International Journal of Forecasting 35 (2019) 1370–1386

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values

David Ardia [a,b], Keven Bluteau [a,c,*], Kris Boudt [c,d,e]

[a] Institute of Financial Analysis, University of Neuchâtel, Neuchâtel, Switzerland
[b] Department of Decision Sciences, HEC Montreal, Canada
[c] Solvay Business School, Vrije Universiteit Brussel, Belgium
[d] School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands
[e] Department of Economics, Ghent University, Belgium

ARTICLE INFO

Keywords:
Elastic net
Sentiment analysis
Time series aggregation
Topic-sentiment
US industrial production
Sentometrics

ABSTRACT

The modern calculation of textual sentiment involves a myriad of choices as to the actual calibration. We introduce a general sentiment engineering framework that optimizes the design for forecasting purposes. It includes the use of the elastic net for sparse data-driven selection and the weighting of thousands of sentiment values. These values are obtained by pooling the textual sentiment values across publication venues, article topics, sentiment construction methods, and time. We apply the framework to the investigation of the value added by textual analysis-based sentiment indices for forecasting economic growth in the US. We find that the additional use of optimized news-based sentiment values yields significant accuracy gains for forecasting the nine-month and annual growth rates of the US industrial production, compared to the use of high-dimensional forecasting techniques based on only economic and financial indicators.

1. Introduction

Understanding the current and future states of the economy is crucial for timely and efficient economic policy and business decision-making. Forecasts of economic variables such as the country's gross domestic product, industrial production, consumer spending, and the unemployment rate are followed closely by policymakers in order to assess the state of the economy. It seems self-evident that not only the readily-available quantitative information but also the qualitative information in news reports is useful for obtaining this assessment.

In practice, however, the dominant approach is to use exclusively the available quantitative information for economic growth prediction. In fact, most often, the macroeconomic variables are forecast using a large panel of macroeconomic indicators that reflects the economic environment; see Stock and Watson (2002). In addition, surveys such as the University of Michigan Consumer Sentiment Index or the Conference Board's Consumer Confidence Index for the US, and the European Economic Sentiment Index (ESI) for Europe, can contain information about the current and future economic growth. The US survey-based sentiment indices are used by Bram and Ludvigson (1998) and Ludvigson (2004) for forecasting US household expenditure and consumer spending, while the ESI is used by Gelper and Croux (2010) for forecasting national and aggregated European industrial production growth rates. Finally, financial indicators that reflect economic and financial expectations, as well as credit conditions, are used by Espinoza, Fornari, and Lombardi (2012) for forecasting long-term US and Euro area GDP growth.

This paper complements the readily-available quantitative information (i.e., macroeconomic, financial, and survey-based indicators) with predictors obtained from a

* Correspondence to: University of Neuchâtel, Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland.
E-mail addresses: david.ardia@unine.ch (D. Ardia), keven.bluteau@unine.ch (K. Bluteau), kris.boudt@vub.be (K. Boudt).

https://doi.org/10.1016/j.ijforecast.2018.10.010

large set of sentiment values expressed by authors of news discussing a country's economy, with the aim of obtaining timely forecasts of the country's economic growth. The approach starts off with a rich (big) data environment of a virtually infinite number of texts. These texts need to be selected, transformed into sentiment values, and then aggregated. The potential high-dimensionality of the data becomes an issue, as we are interested only in extracting the relevant information from the text and creating informative indices for the prediction of economic growth.

We address this challenge by proposing a methodology that starts by computing thousands of sentiment values which capture the tone expressed by the authors of news items that discuss topics related to the country's economic growth. It then maps the hordes of sentiment values into a single economic growth prediction using aggregation based on (1) a sentiment computation method (e.g., using various lexicons), (2) the topic (e.g., "real estate market" or "job creation") and (3) time (e.g., short and long-term sentiment indices). We then use a data-driven calibration approach based on penalized least squares regression to combine the indices optimally so as to forecast a variable of interest. We refer to the resulting optimized aggregate value of the sentiment as a text-based sentiment index. This index is a linear combination of the original sentiment values. This is a choice of design that allows us to perform an attribution analysis of the sentiment prediction in order to gauge the contributions of the various textual sentiment indices to the prediction.

In addition to being flexible, timely, and data-rich, the proposed methodology has the advantage that its design can be backtested. In a real-time setting, its design adapts itself to the changing forecasting environment; that is, the weights attributed to each component of the final sentiment index change according to the economic environment and the targeted variable to be forecast. Gelper and Croux (2010) find that letting the aggregation weights of each component of the survey-based ESI be data-driven improves its forecasting performance compared to the ad-hoc weights set by the European Commission. This feature is integrated into our textual sentiment index by construction. Furthermore, it also removes to a certain extent most of the subjective decisions that a forecaster has to make before the forecasting exercise. Indeed, the optimization process automatically chooses which sentiment computation methods are used for each topic (topic-specific sentiment calculation), which topic is included in the textual-sentiment index (removal of non-predictive topics), and how past values of each component of the textual-sentiment index are considered (structured lag per component). Thus, this adaptive scheme is more flexible than text-based (sentiment) indices with fixed designs, like the Economic Policy Uncertainty (EPU) index of Baker, Bloom, and Davis (2016). Moreover, the latter is not optimized for forecasting and not aimed at the extraction of sentiment.

This paper contributes to the increasing body of literature on the use of text- and news-based measures as sources of information for forecasting and assessing the economy (see e.g. Baker et al., 2016; Shapiro, Sudhof, & Wilson, 2018; Thorsrud, 2016, 2018; Tobback, Naudts, Daelemans, de Fortuny, & Martens, 2018). We exploit the sentiment information in news articles incrementally to the information included in the macroeconomic indicators. Two approaches exist for dealing with the high-dimensionality of the latter. First, via dimensionality reduction through (dynamic) factor models (for a review, see e.g. Stock & Watson, 2011). In this case, one assumes that a small number of unobserved factors drive the economy. Many methods have been developed for tackling the problem of estimating the latent factors (see Bräuning & Koopman, 2014; Doz, Giannone, & Reichlin, 2011, 2012; Stock & Watson, 2002) and choosing the appropriate number of factors (see Alessi, Barigozzi, & Capasso, 2010; Bai & Ng, 2002). Second, via penalized regression models used as a replacement for or in conjunction with factor models. Bai and Ng (2008) combine penalized regression with factor models for first selecting a set of predictors and then constructing the factors from them. Different variants of this approach are tested by Kim and Swanson (2014, 2018), and Smeekes and Wijler (2018). The proposed optimization of textual sentiment can be applied in conjunction with those traditional methods for a wide set of forecasting problems.

We illustrate the methodology for the case of forecasting economic growth for the United States, and find that, for an out-of-sample evaluation window from January 2001 to December 2016, the text-based sentiment indices computed from the news in major US newspapers have additional predictive power for the nine-month and annual growth rates of the US industrial production index, controlling for the standard use of macroeconomic, sentiment-survey, and financial variables. Moreover, we test the extent to which each dimension of the sentiment index (sentiment calculation method, topic, and time) matters, and find that the optimization of all dimensions is important for achieving a high forecasting accuracy, but that the most relevant is the time dimension, followed by the topic and then the sentiment calculation method. Our result is shown to be robust to various choices of implementation.

In an attempt to disseminate the methodology and render the results reproducible, we have released the R package **sentometrics** (Ardia, Bluteau, Borms, & Boudt, 2017, 2018), which implements all of the steps described in this paper in the R statistical language with efficient C++ code. We hope that this paper and the accompanying package will encourage practitioners such as government institutions and academics to use and test our framework for optimizing the use of textual sentiment for forecasting their variable(s) of interest.

The rest of the paper proceeds as follows. Section 2 introduces the methodology. Section 3 presents the empirical study. Section 4 concludes.

2. Methodology

The variable being predicted is the h-period logarithmic change in the variable Y_t, expressed in percentage points:

$$y_t^h \equiv 100 \times (\ln Y_{t+h} - \ln Y_t), \tag{1}$$

where $t = 1, 2, \ldots, T$ is a time index. We require y_t^h to be covariance stationary, which is typically the case when

Table 1
Total number of documents related to a given topic.

Topic	#	Cluster	Topic	#	Cluster
ECONOMIC CONDITIONS	25,522	1	IMPORT TRADE	15,709	3
COMPANY EARNINGS	20,116	1	INTEREST RATES	14,018	3
RECESSION	15,907	1	PRICE INCREASES	12,233	3
COMPANY PROFITS	11,075	1	INFLATION	11,841	3
SALES FIGURES	8,051	1	CURRENCIES	10,281	3
ECONOMIC GROWTH	7,904	1	PRICE CHANGES	9,363	3
BUDGET DEFICITS	6,656	1	ECONOMIC POLICY	7,270	3
OUTPUT & DEMAND	6,200	1	BOND MARKETS	4,027	3
MANUFACTURING OUTPUT	4,924	1	COMMODITIES PRICES	1,264	3
ECONOMIC STIMULUS	3,798	1	DEBT CRISIS	841	3
GROSS DOMESTIC PRODUCT	3,541	1	HOUSING MARKET	14,296	4
ECONOMIC DECLINE	2,818	1	REAL ESTATE DEVELOPMENT	11,144	4
CONSUMPTION	530	1	HOME PRICES	10,133	4
WAGES & SALARIES	37,157	2	CONSUMER CONFIDENCE	3,623	5
EMPLOYMENT	23,993	2	ECONOMIC SURVEYS	963	5
EMPLOYMENT GROWTH	11,708	2	BUSINESS CLIMATE & CONDITIONS	790	5
UNEMPLOYMENT RATES	10,070	2	BUSINESS CONFIDENCE	75	5
JOB CREATION	7,846	2	RETAILERS	32,695	6
PRICES	49,207	3	OIL & GAS INDUSTRY	20,384	6
EXPORT TRADE	19,390	3	MANUFACTURING FACILITIES	12,889	6
OIL & GAS PRICES	17,784	3	UTILITY RATES	3,215	6
INTERNATIONAL TRADE	17,029	3	RETAIL SECTOR PERFORMANCE	896	6
Number of topics			44		
Number of articles			338,408		
Average number of topics per article			1.50		

Notes: The table presents the numbers of articles in the corpus from major US newspapers that are related to a given topic. The list of topics is selected manually from the full list of topics identified by the *LexisNexis* SmartIndexing™ classifier, which provides a set of topics to each article in the database. Non-economics-related topics have been removed, resulting in a corpus that focuses exclusively on the US economy. Documents with fewer than 200 words are removed. Note that each article may be related to multiple topics. Topics are also organized into clusters of topics. The clusters are constructed manually and identified as: 1: GDP output, 2: Job market, 3: Prices & interest rate, 4: Real estate, 5: Surveys, 6: Others.

Y_t represents a country's economic activity (e.g., its gross domestic product or industrial production) or price level (e.g., the consumer price index or the exchange rate), and similarly for corporate variables, like a firm's sales or stock price. In our application, y_t^h is the logarithmic growth in industrial production of the US over horizons ranging from one to twelve months. Note that, due to the publication lag, Y_t may not be known at time t.

Let T be the day for which we need a prediction of y_T^h. Specifically, we want to estimate the expected value of y_T^h given the information available at time T; that is, $\mathbb{E}(y_T^h \mid \mathcal{I}_T)$. This is a common problem in time series forecasting, where the information set $\mathcal{I}_T$ typically consists of the usual available quantitative information, such as past values of Y_t, as well as macroeconomic and financial metrics (see e.g. Espinoza et al., 2012; Stock & Watson, 2002). However, we expand the information set by also including various sentiment values extracted from a corpus of texts published up to date T. We describe the methodology, as depicted in Fig. 1, below.

2.1. Data preparation

Step 1: *Classify texts by topic and use expert opinion to choose a subset of topics in order to select the potentially relevant texts.* We assume that all texts are categorized by a set of topic-markers. These topic-markers are usually provided by the publishers of the texts or extracted from the texts directly. In our application, we use the corpus of major US newspapers from *LexisNexis* for which topics are readily available using *LexisNexis*' proprietary SmartIndexing™ technology. Alternative techniques for topic identification include the use of likelihood-based techniques using probabilistic models such as the latent Dirichlet allocation (for a recent review, see Liu, Tang, Dong, Yao, & Zhou, 2016). For example, the latent Dirichlet allocation has been used recently by Thorsrud (2016) in conjunction with the dynamic factor model developed by Thorsrud (2018) for nowcasting the Norwegian GDP growth. It also includes keywords-based identification such as the keywords that Baker et al. (2016) used for identifying EPU-related texts, or, if topic-labelled news are available for a training set, identification via a support vector machine classifier such as per Tobback et al. (2018).[1] Expert opinion is then used to exclude the topics that can be qualified beforehand as being irrelevant for forecasting the variable of interest y_T^h. The resulting topic-markers for our application to forecasting economic growth are reported in Table 1. The corpus consists of the texts that discuss at least one of the selected topics. The corpus is organized in terms of the publication date t, with $t = 1, \ldots, T$, where N_t is the number of texts in the corpus of texts that were published at time t. We use n to index the texts available at time t, with $n = 1, \ldots, N_t$.

Step 2: *Compute the sentiment for each text n of corpus t using L methods.* For each text, we compute the underlying sentiment using L different textual sentiment computation methods. For a general review of the available methods, we refer the reader to Ravi and Ravi (2015). The methods

[1] Other high-accuracy machine-learning classification methods are, of course, viable.

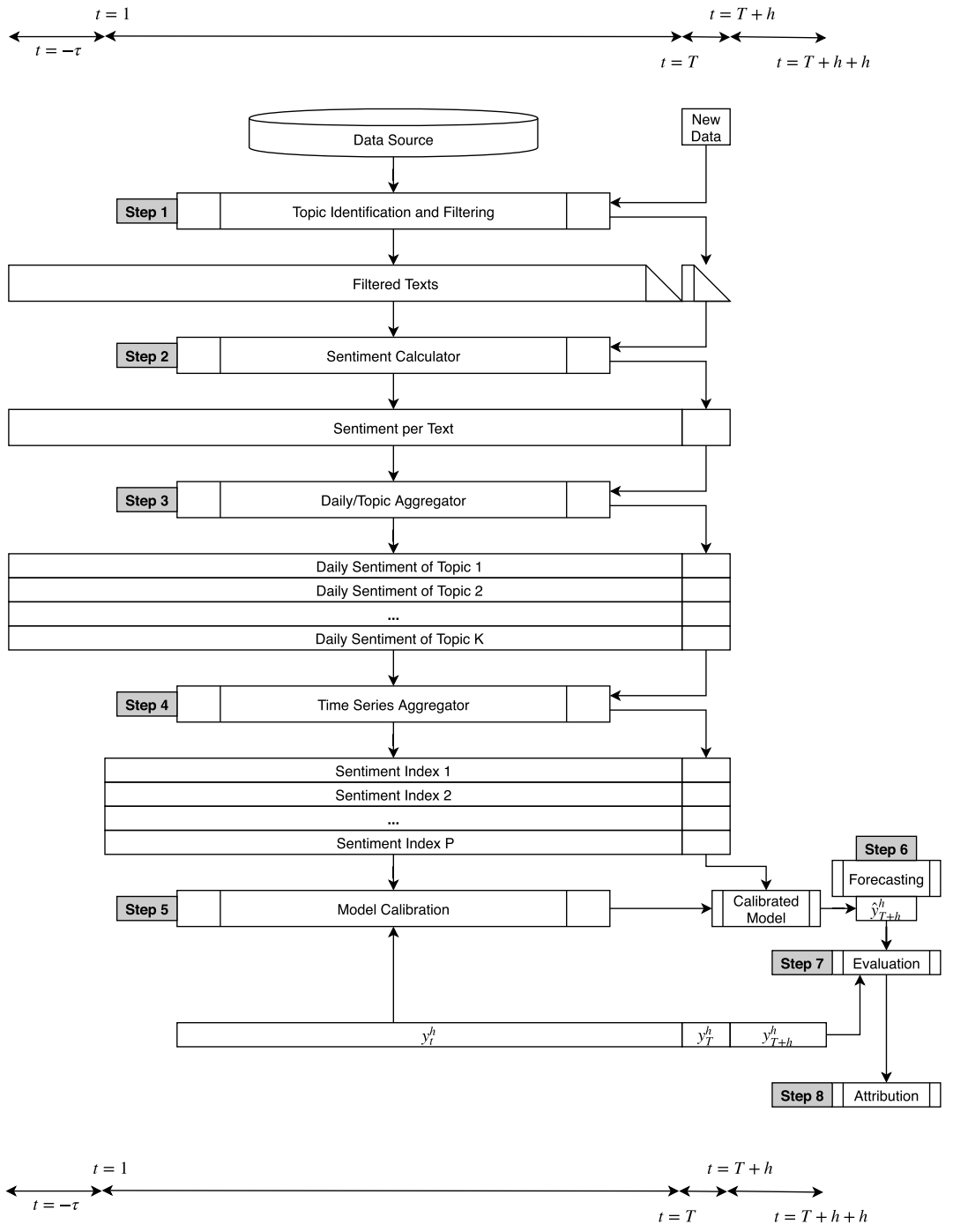

Fig. 1. Methodology. Notes: The figure displays the nine steps of the methodology in diagram form.

can differ from each other in terms of the item classified (e.g., word, sentence, paragraph), the method of classification (e.g., supervised or unsupervised), and the aggregation method used to obtain a single value per text (e.g., equal-weighting, inverse frequency weighting), among others. In our application, we use the simple bag-of-words approach to compute the net sentiment using L different lexicons to classify the words as positive, negative, or neutral, and thus obtain L different sentiment values for each text document $n = 1, \ldots, N_t$, published at time $t = 1, \ldots, T$, which we denote by $s_{n,t,l}$, for $l = 1, \ldots, L$.

2.2. Aggregating sentiment into a prediction

At this stage, we have L textual sentiment computation methods, and thus L vectors $\mathbf{s}_{t,l} \equiv (s_{1,t,l}, \ldots, s_{N_t,t,l})'$ of

size $N_t \times 1$, for each day t and each of the N_t texts. The next steps aim to reduce the dimensionality of the available texts (i.e., the total number of texts is $N_1 + \cdots + N_T$). To that end, we first compute the daily sentiment per topic-marker by aggregating across the sentiment of texts published on a given day, then aggregate over time. We choose a linear mapping, as this allows us to perform sentiment attribution. We do not use aggregation to reduce the dimensionality of the number of methods L, as it is small compared to the cross-section and time series dimensions, and can be handled at the estimation stage through penalized regression.

Step 3: *Obtain K topic-based sentiments for each corpus n and method l.* We compute sentiment values for each topic-marker by aggregating across the sentiment values of the texts associated with each topic-marker. Formally, we define the text-to-topic aggregation matrix $\mathbf{W}_t$ of dimension $K \times N_t$ for each day t such that the L vectors $\mathbf{W}_t\mathbf{s}_{t,l}$ ($l = 1, \ldots, L$) of dimension $K \times 1$ capture the daily sentiment for each of the K topics. In the application, each row of $\mathbf{W}_t$ is divided by its total sum, which corresponds to equally weighting the texts for each topic. The equal-weighting approach has the advantage of simplicity. An alternative approach for calibrating the text-to-topic aggregation matrix $\mathbf{W}_t$ could be to use expert opinion or a data-driven procedure to overweight the sources of news (i.e., type of journal or publisher) that are deemed more informative for predicting economic growth.

Step 4: *Obtain time series aggregated values for each topic k and method l.* Next, we aggregate over time. We take a maximum time-aggregation lag τ ($0 \leq \tau < T$), and, for a given l, stack the vectors column-by-column into $K \times (\tau+1)$ matrices as follows:

$$\mathbf{V}_{t,l} \equiv \begin{bmatrix} | & & | \\ \mathbf{W}_{t-\tau}\mathbf{s}_{t-\tau,l} & \cdots & \mathbf{W}_t\mathbf{s}_{t,l} \\ | & & | \end{bmatrix}. \tag{2}$$

We do this for $l = 1, \ldots, L$, and then stack the matrices row-by-row into a $LK \times (\tau + 1)$ matrix:

$$\mathbf{V}_t \equiv \begin{bmatrix} \mathbf{V}_{t,1} \\ \vdots \\ \mathbf{V}_{t,L} \end{bmatrix}. \tag{3}$$

Given $\mathbf{V}_t$ and a suitable time aggregation matrix $\mathbf{B}$ of size $(\tau+1) \times B$, we then construct the final vector of size $LKB \times 1$ of textual sentiment predictors $\mathbf{s}_t$ as:

$$\mathbf{s}_t \equiv \text{vec}(\mathbf{V}_t\mathbf{B}), \tag{4}$$

where $\text{vec}(\cdot)$ is the vectorization operator.[2]

We use a data-driven calibration of the aggregation matrix $\mathbf{B}$ to strike a balance between a strong decay in weights for obtaining timeliness on the one hand, and, on the other hand, an equal-weighting approach for obtaining efficiency when all time-lags are equally informative. To do so, we rely on the Beta weighting function, which is often used in the mixed-data sampling literature (see Ghysels, Sinko, & Valkanov, 2007). The approach requires two parameters $a > 0$ and $b > 0$:

$$c(i; a, b) \equiv \frac{f(\frac{i}{\tau}; a, b)}{\sum_{i=1}^{\tau} f(\frac{i}{\tau}; a, b)}, \tag{5}$$

where $f(x; a, b) \equiv \frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the Beta density function and $\Gamma(\cdot)$ is the Gamma function.

Given a grid $\{a_i, b_i\}_{i=1}^{B}$, the $(\tau+1) \times B$ aggregation matrix is given by:

$$\mathbf{B} \equiv \begin{bmatrix} c(1; a_1, b_1) & & c(1; a_B, b_B) \\ \vdots & \cdots & \vdots \\ c(\frac{i}{\tau}; a_1, b_1) & & c(\frac{i}{\tau}; a_B, b_B) \\ \vdots & \cdots & \vdots \\ c(0; a_1, b_1) & & c(0; a_B, b_B) \end{bmatrix}. \tag{6}$$

Step 5: *Calibrate to optimize the forecast precision.* The next and final aggregation step is to aggregate these textual sentiment indices optimally, given a variable of interest. To this end, we define the following model:

$$y_t^h = \alpha + \boldsymbol{\gamma}'\mathbf{x}_t + \boldsymbol{\beta}'\mathbf{s}_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{7}$$

where α is an intercept, $\mathbf{x}_t$ is a $M \times 1$ vector of (non-textual sentiment) variables available at time t, $\boldsymbol{\gamma}$ is the corresponding vector of parameters, $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_P)'$ is a vector of parameters associated with the P textual-sentiment indices ($P = LKB$), and ε_t is an error term at time t. Typically, $\mathbf{x}_t$ includes y_s, where y_s is the dependent variable up to time t, that is $s \leq t$. In practice, in economics we often have $s < t$ due to the release lag faced by economic indicators. It is also common to include macroeconomic and financial metrics, or the information obtained from surveys.

We use a penalized least squares criterion to estimate the regression in Eq. (7). Penalization is needed in order to regularize the estimation of the high-dimensional parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Given the high correlation between the sentiment variables, we use the elastic net regularization of Zou and Hastie (2005) to deal with both the high degree of collinearity in the regressors and the need for variable selection.[3]

For ease of presentation, let us define $\mathbf{z}_t \equiv (\mathbf{x}_t', \mathbf{s}_t')'$ and $\boldsymbol{\theta} \equiv (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$, both of size $(M + P) \times 1$. In our context, the optimization problem of the elastic net can then be expressed as:

$$\min_{\widetilde{\alpha}, \widetilde{\boldsymbol{\theta}}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left[y_t^h - \left(\widetilde{\alpha} + \widetilde{\boldsymbol{\theta}}'\widetilde{\mathbf{z}}_t \right) \right]^2 + \lambda_1 \left[\lambda_2 \|\widetilde{\boldsymbol{\theta}}\|_1 + (1 - \lambda_2)\|\widetilde{\boldsymbol{\theta}}\|_2^2 \right] \right\}, \tag{8}$$

where $\|\cdot\|_p$ is the L^p-norm, $\lambda_1 \geq 0$ is the parameter that sets the level of regularization and $0 \leq \lambda_2 \leq 1$ is the

[2] The vectorization operator stacks the columns of a matrix into a vector one on top of another.

[3] All calibrations are performed using the R package **glmnet** (Friedman, Hastie, & Tibshirani, 2010). Various models with sparsity features exist, such as the adaptive elastic net of Zou and Zhang (2009). However, we find that these methods do not improve the forecasting performance significantly in our application to forecasting US growth.

weight between the two types of penalties. The elastic net regularization nests both the Ridge regularization of Hoerl and Kennard (1970) (when $\lambda_2 = 0$) and the LASSO regularization (when $\lambda_2 = 1$) introduced by Tibshirani (1996). The variable $\widetilde{\mathbf{z}}_t$ is the standardized version of $\mathbf{z}_t$ with components $\widetilde{z}_{i,t} \equiv (z_{i,t} - \text{av}_i)/\text{std}_i$, where av_i and std_i are the sample mean and standard deviation of $\{z_{i,t}; t = 1, \ldots, T\}$, respectively. The standardization is crucial in penalized regressions, as the penalty depends on the scale of the components of $\boldsymbol{\theta}$.

Once the estimation is done, $\widetilde{\boldsymbol{\theta}}$ is rescaled to give the corresponding optimal unstandardized vector $\widehat{\boldsymbol{\theta}}$. The unstandardized regression parameter can be recovered by rescaling each component of $\widetilde{\boldsymbol{\theta}}$; $\widehat{\theta}_i \equiv \frac{\widetilde{\theta}_i}{\text{std}_i}$ $(i = 1, \ldots, M + P)$. An additional value must then be subtracted from the regression intercept to account for the centering of the series:

$$\widehat{\alpha} \equiv \widetilde{\alpha} - \sum_{i=1}^{M+P} \frac{\widetilde{\theta}_i}{\text{std}_i} \text{av}_i \,. \tag{9}$$

The implementation of the elastic net in Eq. (8) requires the calibration of the penalty parameters λ_1 and λ_2. We follow Zou, Hastie, and Tibshirani (2007) and minimize the so-called BIC-like criterion, where BIC stands for Bayesian information criterion.[4] Let the vector $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ of size $T \times 1$ be the forecast of $\mathbf{y}^h \equiv (y^h_1, \ldots, y^h_T)'$ obtained by fixing λ_1 and λ_2. The BIC-like criterion is defined as:

$$BIC_{\lambda_1,\lambda_2} \equiv \frac{\|\mathbf{y}^h - \widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}\|_2^2}{T\sigma^2} + \frac{\ln T}{T} \widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}) \,, \tag{10}$$

where σ^2 is defined as the variance of the forecast error given by the largest $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$. In Eq. (10), $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$ is an estimator of the number of degrees of freedom of the elastic net given $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ (see Tibshirani & Taylor, 2012). In the special case where $\lambda_2 = 1$ (i.e., LASSO regularization), $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,1})$ is equal to the number of non-zero parameters.[5]

Step 6: *Forecasting.* As the estimator $\widehat{\boldsymbol{\theta}}$ contains the vectors $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$, our forecast at time T is given by:

$$\widehat{y}^h_T \equiv \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}'\mathbf{x}_T + \widehat{\boldsymbol{\beta}}'\mathbf{s}_T \,. \tag{11}$$

2.3. Forecast precision and attribution

Given the predicted values of y^h_T, it is critical to evaluate whether the computational cost of text-based prediction pays off in terms of a higher out-of-sample precision than when the forecast is obtained using a simpler time series model. Another step in validating the outcome is to attribute the contribution of each topic to the predicted value.

Step 7: *Forecast precision evaluation.* We evaluate the forecasting performance using the root mean squared forecast error (RMSFE) and the mean absolute forecast error (MAFE). Let $e^h_{i,t} \equiv y^h_t - \widehat{y}^h_{i,t}$ be the error term for model i at time t for a horizon h where $\widehat{y}^h_{i,t}$ is the forecast. The RMSFE and MAFE measures of model i at horizon h are defined by:

$$\text{RMSFE}^h_i \equiv \sqrt{\frac{1}{T_F}\sum_{t=T+1}^{T+T_F}(e^h_{i,t})^2}, \quad \text{MAFE}^h_i \equiv \frac{1}{T_F}\sum_{t=T+1}^{T+T_F}|e^h_{i,t}| \,, \tag{12}$$

where T is the size of the estimation sample and T_F is the number of out-of-sample observations.

Statistical techniques like the Diebold and Mariano (1995) (DM) test or the model confidence set (MCS) procedure of Hansen, Lunde, and Nason (2011) can then be used to evaluate the significance of the differences in forecasting precision between models.[6] When comparing nested models, as we do in our application, the p-value of the DM test has a non-standard distribution. We recommend the use of the critical values obtained using the bootstrap approach of Clark and McCracken (2001).

Step 8: *Attribution.* Thus far, our exposition has been a bottom-up story of aggregating the sentiments of individual texts into a prediction of economic growth through cross-sectional, time series, and elastic net weighting. Once this prediction has been obtained, it is important to attribute the obtained prediction to the individual texts from the top down at various granularity levels. In fact, thanks to the linearity of the methodology, it is straightforward to retrieve the forecast as a function of the individual text sentiment $s_{n,t,l}$:

$$\widehat{y}^h_T = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}'\mathbf{x}_T + \sum_{t=(T-\tau)}^{T}\sum_{n=1}^{N_t}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}'\mathbf{e}_{l,k,b}\cdot W_{t,k,n}B_{T-t,b}\cdot s_{n,t,l}, \tag{13}$$

where $\mathbf{e}_{l,k,b}$ is a basis vector of size $LKB \times 1$ that extracts the relevant regression parameter from $\widehat{\boldsymbol{\beta}}$ given l, k and b; $W_{t,k,n}$ is the (k, n)-element of $\mathbf{W}_t$; and $B_{T-t,b}$ is the $(T-t, b)$-element of the matrix $\mathbf{B}$. It is easy to see from Eq. (13) that the weight $\omega_{n,t,l}$ that is attributed to the sentiment $s_{n,t,l}$ is equal to:

$$\omega_{n,t,l} = \sum_{k=1}^{K}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}'\mathbf{e}_{l,k,b}\cdot W_{t,k,n}B_{T-t,b} \,, \tag{14}$$

[4] In our study, the small sample size and the cross-correlation generated by the overlapping data when $h > 1$ make the cross-validation calibration methodology unstable. We also test for other BIC-type criteria such as the extended BIC of Chen and Chen (2008) and the high-dimensional BIC of Wang and Zhu (2011). The performance does not improve significantly in our empirical application.

[5] We use a grid-search to find the pair $(\hat{\lambda}_1, \hat{\lambda}_2)$ that minimizes $BIC_{\lambda_1,\lambda_2}$. More specifically, we use the elements of the vector $\boldsymbol{\lambda}_2 \equiv (0, 0.1, 0.3, 0.5, 0.7, 0.9, 1)$ as candidate values of λ_2 and generate a vector $\boldsymbol{\lambda}_{1,\lambda_{2,i}}$ of size 100, where $\lambda_{2,i}$ is the ith element of $\boldsymbol{\lambda}_2$, for each value in $\boldsymbol{\lambda}_2$, using the strategy outlined by Friedman et al. (2010). This gives 100 pairs per candidate λ_2, for a total of 700 pairs ($\boldsymbol{\lambda}_2$ is of size seven). The pair (λ_1, λ_2) that uses the largest number of degrees of freedom to compute σ^2 is found by computing the degrees of freedom given by each pair. Then, the pair $(\hat{\lambda}_1, \hat{\lambda}_2)$ is the pair that minimizes $BIC_{\lambda_1,\lambda_2}$.

[6] In the DM approach, it is standard to implement the test statistic with a heteroscedasticity and autocorrelation robust (HAC) standard error estimator, such as per Andrews (1991) and Andrews and Monahan (1992), while the MCS approach relies on a (block) bootstrap estimator for the variance.

such that:

$$\widehat{y}_T^h = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}' \mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \omega_{n,t,l} \cdot s_{n,t,l} \,. \tag{15}$$

Clearly, it is not feasible to analyze all (n, t, l)-combinations. Thus, we proceed by grouping them by common attributes, such as the time or topic. For example, we can obtain the attribution of topic g $(1 \leq g \leq K)$ by fixing $k = g$ and computing the attribution by integrating the other dimensions:

$$a_g \equiv \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}' \boldsymbol{e}_{l,g,b} \cdot W_{t,g,n} B_{T-t,b} \cdot s_{n,t,l} \,. \tag{16}$$

3. Application to forecasting US economic growth

We illustrate the use of the complete optimized sentiment calibration framework for forecasting economic growth in the United States. Our corpus consists of all articles published in major US newspapers for which documents are available in *LexisNexis*.[7] We quantify the economic value of the sentiment calibration by evaluating the forecasting gains compared to benchmark approaches that use only the readily available quantitative macroeconomic and financial information in the merged datasets of McCracken and Ng (2016) and Goyal and Welch (2008). We begin by introducing the data and the models that we compare, then present our main results and interpret the attribution that we obtain.

3.1. Data and descriptive statistics

3.1.1. Quantitative data

We aim to forecast the log-growth of US industrial production at the one-month ($h = 1$), three-month ($h = 3$), six-month ($h = 6$), nine-month ($h = 9$), and twelve-month ($h = 12$) horizons. We transform the level of industrial production into the h-month log-growth in percentage points: $y_t^h \equiv 100 \times (\ln IP_{t+h} - \ln IP_t)$, where IP_t is the industrial production realized at time t. Fig. 2 presents the industrial production time series from January 1996 to December 2016.

The workhorse approach to the forecasting of economic growth is the factor model proposed by Stock and Watson (2002). It involves predicting economic growth using the most important principal components from a large panel of macroeconomic variables. Thus, we retrieve all economy-related time series from the FRED-MD historical vintage databases for every month from August 1999 to December 2016 (see McCracken & Ng, 2016). For vintages before August 1999, we use the data as of August 1999. FRED-MD is a large publicly available database of economic variables that satisfy the filtering criteria established by Stock and Watson (1996). The number of variables contained in the database ranges from 105 to 128 for our time period. These variables are divided into various categories; see Table A.1 of the Appendix for an example with the FRED-MD 2016–12 dataset. Using past vintages allows us to get rid of the look-ahead bias.[8]

In addition to the macroeconomic variables, we also consider financial indicators. We use the dataset of Goyal and Welch (2008), which consists of 16 financial metrics such as dividend ratios, long/short term yields, stock variances, etc. We add to this dataset the Chicago Board of Exchange's forward-looking volatility index (VIX).[9] Finally, we add to the list of variables the media-attention EPU index and six survey-based Conference Board indices (CB).[10] We apply standard transformations to render the variables stationary; see Table A.1 of the Appendix for details.

3.1.2. Qualitative data: corpus

We compute textual sentiment indices for the US by retrieving the set of news that consists of all English articles from "Major US Newspapers" in the *LexisNexis* database with reference to the US. The *LexisNexis* "Major US Newspapers" source category consists of the Daily News, the Journal of Commerce, Los Angeles Times, Orange County Register, Pittsburgh Post Gazette, St. Louis Post Dispatch, Star Tribune, Tampa Bay Times, Atlanta Journal–Constitution, Christian Science Monitor, Daily Oklahoman, New York Post, New York Times, Philadelphia Daily News, Philadelphia Inquirer, Tampa Tribune, Washington Post, and USA Today. The dates range from January 1, 1994, to December 31, 2016. We apply the following filters:

- We use the geographic location to ensure that we select only news that is relevant to the US (relevance score greater than or equal to 85 in *LexisNexis*).
- We use the topic filter and filter out non-economics-related topics.
- To be assigned to a topic, the news must have a major reference to the topic (relevance score greater than or equal to 85 in *LexisNexis*).
- Each article must have at least 200 words.

Table 1 presents the topics selected, the number of documents associated with each and a cluster categorization of each topic for a cluster-based attribution analysis. The final corpus amounts to a total of 338,408 articles and 44 topics over six clusters. The six clusters of topics, which have been constructed manually by identifying economic concepts that are closely related, are: "GDP output", "Job market", "Prices & interest rate", "Real estate", "Surveys",

[7] *LexisNexis* provides an easy way of searching for and collecting relevant news from over 26,000 news sources, including online content. Their SmartIndexing™ technology classifies each text by a wide range of meta-information, such as subject, company, person, and country, thus simplifying the collection process and reducing the chance of a false positive inclusion of news in the dataset or in a particular subject. More information can be found at https://www.nexis.com.

[8] Macroeconomic FRED-MD data are available from Michael McCracken's website at https://research.stlouisfed.org/econ/mccracken/fred-databases.

[9] Financial data are available from Amit Goyal's website at http://www.hec.unil.ch/agoyal and VIX data from the Federal Reserve Bank of St. Louis at https://fred.stlouisfed.org/series/VIXCLS.

[10] EPU data are available from http://www.policyuncertainty.com and CB data from https://www.conference-board.org/data/consumerconfidence.cfm. The CB data include the leading economic index, the coincident economic index, the lagging economic index, the employment trend index, the consumer confidence: present situation index, and the consumer confidence: expectations index.

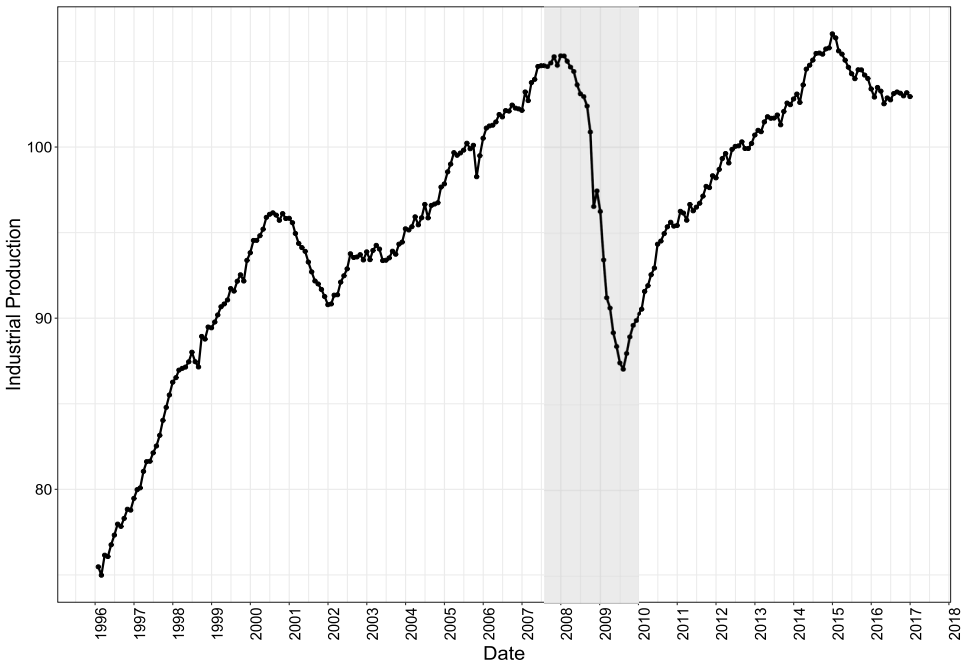

Fig. 2. US industrial production. Notes: The figure presents the US industrial production from January 1996 to December 2016 (192 monthly observations). The gray zone indicates the crisis period, which spans the period July 2007 to December 2009 (30 months).

and "Others". The last consists of all remaining topics. Note that a news article may refer to more than one topic, as the average number of topics per article is 1.50.[11]

3.1.3. Qualitative data: sentiment calculation

We measure the textual sentiment using standard lexicon-based sentiment analysis. The fundamental idea of lexicon-based sentiment analysis (also referred to as the bag-of-words approach) is the qualification of linguistic patterns (e.g., words or sentences) as positive, negative, or neutral using predefined lists called lexicons. Most studies use the Harvard General Inquirer lexicon (2550 positive words and 3695 negative words).[12] This dictionary is built independently of any particular narrative text and may not be the most suitable choice for text analysis of the economic domain. Thus, this implies the need to use specialized financial dictionaries for the analysis of financial and economic discourses, such as those developed by Henry (2008) (105 positive words and 85 negative words) and Loughran and McDonald (2011) (354 positive words and 2355 negative words).[13] We also use four lexicons that are popular in the sentiment analysis literature: (i) the SentiWordNet lexicon of Baccianella, Esuli, and Sebastiani (2016) (8898 positive words and 11,029 negative words), (ii) the SenticNet lexicon of Cambria, Poria, Bajpai, and Schuller (2016) (11,775 positive words and 11,852 negative words), (iii) the SO-CAL lexicon of Taboada, Brooke, Tofiloski, Voll, and Stede (2011) (1643 positive words and 1647 negative words), and (iv) the NRC lexicon of Mohammad and Turney (2010) (2227 positive words and 3241 negative words).[14]

Another aspect of sentiment analysis is valence-shifting words (see Polanyi & Zaenen, 2006). Valence-shifting words are words such as *"very"* or *"barely"*, that affect the context of nearby words. We only consider words that deal with negativity by inverting the sentiment of the first word following it from positive to negative and vice versa.[15]

Once the list of positive and negative words has been established, we then calculate the (net) sentiment of each text document as the relative spread between the numbers of positive and negative words:

$$s_{n,t,l} \equiv \frac{N^{+}_{n,t,l} - N^{-}_{n,t,l}}{N^{+}_{n,t,l} + N^{-}_{n,t,l} + N^{0}_{n,t,l}}, \tag{17}$$

where $N^{+}_{n,t,l}$ is the number of positive words in document n on day t for lexicon l, $N^{-}_{n,t,l}$ is the number of negative words, and $N^{0}_{n,t,l}$ is the number of neutral words. $N^{+}_{n,t,l}$ and $N^{-}_{n,t,l}$ can also be defined as the sum of the positive and negative scores, respectively, in the case where the lexicon weights the words according to the degree of positiveness

11 *LexisNexis* does not provide within-text topic identification, making it impossible to identify which part of the text discusses which topic. Ideally, one would have a single topic per text, to avoid contaminated sentiment indices.

12 The Harvard General Inquire lexicon is available at http://www.wjh.harvard.edu/~inquirer/homecat.htm.

13 The Loughran & McDonald lexicon is available at https://sraf.nd.edu/textual-analysis/resources.

14 These four lexicons are available through the R package **lexicons** (Rinker, 2018). SentiWordNet, SenticNet, and SO-CAL are weighted lexicons, where words are weighted according to their degree of positiveness or negativeness.

15 The list of negative valence-shifting words considered is: *ain't, aren't, can't, couldn't, didn't, doesn't, don't, hasn't, isn't, mightn't, mustn't, neither, never, no, nobody, nor, not, shan't, shouldn't, wasn't, weren't, won't, wouldn't.*

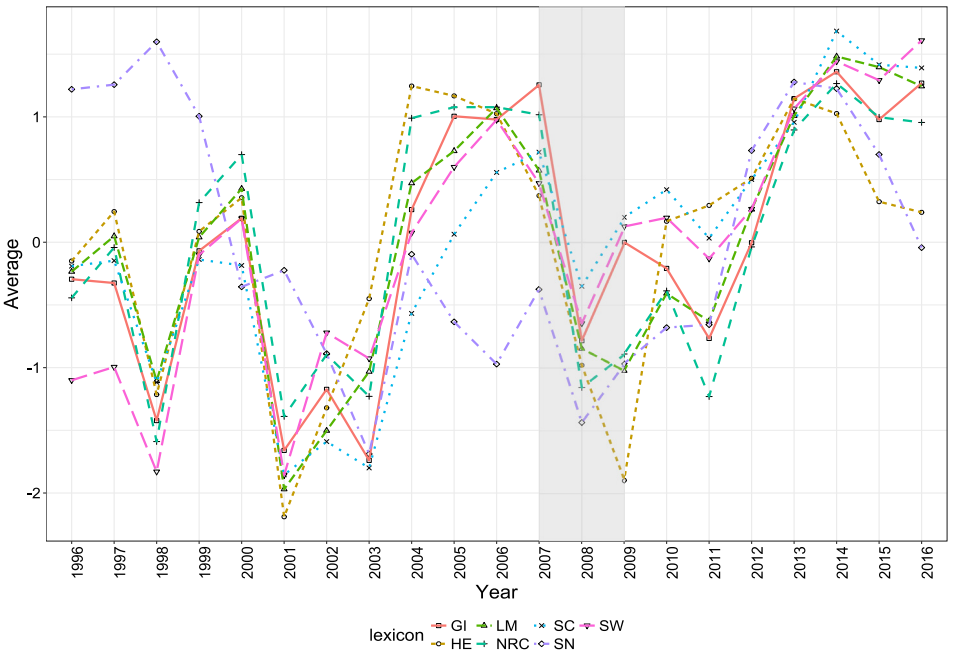

Fig. 3. Yearly lexicon-based averages of the individual news articles' sentiments. Notes: The figure presents the seven lexicon-based yearly averages of the individual news articles' sentiment for the period from 1994 to 2016. Sentiment values are standardized for readability purposes. The gray zone indicates the 2007–2009 crisis period.

and negativeness, in contrast to classifying them as either positive or negative.[16] This use of the net sentiment measure, computed as the difference between the frequencies of positive words (positive sentiment) and negative words (negative sentiment) normalized by the total number of words, is widespread in the literature (see e.g. Arslan-Ayaydin, Boudt, & Thewissen, 2016, and the references therein). Our application uses the net sentiment measure from seven lexicons, thus leading to $L = 7$ sentiment calculation methods.

Fig. 3 presents the yearly (standardized) averages of the individual news article sentiments computed using each of the seven lexicons in turn.[17] First, we see that the time-variation in the seven lexicon-based sentiment averages coincides with the economic cycle. In particular, we observe large common drops during the dot-com bubble burst of 2001 and the financial crisis of 2008. These events are preceded by large, almost linear, increases in the yearly average. In addition to the common behaviors of the seven lexicon-based indices, we also observe cross-sectional variability. The cross-sectional variability is to be expected, as no single lexicon offers a perfect estimate of the sentiment embedded in the text, and the words classified as positive and negative in each lexicon differ. Thus, we are reducing the risk of selecting the wrong lexicon simply by the reasoning that the choice of the lexicon would be irrelevant if no cross-sectional variation was observed.

[16] This is the case for the SentiWordNet, SenticNet, and SO-CAL lexicons, for example.

[17] The sentiment values are standardized (i.e., we subtracted the means and divided the series by their standard deviations) for readability purposes.

3.1.4. Qualitative data: aggregation of sentiment

We build the aggregation matrices $\mathbf{W}_t$ $(t = 1, \ldots, T)$ such that each of the 44 topics is summarized by a sentiment index. The time series aggregation matrix $\mathbf{B}$ contains Beta weights generated from the grid $\{1, 3, 4, 7\} \times \{1, 3, 4, 7\}$ for a total of 16 time-aggregation weights; see Fig. 4. We set the value $\tau = 180$ days. This gives a total of $P = LKB = 7 \times 44 \times 16 = 4928$ sentiment indices.

Fig. 5 presents the yearly average of the 44 topic-based sentiment indices calculated using the Loughran & McDonald lexicon.[18] Similarly to the yearly average of the non-aggregated sentiment shown in Fig. 3, we see a general decrease in all sentiment indices in the years 2001 and 2008. We also note a significant degree of variability in the cross-section of the yearly averages. This indicates that different topics may have different informational contents. This suggests that failing to consider the topic dimension and simply letting all news be part of an overarching topic could be sub-optimal, as we would lose important cross-sectional information.

3.2. Models

The forecasting models that we consider are nested in the linear framework in Eq. (7). The benchmark models $\mathcal{M}_{1a}$ and $\mathcal{M}_{2a}$ include the lagged value of the dependent variable and the macroeconomic, survey-based, and financial indicators ($\mathbf{x}_t$), or factors derived from those variables ($\mathbf{f}_t$). In addition, the alternative specifications $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ also include the 4928 textual-based sentiment indices ($\mathbf{s}_t$).

[18] We observe the same pattern for other lexicons.

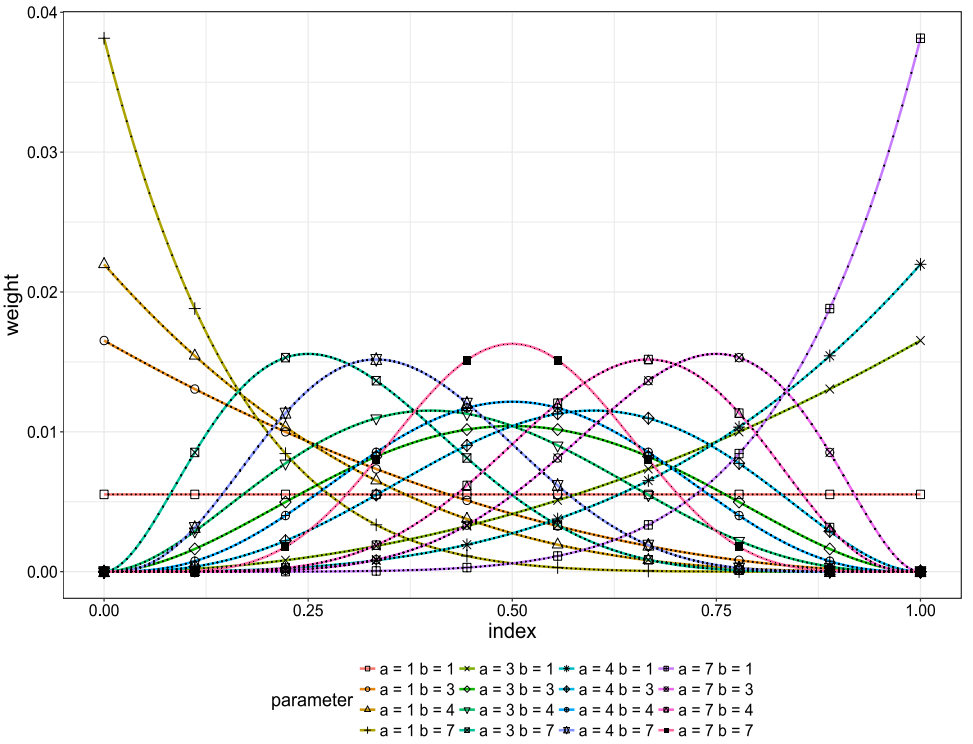

Fig. 4. Beta weights. Notes: The figure presents the time-aggregation weights of the Beta function for the grid $\{1, 3, 4, 7\} \times \{1, 3, 4, 7\}$ for a total of 16 weighting schemes.

More precisely, we study the following specifications:

$$\mathcal{M}_{1a}: \quad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)'\mathbf{x}_t + \varepsilon_t^h \tag{18}$$

$$\mathcal{M}_{1b}: \quad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)'\mathbf{x}_t + (\boldsymbol{\beta}^h)'\mathbf{s}_t + \varepsilon_t^h \tag{19}$$

and:

$$\mathcal{M}_{2a}: \quad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)'\mathbf{f}_t + \varepsilon_t^h \tag{20}$$

$$\mathcal{M}_{2b}: \quad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)'\mathbf{f}_t + (\boldsymbol{\beta}^h)'\mathbf{s}_t + \varepsilon_t^h \tag{21}$$

for $t = 1, \ldots, T$ months, where $\mathbf{f}_t$ are factors extracted from $\mathbf{x}_t$ using the IC_{p1} criterion of Bai and Ng (2002). This criterion performs well compared to the other candidate information criteria in the various Monte Carlo experiments of Bai and Ng (2002). More detail about the construction of the factors is provided in Appendix A.1.[19] Note that we are now dealing with a monthly frequency, as opposed to the daily frequency used in the construction of the sentiment indices.

All models are estimated using the elastic net procedure in Eq. (8). We enforce the inclusion of the lagged dependent variable in the model specification, and therefore exclude it from the penalization of the elastic net. Each model is estimated on a rolling window basis of 60 months.

Because of the overlapping nature of y_t^h when $h > 1$, we evaluate each model using the h-month-ahead observations. That is, if the sample window ranges from months $t = 1$ to $t = 60$, we evaluate the out-of-sample performance using the observation for month $t = 60 + h$.

The out-of-sample forecasting performance is evaluated using the RMSFE and MAFE measures. We evaluate $\mathcal{M}_{1b}$ ($\mathcal{M}_{2b}$) against $\mathcal{M}_{1a}$ ($\mathcal{M}_{2a}$) using the Diebold and Mariano (1995) test with the approach of Clark and McCracken (2001) for nested models at the 5% significance level.[20] We account for possible changes in out-of-sample forecasting performances over time by analyzing the full out-of-sample period and three sub-periods: pre-crisis, crisis, and post-crisis. The complete sample is from January 2001 (January 2003 for $h = 12$) to December 2016 (192 observations for $h = 1$ and 168 for $h = 12$). The pre-crisis period is from January 2001 (January 2003 for $h = 12$) to June 2007 (78 observations for $h = 1$ and 54 for $h = 12$). The crisis period is from July 2007 to December 2009 (30 observations), and finally, the post-crisis period is from January 2010 to December 2016 (84 observations).

3.3. Main results

3.3.1. Model's forecasting performance comparison

Table 2 presents the RMSFE and MAFE measures for the four model specifications and the five forecasting horizons

[19] We justify the use of principal components in conjunction with the Bai and Ng (2002) information criterion by noting that Smeekes and Wijler (2018) showed this method to perform well at forecasting the growth in the US industrial production relative to more complex factor and penalized regression models.

[20] The bootstrapped distribution is computed using 5000 block bootstrap samples, with the optimal block length determined from the fit of an autoregressive model. The variance of the mean loss difference is computed using the HAC standard error estimator of Andrews (1991) and Andrews and Monahan (1992).

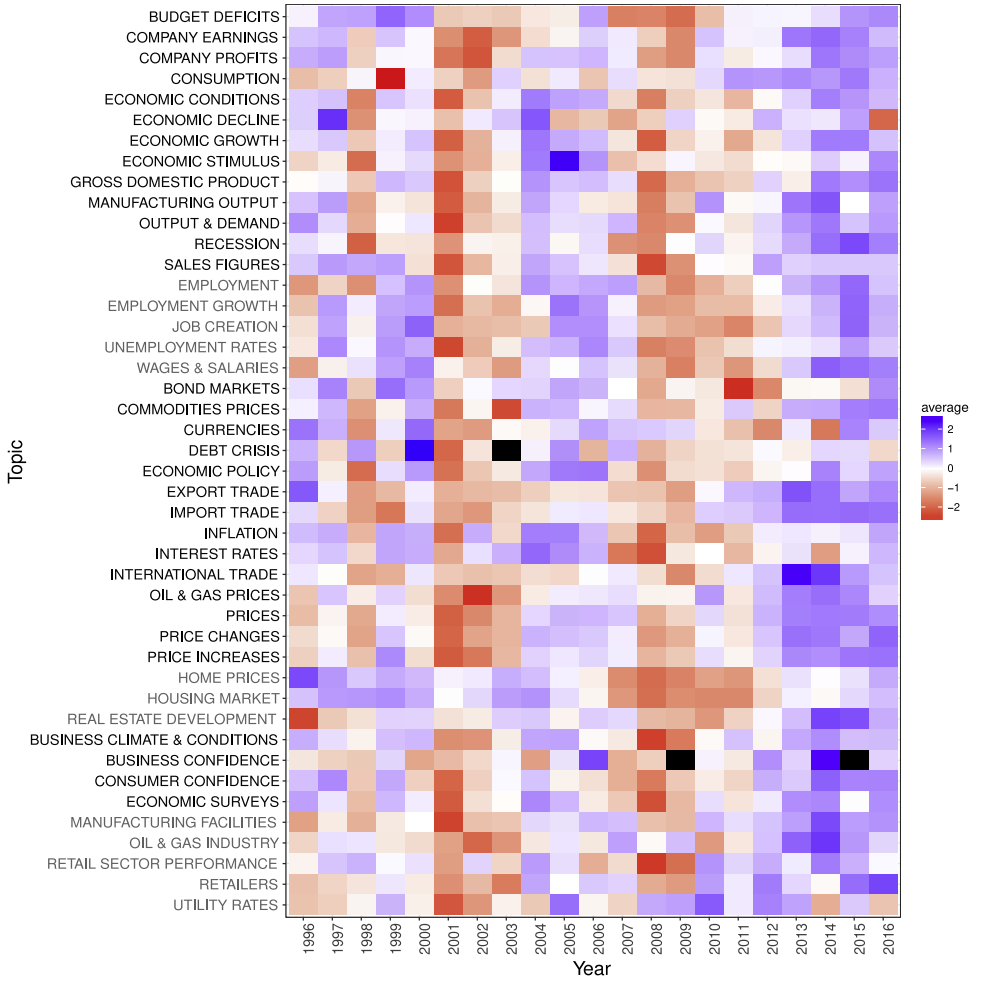

Fig. 5. Yearly average of the 44 topic sentiment indices. Notes: The figure presents the yearly average of 44 sentiment indices for the period from 1996 to 2016. Sentiment values are computed using the Loughran and McDonald (2011) lexicon. Each time series is standardized for the sake of comparability across topics. The topics are organized into clusters on the *y*-axis and delimited by black and gray text labeling. Black rectangles indicate that there is no news for that particular topic during that year.

over the four time windows. We focus our analysis on comparing the value added by using sentiment information, either as raw inputs (i.e., $\mathcal{M}_{1b}$ vs. $\mathcal{M}_{1a}$) or through factors ($\mathcal{M}_{2b}$ vs. $\mathcal{M}_{2a}$), when forecasting economic growth, controlling for readily available predictors. A gray cell indicates that the outperformance is statistically significant at the 5% significance level according to the DM test.

For the full sample, we see that textual sentiment-related specifications do not add forecasting power beyond that contained in the macroeconomic, survey-based, and financial indicators at the one- to six-month horizons. However, at the nine- to twelve-month horizons, they exhibit the best performances, and the results are significant according to the DM test for both the RMSFE and MAFE measures.

This gain in outperformance as the forecasting horizon grows was also observed by Ulbricht, Kholodilin, and Thomas (2017) for news-derived economic sentiment indices in the context of forecasting German industrial production, and is consistent with the "time-lag" effect in economics. While financial markets can react (quasi) instantaneously to the sentiment expressed in economic news, it takes time for that sentiment to affect economic behaviors (consumption, production, investments), and thus to become visible in the published economic growth figures (see George et al., 1999). This may explain why the sentiment becomes more predictive of economic growth over longer horizons.

Looking at the pre-crisis period, we can see that the textual-sentiment-related specifications outperform their benchmark according to the DM test at the twelve-month horizon. The post-crisis period, however, shows outperformance for the nine- and twelve-month horizons. Finally, unlike the other periods, during the crisis period the sentiment-related specification only shows outperformance at the six-month and nine-month horizons, and even then, only according to the RMSFE measure.

Overall, we observe that textual-sentiment-related specifications provide additional forecasting power over traditional macroeconomic, financial, and survey indicators at long horizons.

Table 2
Forecasting results.

		RMSFE				MAFE			
Period	h	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$
Full sample	1	0.68	0.70	0.64	0.70	0.49	0.49	0.45	0.49
	3	1.52	1.54	1.59	1.52	0.96	1.01	1.02	1.01
	6	4.86	3.93	5.01	3.14	2.36	2.35	2.85	2.14
	9	7.01	4.95	8.36	4.58	3.71	3.28	4.89	3.19
	12	6.39	5.19	8.69	5.14	4.25	3.41	6.03	3.32
Pre-crisis	1	0.55	0.57	0.56	0.56	0.43	0.42	0.43	0.44
	3	0.99	0.93	1.21	0.93	0.72	0.70	0.87	0.70
	6	1.67	1.65	2.62	1.62	1.31	1.36	1.80	1.32
	9	2.41	2.42	4.67	2.53	1.96	1.93	3.00	1.98
	12	3.27	2.00	6.07	1.90	2.72	1.67	3.73	1.57
Crisis	1	1.19	1.27	1.08	1.27	0.81	0.87	0.69	0.88
	3	3.20	3.19	3.17	3.04	2.46	2.52	2.31	2.29
	6	11.30	8.54	10.64	6.20	7.63	6.44	7.45	4.99
	9	8.58	7.94	9.92	7.94	6.67	6.20	7.67	6.20
	12	10.43	10.14	9.42	10.12	8.34	7.84	7.49	7.70
Post-crisis	1	0.53	0.50	0.49	0.50	0.42	0.40	0.40	0.41
	3	0.78	0.93	0.89	1.03	0.62	0.74	0.70	0.82
	6	1.72	2.26	2.86	2.32	1.32	1.68	2.05	1.77
	9	8.47	4.93	9.72	4.07	3.93	3.22	5.27	3.00
	12	6.02	3.85	9.81	3.78	3.80	2.98	7.01	2.90

Notes: The table presents the root mean squared forecast errors (RMSFE) and the mean absolute forecast errors (MAFE) for models $\mathcal{M}_{1a}$ (benchmark model with raw variables), $\mathcal{M}_{1b}$ ($\mathcal{M}_{1a}$ augmented by textual sentiments), $\mathcal{M}_{2a}$ (benchmark model with factors), and $\mathcal{M}_{2b}$ ($\mathcal{M}_{2a}$ augmented by textual sentiments). Lower RMSFE and MAFE values are preferred. We consider the one- ($h = 1$), three- ($h = 3$), six- ($h = 6$), nine- ($h = 9$), and twelve-month ($h = 12$) log-growth of the US industrial production. The full out-of-sample period is from January 2001 (January 2003 for $h = 12$) to December 2016 (192 observations for $h = 1$ and 168 for $h = 12$). The out-of-sample pre-crisis period is from January 2001 to June 2007 (78 observations for $h = 1$ and 54 for $h = 12$). The out-of-sample crisis period is from July 2007 to December 2009 (30 observations). The out-of-sample post-crisis period is from January 2010 to December 2016 (84 observations). A gray cell indicates that the extended model is superior to the benchmark model (i.e., $\mathcal{M}_{1b}$ against $\mathcal{M}_{1a}$ and $\mathcal{M}_{2b}$ against $\mathcal{M}_{2a}$) for a given horizon at the 5% significance level. Testing is based on the Diebold and Mariano (1995) test statistic implemented with the heteroscedasticity and autocorrelation robust (HAC) standard error estimators of Andrews (1991) and Andrews and Monahan (1992), and with p-values computed using the bootstrap, following Clark and McCracken (2001).

3.3.2. Attribution

A common criticism of big data approaches to economic forecasting is that their results seem to come from a "black box". This criticism is easy to counter in our setting, since the attribution analysis described in Step 8 of Section 2 allows us to pinpoint the contribution of each sentiment value to the growth prediction. Given a large number of sentiment values, we can analyze the attribution at the intermediate level of the grouping per cluster of topics from the categorization shown in Table 1.

Fig. 6 presents the normalized attributions of these clusters for the twelve-month forecasts obtained with model $\mathcal{M}_{1b}$, where we divide each of its elements by the L^2-norm of the attribution vector at that date.[21] This procedure makes it easier to perform comparisons across different dates. Note first that there is a persistence in the attribution of each cluster over time. This is consistent with the presence of stable information value in the selection and weighting used when engineering the textual sentiment index for predicting economic growth. Over the full sample, we find that "GDP output", "Price & interest rates" and "Survey" clusters contribute the most to the predicted growth, though they dominate the predictions at different times. In the pre-crisis period, texts published about "GDP output" are the main predictors. During the crisis, those discussing the surveys are selected and weighted to have the biggest impact on the predictions. Finally, post-crisis, the "Price & interest rates"-related texts dominate the predictions.

3.4. Importance of the optimization of each dimension

We now proceed to analyze the impacts of some of the modeling choices employed in our study.

We analyze the extent to which the optimization of the lexicon-, topic- and time-dimensions is relevant in predicting the industrial production growth. To that end, we compare the extended specifications $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ with four alternatives in which we (with equal weights) aggregate: (i) the lexicon dimension (denoted LEX), (ii) the topic dimension (denoted TOPIC), (iii) the time dimension (denoted TIME), and (iv) all dimensions (denoted ALL). Thus, the last approach is the naive way of calculating a

[21] The results for model $\mathcal{M}_{2b}$ are similar, and are available from the authors upon request.

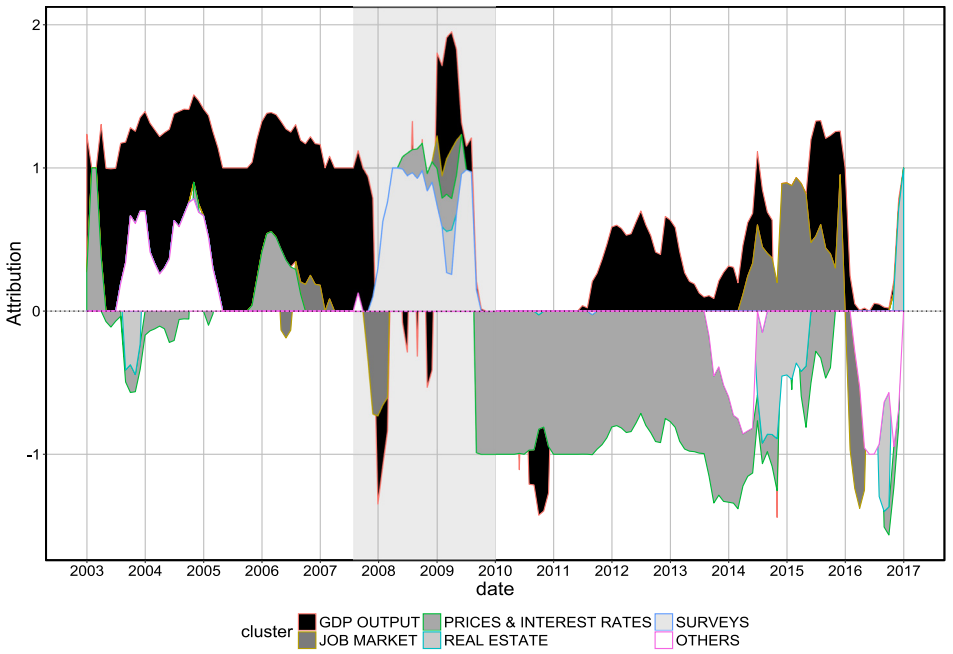

Fig. 6. Forecast attribution. Notes: The figure presents the cluster attribution of model $\mathcal{M}_{1b}$ for the out-of-sample forecasts of the twelve-month US industrial production log-growth. The period ranges from January 2003 to December 2016 (180 monthly observations). The attribution vector for a given date is scaled by dividing each element of the attribution vector by the L^2-norm of the attribution vector for that date. The gray zone indicates the July 2007 to December 2009 crisis period. A positive (negative) value indicates that the topic contributes positively (negatively) to the forecast, and therefore increases (decreases) the forecast of the US industrial production log-growth.

Table 3
Robustness results: aggregation of dimensions.

		RMSFE					MAFE				
	h	$\mathcal{M}$	LEX	TOPIC	TIME	ALL	$\mathcal{M}$	LEX	TOPIC	TIME	ALL
$\mathcal{M}_{1b}$	1	0.70	0.69	0.68	0.68	0.64	0.49	0.48	0.48	0.49	0.46
	3	1.54	1.50	1.41	1.52	1.58	1.01	0.98	0.93	0.96	0.99
	6	3.93	4.51	4.52	4.86	5.24	2.35	2.42	2.32	2.36	2.55
	9	4.95	5.91	5.57	7.01	8.37	3.28	3.43	3.28	3.71	4.17
	12	5.19	5.85	6.11	6.39	8.24	3.41	4.01	4.09	4.25	5.02
$\mathcal{M}_{2b}$	1	0.70	0.69	0.68	0.65	0.68	0.49	0.49	0.48	0.46	0.48
	3	1.52	1.53	1.50	1.39	1.31	1.01	1.06	1.07	0.95	0.92
	6	3.14	3.72	3.23	3.62	3.34	2.14	2.39	2.17	2.25	2.20
	9	4.58	5.65	5.36	6.99	6.23	3.19	3.74	3.42	4.16	4.06
	12	5.14	6.79	7.14	8.18	7.82	3.32	4.81	5.04	5.30	5.34

Notes: The table presents the forecasting results when the various dimensions (lexicon, topic, and time) are aggregated. We compare the results of the extended models $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ with those of four alternative approaches in which we (with equal weights) aggregate: (i) the lexicon-dimension (denoted LEX), (ii) the topic-dimension (denoted TOPIC), (iii) the time-dimension (denoted TIME), and (iv) all dimensions (denoted ALL). A light (dark) gray cell indicates that the extended model ($\mathcal{M}_{1b}$ or $\mathcal{M}_{2b}$) is superior (inferior) at the 5% significance level according to the Diebold and Mariano (1995) test statistic. See Table 2 for details.

sentiment index and adding it to the set of macroeconomic, survey, and financial variables. Note that these dimension reductions are only special cases of the methodology. The results for the full out-of-sample period are reported in Table 3. We can observe that, according to the lowest RMSFEs and MAFEs, the optimization of all dimensions is preferable. This is principally the case at the nine-month and particularly the twelve-month horizons. The time dimension seems to be the most important one to optimize, followed by the topic and lexicon dimensions.

4. Conclusion

Do textual sentiment indices provide any added value to the prediction accuracy of economic growth relative to the use of the information contained in macroeconomic, financial, or survey-based variables? Answering this question requires one first to capture the relevant sentiment-based growth prediction from a textual analysis of news releases. The latter is a big data problem, given the large number of texts that are published every day, the number

Table A.1
List of variables.

Index	Code	Variable	Description
Group 1: Output and income			
1	5	RPI	Real Personal Income
2	5	W875RX1	Real Personal Income ex Transfer Receipts
3	5	INDPRO	IP Index
4	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies
5	5	IPFINAL	IP: Final Products (Market Group)
6	5	IPCONGD	IP: Consumer Goods
7	5	IPDCONGD	IP: Durable Consumer Goods
8	5	IPNCONGD	IP: Nondurable Consumer Goods
9	5	IPBUSEQ	IP: Business Equipment
10	5	IPMAT	IP: Materials
11	5	IPDMAT	IP: Durable Materials
12	5	IPNMAT	IP: Nondurable Materials
13	5	IPMANSICS	IP: Manufacturing (SIC)
14	5	IPB51222s	IP: Residential Utilities
15	5	IPFUELS	IP: Fuels
16	1	NAPMPI	ISM Manufacturing: Production Index
17	2	CUMFNS	Capacity Utilization: Manufacturing
Group 2: Consumption and order			
1	2	HWI	Help–Wanted Index for United
2	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
3	5	CLF16OV	Civilian Labor Force
4	5	CE16OV	Civilian Employment
5	2	UNRATE	Civilian Unemployment Rate
6	2	UEMPMEAN	Average Duration of Unemployment (Weeks)
7	5	UEMPLT5	Civilians Unemployed – Less Than 5 Weeks
8	5	UEMP5TO14	Civilians Unemployed for 5–14 Weeks
9	5	UEMP15OV	Civilians Unemployed – 15 Weeks & Over
10	5	UEMP15T26	Civilians Unemployed for 15–26 Week
11	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over
12	5	CLAIMSx	Initial Claims
13	5	PAYEMS	All Employees: Total Nonfarm
14	5	USGOOD	All Employees: Goods–Producing Industries
15	5	CES1021000001	All Employees: Mining and Logging: Mining
16	5	USCONS	All Employees: Construction
17	5	MANEMP	All Employees: Manufacturing
18	5	DMANEMP	All Employees: Durable goods
19	5	NDMANEMP	All Employees: Nondurable goods
20	5	SRVPRD	All Employees: Service-Providing Industries
21	5	USTPU	All Employees: Trade, Transportation & Utilities
22	5	USWTRADE	All Employees: Wholesale Trade
23	5	USTRADE	All Employees: Retail Trade
24	5	USFIRE	All Employees: Financial Activities
25	5	USGOVT	All Employees: Government
26	1	CES0600000007	Avg Weekly Hours: Goods-Producing
27	2	AWOTMAN	Avg Weekly Overtime Hours: Manufacturing
28	1	AWHMAN	Avg Weekly Hours: Manufacturing
29	1	NAPMEI	ISM Manufacturing: Employment Index
30	6	CES0600000008	Avg Hourly Earnings: Goods–Producing
31	6	CES2000000008	Avg Hourly Earnings: Construction
32	6	CES3000000008	Avg Hourly Earnings: Manufacturing
Group 3: Housing			
1	4	HOUST	Housing Starts: Total New Privately Owned
2	4	HOUSTNE	Housing Starts, Northeast
3	4	HOUSTMW	Housing Starts, Midwest
4	4	HOUSTS	Housing Starts, South
5	4	HOUSTW	Housing Starts, West
6	4	PERMIT	New Private Housing Permits (SAAR)
7	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)
8	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)
9	4	PERMITS	New Private Housing Permits, South (SAAR)
10	4	PERMITW	New Private Housing Permits, West (SAAR)
Group 4: Orders and inventories			
1	5	DPCERA3M086SBEA	Real Personal Consumption Expenditures
2	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
3	5	RETAILx	Retail and Food Services Sales
4	1	NAPM	ISM: PMI Composite Index
5	1	NAPMNOI	ISM: New Orders Index
6	1	NAPMSDI	ISM: Supplier Deliveries Index
Group 4: Orders and inventories			
7	1	NAPMII	ISM: Inventories Index
8	5	ACOGNO	New Orders for Consumer Goods
9	5	AMDMNOx	New Orders for Durable Goods
10	5	ANDENOx	New Orders for Non-defense Capital Goods
11	5	AMDMUOx	Unrolled Orders for Durable Goods
12	5	BUSINVx	Total Business Inventories
13	2	ISRATIOx	Total Business: Inventories to Sales Ratio
14	2	UMCSENTx	Consumer Sentiment Index
Group 5: Money and credit			
1	6	M1SL	M1 Money Stock
2	6	M2SL	M2 Money Stock
3	5	M2REAL	Real M2 Money Stock
4	6	AMBSL	St. Louis Adjusted Monetary Base
5	6	TOTRESNS	Total Reserves of Depository Institutions
6	7	NONBORRES	Reserves Of Depository Institutions
7	6	BUSLOANS	Commercial and Industrial Loans
8	6	REALLN	Real Estate Loans at All Commercial Banks
9	6	NONREVSL	Total Nonrevolving Credit
10	2	CONSPI	Nonrevolving Consumer Credit to Personal Income
11	6	MZMSL	MZM Money Stock
12	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding
13	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding
14	6	INVEST	Securities in Bank Credit at All Commercial Banks
Group 6: Interest rate and exchange rates			
1	2	FEDFUNDS	Effective Federal Funds Rate
2	2	CP3Mx	3–Month AA Financial Commercial Paper Rate
3	2	TB3MS	3–Month Treasury Bill
4	2	TB6MS	6–Month Treasury Bill
5	2	GS1	1–Year Treasury Rate
6	2	GS5	5–Year Treasury Rate
7	2	GS10	10–Year Treasury Rate
8	2	AAA	Moody's Seasoned AAA Corporate Bond Yield AAA Bond
9	2	BAA	Moody's Seasoned BAA Corporate Bond Yield BAA Bond
10	1	COMPAPFFx	3–Month Commercial Paper Minus FEDFUNDS CP–FF Spread
11	1	TB3SMFFM	3–Month Treasury C Minus FEDFUNDS 3 mo–FF Spread
12	1	TB6SMFFM	6–Month Treasury C Minus FEDFUNDS 6 mo–FF Spread
13	1	T1YFFM	1–Year Treasury C Minus FEDFUNDS 1 yr–FF Spread
14	1	T5YFFM	5–Year Treasury C Minus FEDFUNDS 5 yr–FF Spread
15	1	T10YFFM	10–Year Treasury C Minus FEDFUNDS 10 yr-FF Spread
16	1	AAAFFM	Moody's AAA Corporate Bond Minus FEDFUNDS AAA–FF Spread
17	1	BAAFFM	Moody's BAA Corporate Bond Minus FEDFUNDS BAA–FF Spread
18	5	TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies Ex rate: avg
19	5	EXSZUSx	Switzerland/U.S. Foreign Exchange Rate
20	5	EXJPUSx	Japan/U.S. Foreign Exchange Rate
21	5	EXUSUKx	U.S./U.K. Foreign Exchange Rate
22	5	EXCAUSx	Canada/U.S. Foreign Exchange Rate
Group 7: Prices			
1	6	WPSFD49207	PPI: Finished Goods
2	6	WPSFD49502	PPI: Finished Consumer Goods
3	6	WPSID61	PPI: Intermediate Materials
4	6	WPSID62	PPI: Crude Materials
5	6	OILPRICEx	Crude Oil, Spliced WTI and Cushing
6	6	PPICMM	PPI: Metals and Metal Products:
7	1	NAPMPRI	ISM Manufacturing: Prices Index
8	6	CPIAUCSL	CPI: All Items
9	6	CPIAPPSL	CPI: Apparel
10	6	CPITRNSL	CPI: Transportation
11	6	CPIMEDSL	CPI: Medical Care
12	6	CUSR0000SAC	CPI: Commodities
13	6	CUSR0000SAD	CPI: Durables
14	6	CUSR0000SAS	CPI: Services
15	6	CPIULFSL	CPI: All Items Less Food
16	6	CUSR0000SA0L2	CPI: All Items Less Shelter
17	6	CUSR0000SA0L5	CPI: All Items Less Medical Care
18	6	PCEPI	Personal Cons. Exp: Chain Index
19	6	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods
20	6	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods
21	6	DSERRG3M086SBEA	Personal Cons. Exp: Services

(*continued on next page*)

Table A.1 (*continued*).

Index	Code	Variable	Description
Group 8: Stock market			
1	5	S&PIDX	S&P 500 Common Stock Price Index: Composite
2	5	S&PINDUS	Indust S&P's Common Stock Price Index: Industrials
3	2	S&PDIV	Div yield S&P's Composite Common Stock: Dividend Yield S&P Div Yield
4	2	S&PPE	PE ratio S&P' Composite Common Stock: Price-Earnings Ratio S&P PE Ratio
5	1	VXOCLSx	VXO
Group 9: Goyal and Welch (2008) ***financial variables***			
1	1	SR	S&P 500 Return
2	1	RF	Risk Free Rate
3	1	DP	Log Dividend on S&P 500 Minus log S&P 500
4	1	DY	Log Dividend on S&P 500 Minus log Lagged S&P 500
5	1	EP	Log Earnings on S&P 500 Minus log S&P 500
6	1	DP	Log Dividend on S&P 500 Minus log Earnings
7	1	SVAR	Sum of Square Return on the S&P 500
8	1	BM	Average Book Value of Dow Jones over the Down Jones industrial
9	1	NEE	12-Month Sum of net Issue on NYSE over Capitalization of NYSE
10	1	TB	Treasury Bills
Group 9: Goyal and Welch (2008) ***financial variables***			
11	1	LTY	Long Term Government Yield
12	1	LTR	Long Term Government Bond Rate of Return
13	1	TMS	Term Spread
14	1	DYS	Difference Between BAA and AAA-rated Corporate Bond Yield
15	1	DRS	Difference Between the Rate of Return of BAA and AAA–rated Corporate Bond
16	1	INF	Inflation
Group 10: Others			
1	2	VIX	VIX index
2	2	EPU	Economic Policy Uncertainty Index for the US
3	2	LEI	Conference Board Leading Economic Index
4	2	CEI	Conference Board Coincident Economic Index
5	2	LAG	Conference Board Lagging Economic Index
6	2	CCI	Conference Board Consumer Confidence Index
7	2	PSI	Conference Board Present Situation Index
8	2	EXI	Conference Board Expectations Index

This table summarizes the macroeconomic, financial, and additional media–attention and survey–based variables used in our study. The column "Code" refers to one of the following data transformations for a time series: 1: no transformation, 2: level–difference, 3: second level–difference, 4: log, 5: log–difference, 6: second log–difference, 7: growth rate. FRED–MD vintage datasets (Groups 1–8) are available from https://research.stlouisfed.org/econ/mccracken/fred-databases, financial variables (Group 9) from http://www.hec.unil.ch/agoyal, VIX from https://fred.stlouisfed.org/series/VIXCLS, EPU index from http://www.policyuncertainty.com, and Chicago conference board indices from https://www.conference-board.org/data/consumerconfidence.cfm (all in Group 10).

of possible historical dates on which news releases may have predictive value for the future economic activity, and the various methods of calculating sentiment. We show how to overcome this dimensionality issue by introducing a framework that optimizes sentiment aggregation for the prediction of economic growth using topics-based aggregation, time series aggregation, and predictive regressions by means of the elastic net regularization.

We test the predictive power of text-based sentiment indices by forecasting the growth in US industrial production using major newspapers from the news database *LexisNexis* over the period January 2001 to December 2016. We find that the proposed optimized text-based sentiment analysis can improve the forecasting performance for predicting the nine-month and annual growth rates significantly.

To help practitioners and academics to implement our methodology in practice, we have released the open-source R package **sentometrics** (Ardia et al., 2017, 2018). The package is designed in such a way that each step of the methodology, from sentiment calculation to time series aggregation, can be configured for specific needs. Thus, it not only allows one to replicate the configuration used in our empirical application, but also allows for extensions and modifications.

The potential scope of applications of the proposed optimized textual sentiment analysis framework goes far beyond the forecasting of economic growth. In future work, we will consider applying the framework to the quantification of brand reputation when forecasting firm sales, and the study of spillover effects between types of news media.

Acknowledgments

We are grateful to Andreas Alfons, Nabil Bouamara, Samuel Borms, Dries Cornilly, William Doehler, Siem Jan Koopman, and Wouter Torsin, as well as participants at the CFE 2017 conference, the R/Finance 2018 conference, and Brussels SoFiE summer School 2018 for helpful comments. We thank Swiss National Science Foundation (Grant #179281).

Appendix

A.1. Factor model

The high dimensionality of $\mathbf{x}_t$ means that it is practical to reduce the dimensionality of $\mathbf{x}_t$ by assuming that they are driven by a small number of common factors; see for instance Stock and Watson (2011). Let $\mathbf{X}_T \equiv \left[\mathbf{x}_1 \,|\, \cdots \,|\, \mathbf{x}_T\right]'$ be a $T \times M$ matrix of covariates and $\mathbf{F}_T \equiv \left[\mathbf{f}_1 \,|\, \cdots \,|\, \mathbf{f}_T\right]'$ be the $T \times R$ matrix of latent common factors of $\mathbf{X}_T$. We have the following regression problem:

$$\mathbf{X}_T = \mathbf{F}_T \boldsymbol{\Lambda} + \varepsilon_t \,, \tag{A.1}$$

where $\boldsymbol{\Lambda}$ is the $R \times M$ matrix of loadings and ε_t is an error term at time t. We estimate the latent factors by minimizing the following expression:

$$V(\mathbf{F}_T, \boldsymbol{\Lambda}) \equiv \frac{1}{MT} \sum_{i=1}^{M} \sum_{t=1}^{T} (x_{i,t} - \boldsymbol{\lambda}_i \mathbf{f}_t)^2 \,, \tag{A.2}$$

where $\boldsymbol{\lambda}_i$ is the ith row of $\boldsymbol{\Lambda}$. Under some assumptions, principal component (PC) analysis provides us with estimates of $\boldsymbol{\Lambda}$ and $\mathbf{F}_T$ with $R = \min\{M, T\}$. However, with PC, some factors can be considered as pure noise. We estimate the optimal number of factors R by minimizing the information criterion proposed by Bai and Ng (2002):

$$IC_{p1}(k) \equiv \ln\left(V(\widehat{\mathbf{F}}_T^k, \widehat{\boldsymbol{\Lambda}}^k)\right) + k \left(\frac{M+T}{MT}\right) \ln\left(\frac{MT}{M+T}\right) \,, \tag{A.3}$$

where $\widehat{\mathbf{F}}_T^k$ and $\widehat{\boldsymbol{\Lambda}}^k$ are the first k columns of the PC estimator of $\mathbf{F}_T$ and the first k rows of the PC estimator of $\boldsymbol{\Lambda}$. The value $k \in \{1, \ldots, k_{\max}\}$ that leads to the minimal IC_{p1} gives us the number of factors to use in the forecasting models $\mathcal{M}_{2a}$ and $\mathcal{M}_{2b}$ in Eqs. (20)–(21). We follow Bai and Ng (2002) and set $k_{\max} = 8$. Other values were tested but led to qualitatively similar results.

References

Alessi, L., Barigozzi, M., & Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, *80*(23–24), 1806–1813.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*(3), 817–858.

Andrews, D. W. K., & Monahan, J. D. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *60*(4), 953–966.

Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2017). The R package sentometrics to compute, aggregate and predict with textual sentiment. Working paper.

Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2018). *Sentometrics: An integrated framework for textual sentiment time series aggregation and prediction, version 0.4*. URL https://CRAN.R-project.org/package=sentometrics.

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking and Finance*, *72*(72), S132–S147.

Baccianella, S., Esuli, A., & Sebastiani, F. (2016). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC (Vol. 10)* (pp. 2200–2204).

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

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

Bram, J., & Ludvigson, S. (1998). Does consumer confidence forecast household expenditure? A sentiment index horse race. *Economic Policy Review*, *4*(2), 59–78.

Bräuning, F., & Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, *30*(3), 572–584.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2666–2677).

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024.
Espinoza, R., Fornari, F., & Lombardi, M. J. (2012). The role of financial variables in predicting economic activity. *Journal of Forecasting*, *31*(1), 15–46.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.
Gelper, S., & Croux, C. (2010). On the construction of the European economic sentiment indicator. *Oxford Bulletin of Economics and Statistics*, *72*(1), 47–62.
George, E., King, M., Clementi, D., Budd, A., Buiter, W., Goodhart, C., et al. (1999). *The transmission mechanism of monetary policy*. Bank of England.
Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90.
Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508.
Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.
Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, *45*(4), 363–407.
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
Kim, H. H., & Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, *178*, 352–367.
Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354.
Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1608), 1–22.
Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65.
Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives*, *18*(2), 29–50.
McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.
Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Association for Computational Linguistics.
Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In J. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: theory and applications*. In *The information retrieval series*: *Vol. 20*, (pp. 1–10). Springer-Verlag.
Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46.
Rinker, T. W. (2018). *Lexicon: Lexicon data, version 1.0.0*. URL http://github.com/trinker/lexicon.
Shapiro, A. H., Sudhof, M., & Wilson, D. (2018). Measuring news sentiment. Working paper.
Smeekes, S., & Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, *34*(3), 408–430.
Stock, J. H., & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, *14*(1), 11–30.
Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.
Stock, J. H., & Watson, M. (2011). Dynamic factor models. In M. P. Clements, & D. F. Hendry (Eds.), *Oxford handbook on economic forecasting*. Oxford University Press.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.
Thorsrud, L. A. (2016). Nowcasting using news topics. Big data versus big bank. Working paper.
Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics* (in press). http://dx.doi.org/10.1080/07350015.2018.1506344.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B.*, *58*(1), 267–288.
Tibshirani, R. J., & Taylor, J. (2012). Degrees of freedom in LASSO problems. *The Annals of Statistics*, *40*(2), 1198–1232.
Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, *34*(2), 355–365.
Ulbricht, D., Kholodilin, K. A., & Thomas, T. (2017). Do media data help to predict German industrial production? *Journal of Forecasting*, *36*(5), 483–496.
Wang, T., & Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, *102*(7), 1141–1151.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B.*, *67*(2), 301–320.
Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the LASSO. *The Annals of Statistics*, *35*(5), 2173–2192.
Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*(4), 1733–1751.

David Ardia is Assistant Professor of Finance at the University of Neuchâtel, Switzerland. Previously he was senior analyst at aeris CAPITAL AG and head of research at Tolomeo Capital AG, two Swiss-based asset managers. In 2008 he received the Chorafas Prize for his book *Financial Risk Management with Bayesian Estimation of GARCH Models*, published by Springer. He is the author of several scientific articles and statistical packages in R. He holds an M.Sc. in applied mathematics, an MAS in quantitative finance, and a Ph.D. in financial econometrics.

Keven Bluteau is Ph.D. student in finance at the University of Neuchâtel and Vrije Universiteit Brussel. He obtained his B.Sc. and MBA from Laval University. His research is centered on risk management and sentiment analysis applied to finance. He is co-author of the R statistical packages MSGARCH and NSE.

Kris Boudt is Associate Professor in Finance at Vrije Universiteit Brussel and Amsterdam. He is a research partner of Finvex, instructor at Datacamp, affiliated researcher at KU Leuven and guest lecturer at the University of Illinois at Chicago, Southwestern University of Finance and Economics and the University of Aix-Marseille. Kris Boudt has published in leading international finance and statistics journals including the Journal of Econometrics, International Journal of Forecasting, Journal of Financial Econometrics, Journal of Financial Markets, Journal of Portfolio Management, Review of Finance and Statistics and Computing, among others. Kris Boudt has a passion for developing financial econometrics tools in R.