Real-Time Macro Information and Bond Return Predictability: A Weighted Group Deep Learning Approach*

Yinghua Fan Guanhao Feng Andras Fulop Junye Li

First version: October 2019; This version: April 2022

Abstract

This paper proposes a weighted group neural network model and reexamines whether treasury bond returns are predictable when real-time, instead of fully-revised, macro information is used. Two types of real-time macro information are taken into account: real-time macro vintage data and news-based topic attention. We find that news contains rich information on future bond returns beyond traditional macro variables. When both types of real-time data are used as predictors, our proposed model can help find significant statistical evidence for forecasting *non-overlapping* short-term bond returns and for forecasting *overlapping* bond returns with maturities of 2 to 10 years. Furthermore, the statistical evidence of overlapping bond return predictability can be translated into investors' economic gains for long-term bonds when investors are allowed to leverage their investments.

Keywords: Deep Learning, Machine Learning, Bond Return Predictability, Real-Time Macro Data, News Topic Attention.

JEL Classification: C45, C53, G11, G12, G17

*We appreciate insightful comments from Dashan Huang, Bryan Kelly, and Junbo Wang. We are also grateful for helpful comments from seminar and conference participants at ESSEC Business School. Fan (E-mail: yinghufan2-c@cityu.edu.hk) and Feng (E-mail: gavin.feng@cityu.edu.hk) are at the City University of Hong Kong, Fulop (E-mail: fulop@essec.fr) is at ESSEC Business School, and Li (E-mail: li_junye@fudan.edu.cn) is at Fudan University.

1. Introduction

The expectations hypothesis (EH) of the term structure of interest rates asserts that the long-term rate equals the average expected future short rates plus a constant risk premium. A standard way to test the EH is to examine whether bond risk premia are time-varying or excess bond returns are predictable. One strand of the literature argues that macroeconomic variables have strong predictive power for future excess bond returns beyond information in yield curve. For example, several studies have shown that individual macro variables or macro factors extracted from a large panel of macro variables can predict excess bond returns (Ludvigson and Ng, 2009, 2011; Cooper and Priestley, 2009; Cieslak and Povala, 2015; Gargano, Pettenuzzo, and Timmermann, 2019; Bianchi, Büchner, and Tamoni, 2021). Besides, Wright (2011), Joslin, Priebsch, and Singleton (2014), and Li, Sarno, and Zinna (2021) apply unspanned macro term structure models and find bond risk premia are time-varying along with the real economy and inflation.

Bond risk premia should be conditioned on information available to investors in real-time. If any macro information affects bond prices, that should be the real-time one when bond prices are determined, instead of that revised with future information. However, almost all of the above studies use fully-revised macro variables in their empirical analysis, not available to bond investors in real-time due to data revisions and publication delays. Ghysels, Horan, and Moench (2018) find in the standard linear predictive models that the predictive power of macro variables for future overlapping excess bond returns is mostly from data revisions and argue the real-time macro data should be used in bond return predictability to avoid any hindsight problem. Wan, Fulop, and Li (2021) use different types of real-time macro data to implement empirical analysis based on linear predictive models with and without stochastic volatility. They find no statistical or economic evidence for forecasting *non-overlapping* one-month holding period excess bond returns whenever real-time, instead of fully-revised, macro factors are used as predictors. In contrast, Huang et al. (2021) show annual holding period *overlapping* excess bond returns can be predicted by real-time macro vintage data when a data-driven scaled

sufficient forecasting method is used.

A recent strand of the literature on asset return predictability has emphasized the usefulness of modern machine learning techniques that allow for flexible nonlinear predictive relationships. In this paper, instead of relying on standard linear predictive models, we focus, in particular, on deep learning predictive methods. Deep learning is a form of nonlinear supervised machine learning that employs deep neural networks for implementing prediction through a series of nonlinear transformations of a large number of predictors.[1] Gu, Kelly, and Xiu (2020) reinvestigate equity return predictability by using various deep/machine learning models and confirm the predictive power of deep learning as the best. Similarly, using a large panel of fully-revised macro variables as predictors, Bianchi, Büchner, and Tamoni (2021) and its Corrigendum (Bianchi et al., 2021) find that the deep learning models can generate strong statistical and economic evidence of predictability for excess bond returns. Feng, Polson, and Xu (2019) show the characteristics-sorted factor models can be trained as a deep neural network with the economic-driven loss function pricing errors. Chen, Pelger, and Zhu (2020) provide a deep learning model to generate the stochastic discount factor.

However, the panel of macroeconomic variables that can potentially be used for bond return forecasting is very large, and most of those variables are highly correlated. Simple off-the-shelf application of deep learning may hardly make significant improvement in forecasting. Motivated by group features of macroeconomic variables, in this paper, we develop neural networks with a group structure that compress similar information (i.e., variables with high correlations) into the same categories and hence can help alleviate issues related to highly correlated predictor variables. Furthermore, we aim to forecast bond returns with different maturities using the same factors trained from the same neutral network. Having noticed that bond returns with different maturities vary with different extents over time, we design a weighted group neutral network (WGNN) that attaches weight to bond returns of each maturity based on its variations in the loss

[1] In the paper, we distinguish between deep learning and machine learning. We use machine learning to refer to algorithms other than the deep neural network.

function.

We revisit the question of bond return predictability and argue that one should be particularly cautious regarding what information to use when using flexible machine learning methods. If the information includes variables unavailable to bond investors in real-time, these methods may overestimate the extent of predictability. Hence, we take a conservative approach and ask whether such nonlinear predictive methods are still helpful when we only condition on macro information available to investors in real-time. For this purpose, we conduct our empirical analyses based on two types of real-time macroeconomic information. The first type is the real-time vintage data of a large panel of macroeconomic variables that are frequently used in literature. We rely on the Archival Federal Reserve Economic Data (ALFRED) and construct a balanced panel of 125 monthly macroeconomic times series. Following the literature, we group those variables into eight economic categories.

The second type is a large panel of real-time news-based macroeconomic data. We take a panel of 180 monthly news topic attention recently introduced by Bybee, Kelly, Manela, and Xiu (2021), who propose an approach to measuring the state of the economy via textual analysis of the full-text content of about 800,000 Wall Street Journal (WSJ) articles. They show that this text-based news attention accurately tracks a wide range of economic activity measures and has incremental forecasting power for macroeconomic outcomes. In addition, Kelly, Manela, and Moreira (2021) find that WSJ news significantly improves macroeconomic predictions, and Ellingsen, Larsen, and Thorsrud (2021) use news articles from the Dow Jones Newswires Archive for macroeconomic forecasting and find that news contains information that is not captured by complex economic indicators. Following Bybee, Kelly, Manela, and Xiu (2021), we divide news topic attention into two groups, Economy and Politics/Cultures.

We construct monthly frequency excess bond returns with maturities of two to ten years using the yield curve dataset constructed by Liu and Wu (2021). We consider two types of bond returns: the first is the commonly used one-year holding-period *overlapping*

returns, and the other is the one-month holding-period *non-overlapping* returns. Putting all data sources together and given the availability of news-based data, our final sample spans the period from January 1984 and June 2018. We set the out-of-sample period ranging from January 2000 to June 2018, which is particularly interesting and may be challenging for out-of-sample tests, as the US economy experiences the 2008 financial crisis, the interest rates enter an era of zero lower bound, and the US treasury bonds seem to become macro hedge assets (Li, Sarno, and Zinna, 2021). We then investigate whether and in which deep/machine learning models any out-of-sample statistical and/or economic evidence of predictability exists for non-overlapping and overlapping excess bond returns.

When forecasting *non-overlapping* excess bond returns, we find that our WGNN can help find statistically significant out-of-sample evidence of bond return predictability for 2- and 3-year bonds using real-time vintage macro data, and such predictability further improves when news-based topic attention data are combined with vintage data, suggesting that news contains important information that is not fully reflected by standard macroeconomic indicators. We also find that the penalized regressions, such as Lasso and Elastic net, also seem to help find some out-of-sample statistical evidence for forecasting 2- and 3-year non-overlapping excess bond returns when macro vintage data and news-based data are used together. However, all evidence we find is much weaker than that found when fully-revised macro data are used as in Gargano, Pettenuzzo, and Timmermann (2019).

When we move to forecast *overlapping* excess bond returns, we find different results. First, when real-time macro vintage data alone are used, our WGNN can generate statistically significant out-of-sample R^2s for bond returns of all maturities that range from 1.69% for 5-year bond returns to 8.21% for 2-year bond returns. Furthermore, we also find that some standard machine learning models, PCA and Elastic Net, in particular, work well in forecasting bond returns with maturities larger than 5 years. Second, when news-based information is included together with macro vintage data, the performance of

our WGNN improves a lot for bond returns of all maturities, with the out-of-sample R^2s ranging from 4.99% for 5-year bond returns to 11.54% for 2-year bond returns. However, nearly all machine learning models lose their ability to generate positive out-of-sample R^2s. Given that when both types of real-time data are combined, the number of predictors becomes very large (in total, 305 variables), and that the variables extracted from the news are highly correlated with those standard macro variables, negligible or negative out-of-sample R^2s from machine learning models just suggest that they can not efficiently handle such a high-dimensional data with high correlations. Third, we further find that in our WGNN, the groups of Labor, Interest, and Housing, and two news-based groups play important roles in forecasting overlapping bond returns. However, the above evidence of overlapping bond return predictability is still weaker than the findings of the literature using fully-revised macro data (see, e.g., Ludvigson and Ng, 2009, 2011; Cooper and Priestley, 2009; Cieslak and Povala, 2015; Gargano, Pettenuzzo, and Timmermann, 2019; Bianchi, Büchner, and Tamoni, 2021).

We then explore whether the above statistical predictability can be translated into a mean-variance investor's economic gains. For this purpose, we consider two cases: the first enables investors to take a full short position but prevents extreme investments (Goyal and Welch, 2008; Campbell and Thompson, 2008; Ferreira and Santa-Clara, 2011; Thorton and Valente, 2012; Sarno, Schneider, and Wagner, 2016), and the second allows investors for leveraging their investments similar to Huang et al. (2021) and Gargano, Pettenuzzo, and Timmermann (2019), such that the portfolio weights on risky bonds range between -1 and 8. Unlike in the equity market, investors could take extreme positions in the bond market, facilitated by repo agreements. Given that the statistical evidence for forecasting non-overlapping bond returns is weak and only exists in short-term bonds, it clearly cannot translate into investors' economic gains no matter which case is considered.

However, for the overlapping bond returns, we find that the statistical predictability from our WGNN based on both macro vintage data and news attention can be translated

into investors' economic gains for long-term bonds when they are allowed to leverage their investments. For example, when we assume a risk-aversion coefficient of 5, the utility gains are 111 bps in 6-year bond, 135 bps in 7-year bond, 134 bps in 8-year bond, and 86 bps in 9- and 10-year bonds. However, when investors are prevented from extreme investments, the utility gains are negligible or even negative, a result that is very different from those obtained using the fully-revised macro data (see, e.g. Gargano, Pettenuzzo, and Timmermann, 2019; Bianchi, Büchner, and Tamoni, 2021).

When both macro vintage and news-based data are used in our WGNN, we may wonder why utility gains for short-term bonds are so small or negative even when investors are allowed to take leverage, given that the out-of-sample R^2s in those bonds are very high. We find that the vast majority of unrestricted portfolio weights on 2-, 3-, and 4-year bonds are huge, way above the upper bound, and this is particularly striking for the 2-year bond. High portfolio weights on short-term bonds may result from the fact that, unlike stocks and long-term bonds, variations of the short-term bonds are much smaller, and those bonds are always regarded as safe assets or macro hedge assets (He, Krishnamurthy, and Milbradt, 2016; Li, Sarno, and Zinna, 2021), whose correlations with stocks are negative in our out-of-sample period.

Our work makes three main contributions to the literature. First and foremost, we note that using real-time versus fully-revised macro information and overlapping versus non-overlapping returns can lead to starkly contrasting results for flexible nonlinear predictive tools. We argue that because risk premia should be conditioned on investors' information set available when bond prices are determined, a conservative approach that uses real-time macro data should be adopted in bond return predictability studies. Second, we propose a weighted group neural network model based on economic motivations, adding to the recent literature on empirical asset pricing with deep/machine learning for forecasting financial asset returns. Gu, Kelly, and Xiu (2020) forecast equity returns with various deep/machine learning algorithms and find deep learning outperforms. Bianchi, Büchner, and Tamoni (2021) also use multiple deep/machine learning models for forecast-

ing excess bond returns; however, they use the fully-revised macro data and ignore issues related to macro data revisions and publication delay. Huang and Shi (2022) propose a two-step Lasso approach that exploits a similar idea of grouping. Huang et al. (2021) aim at the same research questions as we do and propose a data-driven scaled sufficient forecasting method. Third, our paper clearly shows that news contains rich information on future bond returns that is not captured by standard macroeconomic variables and indicators.

The remainder of the paper is organized as follows. Section 2 presents excess bond returns and introduces our weighted group neutral network models. Section 3 discusses statistical and economic evaluation of out-of-sample return predictability. Section 4 presents the data and summary statistics. Section 5 provides main empirical results. Finally, Section 6 concludes the paper.

2. Bond Return Predictability and Weighted Group Deep Learning

2.1. Bond Return Predictability

Following the existing literature, we define the log-yield of an n-year bond as

$$y_t^{(n)} \equiv -\frac{1}{n} p_t^{(n)}, \tag{1}$$

where $p_t^{(n)} = \ln P_t^{(n)}$ and $P_t^{(n)}$ is the nominal price of an n-year zero-coupon bond at time t. The log forward rate at time t between time $t+n-m/12$ and $t+n$ is

$$f_t^{(n)} \equiv p_t^{(n-m/12)} - p_t^{(n)}, \tag{2}$$

where m is the holding period in months. The corresponding forward spread is given by

$$fs_t^{(n)} = f_t^{(n)} - \frac{m}{12} \times y_t^{(m/12)}. \tag{3}$$

The excess return of an n-year bond is computed as the holding period return from

buying an n-year bond at time t and selling it m-periods later in excess of the yield on an m-period risk-free rate at time t:

$$rx_{t+m}^{(n)} = p_{t+m}^{(n-m/12)} - p_t^{(n)} - \tfrac{m}{12} \times y_t^{(m/12)}, \tag{4}$$

where $y_t^{(m/12)}$ is the annualized m-period risk-free rate. In this paper, we construct monthly frequency excess bond returns and consider two types of excess returns: the first is the commonly used one-year holding period *overlapping* excess bond returns, that is, $m = 12$ months, and the other is the one-month holding period *non-overlapping* excess bond returns, that is, $m = 1$ month. We consider bond maturity n that can take the values of 2 to 10 years.

The standard approach to investigating bond return predictability usually takes a linear predictive model of the form

$$rx_{t+m}^{(n)} = \alpha^{(n)} + \beta^{(n)'}\mathbf{X}_t + \epsilon_{t+m}^{(n)}, \tag{5}$$

where $\mathbf{X}_t$ is a set of the pre-determined predictors (a $v \times 1$ vector, the number of predictors is v), $\beta^{(n)}$ is a vector of corresponding coefficients, and ϵ_t is a mean-zero error term, whose variance can be either constant or stochastic. In most studies of bond return predictability, the predictors $\mathbf{X}$ include either the yield-curve-based factors (see, e.g., Fama and Bliss, 1987; Campbell and Shiller, 1991; Cochrane and Piazzesi, 2005) or macro factors extracted from a large panel of macro data (see, e.g., Ludvigson and Ng, 2009, 2011; Cooper and Priestley, 2009; Cieslak and Povala, 2015; Gargano, Pettenuzzo, and Timmermann, 2019; Wan, Fulop, and Li, 2021).

2.2. Group Deep Learning

Differently, deep learning is a form of *nonlinear* supervised machine learning that employs a deep neural network for predicting the output variable, rx, via a large number of predictor variables, $\mathbf{X}$. In this paper, we consider *two types* of real-time macro information

as our predictors, the first includes the real-time version of a large panel of macroeconomic variables commonly used in the literature, and the other contains a large panel of real-time news-based macro-related variables that are recently constructed by Bybee, Kelly, Manela, and Xiu (2021).

Given that the number of predictor variables are quite large and most of them are highly correlated, the off-the-shelf application of deep learning would hardly result in much improvement in bond return forecasting. Motivated by group features of those macro variables, we develop neural networks with the group structure that compress similar information into categories and hence can help alleviate issues related to highly correlated predictor variables.

2.2.1. Group Structure in Neural Networks

We develop a neural network with group structure (GNN) that can categorize information contained in predictors. There are three layers in GNN. At time t, the first layer is the information that is available at that time as input, $\mathbf{X}_t^{(0)} = [\mathbf{X}_1, \cdots, \mathbf{X}_t]'$, which is a $t \times v$ matrix and can be divided into k groups, $\mathbf{X}_t^{(0)} = [\mathbf{X}_{t,1}^{(0)}, \cdots, \mathbf{X}_{t,k}^{(0)}]$, based on some economic restrictions, where t is the length of sample up to time t and v is the total number of predictors. Each component $\mathbf{X}_{t,i}^{(0)}$ for group i, $i = 1, 2, ..., k$, is a $t \times v_i$ matrix, where v_i represents the number of variables in group i, and $\sum_{i=1}^{k} v_i = v$. Furthermore, we assume that groups 1 to k_1 are from the first type of data, and groups $k_1 + 1$ to k from the second type of data.

The second layer is for the nonlinear transformation and dimension reduction for each group $\mathbf{X}_i^{(0)}$, which is conducted as follows,

$$\underset{t\times 1}{\mathbf{X}_{t,i}^{(1)}} = g\big(\underset{t\times 1}{\mathbf{b}_{t,i}^{(1)}} + \underset{t\times v_i}{\mathbf{X}_{t,i}^{(0)}}\underset{v_i\times 1}{\mathbf{W}_{t,i}^{(1)}}\big), \tag{6}$$

where $g(\cdot)$ is a nonlinear activation function, $\mathbf{b}_{t,i}^{(1)}$ is a $t \times 1$ bias vector containing the same bias components for each group, and $\mathbf{W}_{t,i}^{(1)}$ is a $v_i \times 1$ vector of weights. Commonly used activation functions include sigmoidal (e.g., $1/(1 + exp(-x))$ or $tanh(x)$), Heaviside gate

functions (e.g., $\mathbb{I}(x > 0)$), or rectified linear units (ReLU). In the paper, we use $ReLU$ as the activation function, i.e.,

$$g(x) = ReLU(x) = max(0, x). \tag{7}$$

$ReLU$ activation functions is frequently used in the feed-forward networks. The advantage of $ReLU$ is that it allows for faster and more effective training of deep neural architectures on large and complex datasets. Each neuron at this layer is a $t \times 1$ vector $\mathbf{X}_{t,i}^{(1)}$, which is regarded as a latent factor representing for information contained in group i. All outputs at this layer are then stacked together into a $t \times k$ matrix $\mathbf{X}_t^{(1)} = [\mathbf{X}_{t,1}^{(1)} \cdot \cdot, \mathbf{X}_{t,k}^{(1)}]$.

The third layer is a linear combination of k factors, $\mathbf{X}_t^{(1)}$, derived from the previous hidden layer for forecasting excess bond returns, $\mathbf{rx}_{t+m} = [rx_{m+1}, \cdots, rx_{t+m}]'$, as follows,

$$\underset{t\times N}{\widehat{\mathbf{rx}}_{t+m}} = \underset{t\times 1}{\mathbf{b}_t^{(2)}}\mathbf{1}' + \underset{t\times k}{\mathbf{X}_t^{(1)}}\underset{k\times N}{\mathbf{W}_t^{(2)}}, \tag{8}$$

where $\widehat{\mathbf{rx}}_{t+m}$ represents the predicted excess returns matrix for bonds with N different maturities, $\mathbf{1}$ is a $N \times 1$ vector of ones, and $\mathbf{W}_t^{(2)}$ contains the coefficients used for the combination.

2.2.2. Penalties and Loss Function

With enhanced flexibility, however, neural network models come with a high propensity for overfitting. The most common device for guarding against overfitting is to append a penalty to the objective function in order to favor more parsimonious specifications. Thus, to reduce possibility that the model overfits noises, while preserving its fit of signals, we add L_2-norm regularizers to the weights in $\mathbf{W}_{t,i}^{(1)}$ for each group i in Equation (6) and $\mathbf{W}_t^{(2)}$ in Equation (8). We define the penalty function, $\Phi(\mathbf{W}_t)$, as follows,

$$\Phi(\mathbf{W}_t; \lambda_t) = \lambda_{t,1} \sum_{i=1}^{k_1} ||\mathbf{W}_{t,i}^{(1)}||_2 + \lambda_{t,2} \sum_{i=k_1+1}^{k} ||\mathbf{W}_{t,i}^{(1)}||_2 + \lambda_{t,3}||\mathbf{W}_t^{(2)}||_2, \tag{9}$$

where $\lambda_{t,1}$ and $\lambda_{t,2}$ are two penalty parameters for two types of macro predictor variables, respectively, in the first layer, and $\lambda_{t,3}$ is a penalty parameter for the linear combination of groups in last layer. The distinction of penalties for two types of predictors enables us to easily shut down one of them. In our empirical study, we performs the validation procedure to select $\lambda_t = [\lambda_{t,1}, \lambda_{t,2}, \lambda_{t,3}]$.

Given the grouped structure and penalties, the objective function, $\mathcal{L}(\mathbf{W}_t, \mathbf{b}_t; \mathbf{X}_t)$, which minimizes the mean squared errors (MSE) of GNN, becomes

$$\mathcal{L}(\mathbf{W}_t, \mathbf{b}_t; \mathbf{X}_t) = \frac{1}{N \times t} \sum_{s=1}^{t} \sum_{n=2}^{N+1} \left(\widehat{rx}_{s+m}^{(n)} - rx_{s+m}^{(n)} \right)^2 + \Phi(\mathbf{W}_t, \lambda_t). \tag{10}$$

By minimizing the objective function of Equation (10), we estimate the parameters $\mathbf{W}_t$ and $\mathbf{b}_t$ for the multivariate outcome $\mathbf{rx}_{t+m}$. While we use the *ReLU* activation function at the intermediate layers in GNN, a simple linear transformation function is adopted at the output layer to preserve linearity because we want to exploit interpretable contributions of different groups to return predictability of bonds with different maturities. One can interpret this architectural choice as an extension of Cochrane and Piazzesi (2005) insofar as excess bond returns at various maturities are linked to the same common factors.

2.2.3. Weighted GNN

Bond returns with different maturities vary differently over time. Basically, the longer the maturity is, the larger return variation should be. In our GNN, the bond returns with different maturities are simultaneously trained through the neural network. However, given differences of their variations, it may lead to unbalanced training, that is, training of bond returns with large standard deviations may be attached to high weights. Motivated by the philosophy of weighted least squares (WLS), we change the weights of bond returns with different maturities in Equation (10) by normalizing $rx_{t+m}^{(n)}$ with the estimated standard deviation, $\hat{\sigma}_t^{(n)}$, of the training data that is available at time t.

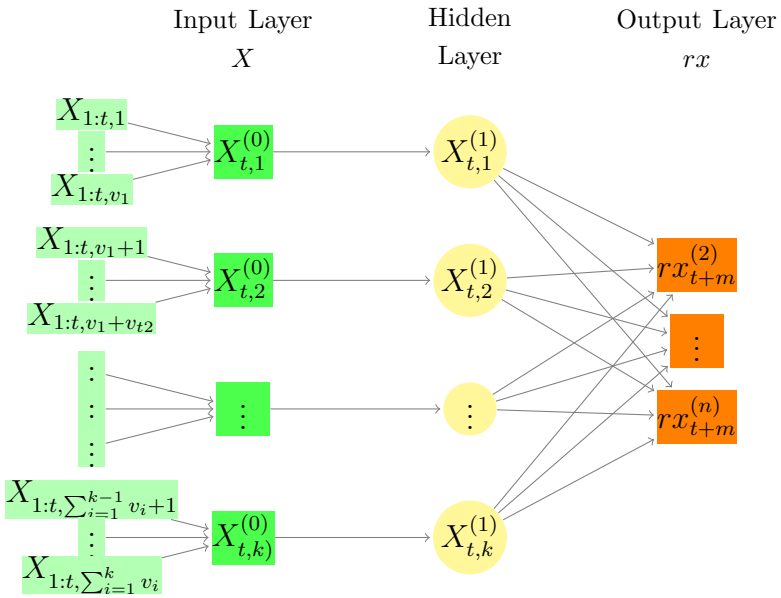

Figure 1: **(Weighted) Group Neural Network with k Groups**

Therefore, the loss function now becomes

$$\mathcal{L}(\mathbf{W}_t, \mathbf{b}_t; \mathbf{X}_t) = \frac{1}{N \times t} \sum_{s=1}^{t} \sum_{n=2}^{N+1} \left(\widehat{rx}_{s+m}^{(n)} - rx_{s+m}^{(n)} / \hat{\sigma}_t^{(n)} \right)^2 + \Phi(\mathbf{W}_t, \lambda_t). \tag{11}$$

In the rest of this paper, we call it weighted group neural network (WGNN). GNN and WGNN are illustrated in Figure 1.

2.3. Machine Learning

For the purpose of comparison, we include in our empirical analysis major machine learning algorithms. Specifically, we consider the following machine learning models: principal component analysis (PCA), partial least squares (PLS), Ridge regression, Lasso regression, Elastic Net, AutoEncoder, and Random Forest, for which we follow the standard setups as in Hastie, Tibshirani, and Friedman (2009) and Goodfellow, Bengio, and Courville (2016). The first five models are linear, whereas the last two are nonlinear. For financial applications of some of these methods, see, for example, Gu, Kelly, and Xiu (2020), Gu, Kelly, and Xiu (2021), and Bianchi, Büchner, and Tamoni (2021), among others.

Unlike deep learning, for these machine learning methods, we perform time-series modeling by putting all predictors together, i.e., forecasting each-maturity excess bond returns separately using all available predictors, and employ the same forward-validation scheme as in Figure 2 to select their respective tuning parameters.

3. Assessing Out-of-Sample Performance

3.1. Model Selection

We adopt an adaptive model selection scheme in training deep/machine learning models to accommodate potential business cycle features in macro and/or yield data. Relying on a large cross-section of equity return data, Gu, Kelly, and Xiu (2020) employ a forward validation scheme, in which they use the recent five-year data for model validation and selection. However, for treasury bond returns, the model selection could be sensitive in different years and may result in highly noisy model selection results from year to year. Therefore, to ensure model selection stability, we introduce a deterministic three-fold cross-validation scheme as presented in Figure 2, which does not randomly split the training samples.

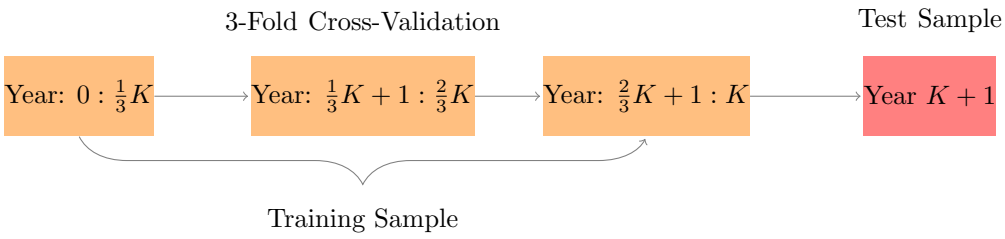

Figure 2: **Out-of-Sample Design**

Specifically, to predict bond returns in Year $K + 1$, we first split the past data up to the end of Year K into three consecutive intervals as three folds (see Figure 2). We then train each model using any two of three folds and validate using the remaining fold, which results in three sets of validation results. Finally, we determine the best tuning parameters according to the average of these three sets of validation errors and refit the model using

all the past data. Our model selection procedure differs from regular randomized cross-validation, designed for independent observations. The advantage of our deterministic three-fold cross-validation scheme is to allow for a high degree of time-series dependence in our data.

For deep learning, we try multiple neural network architectures, different regularization levels, and different learning rates for deep learning model training. For machine learning, we implement a large number of tuning parameters for the variable selection. We update the model selection on an annual basis such that we have additional 12-month data for model training and validation. There is a trade-off between monthly and annual updates. Though the annual update might lose a few months of new data, it provides a relatively stable model interpretation.

3.2. Statistical Evaluation

At each time t in the out-of-sample period, denote the m-months-ahead forecasted value of the n-year excess bond return as $\widehat{rx}_{t+m}^{(n)}$ and define the sum of squared forecast errors (SSE) from the initial time of the out-of-sample period, t_0, to time t as

$$\widehat{SSE}(t) = \sum_{s=t_0}^{t-m} (rx_{s+m}^{(n)} - \widehat{rx}_{s+m}^{(n)})^2. \tag{12}$$

Furthermore, denote the forecast from using the historical average as $\overline{rx}_{t+m}^{(n)} = \frac{1}{t}\sum_{s=1}^{t} rx_s^{(n)}$. Then, its SSE is given by

$$\overline{SSE}(t) = \sum_{s=t_0}^{t-m} (rx_{s+m}^{(n)} - \overline{rx}_{s+m}^{(n)})^2. \tag{13}$$

A natural measure of predictive performance of a model is the out-of-sample R-squared, R_{OS}^2, proposed by Campbell and Thompson (2008). The R_{OS}^2 statistic is computed as

$$R_{OS}^2 = 1 - \frac{\widehat{SSE}(T)}{\overline{SSE}(T)}, \tag{14}$$

where T denotes the end of the out-of-sample period. R^2_{OS} is analogous to the standard R^2 and measures the proportional reduction in prediction errors of the forecast from the predictive model relative to the historical average forecast. We also define an out-of-sample R^2 to measure the overall performance of a model in forecasting bond returns of all maturities as follows

$$R^2_{All,OS} = 1 - \frac{\sum_{s=t_0}^{t-m}\sum_{n=2}^{10}(rx_{s+m}^{(n)} - \widehat{rx}_{s+m}^{(n)})^2}{\sum_{s=t_0}^{t-m}\sum_{n=2}^{10}(rx_{s+m}^{(n)} - \overline{rx}_{s+m}^{(n)})^2}. \tag{15}$$

When $R^2_{OS} > 0$ ($R^2_{All,OS} > 0$), the predictive model clearly statistically outperforms the historical average. We can further test whether this outperformance is statistically significant using the statistic developed by Clark and West (2007). The Clark-West statistic adjusts the well-known Diebold and Mariano (1995) and West (1996) statistic and generates asymptotically valid inference when comparing nested model forecasts. Clark and West (2007) show this statistic performs well in terms of power and size.

3.3. Economic Evaluation

In evaluating economic evidence of bond return predictability in a predictive model, we consider a mean-variance investor who constructs a portfolio consisting of a risk-free zero-coupon bond and a risky bond with maturity of n-year and maximizes her expected utility on the next-period portfolio value.

At each time t in the out-of-sample period, the optimal weight on the risky bond for the mean-variance investor is given by

$$w_t^{(n)} = \frac{1}{\gamma}\frac{E_t\left(rx_{t+m}^{(n)}\right)}{Var_t\left(rx_{t+m}^{(n)} + \frac{m}{12}y_t^{(m/12)}\right)}, \tag{16}$$

where γ measures the investor's relative risk aversion, $E_t\left[rx_{t+m}^{(n)}\right]$ represents the m-month-ahead forecasted value of the n-year excess bond return at time t, and $Var_t\left(rx_{t+m}^{(n)} + \frac{m}{12}y_t^{(m/12)}\right)$ is the conditional variance of the n-year bond return at time t that is estimated

using sample variance from a 36-month rolling window of historical returns similar to Campbell and Thompson (2008). Then, the realized portfolio return at time $t+m$ is given by

$$R_{p,t+m}^{(n)} = w_t^{(n)} \times rx_{t+m}^{(n)} + (1 - w_t^{(n)}) \times \tfrac{m}{12} y_t^{(m/12)}, \tag{17}$$

where $rx_{t+m}^{(n)}$ is the realized excess bond return at time $t+m$, and $y_t^{(m/12)}$ is the annualized m-month risk-free rate at time t.

Over the out-of-sample period, the investor then realizes the average utility as follows:

$$U_{MV} = \mu_p - \frac{1}{2}\gamma\sigma_p^2, \tag{18}$$

where μ_p and σ_p^2 are the sample mean and variance of the portfolio returns over the out-of-sample period. Denote the investor's average utility resulting from using the forecasted values of the deep/machine learning predictive models as $\hat{U}_{MV}$, and denote the investor's average utility resulting from using the historical average forecasts as $\bar{U}_{MV}$. The difference, $\hat{U}_{MV} - \bar{U}_{MV}$, represents the investor's utility gains achieved from using the deep/machine learning forecasts over the historical average forecasts in asset allocation.

4. Data and Summary Statistics

We combine different sources of data for our empirical analyses. Those data include the yield curve data and two types of data on real-time macroeconomic information: macro vintage data and news-based macro data of Bybee, Kelly, Manela, and Xiu (2021). Given availability of the news-based data, we choose our sample that spans over the period from January 1984 to June 2018.

4.1. Yield Data

The commonly used yield data are those from the Fama-Bliss dataset (Fama and Bliss, 1987), which, however, is available only for maturities of one, two, three, four, and five years and also cannot be used to construct the one-month holding-period non-overlapping

bond returns. Recently, Liu and Wu (2021) use a nonparametric smoothing approach to reconstruct the constant-maturity zero-coupon Treasury yield curve with maturities ranging from 1 month to 360 months and show their dataset has much smaller pricing errors than the other commonly used dataset constructed by Gürkaynak, Sack, and Wright (2007), who use a parametric model of Nelson and Siegel (1987) to interpolate/extrapolate a smooth yield curve. Therefore, in this paper, we use the dataset of Liu and Wu (2021) to compute monthly frequency overlapping and non-overlapping excess bond returns.

Panel I of Table 1 presents the summary statistics of annualized non-overlapping excess bond returns. We see that both mean and standard deviation increase with respect to bond maturity. Both skewness and kurtosis have smirk-shapes to bond maturity. The first-order autocorrelation decreases to maturity, ranging from 0.22 in two-year excess bond returns to 0.06 in 10-year excess bond returns.

Panel II of Table 1 presents the summary statistics of overlapping excess bond returns. We find similar term structure patterns of mean and standard deviation to non-overlapping excess bond returns. However, the corresponding mean and standard deviation are smaller in overlapping excess bond returns than in non-overlapping excess bond returns for each maturity. The patterns of skewness and kurtosis in overlapping excess bond returns are different from those in non-overlapping excess bond returns. For each maturity, the skewness is much larger in overlapping excess bond returns than in non-overlapping excess bond returns. In contrast, the kurtosis is smaller in overlapping excess bond returns than in non-overlapping excess bond returns. Furthermore, the overlapping excess bond returns are much more persistent than non-overlapping excess bond returns. First-order autocorrelation ranges from 0.95 in two-year excess bond returns to 0.91 in ten-year excess bond returns.

Figure 3 presents the time series of non-overlapping (thin blue line) and overlapping (thick red line) excess bond returns for maturities of 2, 5, 7, and 10 years. For each maturity, we see that overlapping returns are much more persistent than non-overlapping returns, and their dynamics are very different from those of non-overlapping returns.

4.2. Real-Time Macroeconomic Vintage Data

Our first type of real-time macro information is based on data on macroeconomic variables commonly used in literature. However, most studies on bond return predictability with macro information use fully-revised macro variables in their empirical analysis. Macroeconomic data are subject to possible future revisions and are often released with a delay. If there is any macro information that affects bond prices, that should be the real-time one when bond prices are determined. A recent study by Ghysels, Horan, and Moench (2018) finds the predictive power of macro variables for future excess bond returns is largely from data revisions, and the authors suggest using real-time macro data in bond return predictability.

Therefore, we rely on the Archival Federal Reserve Economic Data (ALFRED), maintained by the Federal Reserve Bank of St. Louis, to construct real-time macro data, which, following Croushore (2011), are collections of macro vintage data, reflecting macro information available at each time without revised with respect to future information.[2]

Following Ludvigson and Ng (2011) and McCracken and Ng (2016), we construct a balanced panel of 125 monthly macroeconomic times series. These macro variables cover major economic categories and are grouped into eight categories: (i) output and income (15 series); (ii) labor market (31 series); (iii) housing sector (10 series); (iv) consumption, orders, and inventories (10 series); (v) money and credit (13 series); (vi) interest rates and foreign exchange rates (21 series); (vii) prices (20 series); and (viii) stock market (5 series). We stationarize these variables following the same methods as those of McCracken and Ng (2016). See the Appendix for the complete list of these macro variables and the transformation methods used to stationarize them.

[2]The use of real-time macro information suggests that econometricians may have less information than economic agents who know the equilibrium relationship between prior macroeconomic information and the yield curve, see, e.g., Atanasov, Moller, and Priestley (2022), and also the discussion in Hansen (2007). The real-time macroeconomic data are also recently used by Huang and Shi (2022) and Huang et al. (2021).

4.3. News-based Real-Time Macro Information

Our second type of real-time macro information is news-based, recently introduced by Bybee, Kelly, Manela, and Xiu (2021) who propose an approach to measuring the state of the economy via textual analysis of the full-text content of about 800,000 Wall Street Journal articles. They estimate a topic model that summarizes business news as topical themes and quantifies the proportion of news attention allocated to each theme. They further show that this text-based news attention accurately tracks a wide range of economic activity measures and has incremental forecasting power for macroeconomic outcomes.

The model used in Bybee, Kelly, Manela, and Xiu (2021) follows the LDA topic modeling approach of Blei, Ng, and Jordan (2003) that treats an individual article as a mixture of topics. The formation of topics is unsupervised and is estimated as clusters of terms that tend to co-occur in articles. We use news attention as our measure of real-time macro information, and there are in total of 180 time series. See the Appendix for the complete list. Following Bybee, Kelly, Manela, and Xiu (2021), we divide news topic attention time series into two groups: (i) economy news (78 series) and (ii) politics and culture news (102 series).

Combining these two types of real-time macroeconomic data, we have a large panel of time series on 305 variables.

5. Empirical Results

We set the out-of-sample period to range from January 2000 to June 2018, 211 months. This period is particularly interesting and maybe challenging for out-of-sample tests, as the US economy experiences the 2008 financial crisis, the interest rates enter an era of zero lower bound, and the US treasury bonds seem to become macro hedge assets and are negatively correlated with stocks (Li, Sarno, and Zinna, 2021). We consider two types of predictors, real-time macro variables and news-based topic attention. For the deep learning models, we consider standard neural networks (NN), group neural networks (GNN), and weighted group neutral networks (WGNN). To reduce stochastic errors when

using deep learning models, we adopt an ensemble learning approach, i.e., for the model selected from each type, we conduct 10 independent runs at each time in the out-of-sample period and take the mean output as its forecast. We only consider one component case for the machine learning models of PCA, PLS, and AutoEncoder. When the number of components becomes larger than one, the model performance generally deteriorates.

5.1. Statistical Evidence

5.1.1. Forecasting Non-overlapping Returns

Table 2 presents the out-of-sample R^2s for forecasting non-overlapping excess bond returns. The Clark-West statistics (Clark and West, 2007) over historical averages are applied only when the out-of-sample R^2s are positive. We first look at the model performance when we only use the real-time macro vintage data in Panel I. First, we find that all three types of deep learning models seem to help find out-of-sample statistical evidence for forecasting two-year-maturity non-overlapping excess bond returns; however, the out-of-sample R^2s from NN and GNN are smaller than 1%, whereas the out-of-sample R^2 of WGNN is about 2.8% for two-year excess bond returns. Differently, for excess bond returns with maturity larger than two years, the out-of-sample R^2s from all three types of deep learning models are either near-zero or negative.

Second, we also consider some commonly used linear and nonlinear machine learning predictive models for comparison. We see that regardless of which machine learning model is used and which bond maturity is under consideration, all out-of-sample R^2s are negative except for random forest in forecasting two-year- and three-year-maturity returns. Wan, Fulop, and Li (2021) show that when real-time macro information is used, it is difficult to find any statistical evidence of non-overlapping bond return predictability in linear predictive models no matter whether stochastic volatility is introduced or not. Our findings are consistent and provide further empirical evidence that ordinary machine learning models cannot help find any statistical evidence of non-overlapping bond return predictability when using real-time macro information.

The literature has found that the news contains useful macroeconomic information. For example, Bybee, Kelly, Manela, and Xiu (2021) show that the text of business news summarizes wide-ranging facets of the state of the economy. Ellingsen, Larsen, and Thorsrud (2021) use news data for macroeconomic forecasting and find that the news data contains information not captured by the complex economic indicators. Therefore, we next examine whether including news-based information can improve the out-of-sample performance of our deep/machine learning predictive models in forecasting non-overlapping excess bond returns.

Panel II of Table 2 presents out-of-sample R^2s resulting from combining real-time macro vintage data and news topic attention (in total, 305 variables) in forecasting non-overlapping excess bond returns. First, the performance of the three types of deep learning models improves in general. For example, the out-of-sample R^2 from WGNN is now 3.92% for forecasting two-year excess bond returns and 2.07% for forecasting three-year excess bond returns. However, the evidence of predictability for bond returns with other maturities is still nonexistent. Second, it seems that the two penalty regressions, Lasso and Elastic Net, can now also find some statistical evidence of out-of-sample predictability for two- and three-year non-overlapping excess bond returns; however, the corresponding out-of-sample R^2s are smaller than those from WGNN: 2.10% and 2.03%, respectively, for 2-year bond returns, and 1.50% and 1.30%, respectively, for 3-year bond returns. The out-of-sample R^2s from the other machine learning models are still negative.

All the above findings suggest that when using real-time macro information, whether news-based information is included, our deep learning models, particularly WGNN, can help find some out-of-sample statistical evidence for forecasting short-term non-overlapping excess bond returns. Furthermore, the shrinkage regressions, such as Lasso and Elastic net, seem also to help find some out-of-sample statistical evidence for forecasting short-term non-overlapping excess bond returns when news-based information is combined with real-time macro variables. However, such evidence is much weaker than that found when fully-revised macro data are used as in Gargano, Pettenuzzo, and Timmermann (2019).

5.1.2. Forecasting Overlapping Returns

We now look at model performance for forecasting overlapping excess bond returns. Table 3 presents out-of-sample R^2s resulting from deep/machine learning predictive models. In Panel I, we only use real-time macro vintage data as our predictors. We find that among the three types of the deep learning models, WGNN performs much better than NN and GNN in general, and such outperformance is much stronger for short-maturity bond returns. The out-of-sample R^2s resulting from WGNN are positive and statistically significant for all-maturity bond returns, ranging from 1.69% for 5-year overlapping excess bond returns to 8.21% for 2-year overlapping excess bond returns. The overall out-of-sample R^2 for forecasting all-maturity returns is 2.61% in WGNN, whereas it is only 1.60% in GNN and is negative in NN.

We then check whether the machine learning models under consideration can help find out-of-sample statistical evidence for forecasting overlapping bond returns. Unlike what we have observed in forecasting non-overlapping returns, we find that although most machine learning models underperform historical averages in forecasting short-maturity overlapping returns, several standard machine learning models can help find statistical out-of-sample evidence for predicting long-maturity overlapping bond returns. In particular, we find that both PCA and Elastic Net work quite well in forecasting overlapping bond returns with maturity larger than or equal to 5 years: the out-of-sample R^2s from PCA range from 3.59% for 5-year bond returns to 9.00% for 10-year bond returns, and the out-of-sample R^2s from Elastic Net range from 3.63% for 5-year bond returns to 10.13% for 10-year bond returns. Given their superior performance in forecasting long-maturity returns, the overall out-of-sample R^2s from PCA and Elastic Net are 6.17% and 6.26%, respectively, larger than that from WGNN. Ridge, Random Forest, and AutoEncoder also perform well for forecasting 8-, 9-, and 10-year bond returns, and their overall out-of-sample R^2s are 3.24%, 2.54%, and 1.20%, respectively.

We further examine whether including news-based information improves the out-of-sample performance of deep/machine learning predictive models for forecasting overlap-

ping bond returns. From Panel II of Table 3 that uses both real-time macro vintage data and news topic attention as our predictors, we find that similar to what we have observed in Table 2 for forecasting non-overlapping returns, the performance of all three types of deep neural networks improves, and this improvement is particularly striking for WGNN. In comparison to those in Panel I, the out-of-sample R^2s resulted from WGNN increase for overlapping bond returns of all maturities, now ranging from 4.99% for 5-year bond returns to 11.54% for 2-year excess bond returns. The overall out-of-sample R^2 of WGNN increases to 6.12%. This result is consistent with what is found by Bybee, Kelly, Manela, and Xiu (2021) who show that news attention accurately tracks a wide range of economic activity measures and that they have incremental forecasting power for macroeconomic outcomes.

However, when a large panel of news-based data is combined with real-time macro vintage data, all those machine learning models that perform well in forecasting long-maturity bond returns in Panel I lose their ability and result in negative out-of-sample R^2s. Given that when both types of real-time macro information are combined, the number of predictors becomes very large (in total, 305 variables), and that the variables extracted from the news are highly correlated with those standard macro variables, negative out-of-sample R^2s from machine learning models just suggest that they can not efficiently handle such a high-dimensional data with high correlations. However, our neural network with a group structure can efficiently work with those high-dimensional data by grouping variables with similar information (high correlation) and reducing their respective dimensions separately. The basic idea of our approach is similar to dimension reduction of Ludvigson and Ng (2009) and variable selection in a group of Huang and Shi (2022).

The above findings suggest that for forecasting overlapping bond returns, our weighted group deep learning model performs quite well for all-maturity bonds, particularly when real-time vintage macro data are combined with news-based data. Several machine learning models, PCA and Elastic Net in particular work well in forecasting long-maturity bond

returns using real-time vintage macro variables alone; however, they can not efficiently handle high-dimensional and highly-correlated data when real-time vintage macro variables are augmented by news-based macro information. However, the evidence we find is relatively weaker compared to what is found using fully-revised macro data (see, e.g., Ludvigson and Ng, 2009, 2011; Cooper and Priestley, 2009; Cieslak and Povala, 2015; Bianchi, Büchner, and Tamoni, 2021).

5.1.3. Importance of Predictors

When both real-time vintage macro data and news-based attentions data are used, there are in total of ten groups inputted into our weighted group neural network (WGNN), of which eight groups are from the vintage macro variables (Output, Labor, Housing, Consumption, Money, Interest, Prices, and Stock), and two groups are from news topic attention (Economy and Politics). We can investigate group importance by extracting the group-wise components from WGNN (see Equation (6)) and then running a regression of excess bond returns of each maturity on them. To alleviate randomness in training deep learning models and to make group importance measure robust, we repeat the above procedure 30 times using the most recent data and report the average squared t-values to measure group importance.

Given that the evidence of bond return predictability in non-overlapping returns is quite weak, only existing in the two-year bond returns, and that such evidence is much stronger in overlapping returns, existing in all-maturity bond returns, we focus on the importance of each group in forecasting overlapping bond returns. Table 4 presents average squared t-values of each group in each regression. We find that the group of Labor is very important for forecasting bond returns with maturities of 2-7 years, and its importance is decreasing with respect to maturity, and the group of Interest plays a critical role in forecasting bond returns with maturities of 4-10 year and its importance is increasing with respect to maturity. Furthermore, the group of Housing and the two news-based groups (Economy and Politics) are very important for forecasting all-maturity

bond returns. These results are generally consistent with Huang and Shi (2022) who find important roles played by Labor and Housing in forecasting annual excess bond returns using a two-step Lasso approach. The importance of news-based groups has already been noticed in Table 3, where we see that the performance of WGNN improves a lot when news-based attentions are introduced.

We have observed important roles played by two news-based group in WGNN in Table 3 and Table 4. We therefore further explore importance of each variable in news topic attention in WGNN. Chen, Pelger, and Zhu (2020) introduce a measure of variable importance in deep learning by relying on the magnitude of gradient values when training the model. The gradient values of a loss function are the slope coefficients in a linear model and perfectly represent variable importance in complex nonlinear deep learning models. For each attention series j, its gradient function is defined as

$$grad(X_{j,t}) = \frac{\partial \mathcal{L}(\mathbf{W}_t, \mathbf{b}_t; \mathbf{X_t})}{\partial X_{j,t}}. \tag{19}$$

A larger absolute gradient means that a variable has a greater effect on loss minimization. When the attention of the news is negative, the higher the attention of the news, the more helpful for the predictability, but when the gradient is positive, the smaller the attention of the news, the greater the help. As above, we repeat model training of WGNN 30 times and present the average gradient values of the last epoch training for the input news topic attention.

Figure 4 and Figure 5 present gradients of Economy news and of Politics news, respectively. In each figure, there are two panels, the left plots 20 news attentions with the most negative gradients, and the right plots 20 news attentions with the largest gradients. For Economy news, we find that in forecasting bond returns, news attention to (i) Steel, (ii) NASD, and (iii) Control stakes in the left panel are of the highest importance, and news attentions to (i) Small changes, (ii) Small caps, and (iii) Competition in the right panel have greatest contributions. For Politics and culture news, news attentions to (i) State politics, (ii) Changes, and (iii) Broadcasting in the left panel, as well as to (i) Courts, (ii)

Fees, and (iii) California in the right panel are most useful in forecasting bond returns. In general, the information contained in these news topics is not covered by traditional macro indicators.

5.2. Economic Evidence

From the previous subsection, we find that the out-of-sample statistical evidence for forecasting non-overlapping excess bond returns using real-time macro information is weak, seemingly only existing in forecasting very short-term bond returns; however, we do find strong out-of-sample statistical evidence for forecasting overlapping excess bond returns of both short- and long-term maturities, in particular, when using the weighted group neural network (WGNN) model based on both macro vintage data and news-based macro data. We then ask whether this statistical evidence of bond return predictability can translate into investors' economic gains. For this purpose, we compute investors' utility gains resulting from using deep/machine learning predictive models over the historical average for a mean-variance investor, as discussed in Subsection 3.3.

In investigating economic evidence, we consider two cases. The first enables investors to take a full short position such that the portfolio weight, $\hat{w}_t^{(n)}$, bounds between -1 and 2 to prevent extreme investments (Goyal and Welch, 2008; Campbell and Thompson, 2008; Ferreira and Santa-Clara, 2011; Thorton and Valente, 2012; Sarno, Schneider, and Wagner, 2016), and the second allows investors for leveraging their investments, such that the portfolio weight, $\hat{w}_t^{(n)}$, could be in the range between -1 and 8. Unlike in the equity market, investors could take extreme positions in the bond market, facilitated by repo agreements. Given that the statistical evidence for forecasting non-overlapping bond returns is weak and we find that it clearly cannot translate into investors' economic gains no matter which case is considered, in this part, we mainly focus on those models whose overall out-of-sample R^2s are larger than 2% in forecasting overlapping bond returns, as shown in Table 3, and see whether such out-of-sample evidence from those models can translate into investors' economic gains.

Table 5 reports the annualized percentage utility gains over the historical average for a mean-variance investor with moderate risk aversion ($\gamma = 5$). Panel I presents the utility gains resulting from those models whose overall out-of-sample R^2s are larger than 2% in Table 3 and where investors are not allowed to take extreme investments. We find that even though some of those models can generate positive utility gains for long-maturity bonds, they are generally quite small. For example, WGNN generates a utility gain of only 10 bps based on both macro vintage data and news-based data for the 10-year bond, and Ridge generates the largest utility gain based on macro vintage data among those models for the 10-year bond, which, however, is still small (32 bps). We then examine whether statistical evidence of overlapping bond return predictability can translate into investors' economic gains when investors could leverage their investments. Panel II presents utility gains from the above models when the portfolio weight bounds between -1 and 8. Now we find that our WGNN can generate positive and relatively large utility gains for bonds with maturities of 6 to 10 years based on both macro vintage data and news-based data: the utility gain ranges from 86 bps (9- and 10-year bonds) to 135 bps (7-year bond); however, when macro vintage data alone are used, the utility gains from WGNN become negative for all maturities, highlighting the importance of news-based macro information in forecasting bond returns and a result consistent with what we have seen in Table 3, where whenever news-based macro information is introduced, the performance of WGNN improves dramatically. Furthermore, we see that the utility gains from all the other models are either very small or negative.

We then explore how the investor's appetite for risk affects her utility gains. For this purpose, we consider the other two types of investors: the first is slightly risk-averse ($\gamma = 3$), and the other is strongly risk-averse ($\gamma = 8$). Table 6 presents utility gains for these two types of investors who are allowed to take leverage for their investments, that is, $\hat{w}_t^{(n)} \in [-1, 8]$. We find that whenever the investor becomes aggressive, the WGNN can still generate positive and relatively large utility gains for bonds with maturities of 8-10 years when both macro vintage data and news-based data are used: the utility gain

is 52 bps for an 8-year bond, 115 bps for a 9-year bond, and 137 bps for the 10-year bond, and the PCA also seems to generate utility gains larger 50 bps for bonds with maturities of 6 and 7 years when macro vintage data are used: 58 bps for 6-year bond and 60 bps for a 7-year bond. However, all the other models generate utility gains that are either negative or close to zero.

When the investor becomes conservative, the WGNN can continue to generate positive and relatively large utility gains for bonds with maturities of 6-10 years when both macro vintage data and news-based data are used: the utility gains are between 52 bps (9- and 10-year bonds) and 90 bps (7-year bond); however, the utility gains from the other models are negative or close to zero.

Perhaps surprisingly, we may wonder why utility gains for short-term bonds are so small or negative, given that in Table 3 we see that the out-of-sample R^2s resulting from WGNN in forecasting short-term bond returns are quite high, in particular, when we use both macro vintage data and news-based data. To investigate this point, we take a deep look at the unrestricted portfolio weights, which represent how much the investor would like to hold on to the risky bonds in the out-of-sample period. The left panels of Figure 6 present the unrestricted portfolio weights on 2-, 3-, and 4-year bonds in WGNN based on macro and news data and PCA and Elastic Net based on macro data. We find that in all three models, the vast majority of the portfolio weights in those three-maturity bonds are huge, way above the upper bound, which is 8 in our study, and this is particularly striking for a 2-year bond. Furthermore, the portfolio weights on those bonds increase dramatically after about 2012. High portfolio weights on bonds and their dramatic increase after the global financial crisis may result from the fact that, unlike stocks and long-term bonds, variations of the short-term bonds are much smaller, and those bonds are always regarded as safe assets (He, Krishnamurthy, and Milbradt, 2016), whose correlation with stocks is negative, especially during market downturns. When computing average utility, we have to cap the portfolio weights at the upper bound, resulting in small or negative utility gains.

However, the issue of high unrestricted portfolio weights is much alleviated for the long-term bonds. The right panels of Figure 6 present the unrestricted portfolio weights on 6-, 8-, and 10-year bonds for the same three models. We find that the unrestricted portfolio weights are much smaller compared to those presented in the left panels. This is particularly true for WGNN with macro and news data, as most of the 6-year bond portfolio weights and all 8- and 10-year bond portfolio weights are below the upper bound we set, explaining its outperformance in generating utility gains for long-term bonds in Tables 5 and 6.

6. Conclusion

Numerous studies have documented that macroeconomic variables have strong predictive power for future excess bond returns. However, most of these works use the fully-revised macro variables in their empirical analysis. Macro data are subject to possible future revisions and are often released with a delay. Ghysels, Horan, and Moench (2018) find that the predictive power of macro variables for future excess bond returns is largely from the data revision and argue that the real-time, instead of fully-revised, macro data should be used in bond return predictability to avoid any hindsight problem. Wan, Fulop, and Li (2021) use different types of real-time macro data to implement a comprehensive analysis based on linear predictive models with and without stochastic volatility and find no statistical and economic evidence for forecasting non-overlapping excess bond returns whenever real-time macro factors are used as predictors.

This paper reexamines whether both non-overlapping and overlapping bond returns are predictable when real-time, instead of fully-revised, macro information relies on a weighted group deep learning model. Deep learning is a form of nonlinear supervised machine learning that employs a deep neural network for implementing prediction through a series of nonlinear transformations using a large number of predictors. A recent study by Gu, Kelly, and Xiu (2020) reinvestigates equity return predictability by using various deep/machine learning models and find deep learning outperforms. Similarly, using a

large panel of fully-revised macro variables as predictors, Bianchi, Büchner, and Tamoni (2021) find the deep learning models can generate strong evidence of overlapping and non-overlapping bond return predictability.

However, the panel of macroeconomic variables potentially used for bond return forecasting is very large, and most of those variables are highly correlated. Simple off-the-shelf application of deep learning may hardly significantly improve forecasting. Motivated by group features of macroeconomic variables, in this paper, we develop weighted neural networks with a group structure that compress similar information into the same categories and hence can help alleviate issues related to highly correlated predictor variables.

The paper considers two types of real-time macro information: real-time macro vintage data and news-based topic attention. We find news contains rich information on future bond returns beyond traditional macro variables. When both types of real-time data are used as predictors, our proposed model can help find significant statistical evidence for forecasting *non-overlapping* short-term bond returns and for forecasting *overlapping* bond returns with maturities of 2 to 10 years. Furthermore, the statistical evidence of overlapping bond return predictability can be translated into investors' economic gains for long-term bonds when investors are allowed to leverage their investments. While some standard machine learning methods can also help find statistical evidence for forecasting long-term overlapping bond returns when macro vintage data alone are used, they cannot efficiently handle high-dimensional predictors with high correlations when both types of real-time data are combined.

References

Atanasov, V., S. V. Moller, and R. Priestley (2022). Consumption fluctuations and expected returns. *Journal of Finance 75*, 1677–1713.

Bianchi, D., M. Büchner, T. Hoogteijling, and A. Tamoni (2021). Corrigendum: Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1090–1103.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2021). Business news and business cycles. Technical report, National Bureau of Economic Research.

Campbell, J. and R. Shiller (1991). Yield spreads and interest rate movement: A bird's eye view. *Review of Economic Studies 58*, 495–514.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies 21*, 1509–1531.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. Technical report, Stanford University.

Cieslak, A. and P. Povala (2015). Expected returns in treasury bonds. *Review of Financial Studies 28*, 2859–2901.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*, 291–311.

Cochrane, J. and M. Piazzesi (2005). Bond risk premia. *American Economic Review 94*, 138–160.

Cooper, I. and R. Priestley (2009). Time-varying risk premiums and the output gap. *Review of Financial Studies 22*, 2801–2833.

Croushore, D. (2011). Frontiers of real-time data analysis. *Journal of Economic Literature 49*, 72–100.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–263.

Ellingsen, J., V. Larsen, and L. A. Thorsrud (2021). News media vs. FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics 37*, 63–81.

Fama, E. and R. Bliss (1987). The information in long-maturity forward rates. *American Economic Review 77*, 680–692.

Feng, G., N. G. Polson, and J. Xu (2019). Deep learning in asset pricing. Technical report, City University of Hong Kong.

Ferreira, M. and P. Santa-Clara (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics 100*, 514–537.

Gargano, A., D. Pettenuzzo, and A. Timmermann (2019). Bond return predictability: Economic value and links to the macroeconomy. *Management Science 65*, 508–540.

Ghysels, E., C. Horan, and E. Moench (2018). Forecasting through the rearview mirror: Data revisions and bond return predictability. *Review of Financial Studies 31*, 678–714.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Goyal, A. and I. Welch (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*, 1455–1508.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*, 429–450.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2007). The US treasury yield curve: 1961 to the present. *Journal of Monetary Economics 54*, 2291–2304.

Hansen, L. (2007). Beliefs, doubts and learning: Valuing macroeconomic risk. *American Economic Review 97*, 1–30.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, Z., A. Krishnamurthy, and K. Milbradt (2016). What makes us government bonds safe assets? *American Economic Review 106*, 519–523.

Huang, D., K. Jiang, Fuwei Li, G. Tong, and G. Zhou (2021). Are bond returns predictable with real-time macro data? Technical report, Singapore Management University.

Huang, J.-Z. and Z. Shi (2022). Machine-learning-based return predictors and the spanning controversy in macro-finance. *Management Science Forthcoming*.

Joslin, S., M. Priebsch, and K. J. Singleton (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *Journal of Finance 69*, 1197–1233.

Kelly, B., A. Manela, and A. Moreira (2021). Text selection. *Journal of Business & Economic Statistics 39*(4), 859–879.

Li, J., L. Sarno, and G. Zinna (2021). Risk and risk premiums in the US treasury market. Technical report, Bank of Italy.

Liu, Y. and J. Wu (2021). Reconstructing the yield curve. *Journal of Financial Economics 142*, 1395–1425.

Ludvigson, S. and S. Ng (2009). Macro factors in bond risk premia. *Review of Financial Studies 22*, 5027–5067.

Ludvigson, S. and S. Ng (2011). *Handbook of Empirical Economics and Finance*, Chapter A factor analysis of bond risk premia, pp. 313–372. New York: Chapman and Hall.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics 34*, 574–589.

Nelson, C. and A. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business 4*, 473–489.

Sarno, L., P. Schneider, and C. Wagner (2016). The economic value of predicting bond risk premia. *Journal of Empirical Finance 37*, 247–267.

Thorton, D. and G. Valente (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies 25*, 3141–3168.

Wan, R., A. Fulop, and J. Li (2021). Real-time bayesian learning and bond return predictability. *Journal of Econometrics Forthcoming*.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*, 1067–1084.

Wright, J. (2011). Term premia and inflation uncertainty: Empirical evidence from an international panel dataset. *American Economic Review 101*, 1514–1534.

Table 1: **Summary Statistics: Excess Bond Returns**

	Panel I: Non-overlapping Excess Bond Returns								
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y
Mean	1.68	2.44	3.01	3.49	3.99	4.33	4.78	5.03	5.55
Std	1.93	3.07	4.22	5.26	6.25	7.19	8.22	9.13	10.03
Skew	0.22	0.02	−0.05	−0.05	−0.04	−0.00	0.06	0.15	0.16
Kurt	4.09	3.73	3.50	3.66	3.75	3.71	4.07	4.48	4.60
ACF	0.22	0.18	0.14	0.12	0.10	0.08	0.07	0.07	0.06
	Panel II: Overlapping Excess Bond Returns								
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y
Mean	0.84	1.55	2.24	2.69	3.25	3.60	4.01	4.32	4.65
Std	1.38	2.61	3.72	4.69	5.74	6.61	7.47	8.40	9.28
Skew	0.55	0.33	0.26	0.20	0.26	0.19	0.26	0.29	0.66
Kurt	3.09	2.95	3.05	3.11	3.40	3.53	3.63	3.87	3.93
ACF	0.95	0.94	0.93	0.92	0.92	0.92	0.91	0.91	0.91

Note: The table reports summary statistics of non-overlapping (one-month holding) and overlapping (annual holding) treasury bond excess returns constructed from the yield curve dataset of Liu and Wu (2021). Means and standard deviations are annualized. The sample includes those bonds with maturities of 2 to 10 years. The data are at the monthly frequency and range from January 1984 to June 2018.

Table 2: **Non-overlapping Bond Return Predictability**

	Panel I: Real-Time Macro Variables									
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y	All
A. Deep Learning Models										
NN	0.12**	0.14***	0.03	0.01	−0.01	0.04	0.01	−0.02	0.04	0.01*
GNN	0.97**	0.16	0.39*	0.11	0.18	0.25*	0.03	0.08	0.14	0.14***
WGNN	2.83**	0.32*	−1.68	−2.56	−3.00	−2.93	−3.57	−3.89	−3.77	−3.32
B. Machine Learning Models										
PCA	−0.76	−1.66	−2.37	−2.43	−2.35	−2.57	−2.84	−3.06	−3.21	−2.82
PLS	−38.85	−38.32	−37.93	−38.74	−39.53	−39.34	−40.36	−40.45	−40.55	−39.96
Lasso	−3.28	−2.52	−2.52	−0.68	−0.95	−1.94	−3.56	−2.57	−4.20	−2.80
Ridge	−2.14	−3.87	−3.77	−4.73	−4.07	−5.22	−4.94	−4.99	−5.44	−4.93
Elastic Net	−4.66	−2.66	−1.42	−1.86	−2.01	−2.23	−3.06	−3.91	−4.19	−3.19
AutoEncoder	−1.73	−2.58	−3.14	−3.06	−2.97	−3.32	−3.54	−3.65	−3.71	−3.44
Boosted Tree	−4.24	−6.18	−6.25	−8.08	−7.99	−11.52	−6.71	−7.52	−11.78	−8.99
Random Forest	1.95**	0.82*	−0.09	−0.47	−0.67	−0.97	−1.35	−1.55	−1.76	−1.21
	Panel II: Real-Time Macro Variables and News Attentions									
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y	All
A. Deep Learning Models										
NN	0.08**	0.05**	0.03**	0.02	0.03***	0.01	0.02**	0.01	0.02**	0.02***
GNN	1.31*	0.47	0.53*	0.53**	0.35**	0.45**	0.13	0.26*	0.05	0.26***
WGNN	3.92**	2.07*	0.23	−0.26	−0.52	−1.02	−1.62	−1.43	−1.50	−1.08
B. Machine Learning Models										
PCA	−1.69	−2.70	−3.66	−3.84	−3.79	−3.87	−4.05	−4.11	−3.98	−3.92
PCA (1+1)	0.00	−1.80	−3.21	−3.59	−3.61	−3.63	−4.00	−4.11	−3.98	−3.80
PLS	−39.22	−41.64	−42.51	−42.46	−42.61	−42.41	−42.42	−41.05	−41.20	−41.79
Lasso	2.10**	1.50*	1.51*	−0.08	−0.67	0.31	−0.67	−2.84	−4.75	−1.88
Ridge	−2.35	−3.57	−4.32	−5.23	−4.77	−4.75	−4.43	−4.44	−4.63	−4.57
Elastic Net	2.03**	1.30*	1.11	−0.29	0.19	0.03	−1.18	−1.97	−2.24	−1.11
AutoEncoder	−0.90	−2.10	−3.13	−3.39	−3.43	−3.52	−3.69	−3.77	−3.66	−3.56
Boosted Tree	−5.08	−8.95	−9.33	−8.30	−9.19	−7.18	−9.03	−15.59	−9.99	−10.38
Random Forest	0.12	−1.01	−1.66	−1.84	−1.88	−1.91	−2.04	−2.09	−2.15	−1.98

Note: This table reports out-of-sample R^2 resulting from the deep/machine learning predictive models for forecasting *non-overlapping* excess bond returns using macro *vintage* data alone (Panel I) and both macro vintage data and news topic attention (Panel II). The last column presents the aggregate performance by combining bond returns of all maturities together to compute out-of-sample R^2. In the table, PCA (1+1) uses the first PC from macro vintage data and the first PC from news attentions data. The out-of-sample R^2 is computed using Equation (15), and its statistical significance is evaluated using the method of Clark and West (2007). *, **, and *** denote significance at the 10%, 5%, and 1% significance levels. The out-of-sample period ranges from January 2000 to July 2017.

Table 3: **Overlapping Bond Return Predictability**

	Panel I: Real-Time Macro Variables									
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y	All
A. Deep Learning Models										
NN	−16.19	−9.46	−8.98	−6.01	−3.44	−0.87	1.06	2.03*	2.69*	−0.44
GNN	0.00	0.00	0.05***	0.25***	0.86**	1.25**	1.85**	2.22*	2.22*	1.60***
WGNN	8.21***	4.69***	1.92***	1.69***	1.77***	2.79***	2.34***	2.12**	3.49***	2.61***
B. Machine Learning Models										
PCA	−5.19	−1.08	0.23***	3.59***	4.26**	5.42***	6.75**	7.39**	9.00***	6.17***
PLS	−95.19	−86.73	−89.45	−77.62	−81.34	−72.31	−69.96	−65.91	−52.43	−68.45
Lasso	−4.41	−1.52	−1.34	−2.16	0.54	−0.76	−0.09	1.50**	4.99***	1.16***
Ridge	0.98**	1.64**	1.06*	1.71*	0.97	2.05*	3.48**	3.72**	5.42**	3.24***
Elastic Net	1.03*	1.24	2.69	3.63*	3.84**	4.08**	6.04**	7.86***	10.13***	6.26***
AutoEncoder	−6.38	−4.16	−4.28	−1.68	−0.89	0.25**	1.61**	2.16**	4.23**	1.20***
Boosted Tree	−18.90	−15.69	−19.33	−12.62	−16.82	−25.99	−15.00	−25.31	−14.65	−18.75
Random Forest	−6.89	−4.94	−3.15	−0.82	0.15	1.71	2.97*	4.04**	5.65**	2.54***
	Panel II: Real-Time Macro Variables and News Topic Attention									
	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y	All
A. Deep Learning Models										
NN	−0.49	−1.13	−2.59	−1.06	0.73**	1.73***	2.39**	3.09*	2.71	1.76***
GNN	0.00	0.00	0.11***	0.38***	1.12***	1.53**	2.15**	2.46*	2.40*	1.82***
WGNN	11.54**	6.64**	5.68*	4.99*	5.64**	5.84**	6.34**	6.50**	6.20**	6.12***
B. Machine Learning Models										
PCA	−26.34	−34.33	−36.30	−38.51	−39.76	−39.75	−38.21	−38.47	−37.08	−38.10
PCA (1+1)	0.63**	−7.21	−10.90	−13.94	−15.57	−14.83	−14.25	−14.98	−14.36	−14.17
PLS	−141.10	−150.05	−146.94	−129.39	−121.28	−107.66	−93.84	−82.77	−68.13	−96.85
Lasso	−58.74	−38.93	−33.66	−38.92	−44.11	−44.79	−37.87	−35.76	−19.43	−36.83
Ridge	−20.68	−22.58	−20.53	−20.29	−20.35	−18.55	−16.04	−13.22	−8.36	−15.48
Elastic Net	−50.37	−36.41	−29.14	−41.17	−45.79	−32.44	−21.12	−14.24	−8.56	−24.36
AutoEncoder	−24.78	−33.00	−35.08	−37.56	−39.01	−39.18	−37.80	−38.20	−36.88	−36.71
Boosted Tree	−14.01	−22.21	−33.31	−29.79	−18.92	−26.06	−20.00	−27.21	−24.01	−21.92
Random Forest	−12.93	−14.68	−12.49	−9.74	−7.25	−5.53	−3.17	−1.49	−0.02	−4.37

Note: This table reports out-of-sample R^2 resulting from the deep/machine learning predictive models for forecasting *overlapping* excess bond returns using macro *vintage* data alone (Panel I) and both macro vintage data and news topic attention (Panel II). The last column presents the aggregate performance by combining bond returns of all maturities together to compute out-of-sample R^2. In the table, PCA (1+1) uses the first PC from macro vintage data and the first PC from news attentions data. The out-of-sample R^2 is computed using Equation (15), and its statistical significance is evaluated using the method of Clark and West (2007). *, **, and *** denote significance at the 10%, 5%, and 1% significance levels. The out-of-sample period ranges from December 2000 to June 2018.

Table 4: **Variable Group Importance for Weighted Group Neural Network (WGNN)**

	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y
Output	1.08	0.85	0.70	0.72	0.75	0.72	0.72	0.75	0.69
Labor	10.78***	10.63***	8.21***	6.59***	4.86**	4.08**	3.12*	2.43	2.47
Housing	5.34***	5.81***	6.97***	7.95***	8.69***	8.76***	8.71***	8.46***	8.44***
Consumption	1.26	1.40	1.66	2.00	2.15	2.31	2.37	2.58	2.67
Money	2.21	2.09	2.17	2.32	2.37	2.37	2.53	2.66	2.65
Interest	2.93*	3.79*	5.84***	8.09***	10.50***	12.97***	15.61***	17.86***	19.28***
Prices	0.37	0.39	0.43	0.49	0.46	0.48	0.52	0.55	0.59
Stock	2.41	2.65	2.79*	2.94*	2.99*	3.12*	3.17*	3.24*	3.30*
Economy news	9.11***	6.57***	6.16***	5.59***	5.97***	5.40***	5.36***	5.28***	4.99**
Politics news	10.61***	8.46***	7.93***	7.50***	7.60***	7.06***	6.88***	6.72***	6.27***

Note: This table reports significance of group importance. We can investigate group importance by extracting the group-wise components from WGNN and then running a regression of excess bond returns of each maturity on them. To alleviate randomness in training deep learning models and to make group importance measure robust, we repeat the above procedure 30 times using the most recent data and report the average squared t-values to measure group importance. *, **, and *** denote significance at the 10%, 5%, and 1% significance levels.

Table 5: **Utility Gains and Bond Return Predictability**

	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y
A. $w \in$ [-1, 2]									
WGNN (macro+news)	0.00	0.00	0.00	0.00	−0.03	0.03	−0.01	0.04	0.10
WGNN (macro)	−0.01	−0.07	−0.15	−0.30	−0.52	−0.70	−0.74	−0.64	−0.50
PCA (macro)	0.00	0.00	0.00	0.00	0.01	0.10	0.18	0.24	0.31
Ridge (macro)	0.00	0.00	0.00	0.00	0.00	0.07	0.14	0.23	0.32
Elastic Net (macro)	0.00	0.00	0.00	−0.01	−0.33	−0.60	−0.74	−0.69	−0.71
Random Forest (macro)	0.00	0.00	0.00	−0.06	−0.13	−0.11	−0.03	0.11	0.20
B. $w \in$ [-1, 8]									
WGNN (macro+news)	−0.22	−0.08	−0.16	0.13	1.11	1.35	1.34	0.86	0.86
WGNN (macro)	−0.43	−0.44	−1.72	−1.99	−2.23	−3.81	−4.90	−6.33	−7.73
PCA (macro)	0.06	0.33	0.21	−0.48	−2.86	−3.43	−2.60	−1.41	−1.55
Ridge (macro)	0.00	−0.23	−0.32	−1.06	−3.21	−5.70	−7.13	−7.06	0.00
Elastic Net (macro)	−0.11	−0.05	−0.34	−0.33	−2.88	−6.66	−13.99	−16.56	−23.04
Random Forest (macro)	−0.07	−0.59	−0.92	−0.96	−3.48	−6.27	−7.42	−6.61	−7.43

Note: This table reports the annualized percentage utility gains for forecasting overlapping excess bond returns for a mean-variance investor with the coefficient of relative risk-aversion equal to 5. We consider two types of investors: the first can only take a short position, that is, $w \in$ [-1, 2], and the other can leverage her investments, i.e., $w \in$ [-1, 8]. Only those deep/machine learning models that can generate the overall out-of-sample R^2s larger than 2% are taken into account. The out-of-sample period ranges from December 2000 to June 2018.

Table 6: **Risk-Aversion and Utility Gains**

	2Y	3Y	4Y	5Y	6Y	7Y	8Y	9Y	10Y
A. $\gamma = 3$									
WGNN (macro+news)	0.00	−0.12	−0.12	−0.04	−0.16	0.05	0.52	1.15	1.37
WGNN (macro)	−0.21	−0.77	−1.20	−1.96	−2.22	−1.31	−0.65	−1.35	−2.05
PCA (macro)	0.00	0.11	0.25	0.38	0.58	0.60	0.14	−0.90	−2.01
Ridge (macro)	0.00	0.01	−0.12	−0.23	−0.01	−0.23	−0.90	−2.39	−3.80
Elastic Net (macro)	0.00	−0.15	−0.35	−0.39	−0.88	−1.88	−2.59	−4.27	−6.96
Random Forest (macro)	0.00	−0.16	−0.46	−0.77	−0.51	−0.46	−0.95	−2.58	−4.01
B. $\gamma = 8$									
WGNN (macro+news)	−0.09	−0.07	−0.12	0.20	0.74	0.90	0.81	0.52	0.52
WGNN (macro)	−0.11	−0.33	−1.26	−1.18	−2.53	−3.82	−4.89	−5.97	−6.50
PCA (macro)	0.17	0.25	0.00	−1.21	−2.31	−2.23	−1.59	−0.86	−0.95
Ridge (macro)	0.04	−0.13	−0.34	−1.27	−2.53	−3.84	−4.77	−4.54	−6.67
Elastic Net (macro)	−0.04	0.01	−0.17	−0.41	−2.18	−5.42	−11.78	−12.72	−15.77
Random Forest (macro)	−0.02	−0.47	−0.60	−1.09	−2.95	−4.16	−4.54	−4.04	−4.55

Note: This table reports the annualized percentage utility gains for forecasting overlapping excess bond returns for a mean-variance investor who can leverage her investments, i.e., $w \in$ [-1, 8]. We consider two types of investors: one is slightly risk averse, that is, $\gamma = 3$, and the other is strongly risk averse, that is, $\gamma = 8$. Only those deep/machine learning models that can generate the overall out-of-sample R^2s larger than 2% are taken into account. The out-of-sample period ranges from December 2000 to June 2018.

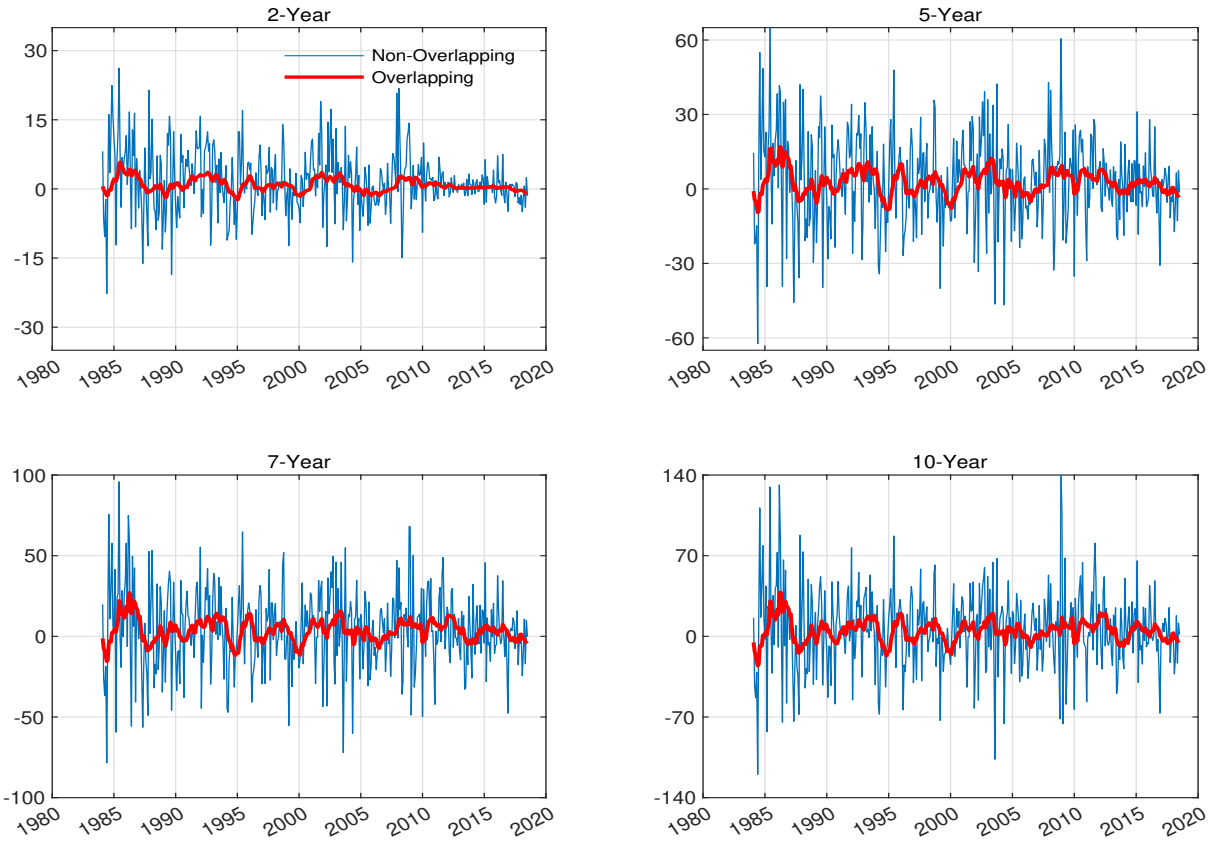

Figure 3: **Time Series of Excess Bond Returns**

Note: The figure presents the time series of overlapping and non-overlapping excess bond returns for maturities of two, five, seven, and ten years. All excess bond returns are computed using the Liu and Wu (2021) yield curve dataset. The data are at the monthly frequency and range from January 1984 to June 2018.

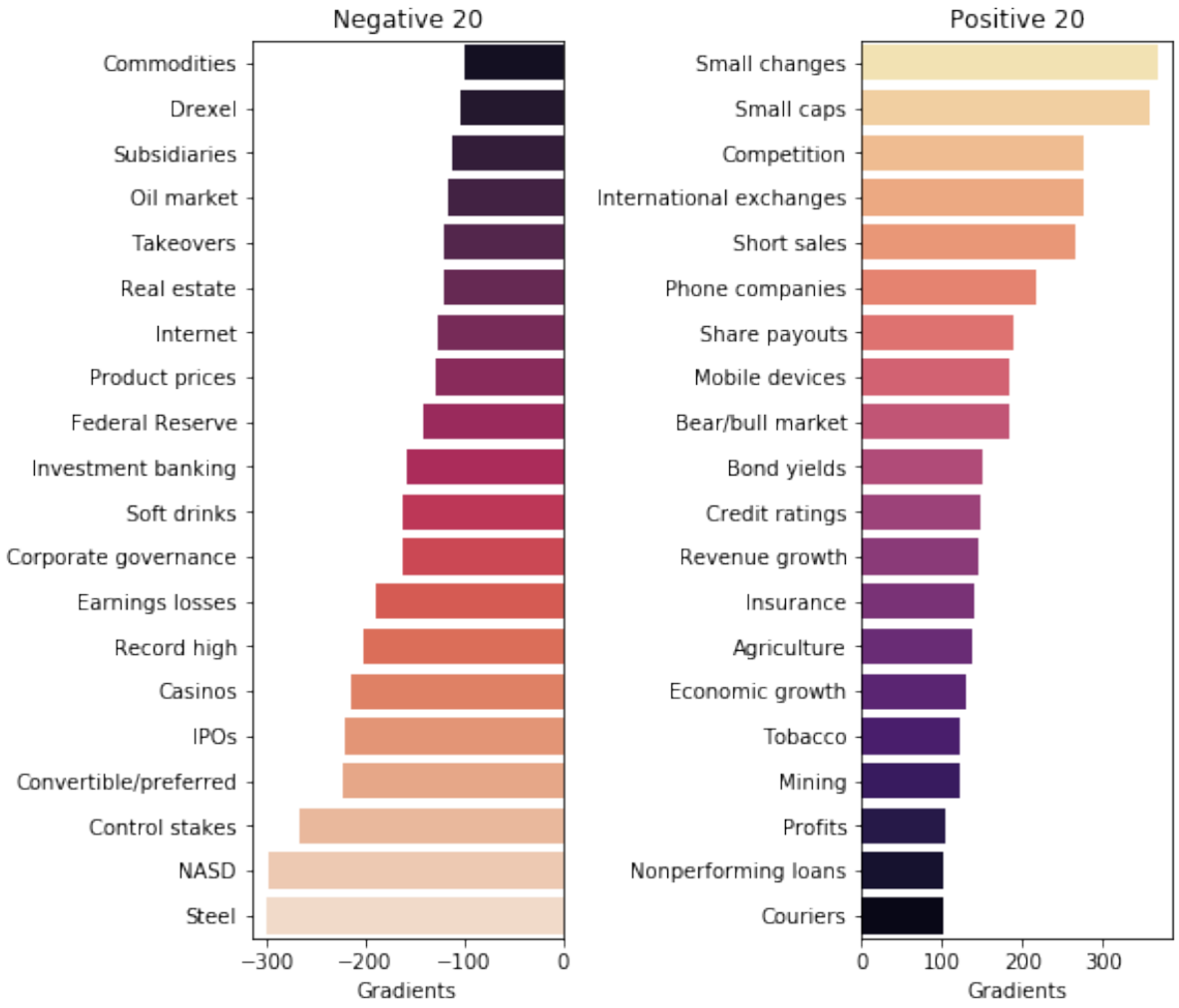

Figure 4: **Importance of Economy News in WGNN**

Note: This figure displays the importance of economy news topic attention trained in the Weighted Group Neural Network (WGNN) for overlapping bond returns.To alleviate randomness in training deep learning models and to make group importance measure robust, we repeat the above procedure 30 times using the most recent data and report the average gradient values.

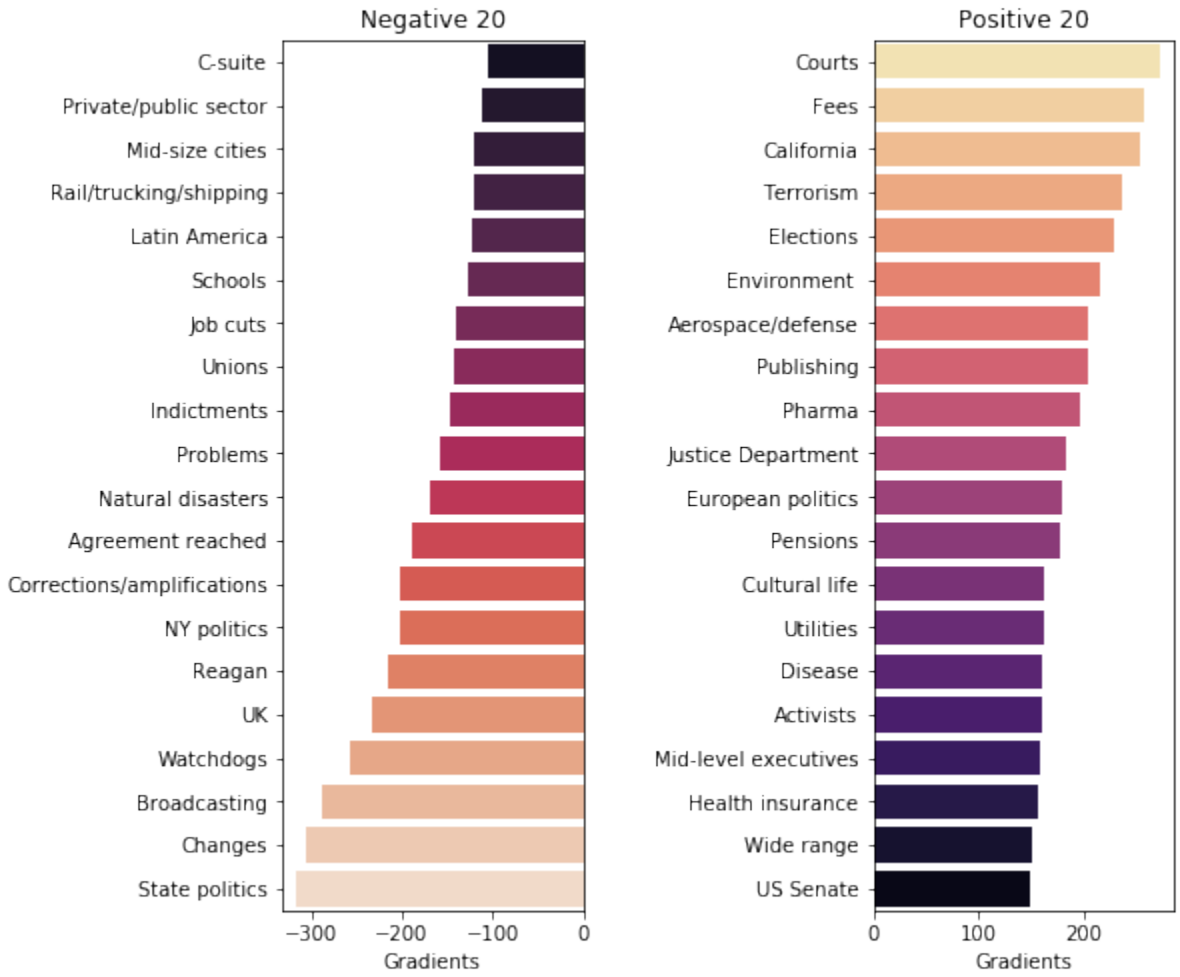

Figure 5: **Importance of Politics and Culture News in WGNN**

Note: This figure displays the importance of political and cultural news topic attention trained in the Weighted Group Neural Network (WGNN) for overlapping bond returns.To alleviate randomness in training deep learning models and to make group importance measure robust, we repeat the above procedure 30 times using the most recent data and report the average gradient values.

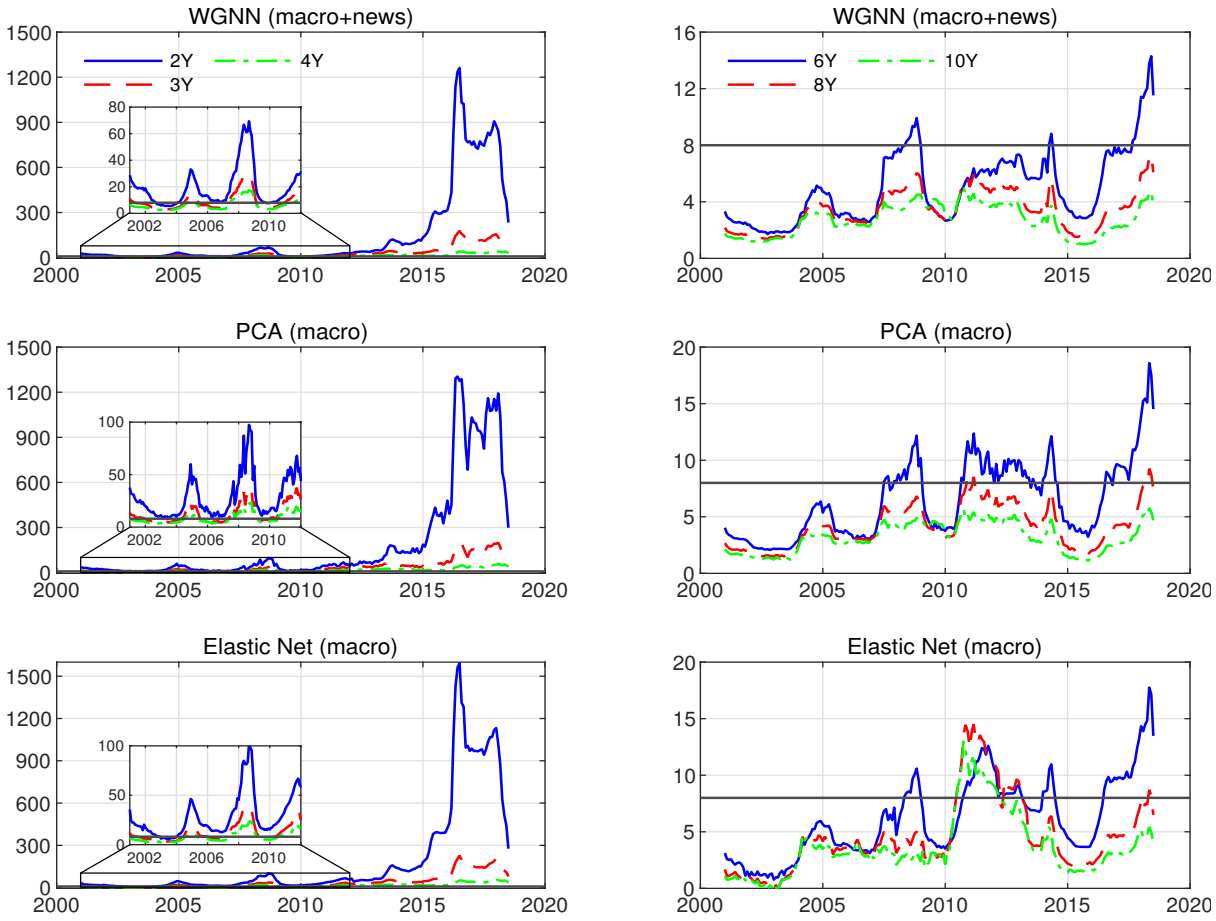

Figure 6: **Time Series of Unrestricted Portfolio Weights**

Note: This figure presents the unrestricted portfolio weights at each time in the out-of-sample period in WGNN, PCA, and Elastic Net for forecasting overlapping bond returns. The bold horizontal lines are the upper bound of 8 of the portfolio weights. The coefficient of relative risk-aversion is equal to 5. The out-of-sample period ranges from December 2000 to June 2018.

Appendix A

Table A1 presents a detailed description of the real-time macro variables used in the paper. The real-time macro data are downloaded from the Archival Federal Reserve Economic Database (ALFRED). For each macro variable, we provide the ALFRED mnemonic, variable description, and the transformation code used to stationarize the data as in McCracken and Ng (2016), that is, 1 for level (no transformation needed), 2 for first difference, 3 for second difference, 4 for natural log, 5 for first difference of natural log, 6 for second difference for natural log, and 7 for first difference of percentage change. All macro variables are classified into eight groups.

Table A1: **List of Macro Variables**

No.	Tcode	Fred	Description	Group
1	5	RPI	Real Personal Income	1
2	5	W875RX1	Real personal income ex transfer receipts	1
3	5	DPCERA3M086SBEA	Real personal consumption expenditures	4
4	5	CMRMTSPLx	Real Manu. and Trade Industries Sales	4
5	5	RETAILx	Retail and Food Services Sales	4
6	5	INDPRO	IP Index	1
7	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies	1
8	5	IPFINAL	IP: Final Products (Market Group)	1
9	5	IPCONGD	IP: Consumer Goods	1
10	5	IPDCONGD	IP: Durable Consumer Goods	1
11	5	IPNCONGD	IP: Nondurable Consumer Goods	1
12	5	IPBUSEQ	IP: Business Equipment	1
13	5	IPMAT	IP: Materials	1
14	5	IPDMAT	IP: Durable Materials	1
15	5	IPNMAT	IP: Nondurable Materials	1
16	5	IPMANSICS	IP: Manufacturing (SIC)	1
17	5	IPFUELS	IP: Fuels	1
18	2	CUMFNS	Capacity Utilization: Manufacturing	1
19	2	HWI	Help-Wanted Index for United States	2
20	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed	2
21	5	CLF16OV	Civilian Labor Force	2
22	5	CE16OV	Civilian Employment	2
23	2	UNRATE	Civilian Unemployment Rate	2
24	2	UEMPMEAN	Average Duration of Unemployment (Weeks)	2
25	5	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	2
26	5	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	2
27	5	UEMP15OV	Civilians Unemployed - 15 Weeks & Over	2
28	5	UEMP15T26	Civilians Unemployed for 15-26 Weeks	2
29	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	2
30	5	CLAIMSx	Initial Claims	2
31	5	PAYEMS	All Employees: Total nonfarm	2
32	5	USGOOD	All Employees: Goods-Producing Industries	2
33	5	CES1021000001	All Employees: Mining and Logging: Mining	2
34	5	USCONS	All Employees: Construction	2
35	5	MANEMP	All Employees: Manufacturing	2
36	5	DMANEMP	All Employees: Durable goods	2
37	5	NDMANEMP	All Employees: Nondurable goods	2
38	5	SRVPRD	All Employees: Service-Providing Industries	2
39	5	USTPU	All Employees: Trade, Transportation & Utilities	2
40	5	USWTRADE	All Employees: Wholesale Trade	2
41	5	USTRADE	All Employees: Retail Trade	2
42	5	USFIRE	All Employees: Financial Activities	2
43	5	USGOVT	All Employees: Government	2
44	1	CES0600000007	Avg Weekly Hours : Goods-Producing	2
45	2	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	2
46	1	AWHMAN	Avg Weekly Hours : Manufacturing	2
47	4	HOUST	Housing Starts: Total New Privately Owned	3
48	4	HOUSTNE	Housing Starts, Northeast	3
49	4	HOUSTMW	Housing Starts, Midwest	3
50	4	HOUSTS	Housing Starts, South	3
51	4	HOUSTW	Housing Starts, West	3
52	4	PERMIT	New Private Housing Permits (SAAR)	3
53	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)	3
54	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)	3
55	4	PERMITS	New Private Housing Permits, South (SAAR)	3
56	4	PERMITW	New Private Housing Permits, West (SAAR)	3
57	5	ACOGNO	New Orders for Consumer Goods	4
58	5	AMDMNOx	New Orders for Durable Goods	4
59	5	ANDENOx	New Orders for Nondefense Capital Goods	4
60	5	AMDMUOx	Unfilled Orders for Durable Goods	4

No.	Tcode	Fred	Description	Group
61	5	BUSINVx	Total Business Inventories	4
62	2	ISRATIOx	Total Business: Inventories to Sales Ratio	4
63	6	M1SL	M1 Money Stock	5
64	6	M2SL	M2 Money Stock	5
65	5	M2REAL	Real M2 Money Stock	5
66	6	TOTRESNS	Total Reserves of Depository Institutions	5
67	7	NONBORRES	Reserves Of Depository Institutions	5
68	6	BUSLOANS	Commercial and Industrial Loans	5
69	6	REALLN	Real Estate Loans at All Commercial Banks	5
70	6	NONREVSL	Total Nonrevolving Credit	5
71	2	CONSPI	Nonrevolving consumer credit to Personal Income	5
72	5	S&P 500	S&P500 Common Stock Price Index: Composite	8
73	5	S&P: indust	S&P500 Common Stock Price Index: Industrials	8
74	2	S&P div yield	S&P500 Composite Common Stock: Dividend Yield	8
75	5	S&P PE ratio	S&P500 Composite Common Stock: Price-Earnings Ratio	8
76	2	FEDFUNDS	Effective Federal Funds Rate	6
77	2	CP3Mx	3-Month AA Financial Commercial Paper Rate	6
78	2	TB3MS	3-Month Treasury Bill:	6
79	2	TB6MS	6-Month Treasury Bill:	6
80	2	GS1	1-Year Treasury Rate	6
81	2	GS5	5-Year Treasury Rate	6
82	2	GS10	10-Year Treasury Rate	6
83	2	AAA	Moody's Seasoned Aaa Corporate Bond Yield	6
84	2	BAA	Moody's Seasoned Baa Corporate Bond Yield	6
85	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS	6
86	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	6
87	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	6
88	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS	6
89	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS	6
90	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS	6
91	1	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	6
92	1	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	6
93	5	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate	6
94	5	EXJPUSx	Japan / U.S. Foreign Exchange Rate	6
95	5	EXUSUKx	U.S. / U.K. Foreign Exchange Rate	6
96	5	EXCAUSx	Canada / U.S. Foreign Exchange Rate	6
97	6	WPSFD49207	PPI: Finished Goods	7
98	6	WPSFD49502	PPI: Finished Consumer Goods	7
99	6	WPSID61	PPI: Intermediate Materials	7
100	6	WPSID62	PPI: Crude Materials	7
101	6	OILPRICEx	Crude Oil, spliced WTI and Cushing	7
102	6	PPICMM	PPI: Metals and metal products:	7
103	6	CPIAUCSL	CPI : All Items	7
104	6	CPIAPPSL	CPI : Apparel	7
105	6	CPITRNSL	CPI : Transportation	7
106	6	CPIMEDSL	CPI : Medical Care	7
107	6	CUSR0000SAC	CPI : Commodities	7
108	6	CUSR0000SAD	CPI : Durables	7
109	6	CUSR0000SAS	CPI : Services	7
110	6	CPIULFSL	CPI : All Items Less Food	7
111	6	CUSR0000SA0L2	CPI : All items less shelter	7
112	6	CUSR0000SA0L5	CPI : All items less medical care	7
113	6	PCEPI	Personal Cons. Expend.: Chain Index	7
114	6	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	7
115	6	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	7
116	6	DSERRG3M086SBEA	Personal Cons. Exp: Services	7
117	6	CES0600000008	Avg Hourly Earnings : Goods-Producing	2
118	6	CES2000000008	Avg Hourly Earnings : Construction	2
119	6	CES3000000008	Avg Hourly Earnings : Manufacturing	2
120	2	UMCSENTx	Consumer Sentiment Index	4
121	6	MZMSL	MZM Money Stock	5
122	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	5
123	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding	5
124	6	INVEST	Securities in Bank Credit at All Commercial Banks	5
125	1	VXOCLSx	VXO	8

Table A2 contains the description of the news attention variables used in the paper. Bybee et al. (2021) propose an approach to measuring the state of the economy using the full-text content of Wall Street Journal articles from 1984 to 2017. The monthly data of topic attention is downloaded from their website <http://structureofnews.com/>.

Table A2: **List of News Topic Attention Variables**

No.	Topic label	Metatopic label	Group
1	IPOs	Financial Markets	9
2	Bond yields	Financial Markets	9
3	Short sales	Financial Markets	9
4	Small caps	Financial Markets	9
5	Treasury bonds	Financial Markets	9
6	International exchanges	Financial Markets	9
7	Exchanges/composites	Financial Markets	9
8	Currencies/metals	Financial Markets	9
9	Share payouts	Financial Markets	9
10	Bear/bull market	Financial Markets	9
11	Commodities	Financial Markets	9
12	Trading activity	Financial Markets	9
13	Options/VIX	Financial Markets	9
14	M&A	Buyouts & Bankruptcy	9
15	Control stakes	Buyouts & Bankruptcy	9
16	Drexel	Buyouts & Bankruptcy	9
17	SEC	Buyouts & Bankruptcy	9
18	Bankruptcy	Buyouts & Bankruptcy	9
19	Corporate governance	Buyouts & Bankruptcy	9
20	Takeovers	Buyouts & Bankruptcy	9
21	Real estate	Buyouts & Bankruptcy	9
22	Convertible/preferred	Buyouts & Bankruptcy	9
23	Mutual funds	Asset Managers/I-Banks	9
24	Accounting	Asset Managers/I-Banks	9
25	Investment banking	Asset Managers/I-Banks	9
26	Acquired investment banks	Asset Managers/I-Banks	9
27	Private equity/hedge funds	Asset Managers/I-Banks	9
28	NASD	Asset Managers/I-Banks	9
29	Savings & loans	Banks	9
30	Nonperforming loans	Banks	9
31	Credit ratings	Banks	9
32	Financial crisis	Banks	9
33	Bank loans	Banks	9
34	Mortgages	Banks	9
35	Record high	Economic Growth	9
36	Economic growth	Economic Growth	9
37	Federal Reserve	Economic Growth	9
38	European sovereign debt	Economic Growth	9
39	Recession	Economic Growth	9
40	Product prices	Economic Growth	9
41	Optimism	Economic Growth	9
42	Macroeconomic data	Economic Growth	9
43	Steel	Oil & Mining	9
44	Mining	Oil & Mining	9
45	Machinery	Oil & Mining	9
46	Oil drilling	Oil & Mining	9
47	Agriculture	Oil & Mining	9
48	Oil market	Oil & Mining	9
49	Profits	Corporate Earnings	9
50	Revised estimate	Corporate Earnings	9

No.	Topic label	Metatopic label	Group
51	Earnings losses	Corporate Earnings	9
52	Small changes	Corporate Earnings	9
53	Financial reports	Corporate Earnings	9
54	Earnings forecasts	Corporate Earnings	9
55	Earnings	Corporate Earnings	9
56	Soft drinks	Industry	9
57	Small business	Industry	9
58	Cable	Industry	9
59	Fast food	Industry	9
60	Competition	Industry	9
61	Chemicals/paper	Industry	9
62	Venture capital	Industry	9
63	Tobacco	Industry	9
64	Subsidiaries	Industry	9
65	Credit cards	Industry	9
66	Couriers	Industry	9
67	Foods/consumer goods	Industry	9
68	Insurance	Industry	9
69	Luxury/beverages	Industry	9
70	Casinos	Industry	9
71	Revenue growth	Industry	9
72	Internet	Technology	9
73	Mobile devices	Technology	9
74	Electronics	Technology	9
75	Phone companies	Technology	9
76	Computers	Technology	9
77	Software	Technology	9
78	Microchips	Technology	9
79	Executive pay	Labor/Income	10
80	Job cuts	Labor/Income	10
81	Unions	Labor/Income	10
82	Health insurance	Labor/Income	10
83	Pensions	Labor/Income	10
84	Government budgets	Labor/Income	10
85	Fees	Labor/Income	10
86	Taxes	Labor/Income	10
87	Connecticut	Management	10
88	C-suite	Management	10
89	Mid-level executives	Management	10
90	Management changes	Management	10
91	Natural disasters	Trans/Defense/Local	10
92	Police/crime	Trans/Defense/Local	10
93	Mid-size cities	Trans/Defense/Local	10
94	NY politics	Trans/Defense/Local	10
95	Rail/trucking/shipping	Trans/Defense/Local	10
96	California	Trans/Defense/Local	10
97	Rental properties	Trans/Defense/Local	10
98	Disease	Trans/Defense/Local	10
99	US defense	Trans/Defense/Local	10
100	Pharma	Trans/Defense/Local	10
101	Aerospace/defense	Trans/Defense/Local	10
102	Automotive	Trans/Defense/Local	10
103	Airlines	Trans/Defense/Local	10
104	Retail	Trans/Defense/Local	10
105	Political contributions	Government	10
106	Regulation	Government	10
107	Environment	Government	10
108	Private/public sector	Government	10
109	State politics	Government	10
110	Watchdogs	Government	10
111	Utilities	Government	10
112	Safety administrations	Government	10
113	National security	Government	10
114	Justice Department	Courts	10
115	Indictments	Courts	10
116	Courts	Courts	10
117	Lawsuits	Courts	10
118	Clintons	Political Leaders	10
119	US Senate	Political Leaders	10
120	Reagan	Political Leaders	10

No.	Topic label	Metatopic label	Group
121	Bush/Obama/Trump	Political Leaders	10
122	Elections	Political Leaders	10
123	European politics	Political Leaders	10
124	Middle east	Terrorism/Mideast	10
125	Nuclear/North Korea	Terrorism/Mideast	10
126	Terrorism	Terrorism/Mideast	10
127	Iraq	Terrorism/Mideast	10
128	Russia	International Affairs	10
129	Trade agreements	International Affairs	10
130	Latin America	International Affairs	10
131	Japan	International Affairs	10
132	Canada/South Africa	International Affairs	10
133	China	International Affairs	10
134	Southeast Asia	International Affairs	10
135	Germany	International Affairs	10
136	France/Italy	International Affairs	10
137	UK	International Affairs	10
138	Music industry	Entertainment	10
139	Broadcasting	Entertainment	10
140	Publishing	Entertainment	10
141	Marketing	Entertainment	10
142	Movie industry	Entertainment	10
143	Economic ideology	Social/Cultural	10
144	Schools	Social/Cultural	10
145	Sales call	Social/Cultural	10
146	Cultural life	Social/Cultural	10
147	Arts	Social/Cultural	10
148	Immigration	Social/Cultural	10
149	Positive sentiment	Social/Cultural	10
150	Humor/language	Social/Cultural	10
151	Gender issues	Social/Cultural	10
152	Changes	Challenges	10
153	Key role	Challenges	10
154	Problems	Challenges	10
155	Challenges	Challenges	10
156	Small possibility	Challenges	10
157	Spring/summer	Challenges	10
158	Long/short term	Challenges	10
159	Research	Science/Language	10
160	Scenario analysis	Science/Language	10
161	Programs/initiatives	Science/Language	10
162	Biology/chemistry/physics	Science/Language	10
163	Space program	Science/Language	10
164	Systems	Science/Language	10
165	Size	Science/Language	10
166	Wide range	Science/Language	10
167	Activists	Activism/Language	10
168	Announce plan	Activism/Language	10
169	Major concerns	Activism/Language	10
170	Futures/indices	Activism/Language	10
171	Corrections/amplifications	Activism/Language	10
172	Buffett	Activism/Language	10
173	Mexico	Activism/Language	10
174	Restraint	Negotiations	10
175	News conference	Negotiations	10
176	Company spokesperson	Negotiations	10
177	People familiar	Negotiations	10
178	Agreement reached	Negotiations	10
179	Committees	Negotiations	10
180	Negotiations	Negotiations	10