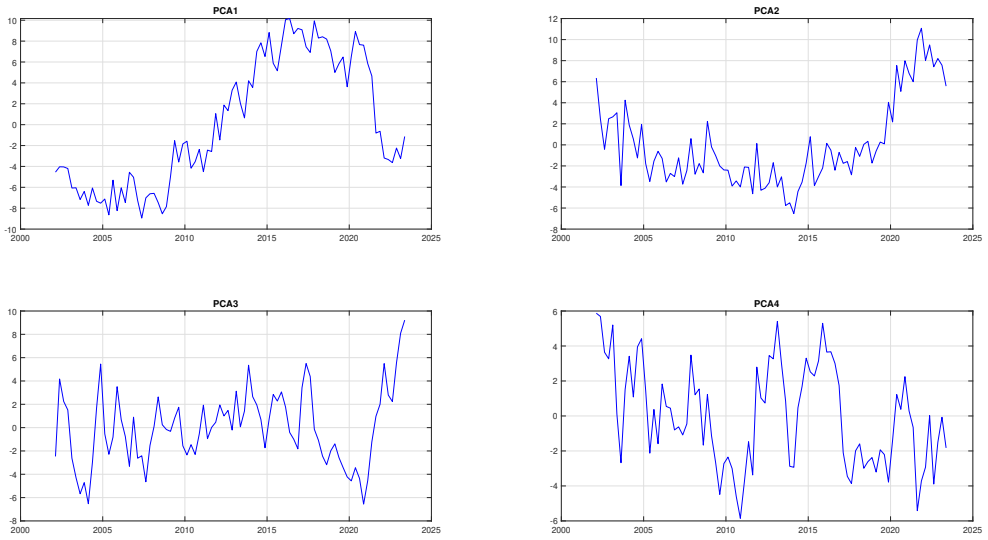


Figure 5: In-sample estimates of first four principal components - Word2Vec/BERT/OpenAI

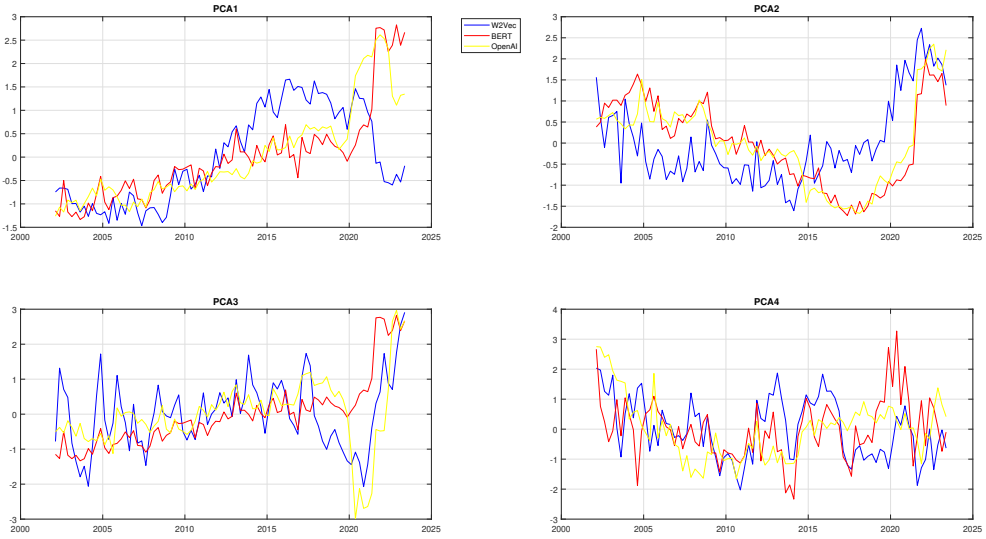
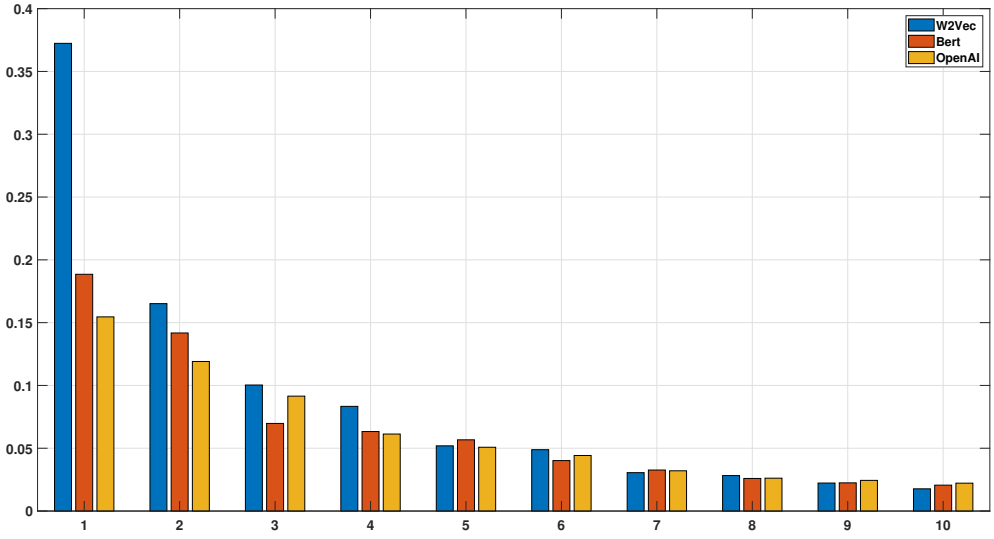


Figure 6: Variance of embeddings explained by first ten principal components



Note: Horizontal Axis: ranked PCAs. Vertical axis: Percentage of total variance of embeddings explained. Blue: Word2Vec. Orange: BERT. Yellow: OpenAI.

A significant share of the variance of our 100 time-series (produced by the evolution over time

of the 100 embedding vector elements) can be explained by a very small number of principal components of the vectors (Fig. 6). This implies that we can replace the 100 time-series with, say, four PCAs, retaining a large part of the information included in the 100 time-series. Our m_t variables for the embeddings will be captured by the first four PCAs. These series are computed recursively to maintain its out-of-sample characteristic.

4.3 Pseudo out-of-sample estimates

In addition to Word2Vec model described above, we use large language models (LLMs) to obtain the embeddings for the same texts. As discussed above, these models trade-off the possibility for pure out-of-sample forecast evaluations for the more nuanced representation of the texts that results from the massively larger training datasets

Our focus is on two different alternatives, BERT (Devlin et al. (2018))(768-dimensional vector) and OpenAI (1536-dimensional vector).

4.4 Alternative indicators for the bivariate models

A legitimate question is whether the textual signals that could be related to inflation, if any, could be extracted in alternative ways other than through embeddings. To test this, we use two sets of series with considerably simpler textual transformations.

The first group is deemed as placebo because they represent metrics derived from each text but without any theoretical reason why they should represent meaningful information. The first placebo variable we calculate is the count of the word "inflation" in each introductory statement. The second placebo is the length, in words, of each statement. The next group of variables does contain economic signal, albeit extracted from texts in a different way as the embeddings. Following Gardner et al. (2022), we use dictionary-based sentiment metrics to measure the ECB's sentiment on inflation, as well as the Governing Council's sentiment. (**Note: we need to explain better what that is**). In addition, to check whether the performance of the embeddings is simply capturing specific sentiments about macroeconomic developments of importance to future inflation, we also include a similar measure focused on labour markets, monetary policy and output.

(**We should also plot these four series here for the sake of completeness.**)

5 Empirical results

5.1 Relative MSE Core Inflation

The out-of-sample MSE for core inflation measure is reported in Table 1. Its first row presents the results using the recursive embeddings calculated with the Word2Vec methodology to obtain true out-of-sample forecasts. While at shorter horizons these embeddings do not contribute meaningfully to improve inflation forecasts, more distant forecasts are significantly better when the signal from the ECB introductory statements are included. For example, the forecast error four quarters ahead is more than 16% lower than the AR model. The results are even better when more advanced embedding methodologies are considered, as seen in the second and third rows: inflation forecasting four quarters ahead is a quarter to a third more precise, although it is not possible to exactly attribute this gain to the models' sophistication or to the (partial) in-sample nature of the exercise.

Table 1: Results for the full sample

	H=1	**H=2**	**H=3**	**H=4**
Word2Vec	0.9705	0.9512	0.8606	0.8374
BERT	0.8075	0.7728	0.6756	0.6440
OpenAI	0.7746	0.7479	0.6714	0.7425
Count Inflation	1.0336	1.0835	1.1016	1.1091
Statement length	1.0157	1.0195	1.0049	1.0030
Sent. Inflation	0.9408	0.9639	0.9389	0.9627
Sent. Gov. Council	0.9820	0.9805	0.9621	0.9695
Sent. Labour	1.0219	0.9993	0.9804	0.9913
Sent. Monetary	1.0542	1.0211	1.0003	0.9994
Sent. Output	1.0675	1.0706	1.0402	1.0232

Interestingly, the alternative measures in built using the same original textual data do not improve inflation forecast at any horizon. The simple metrics presents in the fourth and fifth rows, respectively the count of "inflation" mentions in each text and the statement's length, contribute if anything to worse predictions (values higher than one). The bottom four rows, based on sentiment metrics, all present values that hover around one, demonstrating that at best their information content is neutral for inflation forecasting. This observation strongly suggests embeddings are able to capture sufficient nuance in the texts that sets it apart in predictive power from other text-based metrics.

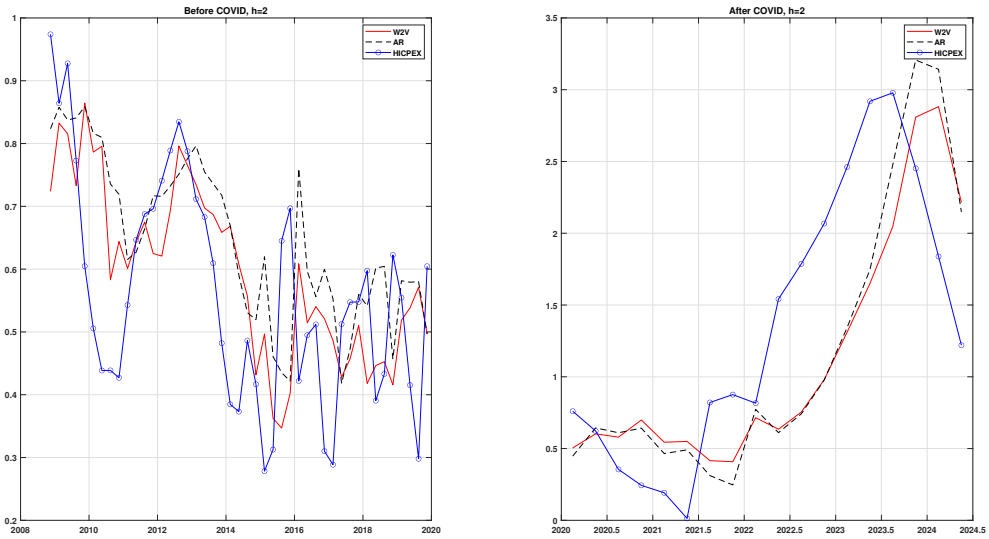
Focusing now only on the pre-Covid sample to evaluate a period where inflation readings were

considerably less volatile, confirms the signal contained in the introductory statement, and more particularly the value of embeddings as a technique that is able to extract this signal compared to alternatives. The first three rows of Table 2, which as in Table 1 represent the three embedding methods studied in this paper, demonstrate a marked reduction in forecast error, this time even for shorter horizons. And similar to before, most other alternative techniques fail to Granger-cause inflation prediction, save for sentiment on inflation (sixth row) and at a lower scale, the Governing Council sentiment metric for longer horizons.

Table 2: Results for the pre-Covid sample

	H=1	**H=2**	**H=3**	**H=4**
Word2Vec	0.8630	0.7699	0.7446	0.7428
BERT	0.8799	0.8781	0.8905	0.9098
OpenAI	0.8623	0.7950	0.7637	0.7595
Count Inflation	1.0280	1.0583	1.0916	1.1345
Statement length	1.0238	1.0597	1.0764	1.1048
Sent. Inflation	0.9527	0.8862	0.8901	0.8987
Sent. Gov. Council	0.9799	0.9259	0.9251	0.9378
Sent. Labour	1.0623	1.0791	1.0916	1.0965
Sent. Monetary	1.2128	1.2681	1.2976	1.3316
Sent. Output	1.0426	1.0413	1.0626	1.0805

Figure 7: HICPex AR and VAR forecasts before and after Covid, H=2



Note: Left panel: 2008Q2-2019Q4. Right panel: 2020Q1-2023Q4.

The results presented are strong evidence that central bank text embeddings Granger-cause in-

flation, agreeing with other findings (also for Europe) that central bank monetary policy communication *surprises* influences future inflation (Bank (2024)). But one important observation is that the methodology to extract the signals matter. In particular, the flexibility and information density afforded by the embedding techniques clearly demonstrate better out-of-sample (or pseudo-out-of-sample for the BERT (Devlin et al. (2018)) and OpenAI models) performance compared with other methogologies. In fact, embeddings are not only an *effective* way to extract signal from central bank texts, they are also *efficient*: the results above obtain with only part of the signal, through the largest principal components.

Other methods allow a more refined use of the richness in the dimensionality of embeddings. While the PCA reduction of the embedding series restrained its dimensionality for the VAR in Eq. 2 and facilitated interpretation of the results, the results themselves do not depend on such dimensionality-reduction devices. We show this by regressing future inflation on all 100 time series $\{\phi_t^{(1)}\}, \{\phi_t^{..}\}, \{\phi_t^{(100)}\}$ of recursively calculated quarterly embeddings. The regression is estimated with a spike-and-slab prior (Mitchell and Beauchamp (1988)), with a hierarchical prior on the degree of shrinkage and model size as in Giannone et al. (2021). This model nests the case of lasso (Tibshirani (1996)), which restricts the L1-norm of coefficients (effectively leading most to be exactly zero) and ridge regression, where the L2-norm is restricted, and all in-between cases. In other words, our regression estimates a linear model that nests the full spectrum of sparse and dense linear regression models with the full set of embedded dimensions.[13]

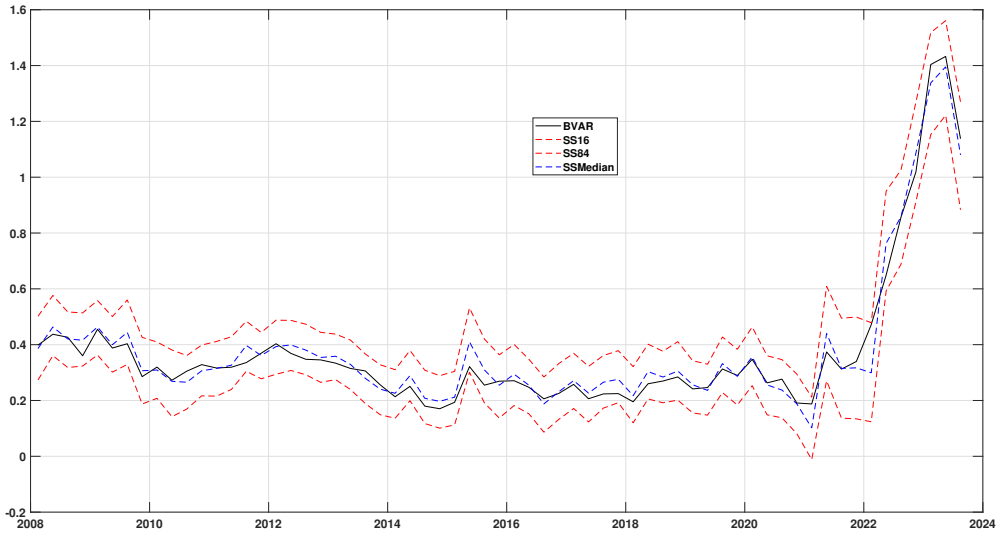
The spike-and-slab regression results in broadly the same data as Eq. 2, confirming the potential for multi-dimensional representation of the textual data to convey predictive signal. as shown in Fig. 8, the estimates for the VAR are contained in the interval with highest posterior density of the spike-and-slab regression, and the VAR's point estimates coincide with the median of the latter regression. As above, note that these results are completely out-of-sample. [**Q: what is the scale of this graph? What does it represent? It wasn't clear from the text or the slides.**]

5.2 Comparisong with Eurosystem inflation forecasts

One possible reason that could explain the inflation forecasting gains from including signal from introductory statements is that these texts might reflect the ECB's own forecast of inflation, itself a reflection of economic developments as well as other information that provides signals on forthcoming inflation. One way to test if this is the case is to substitute m_t in Eq. 2 for an optimal combination of the predicted inflation from the embeddings (π_t^m) and the ECB's own

[13] See Appendix B for a formal description of the spike-and-slab regression prior.

Figure 8: Comparison of bivariate VAR with PCA and spike-and-slab regression



Note: 2008Q1-2023Q3

inflation forecast Broad Macroeconomic Projection Exercises (BMPE) of each period, π_t^{BMPE} as in Eq. 4 below. The BMPE is an Eurosystem forecasting exercise based on all euro area central banks; they are part of the briefing that is in the background of the Introductory Statement. For each period, we use the most recent BMPE inflation forecast. The optimal inflation forecast combination is then:

$$\pi_{t+h}^{\text{opt}} = \theta + \lambda \pi_{t+h}^{m} + (1-\lambda)\pi_{t+h}^{BMPE}. \tag{4}$$

To assess the degree to which the text embeddings provide different forecasting signal compared to the BMPE, we estimate λ for the two sampling periods (pre-Covid and full sample, as before) and for each of the four horizons. Values of λ closer to zero would indicate that the embeddings would not contribute any signal on top of the BMPE for inflation forecasting; values closer to one would indicate that embeddings actually provide most or all of the signal.[14]

Considering the sample before the COVID pandemic (top row of Table 3), the embedding-based forecasts provide signal that are, generally, not encompassed by the BMPE. This is stronger for

[14]Negative values are theoretically possible but would represent an anomaly, since π_t^m is either orthogonal to m_t or it responds to the embedding, which can be arbitrarily multiplied by -1.

Table 3: Optimal weight of embeddings signal compared with ECB inflation expectation in inflation forecasts (Note: HAC standard errors in parenthesis.)

	H=1	**H=2**	**H=3**	**H=4**
Pre-COVID	0.37	0.55	0.56	0.52
	(0.22)	(0.21)	(0.23)	(0.21)
Full Sample	0.22	0.04	0.26	0.68
	(0.24)	(0.26)	(0.32)	(0.31)

the second to the fourth quarter ahead. When the sample includes the turbulent post-Covid period (bottom row), the result is attenuated for short-term forecasts but remains robust at the four-quarter mark. This is consistent with Rosa and Verga (2007)'s result that ECB introductory statements and macroeconomic variables act as complementary, not substitutable, information for the prediction of the policy rate.

6 Conclusions

In this paper, we document how central bank texts can offer predictive signal on inflation. This finding is surprising: unlike the monetary policy rate, which is under direct central bank control, or financial market variables, which are often priced with the policy rate as a reference, the actual inflation reading results from multiple prices across different real economy sectors that rely on the decisions and interactions of a myriad economic agents. This result obtains even in a completely out-of-sample exercise where even the embeddings are learned in periods before the actual forecast, and when inflation forecasts by ECB staff are also considered.

Our main contribution is to show how a relatively simple way to process central banking text can contribute to improve forecasting performance in a "policy-relevant" exercise. We suggest a simple way to use word embeddings. Validation: assess whether it helps to extract relevant information for future inflation from Central Bank texts. Our measure is derived by using a relatively simple NLP model, which can be re-estimated with little cost and, hence, allows an appropriate out-of-sample analysis. The simple setting, while clearly not capturing all of the signal and the nuance in the texts, offers the desirable possibility to conduct a rigorous out-of-sample analysis, which we emphasize as the only way to ensure the robustness of forecasting power comparisons.

There are remaining challenges. Despite these promising results, certain challenges and open questions remain. One challenge is related to the interpretability of the embeddings measure, which can be inescrutable in its raw format. Another important consideration is how to assess the uncertainty of the predictions. And while our results survive a battery of tests that confirms the

predictive power of embeddings of introductory statements in ECB press conferences, a question remains about the extent to which other central bank texts are also as meaningful. In addition, communication varies significantly across central banks (eg, Bulíř et al. (2013)); assessing whether these results hold for other institutions and economies would require careful consideration of the policy and communication practices in each jurisdiction.

The facts that our measure based on word embeddings improves on classical sentiment analysis and its signal is not encompassed by the BMPE hint at potentially interesting questions for future research. A first thread is to explore the connection between Granger-causality, which we document, and actual economic causality: what could be potential channels, eg impacting agents' expectations of policy, that would influence price changes at the horizons we study. Such studies would require equilibrium models that are outside the scope of this paper. A second line of question is how to identify the aspects of central bank texts that matter for inflation prediction. Intepreting embeddings can be a challenging task, but comparing more precisely the texts and the resulting embeddings could offer interesting insights and could eventually help tailor optimal communication policy.

Central bank texts are increasingly available in ways that are amenable for empirical applications, such as the downloadable collection of central bank speeches and monetary policy statements available through the open source `gingado` library (Araujo (2023b)). Some institutions, as the ECB, make them available for download as well in easy-to-process formats (such as CSV files). Macroeconomists and forecasters seem well-posed to use these types of datasets in increasingly sophisticated ways, leveraging the textual processing capabilities powered by ever-advancing machine learning techniques.

A third area is to study whether there are there other policy communications that can Granger-cause inflation. Central banks produce a wide range of communication pieces ranging from more structured and formal inflation reports to ad hoc speeches by officials. In addition to timing, their focus also include other areas of central bank action, such as financial stability and prudential regulation. Would current embedding techniques, such as Word2Vec, also result in meaningful signal also when texts are not explicitly focused on inflation? Understanding how the content and the timing of such communications influence future inflation appears to be a fruitful avenue for future investigations.

A Prior distributions

Assume that $y_t = [\pi_t \;\; te_t]'$. We adopt the following VAR setting:

$$\begin{aligned} y_t &= C + B_1 y_{t-1} + \cdots + B_p y_{t-p} + \epsilon_t, \\ \epsilon_t &\sim \mathcal{N}(0, \Sigma), \end{aligned}$$

where y_t is an N-dimensional vector of time-series, $B_1, \ldots, B_p$ are $N \times N$ matrices of coefficients on the p lags of the variables, C is an N-dimensional vector of constants and Σ is the covariance matrix of the errors. We have $N = 1$ or $N = 2$ and $p = 8$. The largest available estimation sample ranges from 2002Q1 to 2023Q2.

Potentially, this model may be subject to the "curse of dimensionality" due to the large number of parameters to be estimated, relative to the available sample. In such circumstances, the estimation via classical techniques would very likely result in overfitting the data and large estimation uncertainty. De Mol et al. (2008) and Banbura et al. (2010) showed that imposing informative priors which push the parameter values of the model toward those of naïve representations (such as, for example, the random walk model) reduces estimation uncertainty without introducing substantial bias in the estimates, thanks to the tendency for most macroeconomic and financial variables to co-move. In fact, in presence of comovement, the information in the data strongly "conjures" against the prior, so that the parameter estimates reflect sample information even if very tight prior beliefs are enforced.

For this reason, we adopt a Bayesian estimation technique. The prior for the covariance matrix of the residuals Σ is Inverse Wishart, while the prior for the autoregressive coefficients is normal (conditional on Σ). More in details, the prior distributions in our Bayesian VAR are specified as follows. For the prior on the covariance matrix of the errors Σ, we set the degrees of freedom of the Inverse Wishart distribution equal to $N + 2$, the minimum value that guarantees the existence of the prior mean, and we assume a diagonal scaling matrix Ψ, which we parameterize by setting the diagonal values equal to the variance of the residual from an AR(1) model for each individual variable. The baseline prior on the model coefficients is a version of the Minnesota prior (see Litterman, 1979). This prior is centered on the assumption that each variable follows an independent random walk process in levels, possibly with drift. We difference the inflation data before entering the models and we entertain the prior belief that the text embeddings are stationary. For this reason, our prior distribution for all the VAR coefficients is centered on zero.

The prior second moments for the VAR coefficients are:

$$\operatorname{cov}\big((B_s)_{ij}, (B_r)_{hm} \big| \Sigma \big) = \begin{cases} \lambda^2 \frac{1}{s^2} \frac{\Sigma_{ih}}{\psi_j/(d-n-1)} & \text{if } m = j \text{ and } r = s \\ 0 & \text{otherwise} \end{cases}.$$

Notice that the variance of the prior is lower for the coefficients associated with more distant lags and that coefficients associated with the same variable and lag in different equations are allowed to be correlated. The lower prior for longer lags captures the idea that autocorrelation is more likely to die out the longer the lags we are considering. This feature allows us to specify models with long lags. The terms Σ_{ih}/Ψ_j account for the relative scale of the variables. The prior for the intercept C is non-informative.

The setting of the prior distributions depends on the hyperparameter λ, which effectively determines the overall tightness of the prior distributions for the model coefficients because it controls the scale of all variances and covariances. In setting this parameter, we follow the theoretically grounded approach proposed by Giannone et al. (2015), who suggest to treat the hyperparameters as additional parameters, in the spirit of hierarchical modelling. As hyper-prior (i.e., prior distribution for the hyperparameter), we use a proper but almost flat distribution.

B Spike-and-Slab prior - more formally

$$\pi_t = \alpha(L)\pi_{t-1} + \beta_1 m_{1,t-1} + \cdots + \beta_{100} m_{100,t-1} + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d.} \mathcal{N}(0, \sigma^2)$$

Prior

$$\beta_i \mid \sigma^2, \gamma^2, q \sim \text{i.i.d.} \begin{cases} \mathcal{N}(0, \sigma^2\gamma^2) & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases}$$

- Inflation lags always included
- Mitchell and Beauchamp (1988)

- Giannone et al. (2021): conduct posterior inference on prior tightness for degree of shrinkage and model size

References

Ahrens, M., D. Erdemlioglu, M. McMahon, C. J. Neely, and X. Yang (2023): "Mind Your Language: Market Responses to Central Bank Speeches," *Available at SSRN 4471242*.

Ahrens, M. and M. McMahon (2021): "Extracting Economic Signals from Central Bank Speeches," in *Proceedings of the Third Workshop on Economics and Natural Language Processing*, ed. by U. Hahn, V. Hoste, and A. Stent, Punta Cana, Dominican Republic: Association for Computational Linguistics, 93–114.

Anderes, M., A. Rathke, S. Streicher, and J.-E. Sturm (2021): "The role of ECB communication in guiding markets," *Public Choice*, 186, 351–383.

Araujo, D. K. G. (2023a): "Facilitating the use of machine learning by central banks with transfer learning," in *IFC Bulletin No. 57*, Bank for International Settlements, Box 1.

——— (2023b): "gingado: a machine learning library focused on economics and finance," BIS Working Papers No 1122, Bank for International Settlements.

Araujo, D. K. G., G. Bruno, J. Marcucci, R. Schmidt, and B. Tissot (2023): "Machine learning applications in central banking," *Journal of AI, Robotics & Workplace Automation*, 2, 271–293.

Ash, E. and S. Hansen (2023): "Text algorithms in economics," *Annual Review of Economics*, 15, 659–688.

Atkeson, A. and L. E. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?" *Quarterly Review*, 25, 2–11.

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Bank, E. C. (2024): "How central bank communication affects the economy," Accessed: 2024-08-16.

Baranowski, P., H. Bennani, and W. Doryń (2021): "Do the ECB's introductory statements help predict monetary policy? Evidence from a tone analysis," *European Journal of Political Economy*, 66, 101964.

Barbaglia, L., S. Consoli, and S. Manzan (2023): "Forecasting with economic news," *Journal of Business & Economic Statistics*, 41, 708–719.

Bengio, Y., R. Ducharme, and P. Vincent (2000): "A neural probabilistic language model," *Advances in neural information processing systems*, 13.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent dirichlet allocation," *Journal of machine Learning research*, 3, 993–1022.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. De Haan, and D.-J. Jansen (2008): "Central bank communication and monetary policy: A survey of theory and evidence," *Journal of economic literature*, 46, 910–945.

Bulíř, A., M. Čihák, and D.-J. Jansen (2013): "What drives clarity of central bank communication about inflation?" *Open Economies Review*, 24, 125–145.

Carriero, A., D. Pettenuzzo, and S. Shekhar (2024): "Macroeconomic Forecasting with Large Language Models," *arXiv preprint arXiv:2407.00890*.

Chahad, M., A.-C. Hofmann-Drahonsky, B. Meunier, A. Page, and M. Tirpák (2022): "What explains recent errors in the inflation projections of Eurosystem and ECB staff?" https://www.ecb.europa.eu/press/economic-bulletin/focus/2022/html/ecb.ebbox202203_05~6d1fb8f5b0.en.html, accessed: 2024-08-21.

Cieslak, A. and A. Schrimpf (2019): "Non-monetary news in central bank communication," *Journal of International Economics*, 118, 293–315.

Cour-Thimann, P. and A. Jung (2021): "Interest-rate setting and communication at the ECB in its first twenty years," *European Journal of Political Economy*, 70, 102039.

de Bandt, O., J.-C. Bricongne, J. Denes, A. Dhenin, A. D. Gaye, and P.-A. Robert (2023): "Using the Press to Construct a New Indicator of Inflation Perceptions in France," Working Paper 921, Banque de France.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, 146, 318–328.

Dell, M. (2024): "Deep Learning for Economists," Working Paper 32768, National Bureau of Economic Research.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, abs/1810.04805.

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

Ehrmann, M. and A. Wabitsch (2022): "Central bank communication with non-experts–A road to nowhere?" *Journal of Monetary Economics*, 127, 69–85.

Faust, J. and J. H. Wright (2013): "Chapter 1 - Forecasting Inflation," in *Handbook of Economic Forecasting*, ed. by G. Elliott and A. Timmermann, Elsevier, vol. 2 of *Handbook of Economic Forecasting*, 2–56.

Fraccaroli, N., A. Giovannini, J.-F. Jamet, and E. Persson (2022): "Ideology and monetary policy. The role of political parties' stances in the European Central Bank's parliamentary hearings," *European Journal of Political Economy*, 74, 102207.

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

Gáti, L. and A. Handlan (2022): "Monetary communication rules," Tech. rep., ECB Working Paper.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as data," *Journal of Economic Literature*, 57, 535–574.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior Selection for Vector Autoregressions," *The Review of Economics and Statistics*, 97, 436–451.

——— (2021): "Economic Predictions With Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437.

Grand, G., I. A. Blank, F. Pereira, and E. Fedorenko (2022): "Semantic projection recovers rich human knowledge of multiple object features from word embeddings," *Nature human behaviour*, 6, 975–987.

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

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

Hansen, S., T. Ramdas, R. Sadun, and J. Fuller (2021): "The demand for executive skills," Tech. rep., National Bureau of Economic Research.

Hirschbühl, D., L. Onorante, and L. Saiz (2021): "Using machine learning and big data to analyse the business cycle," *Economic Bulletin Articles*, 5.

Jansen, D.-J. and J. De Haan (2007): "The importance of being vigilant: Has ECB communication influenced Euro area inflation expectations?" Tech. Rep. 2134.

Jarociński, M. and P. Karadi (2020): "Deconstructing monetary policy surprises—the role of information shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2022): "Making text count: Economic forecasting using newspaper text," *Journal of Applied Econometrics*, 37, 896–919.

Kogan, L., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2021): *Technology-skill complementarity and labor displacement: Evidence from linking two centuries of patents with occupations*, National Bureau of Economic Research.

Litterman, R. B. (1979): "Techniques of forecasting using vector autoregressions," Tech. rep.

Masciandaro, D., O. Peia, and D. Romelli (2024): "Central bank communication and social media: From silence to Twitter," *Journal of Economic Surveys*, 38, 365–388.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013): "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian variable selection in linear regression," *Journal of the american statistical association*, 83, 1023–1032.

Muennighoff, N., N. Tazi, L. Magne, and N. Reimers (2023): "MTEB: Massive Text Embedding Benchmark," .

Mumtaz, H., J. Saleheen, and R. Spitznagel (2023): "Keep it simple: Central bank communication and asset prices," Tech. rep., Working Paper.

Pavelkova, A. (2022): "The ECB press conference: a textual analysis," .

Pennington, J., R. Socher, and C. D. Manning (2014): "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. (2018): "Deep Contextualized Word Representations," *arXiv preprint arXiv:1802.05365*.

Rosa, C. and G. Verga (2007): "On the consistency and effectiveness of central bank communication: Evidence from the ECB," *European Journal of Political Economy*, 23, 146–175.

Russell, S. J. and P. Norvig (2022): "Artificial intelligence: A modern approach, global edition," *Harlow: Pearson*.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.

Tobback, E., S. Nardelli, and D. Martens (2017): "Between hawks and doves: measuring central bank communication," Tech. rep., ECB Working paper.

Van der Maaten, L. and G. Hinton (2008): "Visualizing data using t-SNE." *Journal of machine learning research*, 9.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017): "Attention is All you Need," in *Advances in Neural Information Processing Systems*, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., vol. 30.