CENTER FOR ADVANCED RESEARCH IN FINANCE

CARF Working Paper

CARF-F-458

Forecasting Japanese inflation with a news-based leading indicator of economic activities

Keiichi Goshima
Waseda University and Bank of Japan

Hiroshi Ishijima
Chuo University

Mototsugu Shintani
The University of Tokyo and Bank of Japan

Hiroki Yamamoto
The University of Tokyo

May, 2019

CARF is presently supported by Dai-ichi Mutual Life Insurance Company, Nomura Holdings, Inc., Sumitomo Mitsui Banking Corporation, MUFG Bank, Finatext Ltd., The Norinchukin Bank and The University of Tokyo Edge Capital Co., Ltd. This financial support enables us to issue CARF Working Papers.

Forecasting Japanese inflation with a news-based leading indicator of economic activities

Keiichi Goshima*, Hiroshi Ishijima†,
Mototsugu Shintani‡, Hiroki Yamamoto§

Abstract

We construct business cycle indexes based on the daily Japanese newspaper articles and estimate the Phillips curve model to forecast inflation at a daily frequency. We find that the news-based leading indicator, constructed from the topic on future economic conditions, is useful in forecasting the inflation rate in Japan.

*Waseda University and Bank of Japan
†Chuo University
‡Corresponding author, The University of Tokyo and Bank of Japan
§The University of Tokyo

1 Introduction

With increasing attention to monitoring current inflation in real time, inflation series are now available at a daily frequency. For example, the Federal Reserve Bank of Cleveland has been providing a daily nowcasting series of the personal consumption expenditures (PCE) and the consumer price index (CPI), while the Billion Prices Project at MIT (Cavallo and Rigobon, 2016) uses prices collected from hundreds of online retailers around the world on a daily basis to measure inflation. In Japan, the *Nikkei CPINow* data, which is constructed from scanner or Point-of-Sale (POS) data, can be viewed as a daily inflation series. Stock and Watson (1999) have emphasized the advantage of using a real economic activity index in forecasting future inflation at a monthly frequency. Their forecasting model is motivated by the Phillips curve, which describes the positive correlation between inflation and real economic activity. In order to examine the usefulness of a Phillips curve model in forecasting inflation at a daily frequency, however, a daily index series of real economic activity needs to be constructed.

In this paper, we first extract text information from the daily newspaper articles, the *Nikkei*, from 1989 to 2017, in constructing an index of real economic activities in Japan at a daily frequency. We then investigate whether the Phillips curve model with the daily news-based business cycle index could improve over the univariate benchmark model in forecasting the daily inflation series (the *Nikkei CPINow*) in Japan. In general, one may also employ a more direct approach to exploring the linear or non-linear relationship between the target inflation series and the relevant text data, rather than summarizing all the text information in a single indicator of economic activity. However, identifying the source of forecast improvement is difficult in such an approach since forecasts are generated inside the black box. In contrast, our approach, based on the Phillips curve, has an advantage in the sense that its outcome can be interpreted through standard economic theory[1].

[1] In the Nobel Prize lecture, Akerlof (2002) states that "probably the single most important macroeconomic relationship is the Phillips curve."

The idea of constructing the business cycle index from text data is not new. For example, Shapiro, Sudhof, and Wilson (2018) develop a US news sentiment index, based on 16 major US newspapers from 1980 to 2015, which is found to be strongly correlated with the state of contemporaneous business cycles. In a similar vein, by combining quarterly GDP data and the daily news topic variables in mixed-frequency dynamic factor models, business cycle indexes are constructed by Thorsrud (2018) for Norway and by Larsen and Thorsrud (2018) for the US, Japan and Europe (the euro area), respectively. While we share the motivation behind these studies, our methodology of constructing business cycle indexes has several distinct features that diverge from their analysis.

In economic applications of sentiment analysis, sentiment indexes are typically computed by using pre-defined dictionaries. Indexes are then investigated to explain the outcome of economic activities, such as the financial performance of the firm and asset prices[2]. Instead of relying on sentiment dictionaries, we take advantage of our unique training data set, the *Economy Watchers Survey*, which is composed of a five-point rating scale through the Japanese workforce assesses the conditions of the Japanese economy and provides an itemized short description of the reasons for their assessment choices. We utilize a machine learning method and train a text classification model using the former (quantitative data) as output and the latter (text data) as input. Shapiro, Sudhof, and Wilson (2018) do not consider only the dictionary-based approach but also the machine learning approach in their analysis. However, their model is trained by a corpus with emotional labels obtained from a social network website, and thus the outcome may not be directly related to economic conditions. In contrast, our training data set is based on a government survey directly questioning people who are engaged in the regional economic activities, including retail, food and beverage, services, housing, manufacturing, non-manufacturing, and employment categories. Therefore,

[2]For example, Tetlock (2007) and Tetlock, Saar-Tsechansky, and Macskassy (2008) use the Harvard IV-4 psychosocial dictionary. Loughran and McDonald (2011) claim that a dictionary based on expert judgement to select words specific to financial topics serves better than the Harvard dictionary. Baker, Bloom, and Davis (2016) also made an expert judgement to select words related to economic policy uncertainty and use them to construct an uncertainty index from newspaper articles.

this corpus has a more specialized economic vocabulary than the more general corpus. In addition, our training data set is based on a public survey conducted by the Cabinet Office of the Japanese government, and it is therefore easily accessible to any researcher.[3]

Since the goal of our analysis is to forecast future inflation using the index of real economic activities, it is important to determine if the sentences in the *Nikkei* newspaper articles are referring to current or future economic conditions. In the *Economy Watchers Survey*, our training data set, respondents of the Japanese workforce, are asked to evaluate the current economic conditions in comparison to the conditions of the past three-month mark, according to the following categories: Worse, Slightly worse, Unchanged, Slightly better, and Better. These correspond to scores of 0, 0.25 ,0.5, 0.75 and 1.[4] We can directly use these scores to learn a text classification model if one is interested in the analysis of the contemporaneous economic conditions (the learner 1 in the main text). Fortunately, in addition to the current status, the survey also requests a 2 to 3-month prognosis of the Japanese economy from the following categories: Will get worse, Will get slightly worse, Will remain unchanged, Will get slightly better, and Will get better. Respondents are asked to provide separate text descriptions of their evaluation reasons for current and future economic conditions. We make use of this survey structure and consider the supervised learning of topics to determine whether the sentences describe either current or future economic conditions. Our use of supervised learning to estimate the sentence topic is in contrast with the analyses of Thorsrud (2018) and Larsen and Thorsrud (2018) that are based on the Latent Dirichlet Allocation (LDA) model, one of the most popularly used unsupervised topic models. For our purposes, supervised learning seems more appropriate since the topic of our interest, namely, future

[3]For this reason, the *Economy Watchers Survey* data set has been used in several text analyses of the Japanese economy. For example, Okazaki and Tsuruga (2015) first employed a dictionary-based method. Later, Suimon, Kinoshita, and Yamamoto (2015) employed a machine learning method to construct a monthly index. However, to the best of our knowledge, our study is the first one to use this data to construct a business cycle index at a daily frequency.

[4]In general, the category of emotions is not limited to the positive or negative (hawkish or dovish, in other words). Bollen, Mao, and Zeng (2011) use the Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions including calm, alert, sure, vital, kind, and happy. Since our classification is ordered, we treat our scores in the analysis below as a continuous variable rather than as a classification problem.

economic conditions, is difficult to discover by means of unsupervised topic model such as LDA.[5] Once the topic is learned (the learner 3 in the main text), we can compute the topic probability of each sentence in *the Nikkei* newspaper articles being a description of future economic conditions. We then use this probability as a weight on the scores obtained from the text classification model for the five-point rating scale of future economic conditions (the learner 2 in the main text). We refer to the resulting business cycle index, the *News-based Leading Indicator* and use this index for the purpose of forecasting inflation.[6]

To evaluate the empirical performance of the inflation forecasting at daily frequency, by the Phillips curve combined with the news-based leading indicator, we employ a simulated out-of-sample forecasting methodology in which competing models are repeatedly estimated in each period to compute forecast errors. We use an autoregressive (AR) model as a benchmark and investigate whether mean squared errors (MSEs) of the Phillips curve forecast are less than those of the benchmark forecast. As our forecasting models of interest are nested, we employ an out-of-sample F-test statistic proposed by McCracken (2007) to test the null hypothesis of equal predictive accuracy in nested models. We find that our News-based Leading Indicator can help forecast the future inflation rate in Japan at a daily frequency.

The paper is organized as follows. Section 2 elaborates on the methodology of constructing our business cycle indexes in detail. Section 3 develops a forecasting model for daily inflation and presents the main results. Section 4 concludes.

[5] Other economic applications of unsupervised topic models, specifically the LDA, include Hansen and McMahon (2016) and Hansen, McMahon and Prat (2018). Our approach to the supervised learning of topics is closer to the idea used in Gentzkow and Shapiro (2010), who estimate the political slants of U.S. daily newspapers by using the speech record of Democrat and Republican members of Congress as a training data set.

[6] An alternative business cycle index based on the topic-weighted scores obtained from learner 1 may be useful for nowcasting purposes rather than forecasting purposes. We refer to the resulting business cycle index, *News-based Coincident Indicator* (NCI).

2 Building news-based business cycle indexes

2.1 Text data: The *Nikkei* and the *Economy Watchers Survey*

To construct the daily index of real economic activities, we utilize daily news articles in the *Nikkei*, the leading economic newspaper in Japan, with a circulation of 2.8 million as of June 2018. The *Nikkei* is a newspaper that specializes in financial, business, and industry sectors. We use news articles, from both morning and evening papers from April 1, 1989 to December 31, 2017. Table 1 shows the summary statistics of our text data from the *Nikkei*. Our main data consists of about 3.8 million articles with more than 30 million sentences in total. All texts in our data are written in Japanese. Unlike English texts, Japanese texts do not have spaces between words. Hence, we split sentences into words in advance using MeCab, which is a morphological analysis library for Japanese texts.[7] After we divide articles into individual sentences, we estimate the news sentiment score on a sentence basis because our training data set, the *Economy Watchers Survey*, provides annotated scores on a sentence basis.

In what follows, we employ a machine learning method to learn text classification models from our training data set for the purpose of building two types of news-based business cycle indexes, called the News-based Coincident Indicator (hereafter, NCI) and the News-based Leading Indicator (hereafter, NLI). NCI and NLI are designed to capture, on a daily basis, current and future economic conditions, respectively. We suppose that the sentences in the *Nikkei* on Japanese economic activities have the same (and time-invariant) structure as those in the *Economy Watchers Survey*. In other words, from the view point of natural language processing (NLP), the domain of our training data set, the *Economy Watchers Survey*, is same as that of our input data of the *Nikkei*. Figure 1 shows an outline of the procedure we use to construct our business cycle indexes. In the first step, we train text classification models from the *Economy Watchers Survey* using a supervised learning. In the second step,

[7] For details of MeCab, see the web page (https://taku910.github.io/mecab/).

we estimate the sentiment scores of sentences in the *Nikkei* articles using trained models. In the last step, we aggregate news sentiment scores to build news-based business cycle indexes.

To train our model, we utilize descriptions for the assessment of the economy in the *Economy Watchers Survey*, which is published by Cabinet Office of the government of Japan.[8] The purpose of this monthly survey is to promptly grasp current and future economic conditions by consulting people involved in regional economy activity. The assessment of current conditions refers to the direction of change in the economic conditions compared to previous conditions from 3 months before on a five-point scale. Similarly, the assessment of future conditions refers to the expected direction of change within the next 2 or 3 months on a five-point scale. In the *Economy Watchers Survey*, an assessment by each respondent is accompanied by his or her description. Our supervised learning utilizes descriptions as input data and the assessments as output labels. Table 2 shows selected example descriptions of the training data set from the *Economy Watchers Survey*. We employ the survey data from January 2011 to June 2018. The number of total sentences in descriptions for the assessment of current and future economic conditions are 112,214 and 125,055, respectively. Table 3 shows the breakdown of descriptions in the training data set. Descriptions for the assessment of current economic conditions tend to use the present tense and the present progressive tense. On the other hand, descriptions for the assessment of future economic conditions include auxiliary verbs related to the future tense, such as "will," and words to anticipate the future, such as "expect."

2.2 Text classification model

In economic applications of sentiment analysis, the two most frequently used approaches are the dictionary-based approach and the machine learning approach. The advantage of the dictionary-based approach is its tractability and an easy implementation. The method quan-

[8] For details of the *Economy Watchers Survey*, see the webpage of Cabinet Office (https://www5.cao.go.jp/keizai3/watcher-e/index-e.html).

tifies text sentiments by simply counting the number of positive and negative words using a pre-defined dictionary. This approach, however, can fail to judge whether a sentence has good or bad information because word meaning often depends on the context. For example, the sentence, "The firm's performance is not good," can wrongly be taken as a good signal on economic conditions because it contains the word "good." Therefore, it takes considerable time to incorporate all the possible patterns into the pre-defined dictionary. In addition, no established Japanese sentiment dictionary specializing in economic fields is available in the literature. Unlike the dictionary-based approach, the machine learning approach automatically recognizes patterns from the text data (input) and annotated scores (output). Among machine learning methods, neural networks in recent years have achieved high performance in text classification tasks. They excel in utilizing syntax information, such as sequence alignment and word co-occurrence.

Over the years, various neural network-based models have been developed in the field of computer science, including the recurrent neural network (Chung *et al.,* 2014), the recursive neural network (Socher *et al.,* 2013), the convolutional neural network (CNN, Kim, 2014, Zhang, Zhao, and LeCun, 2015) and the self-attention network (Lin *et al.,* 2017). In our analysis, we employ a text classification model based on neural networks called fastText, which was developed by Joulin *et al.* (2017).[9] According to Joulin *et al.* (2017), fastText is a simple and computationally efficient network architecture that at the same time performs as well as classifiers based on other neural network models, such as the CNN and long short-term memory (LSTM) in terms of accuracy. Joulin *et al.* (2017) compared the classification accuracy of test data sets using eight corpora in comparison with six models, including three (multinomial) logistic regression-based models and three neural network-based models: the character-level CNN (char-CNN) of Zhang, Zhao, and LeCun (2015), the character-level convolutional recurrent neural network (char-CRNN) of Xiao and Cho (2016) and the very deep convolutional neural network (VDCNN) of Conneau *et al.* (2016). In the panel (a)

[9]For details of fastText and its setting in our analysis, see the appendix.

of Table 4, we summarize the performance of fastText and other competing models from the experiment of Joulin *et al.* (2017). In addition, we also use our training data set, the *Economy Watchers Survey*, and compare the binary classification accuracy of fastText with those of support vector machine (SVM), CNN, char-CNN and LSTM. In this experiment, the binary classification task is classifying texts into two topic classes, current and future economic conditions. The results of our own experiment in terms of the accuracy of the test data set are reported in panel (b) of Table 4.[10] Based on the results of experiments conducted by Joulin *et al.* (2017) and by ourselves, it seems fair to say that the performance of fastText is comparable to alternative machine learning models.

Finally, in terms of computation time, Joulin *et al.* (2017) report that training and evaluation of sentiment analysis data sets using fastText are many orders of magnitude faster than char-CNN and VDCNN. In our own experiment with the *Economy Watchers Survey*, we also confirm that fastText can be trained much more quickly than other models. Therefore, the method seems to be effective in assigning sentiment scores to our large-scale news article data set without losing much of accuracy.

2.3 Estimation and aggregation

In order to compute NCI and NLI from sentences in the *Nikkei*, we employ three types of learners based on fastText, two of which are regression learners and the other, a classification learner. The first learner (learner 1) is trained from the assessment of the current economy on a five-point scale and its descriptions from the training data set, the *Economy Watchers Survey*. This learner assigns continuous sentiment scores, $Score\text{-}CI$, to sentences in the *Nikkei*. The second learner (learner 2) is trained from the assessment of the future economy on a five-point scale and its descriptions. This learner assigns continuous sentiment scores, $Score\text{-}LI$. Higher $Score\text{-}CI$ and $Score\text{-}LI$ indicate that texts contain information about better current

[10] We divide the *Economy Watchers Survey* into three parts: 80% for training, 10% for validation and 10% for testing.

and future economic conditions, respectively. The third learner (learner 3) is trained using assessments of both the current and future economic conditions as well as their descriptions. This learner assigns topic probabilities, W, to sentences in the *Nikkei*. We regard outputs from a sigmoid function as topic probabilities. The topic probability W takes a value near zero if a sentence is similar to descriptions for the assessment of the current economy, and takes a value near one if a a sentence is similar to descriptions for the assessment of the future economy. Table 5 shows the summary of three types of our learners.

We give sentiment scores and topic probabilities to all sentences in news articles with three learners. When a trained model assigns a sentiment score to each sentence in news articles, out-of-vocabulary words (words or tokens not included in a training data set) are replaced by a common special character, such as <UNKNOWN>. Sentences in news articles longer than the longest sentence in the training data set are truncated. In general, even if exactly the same neural network model is employed, slightly different outputs can be obtained on each run depending on initialization. Therefore, we train the models ten times, and use average scores from all cases.

In the next step, we construct two news-based business cycle indexes using the average of scores weighted by the topic probabilities. In particular, NCI and NLI are respectively defined by

$$NCI_t = \frac{\sum_{k=1}^{N_t} Score\text{-}CI_{t,k} * (1 - W_{t,k})}{\sum_{k=1}^{N_t} (1 - W_{t,k})}, \tag{1}$$

$$NLI_t = \frac{\sum_{k=1}^{N_t} Score\text{-}LI_{t,k} * W_{t,k}}{\sum_{k=1}^{N_t} W_{t,k}}, \tag{2}$$

where N_t is the number of sentences in day t, $Score\text{-}CI_{t,k}$ is a score of k-th sentence assigned by learner 1 in day t, $Score\text{-}LI_{t,k}$ is a score of k-th sentence assigned by learner 2 in day t and $W_{t,k}$ is a topic probability of k-th sentence assigned by learner 3 in day t. Figure 2 plots our constructed news-based business cycle indexes, NCI and NLI. Table 6 shows their summary statistics. Overall, two indicators move closer to each other. However, there

are some differences between the two, reflecting the fact that the NCI captures the current economic conditions while the NLI captures future economic conditions. In particular, the median of the NLI seems to be higher than that of the NCI, which may indicate that the typical newspaper article tends to be relatively optimistic about future economic conditions. This tendency becomes clearer during the post-financial crisis period.

For the purpose of comparing our new business cycle indexes with other indicators of economic activity, we converted the daily series into a monthly series by using monthly averages. Table 7 reports the correlations of our business cycle indexes and official business cycle indicators, namely two diffusion indexes (DIs) from the summary results of the *Economy Watchers Survey* and three composite indexes (CIs) of business conditions from the Economic Social Research Institute (ESRI) of the Cabinet Office. Both our business cycle indexes, the NCI and the NLI, turn out to be highly correlated with the current and future DIs in the *Economy Watchers Survey*. This outcome may not be very surprising given the fact that our indexes are calculated from the same survey information as the DIs in the *Economy Watchers Survey*. However, among three official CIs, namely, (i) the leading index, (ii) the coincident index, and (iii) the lagging index, the NLI is highly correlated with the leading index, which is computed without using newspaper articles or the *Economy Watchers Survey*. Figure 3 plots NCI, NLI, and ESRI's leading index along with official recession episodes. The figure shows that both the NCI and the NLI tend to decrease during economic downturns. In particular, all indexes clearly dropped during the financial crisis of 2008. Overall, it seems fair to say that our news-based business cycle indexes capture well the business cycle properties of the Japanese economy.[11] In the following sections, we mainly focus on using the NLI in our simulated out-of-sample forecasting exercise of daily inflation series.

[11] The only exceptions are that the value of NCI becomes smaller than that in 2008 and that NCI deviates from NLI and ESRI's leading index during 2010s.

3 Forecasting performance of news-based leading indicator

3.1 Phillips curve inflation forecast

Motivated by the well-known Phillips curve model, which describes the correlation between the inflation and unemployment rate, Stock and Watson (1999) conducted a simulated out-of-sample forecasting analysis of US inflation at the 12-month horizon. They claimed that the inflation forecast could be improved by replacing the unemployment rate with a single real economic activity index, especially with one constructed from 168 economic indicators. Atkeson and Ohanian (2001) reviewed the literature of the 1990s and challenged the belief that the conventional Phillips curve model is useful tool for inflation forecasting. They revisited the forecast performance of the Phillips curved model adopted in Stock and Watson (1999) and claimed that their naive forecasts outperformed the Phillips curve forecast. Later, Stock and Watson (2009) provided a comprehensive literature review on inflation forecasting and pointed out that the results against the Phillips curve forecast obtained by Atkeson and Ohanian (2001) could disappear, depending on the forecast horizon and sample period.[12] In summary, it seems fair to say that no consensus has been reached regarding the validity of the Phillips curve relationship in forecasting inflation. However, almost all the existing studies utilized monthly or quarterly data to examine the performance of the Phillips curve inflation forecast. Here, we use the daily data to evaluate the usefulness of the Phillips curve in forecasting Japanese inflation.

In our analysis, we employ a Phillips curve model similar to the one considered by Stock and Watson (1999), who replaced unemployment rate with a single real economic activity index. In particular, we consider the daily Japanese inflation series called *Nikkei CPINow*

[12] The results of Faust and Wright (2013) on the Phillips curve inflation forecast are not supportive. In a more recent study, Tallman and Zaman (2017) find that the Phillips curve is useful in forecasting a sub-aggregate measure of inflation.

and investigate its relationship with our news-based business cycle index. *Nikkei CPINow* is known for the first real time price index in Japan. NOWCAST, Inc. releases two CPINow series called the CPINow-T index and the CPINow-S index. The CPINow-T index is a daily series calculated from POS data and has been available since April 1, 1989. More than 800 stores and 300,000 different products such as food and daily necessities items are covered in the survey. Unlike the official CPI, shares of the items are taken into account everyday by taking the advantage of the POS data. On the other hand, the CPINow-S index is a monthly series designed to closely match the official CPI by selecting the same representative items and by using the same index formula. The CPINow-S index has been available since January 2015.

The summary statistics of the two CPINow series are provided in Table 8.[13] Table 9 shows the correlations between the two CPINow series and the official CPI at a monthly frequency. Here, the daily series of the CPINow-T index series is transformed by using the monthly average. In the table, a remarkably high correlation stands out between the CPINow-S index and the official CPI for all items fresh food and energy. This result suggests that the CPINow series can serve as a good proxy for the official CPI at a higher frequency. Since the CPINow-T index is released two days after actual transactions, it is also useful for nowcasting purposes. A direct comparison between CPINow-T index and official CPI inflation series is provided in Figure 4.

In what follows, we transform the CPINow-T index to construct 1-month to 12-month inflation at an annual rate, which is the target variable in our forecasting analysis. Following the previous studies, including Atkeson and Ohanian (2001) and Stock and Watson (1999, 2009), our target variable is m-period inflation defined by

$$\pi_t^m = \frac{1}{m} \sum_{j=0}^{m-1} \pi_{t-j} \tag{3}$$

[13] In our analysis, missing observations of CPINow series in November and December of 2003 and 2004 are replaced by the fitted values from the regresson of the monthly average of CPINow on the official CPI from April 1989 to December 2017.

where m is a window size and π_t is the CPINow-T index series in the form of the daily inflation rate at an annual rate. For the forecast horizon h, we consider approximately one month to one year by setting $h = 30 \times k$ for $k = 1, 2, ..., 12$.

The h-period ahead forecast is constructed from the Phillips curve model given by

$$\pi^h_{t+h} = \alpha + \beta NLI_t + \phi_h(L)\pi_t + e_{t+h}, \tag{4}$$

where NLI_t is our news-based leading indicator (2) designed to capture the future economic conditions, $\phi_h(L) = \sum_{j=1}^{h} \phi_j L^{j-1}$ and e_{t+h} is the forecast error. It should be noted that if $\beta = 0$, (4) reduces to an AR(h) forecast with no restriction on the AR coefficients. In our exercise, we consider an AR model as a benchmark model and investigate whether the Phillips curve model combined with the news-based leading indicator can outperform the benchmark model. However, if no restriction is imposed on AR coefficients, the number of unknown parameters in the AR(h) model can be as large as 360. To avoid the issue of overfitting with too many parameters in the AR model, we consider several parsimonious specifications.

First, as in Atkeson and Ohanian (2001), a naive forecast of h-period inflation can be constructed using the current value of h-period inflation or $\phi_h(L)\pi_t = \pi^h_t = (1/h)\sum_{j=0}^{h-1} \pi_{t-j}$. It should be noted that this specification can be obtained as the h-period moving average of the random walk forecast and only α needs to be estimated. Since it imposes a non-stationary (NS) restriction on AR coefficients, we refer this specification to AR-NS. Second, we can combine several values of m-period inflation with different window size m and employ

$$\phi_h(L)\pi_t = \phi_1\pi_t + \phi_7\pi^7_t + \sum_{k=1}^{h/30} \phi_{30k}\pi^{30k}_t.$$

For example, in this specification, a one-month ahead inflation forecast is computed by combining daily inflation, weekly inflation and monthly inflation. A similar parsimonious specification has also been employed by Ito and Yabu (2007) and Fatum and Hutchison (2010) in

their government intervention analysis of daily foreign exchange rates, and by Corsi (2009) in his forecasting analysis of realized volatilities.[14] It should be noted that this specification resembles a mixed-frequency (MF) approach employed in the MIDAS regression of Ghysels, Santa-Clara, and Valkanov (2006) and Ghysels, Sinko, and Valkanov (2007). For this reason, we refer this specification to AR-MF. Third, the lag length can be selected using information criteria such as AIC and BIC and we refer these specifications to AR-AIC and AR-BIC, respectively. Using the simulated out-of-sample forecasting methodology explained below, we select a benchmark AR forecast from these alternative specifications of AR coefficients (namely, AR(h), AR-NS, AR-MF, AR-AIC, and AR-BIC).

3.2 Simulated out-of-sample forecasting

We employ the simulated out-of-sample forecasting methodology in evaluating the forecasting model. In this approach, out-of-sample forecasts are computed as if a real-time forecaster were estimating the model using only the data available at the time of the past forecast. In particular, by using the sample only through the period t, the h-period ahead forecast of inflation, $\pi^h_{t+h|t}$ is obtained. We then compare the forecast value $\pi^h_{t+h|t}$ with a realized value π^h_{t+h} to compute the forecast error at the period $t+h$, namely, $\hat{e}_{t+h}$. Next, we follow the same procedure by using the sample through the period $t+1$, estimate the model and compute $\hat{e}_{t+h+1}$. We repeat this process $P-h+1$ times to obtain $P-h+1$ forecast errors.[15] Then, the MSE of h-period ahead inflation forecast, defined as $\sigma^2 = E(e^2_{t+h})$, can be estimated by $\hat{\sigma}^2 = (P-h+1)^{-1}\sum_{t=R}^{P+R-h}\hat{e}^2_{t+h}$ where R is the sample size used to estimate the forecast model at the beginning of the forecast evaluation.

We consider two estimation schemes to evaluate the forecasting performance. The first is the *rolling scheme* where the model is estimated using a moving data window of the length R.

[14] The technical trading rule regression considered in Shintani, Yabu and Nagakura (2012) is also a special case of this specification, which combines moving averages of different windows in forecasting.

[15] It should be also noted that this simulated out-of-sample forecasting exercise assumes that our news-based business cycle index is available at each time the forecast is computed. However, in reality, the fact that our training data is taken from 2011 to 2018 makes it impossible to utilize the index at the time of the forecast.

The second is the *recursive scheme* where the data with an increasing number of observations is used each time the new model is estimated. In the recursive scheme, R represents the sample size used in the initial step. For the sake of robustness, we conduct simulated out-of-sample forecasting experiments with $P/R = 0.4$ and 1.0. The exact numbers of P and R are selected using the identity $P + R = T - h + 1$ where T, the full sample size, is 10,413.

We are interested in determining if the MSE of the Phillips curve model is less than the benchmark AR models without using the news-based leading indicator. In the first step, we investigate an appropriate benchmark AR model by estimating the MSE of inflation forecast using the simulated out-of-sample forecasting methodology described above. The results of the MSE estimates for various forecast horizons h are summarized in Table 10. The performance of AR-NS is uniformly worse than that of an unrestricted AR(h) model. The AR-MF specification performs almost the same as an unrestricted AR(h) model. Both AR-AIC and AR-BIC perform well in shorter horizons but not in longer horizons. On balance, we select AR-MF as our preferred specification for the benchmark AR model. Once the benchmark model is selected, we can estimate the Phillips curve model (4) in the next step. For the case when the full sample period from April 1, 1989 to December 31, 2017 is used, estimated coefficients are provided in Table 11. This table demonstrates that when the forecast horizon h becomes 120 or longer, coefficients on the news-based leading indicator (β) become positive and significant. The table also reports the estimates of the sum of AR coefficients, $\phi_h(1) = \sum_{j=1}^{h} \phi_j$. The sum of AR coefficients tends to be decreasing with the forecast horizon h, suggesting the stationarity of inflation in Japan. In summary, for four-month to one-year horizons, signs of the coefficients turn out to be consistent with the notion of the Phillips curve.

The benchmark AR model is clearly nested by the Phillips curve model (4). Since the forecasting model of our interest nests the benchmark model, we employ the out-of-sample F type test statistic proposed in McCracken (2007), which is designed to compare the forecasting

performance of nested models. The out-of-sample F type test statistic is defined as

$$F = (P - h + 1) \times \frac{\hat{\sigma}^2_{AR} - \hat{\sigma}^2_{PC}}{\hat{\sigma}^2_{PC}} \tag{5}$$

where $\hat{\sigma}^2_{AR}$ is the estimator of the MSE of the AR inflation forecast σ^2_{AR} and $\hat{\sigma}^2_{PC}$ is the estimator of the MSE of the Phillips curve inflation forecast σ^2_{PC}.

We use this statistic to test for the null hypothesis of H_0:$\sigma^2_{PC} = \sigma^2_{AR}$ against an alternative hypothesis of H_1:$\sigma^2_{PC} < \sigma^2_{AR}$. McCracken (2007) shows that when $h = 1$, the F statistic asymptotically follows non-standard distribution under the null hypothesis. He provides the critical values which depend on P/R and the difference in the number of regressors. In general, for the case of a longer forecast horizon $h > 1$, the distribution becomes data dependent. However, for the case when the number of additional regressors is exactly one, as in our setting, the critical values in McCracken (2007) can still be valid (see West (2006) for this point more in detail). We compute the relative size of MSEs of the benchmark AR model and the Phillips curve model, then conduct the out-of-sample F test for various h.

Table 12 shows the relative MSE in terms of percentage deviations given by $100 \times (\hat{\sigma}^2_{AR} - \hat{\sigma}^2_{PC})/\hat{\sigma}^2_{PC}$ with out-of-sample F test statistics in parenthesis. Note that the sample period in the analysis depends on forecast horizon (h). For example, the sampling period for estimation in the first step of the rolling scheme with $P/R = 1$ for $h = 360$ is from June 29, 1989 to February 19, 2010. In this case, both R and P are 2,873.

Based on the point estimates of MSEs, the Phillips curve forecast performs better than the AR forecast when h is 120 and longer for the rolling scheme and when h is 90 and longer for the recursive scheme. Furthermore, the results of the out-of-sample F test imply that when the forecasting horizon becomes longer, σ^2_{PC} becomes significantly less than σ^2_{AR}. These results confirm that the Phillips curve forecast outperforms the AR forecast at least for horizons greater than 3 months. The cumulative squared prediction error differences of two models for $h = 120, 240$ and 360 are also shown in Figure 5. This figure implies

that the improvements are clearly observed in both in the beginning and after 2016 in the evaluation period.[16] The significant reduction of the MSE does not change among the out-of-sample simulation designs with $P/R = 0.4$ and 1.0. Our finding that the news-based leading indicator contains valuable information about future inflation is consistent with the fact that the assessment of future economic conditions refers to the 2 to 3 months ahead projected status of the Japanese economy in the *Economy Watchers Survey*.

3.3 Discussions

Our results on inflation forecast improvements by using the Phillips curve model at a daily frequency cast new light on the literature on inflation dynamics in Japan, since no previous studies have used a daily inflation series in estimating the Phillips curve.[17] However, it is still of interest to compare our main result with previous studies on the inflation forecast in Japan. As pointed out by Fukuda and Keida (2001), the Phillips curve in Japan has not been considered very useful in inflation forecasting compared to the U.S. case. As shown in Figure 4, Japan has experienced a long-lasting deflation period since the second half of the 1990s. Unlike the status during the inflation periods of the 1970s and 1980s, the Phillips curve relationship is believed to be weakened during the period of deflation. This weak relationship between inflation and real economic activity has also stood out at the time of global financial crisis because the sharp drop in output gap and the sharp rise in the unemployment did not result in severe deflation. To allow for the possibility of changing coefficients in the Phillips curve, Nishizaki, Sekine and Ueno (2014) estimated the time-varying parameter model of inflation. It should be noted that, in our main analysis in the previous section, the sample period includes both periods of declining inflation until 1995 and of deflation. Our sample period also contains the time of the global financial crisis.

[16] See Welch and Goyal (2008, Figure 1), for example, on the use of the cumulative squared prediction error difference.

[17] The only exception is the working paper version of Abe and Tonogi (2010), which contains some discussion of the correlation of the daily inflation series and the GDP gap.

In response to prolonged deflation, the Bank of Japan (BOJ) introduced the price stability target of a 2 percent annual inflation rate in January 2013 and has been conducting a series of unconventional monetary easing policies, including the quantitative and qualitative easing (QQE) policy of April 2013. Using micro price data, Watanabe and Watanabe (2018) investigated how items whose price remained unchanged contributed to inflation dynamics in Japan under the monetary easing policy since April 2013. They discovered that items with flexible prices contributed to raising the inflation in 2014 while items with sticky prices did not.

To incorporate the different phases of the Japanese economy described above and possible parameter shifts in the Phillips curve, we examine the robustness of our main results by repeating the same out-of-sample forecasting exercise using the following subsamples. The first subsample is from January 1, 1996 to December 31, 2017, so that the declining inflation phase until 1995 is removed from the full sample. The second subsample we consider is the post-global financial crisis era from September 1, 2008 to December 31, 2017. The third subsample is from April 1, 2013 to December 31, 2017, which corresponds to the period when the BOJ conducted the unconventional monetary easing policy. Note that all the subsamples end in 2017, mainly because our training data extends from 2011 to 2018.

Tables 13 shows the resulting relative MSEs from the subsample analysis. Even if the declining inflation period is removed, the results are very similar to the full sample case. It is interesting to note that for the post-crisis subsample, significant reductions of MSE are observed for all forecast horizons, including one to three months. In contrast, much weaker evidence is obtained for the subsample of the unconventional monetary easing policy period. However, even in this case, a significant reduction is still obtained when the horizon approaches around one year.

In the previous subsection, a univariate AR model with some parameter restriction is selected as a benchmark model to evaluate the performance of the Phillips curve forecast. Let us now turn to a different way of checking the robustness of our main results by introducing

alternative benchmark models. In particular, we consider whether our news-based leading indicator is still useful even when the benchmark univariate AR model is replaced by a vector autoregressive (VAR) model with an additional variable x_t. In this case, the Phillips curve model (4) is replaced by its extended version given by

$$\pi_{t+h}^h = \alpha + \beta NLI_t + \phi_h(L)\pi_t + \varphi_h(L)x_t + e_{t+h}, \tag{6}$$

where NLI_t is the news leading indicator, $\varphi_h(L) = \sum_{j=1}^h \varphi_j L^{j-1}$, and x_t is some daily variable possibly representing daily market information. It should be noted that (6) reduces to a VAR(h) forecast if NLI_t has no predictive power or $\beta = 0$. As in the benchmark AR model in the main results, we employ a parsimonious AR-MF specification for $\phi_h(L)$. Since (6) is no longer in the form of the standard Phillips curve model, in what follows we simply refer it to the extended Phillips curve model.

For the choice of the additional variable x_t, we consider (i) the daily percentage change for the Nikkei Stock Average, (ii) the daily percentage change for the dollar to yen exchange rate, (iii) the daily percentage change for West Texas Intermediate (WTI) crude oil spot prices, and (iv) the daily series of spread between 10-year and 1-year Japanese government bonds (JGBs). In addition, we also consider (v) the first principal component of all four series ((i) to (iv)) as the fifth choice. This last choice can be viewed as the unobserved common factor similar to the one employed in Shintani (2005), who claimed that the common factor computed from the principal component analysis is useful in improving monthly inflation forecasts in Japan. Finally, we employ a parsimonious lag specification of the additional variable simply by imposing the h-period moving average restriction $\varphi_h(L)x_t = \varphi x_t^h$ where $x_t^h = (1/h)\sum_{j=0}^{h-1} x_{t-j}$.

Tables 14 and 15 show the results of this additional analysis. Table 14 reports the estimated MSEs of the benchmark VAR model analogues to Table 10. Table 15 shows the relative MSE in terms of percentage deviations given by $100 \times (\hat{\sigma}_{VAR}^2 - \hat{\sigma}_{EPC}^2)/\hat{\sigma}_{EPC}^2$ where

$\hat{\sigma}^2_{VAR}$ is the estimator of the MSE of the VAR inflation forecast and $\hat{\sigma}^2_{EPC}$ is the estimator of the MSE of the extended Phillips curve inflation forecast, with out-of-sample F test statistics in parenthesis. Note that, since the difference in the number of regressors between the VAR model and the extended Phillips curve model is still one, we can use the critical values from McCracken (2007).

The results indicate that the addition of the news-based leading indicator can improve forecasting accuracy even if the benchmark univariate AR model is replaced by the VAR model. This fact suggests that our news-based leading indicator contains information for future inflation that is not included in the additional daily market data.

4 Conclusion

We constructed the news-based business cycle index from the daily newspaper articles and examine its informational content in predicting the future inflation in Japan. Our analysis suggested that our news-based business cycle index captures daily changes in real economic activity and our index is highly correlated with other business cycle indicators in lower frequency. We found that the news-based leading indicator, rather than the news-based coincident indicator, is useful in forecasting the inflation rate in Japan for horizons longer than 3 months. Our finding that the news-based leading indicator contains valuable information about future inflation is consistent with the fact that the assessment of future conditions refers to the projected 3-months ahead economic conditions in the *Economy Watchers Survey*.

Appendix A. Text analysis using fastText

In this appendix, we describe details on our employed text classification model and its setting in our analysis. This text classification model is known for an algorithm in the fastText library from Facebook. However, the official fastText library deals with only a task for classification, not for regression. Thus, we reimplemented the model using Keras, which is a python library, instead of utilizing the official library.[18] According to Joulin *et al.* (2017), fastText is a simple and computationally efficient network architecture, and at the same time, it performs on a par with classifiers based on famous neural network models, such as the CNN and the LSTM in term of accuracy.

In fastText, there are only three layers: (i) the embedding layer; (ii) the average pooling layer; and (iii) the output layer. Sentences that are represented as N-gram features are converted into fixed length vectors within an embedding layer and an average pooling layer. In the embedding layer, each word is converted into a fixed length vector. Let x_w be a $V \times 1$ one-hot vector, where V is the vocabulary size.[19] For the vocabulary, we used all types of words (unigram features) and their bigram features in a training data set. According to Joulin *et al.* (2017), including the bigram as input features, in addition to the unigram, improves the accuracy of the model.

Embedding layer

The operation of the embedding layer is defined as

$$\overline{x}_w = \mathbf{E}x_w, \tag{A.1}$$

where $\mathbf{E}$ is a $d \times V$ embedding matrix and $\overline{x}_w$ is a $d \times 1$ embedded vector. The embedding layer

[18] For details of Keras, see the web page (https://keras.io/).

[19] A one-hot vector is a transformation from categorical variables to a binary vector. The concept is almost the same as a dummy variable in econometrics. In the literature of natural language processing (NLP) and computational linguistics, each element in the vector corresponds to a unique word (token) in the corpus vocabulary. The corresponding element takes a value 1 if a particular word appears in the document, and 0, otherwise.

turns one-hot vectors into dense vectors of fixed size. Here, since the number of words differs among sentences, we pad all sentences to the same sequence length. This length is selected from the longest sequence in the training data set. Figure A1 shows an example of the operation in the embedding layer of how 10-dimensional one-hot vectors x_w's are converted into 5-dimensional embedded vectors $\overline{x}_w$'s. To be more specific, the embedding operation applied to a sentence consisting of seven words is shown as $(\overline{x}_{w1}, \overline{x}_{w2}, ..., \overline{x}_{w7})' = (x_{w1}, x_{w2}, ..., x_{w7})'\mathbf{E}'$ in the figure. For illustrative purposes, original Japanese words in parentheses are translated to English words. Padding is represented by a special character <PAD> in the figure.

Here, d is a hyperparameter that defines the size of the output vectors from the embedding layer for each word. For example, setting $d = 100$ implies that we map each word vector into 100-dimensional vector. Generally, neural network models become more computationally efficient when d become smaller. However, if d is too small, embedded vectors are not enough to represent features of words and the model may not perform well in term of accuracy. For the text classification task, Joulin *et al.* (2017) reported that the model will be effective with $d = 10$. Following their suggestion, we also set $d = 10$.

Average pooling layer

The word representations are then averaged into a text representation in the average pooling layer. The average-pooling is a operation that returns the average of input tensors of rank optional n. In case of fastText, it computes the average of word vectors outputted from the embedding layer. In other words, fastText maps each sentence into fixed length vectors using the average of words vectors in each sentence. In comparison with the CNN and the LSTM, fastText achieves high performance in text classification without using word order information. The average pooling operation is defined as

$$x_s = \frac{1}{L} \sum_{w \in s} \overline{x}_w, \tag{A2}$$

where x_s is a $d \times 1$ averaged vector. Figure A2 depicts an example of the operation in the

average pooling layer showing how 5-dimensional word embedded vectors $\overline{x}_w$'s are converted into a 5-dimensional sentence embedded vector x_s. In (A2), the window size is fixed at the length of the longest sequence in the training data set, namely, L.

Output layer

Finally, text representations feed to the output layer. The output layer is given by

$$y = \phi(b + \mathbf{w}x_s), \tag{A3}$$

where y is a scalar output, $\mathbf{w}$ is a $1 \times d$ weight vector, b is a scalar bias, and ϕ is an activation function (in general, y can be a $n \times k$ output vector where n is the number of outputs and k is the number of classes). The choice of the activation function depends on the problem. For example, a linear function is used for the regression problem while a softmax function is used for the multi-class classification problem. As an activation function in an output layer, we use a linear function for estimating sentiment scores and a sigmoid function for binary-classifying topics of assessments of the current and future economic conditions. We used Adam for optimization with a learning rate of 0.001, a mini-batch size of 100, and 200 epochs.[20] We used a random uniform initializer for the embedding layer and the Glorot normal initializer for the output layer.[21]

[20] Adam is an optimization method that extends the stochastic gradient descent (Kingma and Da, 2014). It estimates parameters by taking advantage of the first and second moments of the gradients as following equations.

$$\begin{aligned} w_{t+1} &= w_t - \alpha \frac{\hat{m_t}}{\sqrt{\hat{v_t}} + \epsilon}, \\ \hat{m_t} &= \frac{m_t}{1 - {\beta_1}^t}, \ \hat{v_t} = \frac{v_t}{1 - {\beta_2}^t}, \\ m_t &= \beta_1 m_{t-1} + (1 - \beta_1)\frac{\partial E(w_t)}{\partial w_t}, \ v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\frac{\partial E(w_t)}{\partial w_t})^2, \end{aligned}$$

where t is an epoch, ϵ is a value to avoid division by zero, α is a learning rate, β_1 and β_2 are hyperparameters.

[21] The Glorot normal initializer is a method of initializing the weight tensor with the stochastic gradient descent as in the following equation (Glorot and Bengio, 2010).

$$w_{glorot} \sim N(0, \sqrt{\frac{2}{fan_{in} + fan_{out}}}),$$

The training data is randomly shuffled at each epoch. We divided the training data set into two parts: 90% for training and 10% for validation. We select the weights of the models to minimize the MSE for regression and the cross entropy for classification on a validation data set out of all epochs. In our setting, two loss functions, the MSE (L_{mse}) and the cross entropy (L_{ce}), are defined as

$$\begin{aligned} L_{mse} &= \frac{1}{N}\sum_{n\subseteq N}(y_n - t_n)^2, \\ L_{ce} &= -\frac{1}{N}\sum_{n\subseteq N}(t_n \ln y_n + (1-t_n)\ln(1-y_n)), \end{aligned}$$

where N is a mini-batch size, t_n is a target variable and y_n is a predictor variable.

Acknowledgments

The authors would like to thank Shigenori Shiratsuka, Kumiko Tanaka, Tomohiro Tsuruga and seminar and conference participants at the University of Tokyo, and the 27th Annual Symposium of the Society for Nonlinear Dynamics and Econometrics in Dallas for useful comments and suggestions. Shintani greatly acknowledges the financial support of RCAST at the University of Tokyo, Grant-in-aid for Scientific Research 17H02510 and the Joint Usage and Research Center Programs of IER at Hitotsubashi University.

References

[1] Abe, N., & Tonogi, A. (2010). Micro and macro price dynamics in daily data. *Journal of Monetary Economics*, 57(6), 716–728.

where fan_{in} is the number of input units in the weight tensor and fan_{out} is the number of output units in the weight tensor.

[2] Akerlof, G. A. (2002). Behavioral macroeconomics and macroeconomic behavior. *American Economic Review*, 92(3), 411–433.

[3] Atkeson, A., & Ohanian, L.E. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2–11.

[4] Baker, S.R., Bloom, N., & Davis, S.J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636.

[5] Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

[6] Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2), 151–178.

[7] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*.

[8] Conneau, A., Schwenk, H., Barrault, L., & LeCun, Y. (2016). Very deep convolutional networks for natural language processing. *arXiv:1606.01781*.

[9] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7 (2), 174–196.

[10] Fatum, R., & Hutchison, M.M. (2010). Evaluating foreign exchange market intervention: Self-selection, counterfactuals and average treatment effects. *Journal of International Money and Finance*, 29 (3), 570–584

[11] Faust, J., & Wright, J.H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting*, vol. 2. North Holland.

[12] Fukuda, S., & Keida, M. (2001). Prospects for empirical analysis on inflation forecasts: The predictive power of Phillips curves in Japan. *Bank of Japan Research and Statistics Department Working Paper*, 01-21. (In Japanese)

[13] Gentzkow, M., & Shapiro, J. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1), 35–71.

[14] Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131, 59–95.

[15] Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1), 53–90.

[16] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 249–256.

[17] Hansen, S. & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(1), S114–S133.

[18] Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within The FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2), 801–870.

[19] Ito, T., & Yabu, T. (2007). What prompts Japan to intervene in the forex market? A new approach to a reaction function. *Journal of International Money and Finance*, 26(2), 193-212.

[20] Joulin, A., Grave, E.,Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 427–431.

[21] Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746-1751.

[22] Kingma, D. P., & Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv:1412.6980*.

[23] Larsen, V.H., & Thorsrud, L.A. (2018). Business cycle narratives. *Norges Bank Working Paper* 2018-03.

[24] Lin, Z., Feng, M., Santos, C.N., Yu, M., Xiang, B., Zhou, B., & Bengio, B. (2017). A structured self-attentive sentence embedding. *Proceedings of the 5th International Conference on Learning Representations*.

[25] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.

[26] McCracken, M. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2), 719–752.

[27] Nishizaki, K., Sekine, T., & Ueno, Y. (2014). Chronic deflation in Japan. *Asian Economic Policy Review* , 9, 20–39.

[28] Okazaki, Y., & Tsuruga, T. (2015) On economic and general price analysis using big data: Survey of research works and text analysis of the Economy Watchers Survey. *BOJ Reports & Research Papers*. (in Japanese)

[29] Schwert, G.W. (1989). Tests for unit roots: A Monte Carlo investigation. *Journal of Business and Economic Statistics*, 7, 147–159.

[30] Shapiro, A.H., Sudhof, M., & Wilson, D. (2018). Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper* 2017-01.

[31] Shintani, M. (2005). Nonlinear forecasting analysis using diffusion indexes: an application to Japan. *Journal of Money, Credit and Banking*, 37 (3), 517–538.

[32] Shintani, M., Yabu, T., & Nagakura, D., (2012). Spurious regressions in technical trading. *Journal of Econometrics*, 169(2), 301–309.

[33] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., & Potts, C. (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

[34] Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44, 293–305.

[35] Stock, J. H., & Watson, M. W. (2009). Phillips curve inflation forecasts. In J. Fuhrer, Y. K. Kodrzycki, J. S. Little, and G. P. Olivei (Eds.), *Understanding Inflation and the Implications for Monetary Policy, a Phillips Curve Retrospective*, MIT Press, 99–184.

[36] Suimon, Y., Kinoshita, T., & Yamamoto, Y. (2015) Indexation of business outlook of government and BOJ by artificial intelligence. *NOMURA Macroeconomic Insight*. (in Japanese)

[37] Tallman, E. W., & Zaman, S. (2017). Forecasting inflation: Phillips curve effects on services price measures. *International Journal of Forecasting*, 33(2), 442–457.

[38] Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168.

[39] Tetlock, P.C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467.

[40] Thorsrud, L.A. (2018). Words are the new numbers: A newsy coincident index of business cycles. *Journal of Business & Economic Statistics*, forthcoming.

[41] Watanabe, K., & Watanabe, T. (2018). Why has Japan failed to escape from deflation? *Asian Economic Policy Review* , 13, 23–41.

[42] Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies* 21(4), 1455-508.

[43] West, K. D. (2006). Forecast evaluation. In *Handbook of Economic Forecasting*, vol. 1. North Holland.

[44] Xiao, Y., & Cho, K. (2016). Efficient character-level document classification by combining convolution and recurrent layers. *arXiv:1602.00367*.

[45] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 649–657.

Figures and Tables

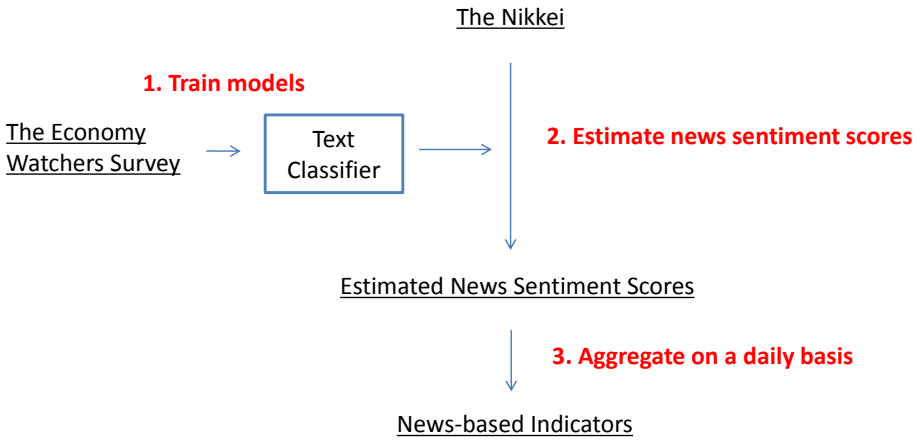


Figure 1: Outline of the procedure

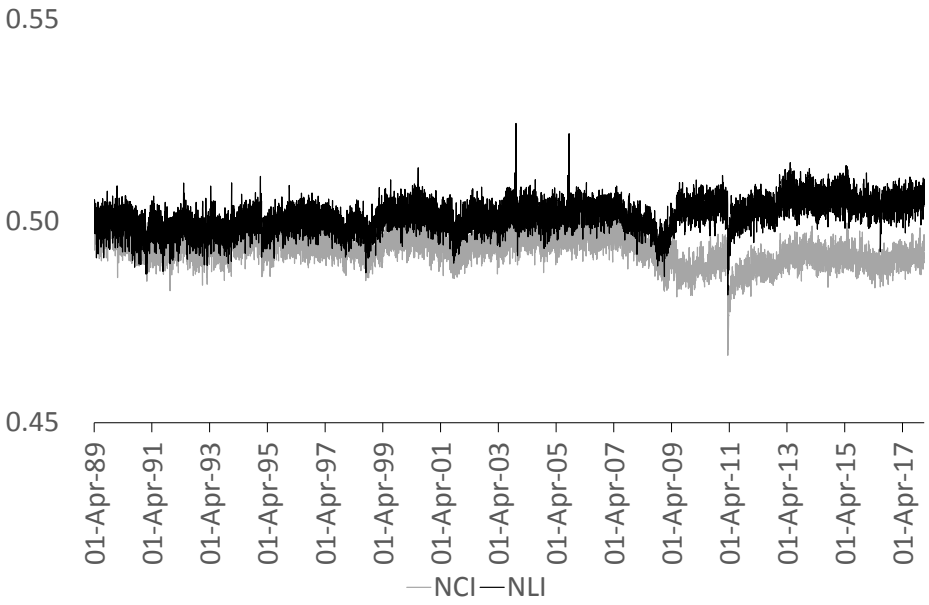


Figure 2: News-based business cycle indexes

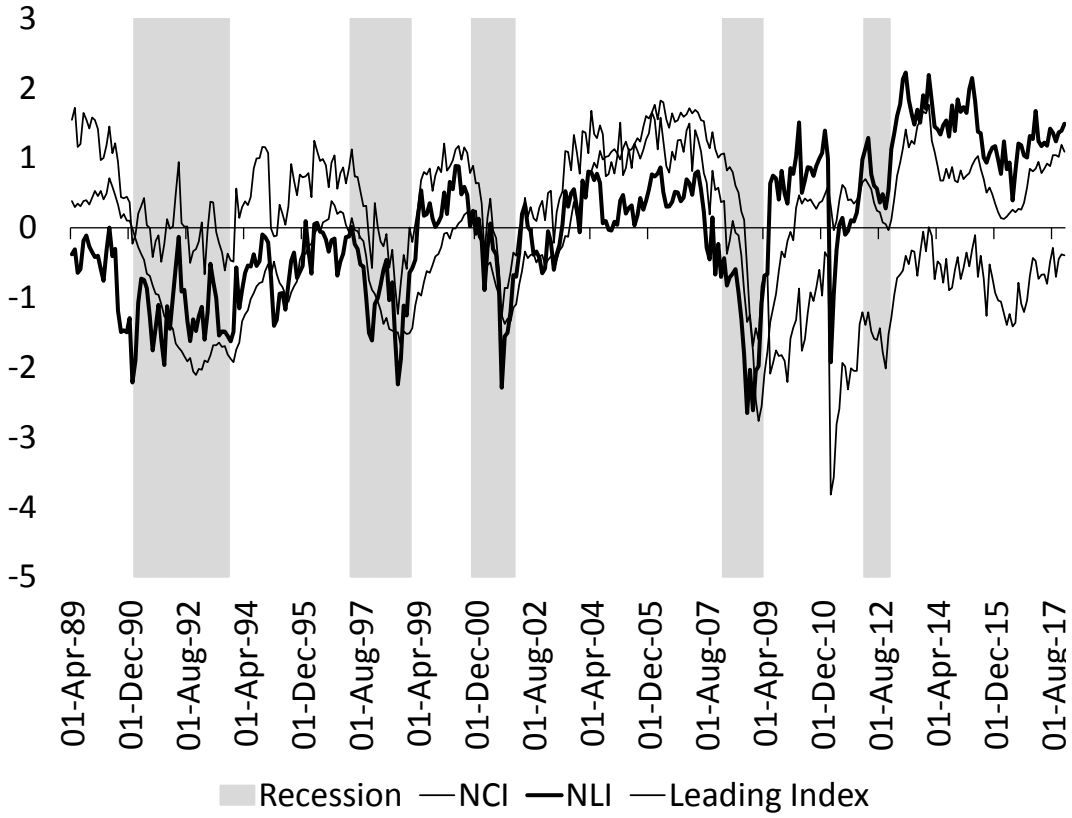


Figure 3: News-based business cycle indexes and official business cycle index

Note: All the series are normalized to have zero mean and unit variance. NCI and NLI are news-based business cycle indexes. The leading index is from the composite indexes of business conditions by ESRI, the Cabinet Office. The shaded area shows the official recession episodes of ESRI, the Cabinet Office.

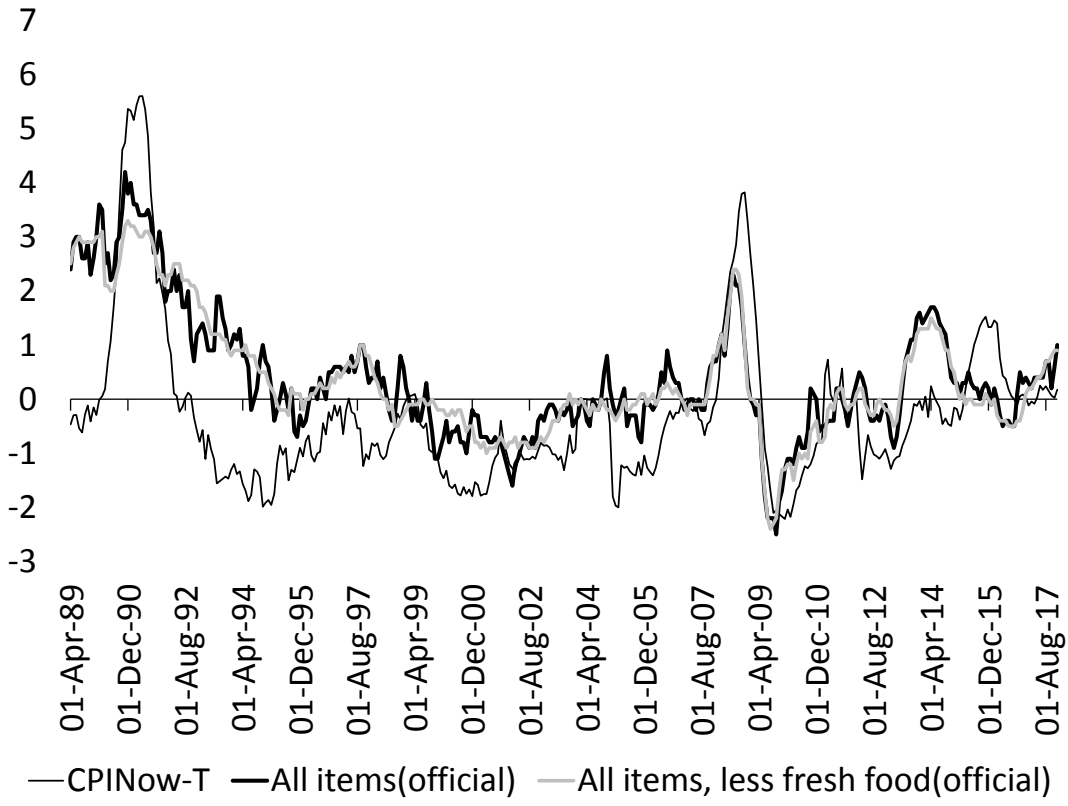
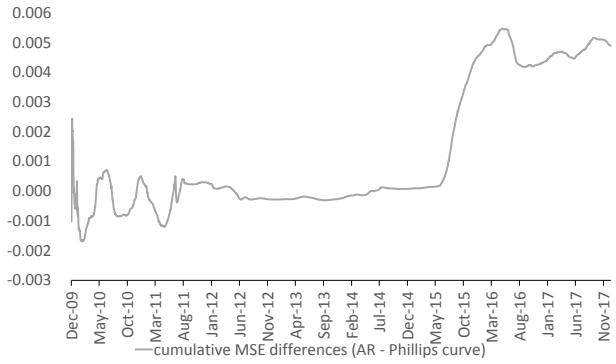
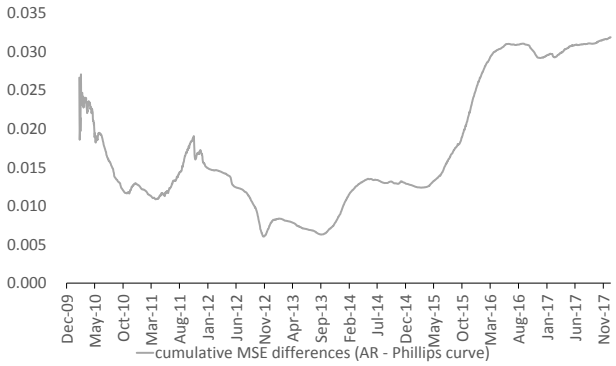


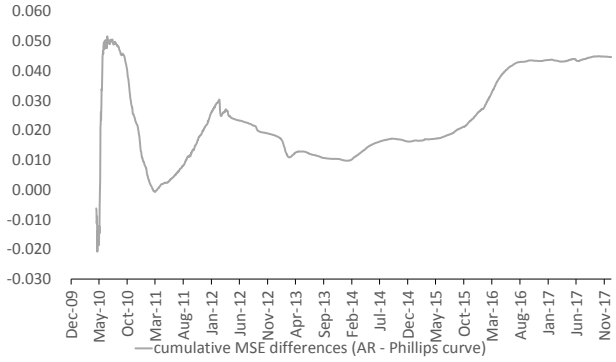
Figure 4: Inflation based on the CPINow-T index and official CPI



(a) h=120



(b) h=240



(c) h=360

Figure 5: Cumulative squared prediction error differences between the AR and the Phillips curve forecasts

Note: MSEs are estimated by the rolling scheme and $P/R = 0.4$.

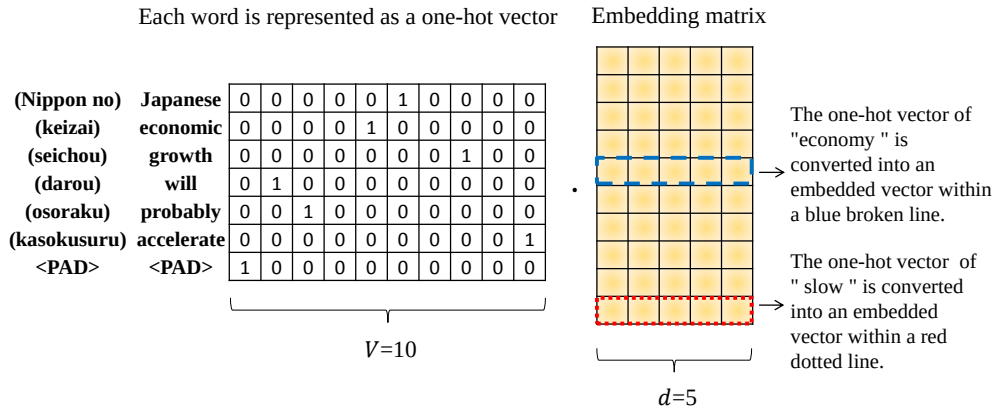


Figure A 1: An example of the operation in the embedding layer

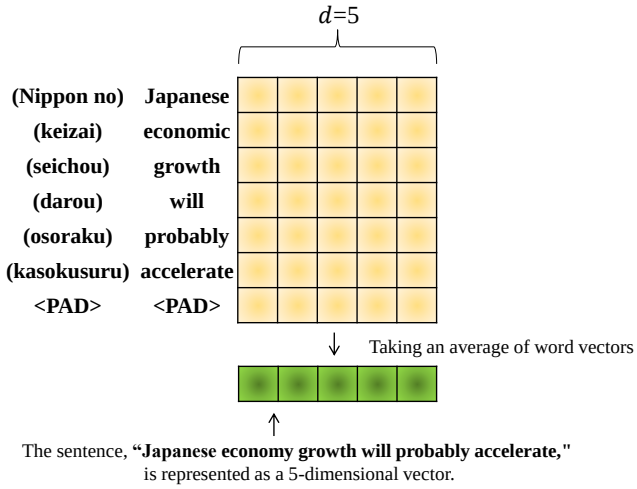


Figure A 2: An example of the operation in the average-pooling layer

Table 1: Summary statistics of *Nikkei*

The number of news articles	3,809,207
The number of sentences	31,797,881
The number of unique words	946,480
The number of total words	812,403,912

Table 2: Example descriptions from *Economy Watchers Survey*

(a) Assessment of current economic conditions

Assessment (score)	Description
Better (1)	The number of switches from dispatched workers to regular workers is increasing.
Slightly Better (0.75)	Financial demand is slightly increasing.
Unchanged (0.5)	It was good at the beginning of the year, but it is now slowing down in latter half of the month.
Slightly Worse (0.25)	It becomes slightly worse due to the effect of the new president of the U.S.
Worse (0)	Our orders decrease after our special busy season.

(b) Assessment of future economic conditions

Assessment (score)	Description
Will Get Better (1)	It will get better thanks to a combination of end of the fiscal year sale and a high season of moving.
Will Get Slightly Better (0.75)	We expect good outcomes as reservations for the new product in March are going well.
Will Remain Unchanged (0.5)	We see no additional orders from existing customers and expect no improvement after the new fiscal year.
Will Get Slightly Worse (0.25)	We believe the economy will improve as foreign political situations become stabilized.
Will Get Worse (0)	It will get worse due to the effect of the new president of the U.S.

Table 3: Breakdown of descriptions in the training data set

(a) Assessment of current economic conditions

Assessment (score)	The number of sentences
Better (1)	2,284
Slightly Better (0.75)	24,927
Unchanged (0.5)	53,352
Slightly Worse (0.25)	25,098
Worse (0)	6,553
Total	112,214
The number of unique words	23,444
The number of total words	3,933,373

(b) Assessment of future economic conditions

Assessment (score)	The number of sentences
Will Get Better (1)	2,488
Will Get Slightly Better (0.75)	29,383
Will Remain Unchanged (0.5)	62,683
Will Get Slightly Worse (0.25)	24,184
Will Get Worse (0)	6,317
Total	125,055
The number of unique words	23,628
The number of total words	4,120,109

Table 4: Relative performance of fastText and other classifiers

(a) Average accuracy of test data set from Joulin *et al.* (2017)

model	accuracy
(multinomial) logistic regression with Bag-of-Words	80.30%
(multinomial) logistic regression with N-grams	81.80%
(multinomial) logistic regression with N-grams and TF-IDF	81.36%
char-CNN	82.81%
char-CRNN	83.31%
VDCNN	84.91%
fastText with unigram	82.09%
fastText with unigram and bigram	84.33%

(b) Accuracy of test data set using the *Economy Watcher Survey*

model	accuracy
LSTM	86.38%
CNN	86.17%
char-CNN	85.90%
Linear Kernel SVM with Bag-of-Words	83.87%
fastText with unigram	84.54%
fastText with unigram and bigram	86.30%

Note: Table 4(a) summarizes fastText's performances from Table 1 of Joulin *et al.* (2017).

Table 5: Summary of learners

	method	training data set	output
learner 1	regression	descriptions of current economy	values ([0,1])
learner 2	regression	descriptions of future economy	values ([0,1])
learner 3	classification	both descriptions	two classes ({*current*, *future*})

Table 6: Summary statistics of news-based business cycle indexes

	NCI	NLI
Mean	49.34	50.15
S.D.	0.40	0.39
Median	49.34	50.16
Min	46.67	48.17
Max	50.92	52.43
Obs	10,474	10,474

Note: The unit is percent. The sample period is April 1, 1989 to December 31, 2017.

Table 7: Correlations of news-based business cycle indexes and other official business cycle indicators

	Economy Watchers Survey		ESRI's CI of business conditions		
	Current DI	Future DI	Leading index	Coincident index	Lagging index
NCI	0.774	0.768	0.202	0.326	0.317
NLI	0.550	0.713	0.682	0.386	-0.145

Note: Current diffusion index (DI) and future DI are from the *Economy Watchers Survey*, Cabinet Office. The leading index, coincident index and Lagging index are from composite indexes (CIs) of business conditions by ESRI, the Cabinet Office.

Table 8: Summary statistics of CPINow

	T-index	S-index
Mean	0.27	1.16
S.D.	1.57	0.81
Median	-0.51	0.81
Min	-5.54	0.15
Max	8.04	3.15
Obs	10,380	36

Note: The sample period is April 1, 1989 to December 31, 2017 for the daily CPINow T-index and is January 2015 to December 2017 for the monthly CPINow S-index.

Table 9: Correlations of CPINow and official CPI inflation

	All items	All items, less fresh food	All items, less fresh food and energy
CPINow-T index	0.649	0.602	–
CPINow-S index	0.591	0.571	0.946

Note: The sample period is April 1989 to December 2017 and January 2015 to December 2017 for the monthly CPINow T-index and CPINow S-index, respectively. The correlation between the CPINow-T index and the CPI inflation for all items, less fresh food and energy, is missing because the CPINow-T index adjusts the effects of the introduction of the consumption tax, while the corresponding adjusted series is not available for the CPI for all items, less fresh food and energy.

Table 10: MSEs of AR forecasts

(a) $P/R = 0.4$

	AR(h)	AR-NS	AR-MF	AR-AIC	AR-BIC
$h = 30$	0.06	0.08	0.06	0.06	0.06
$h = 60$	0.09	0.12	0.09	0.07	0.09
$h = 90$	0.11	0.16	0.11	0.09	0.10
$h = 120$	0.12	0.23	0.12	0.11	0.11
$h = 150$	0.13	0.34	0.13	0.13	0.13
$h = 180$	0.15	0.48	0.15	0.15	0.15
$h = 210$	0.17	0.64	0.17	0.19	0.18
$h = 240$	0.19	0.83	0.19	0.22	0.20
$h = 270$	0.21	1.02	0.21	0.24	0.22
$h = 300$	0.23	1.21	0.23	0.27	0.24
$h = 330$	0.26	1.39	0.25	0.29	0.29
$h = 360$	0.27	1.52	0.27	0.32	0.33

(b) $P/R = 1.0$

	AR(h)	AR-NS	AR-MF	AR-AIC	AR-BIC
$h = 30$	0.10	0.12	0.10	0.07	0.10
$h = 60$	0.16	0.24	0.16	0.11	0.16
$h = 90$	0.23	0.39	0.23	0.15	0.19
$h = 120$	0.25	0.58	0.26	0.20	0.24
$h = 150$	0.28	0.80	0.28	0.25	0.28
$h = 180$	0.33	1.02	0.32	0.30	0.33
$h = 210$	0.39	1.23	0.38	0.36	0.39
$h = 240$	0.44	1.41	0.43	0.41	0.45
$h = 270$	0.49	1.56	0.48	0.46	0.50
$h = 300$	0.54	1.66	0.52	0.51	0.55
$h = 330$	0.58	1.73	0.56	0.55	0.62
$h = 360$	0.61	1.77	0.59	0.59	0.66

Note: MSEs are estimated by the rolling scheme.

Table 11: Full sample coefficient estimates of the Phillips curve model

	β	$\phi_h(1)$
h=30	-1.83	0.97***
h=60	0.4	0.95***
h=90	3.19	0.92***
h=120	5.40*	0.87***
h=150	8.48**	0.83***
h=180	11.00***	0.78***
h=210	13.83***	0.75***
h=240	16.66***	0.72***
h=270	19.56***	0.70***
h=300	22.35***	0.68***
h=330	26.23***	0.66***
h=360	31.11***	0.65***

Note: For HAC standard errors, we follow Schwert (1989) and set the lag length by the integer part of $4 \times ((T - h + 1)/100)^{1/4}$. The sample period is from April 1, 1989 to December 31, 2017. ***, **, * denotes that coefficients are significantly different from zero at the 1%, 5%, and 10% significance levels, respectively.

Table 12: Relative MSEs (percentage deviations) of the AR forecast and the Phillips curve forecast

	Rolling				Recursive			
	$P/R = 0.4$		$P/R = 1.0$		$P/R = 0.4$		$P/R = 1.0$	
h=30	-2.46	(-73.31)	-0.12	(-6.26)	-0.70	(-21.01)	-0.21	(-10.80)
h=60	-2.06	(-61.06)	-0.22	(-11.22)	-0.66	(-19.69)	-0.22	(-11.26)
h=90	-0.09	(-2.55)	-0.04	(-1.83)	0.20***	(5.81)	0.05**	(2.54)
h=120	2.89***	(84.80)	0.47***	(23.92)	1.10***	(32.30)	0.28***	(14.53)
h=150	7.59***	(221.33)	1.38***	(70.25)	2.84***	(82.74)	0.75***	(38.33)
h=180	11.41***	(330.69)	2.33***	(118.26)	4.31***	(124.82)	1.09***	(55.18)
h=210	14.07***	(405.31)	2.89***	(145.50)	5.73***	(164.94)	1.46***	(73.48)
h=240	16.41***	(470.01)	3.38***	(169.24)	7.42***	(212.61)	1.88***	(94.23)
h=270	18.16***	(517.14)	3.72***	(185.38)	9.07***	(258.21)	2.29***	(113.85)
h=300	18.92***	(535.20)	3.85***	(190.70)	10.25***	(290.08)	2.59***	(128.43)
h=330	19.04***	(535.40)	3.85***	(189.35)	11.52***	(323.91)	3.19***	(156.83)
h=360	19.85***	(554.78)	3.87***	(189.20)	15.10***	(421.91)	4.30***	(210.43)

Note: The percentage deviation given by $100 \times (\hat{\sigma}^2_{AR} - \hat{\sigma}^2_{PC})/\hat{\sigma}^2_{PC}$ percent where $\hat{\sigma}^2_{AR}$ is the estimator of the MSE of the AR forecast and $\hat{\sigma}^2_{PC}$ is the estimator of the MSE of the Phillips curve inflation forecast. The numbers in parentheses are out-of-sample Ftest statistics of McCracken (2007). ***, **, * denotes the rejection of the null hypothesis of equal predictability (zero percent deviation of MSEs) using a one-tailed test at the 1%, 5%, and 10% significance levels, respectively.

Table 13: Relative MSEs (percentage deviations) of the AR forecast and the Phillips curve forecast: subsamples

	(1) 1996-2017		(2) 2008-2017		(3) 2013-2017	
h=30	-2.99	(-68.34)	5.92***	(57.10)	-1.63	(-7.94)
h=60	-1.03	(-23.50)	12.10***	(115.79)	-1.32	(-6.33)
h=90	1.86***	(42.13)	15.70***	(148.83)	-0.82	(-3.84)
h=120	5.64***	(127.63)	17.47***	(164.20)	-1.29	(-5.97)
h=150	12.24***	(275.87)	19.48***	(181.40)	-2.28	(-10.31)
h=180	17.84***	(400.54)	19.49***	(179.89)	-2.86	(-12.73)
h=210	19.83***	(443.32)	20.02***	(182.99)	-3.40	(-14.84)
h=240	20.82***	(463.65)	21.77***	(197.02)	-1.08	(-4.60)
h=270	22.50***	(499.31)	20.78***	(186.43)	-0.54	(-2.26)
h=300	23.55***	(520.51)	19.38***	(172.12)	0.24*	(0.97)
h=330	24.58***	(541.20)	18.81***	(165.57)	0.55**	(2.22)
h=360	26.11***	(572.59)	18.47***	(160.83)	0.46**	(1.79)

Note: MSEs are estimated by the rolling scheme and $P/R = 0.4$. See also the note for Table 12.

Table 14: MSEs of VAR forecasts

	Stock Price	FX	WTI	Term Spread	Factor
h=30	0.06	0.06	0.06	0.06	0.06
h=60	0.09	0.09	0.09	0.09	0.09
h=90	0.11	0.11	0.11	0.11	0.11
h=120	0.12	0.12	0.12	0.11	0.12
h=150	0.13	0.13	0.13	0.11	0.12
h=180	0.15	0.15	0.15	0.11	0.14
h=210	0.17	0.17	0.17	0.12	0.16
h=240	0.19	0.19	0.19	0.13	0.18
h=270	0.21	0.21	0.21	0.14	0.21
h=300	0.23	0.23	0.23	0.15	0.23
h=330	0.25	0.25	0.25	0.17	0.25
h=360	0.27	0.27	0.27	0.17	0.25

Note: MSEs are estimated by the rolling scheme and $P/R = 0.4$.

Table 15: Relative MSEs (percentage deviations) of the VAR forecast and the generalized Phillips curve forecast

	Stock Price		FX		WTI		Term Spread		Factor	
h=30	-2.43	(-72.47)	-2.46	(-73.45)	-2.44	(-72.84)	-1.15	(-34.43)	-1.56	(-46.58)
h=60	-2.00	(-59.43)	-2.06	(-60.99)	-2.05	(-60.73)	-1.51	(-44.92)	-0.72	(-21.40)
h=90	-0.01	(-0.38)	-0.08	(-2.25)	-0.08	(-2.30)	-0.92	(-27.23)	0.42***	(12.53)
h=120	3.00***	(87.91)	2.91***	(85.34)	2.90***	(85.02)	-0.10	(-2.84)	1.55***	(45.31)
h=150	7.73***	(225.24)	7.62***	(222.11)	7.59***	(221.39)	0.87***	(25.24)	3.78***	(110.05)
h=180	11.53***	(334.05)	11.44***	(331.62)	11.40***	(330.44)	1.49***	(43.09)	6.48***	(187.81)
h=210	14.18***	(408.42)	14.10***	(406.12)	14.06***	(404.98)	2.12***	(61.17)	9.65***	(277.88)
h=240	16.50***	(472.67)	16.43***	(470.60)	16.39***	(469.54)	2.86***	(81.93)	13.17***	(377.10)
h=270	18.25***	(519.62)	18.19***	(517.75)	18.14***	(516.53)	3.35***	(95.43)	15.89***	(452.41)
h=300	18.99***	(537.32)	18.93***	(535.63)	18.89***	(534.48)	3.55***	(100.40)	16.58***	(469.07)
h=330	19.11***	(537.47)	19.06***	(535.84)	19.01***	(534.64)	3.48***	(97.85)	15.60***	(438.71)
h=360	19.92***	(556.66)	19.86***	(555.09)	19.82***	(554.10)	3.53***	(98.76)	14.10***	(394.13)

Note: The percentage deviation given by $100 \times (\hat{\sigma}^2_{VAR} - \hat{\sigma}^2_{EPC})/\hat{\sigma}^2_{EPC}$ percent where $\hat{\sigma}^2_{VAR}$ is the estimator of the MSE of the VAR forecast and $\hat{\sigma}^2_{EPC}$ is the estimator of the MSE of the extended Phillips curve inflation forecast. MSEs are estimated by the rolling scheme and $P/R = 0.4$. See also note for Table 12.