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]

[a] *Institute of Financial Analysis, University of Neuchâtel, Neuchâtel, Switzerland*
[b] *Department of Finance, Insurance and Real Estate, Laval University, Québec City, Canada*
[c] *Solvay Business School, Vrije Universiteit Brussel, Belgium*
[d] *Department of Econometrics, Vrije Universiteit Amsterdam, The Netherlands*

Abstract

Questionnaire–based indices of economic sentiment are often used to forecast economic growth. We recommend to complement the survey data with the use of textual analysis–based sentiment indices. In contrast with surveys, the design of the latter can be optimized in retrospect to obtain the best possible prediction of economic growth. We introduce a general framework for textual sentiment engineering that includes the use of the elastic–net for sparse data–driven selection and weighting of thousands of sentiment values. These values are obtained by pooling the text–based sentiment values across newspapers and magazines, article topics, sentiment construction methods, and time. We find that the proposed textual analysis approach yields significant accuracy gains in forecasting the semi–annual and annual growth rate of German industrial production, compared to the traditional use of questionnaire–based economic sentiment indices.

Keywords: elastic–net, German industrial production, sentiment analysis, time–series aggregation, topic–sentiment

1. Introduction

Understanding the current and future state of the economy is crucial for timely and efficient economic policy and business decision–making. Economic variables such as the country's gross domestic product, industrial production, consumer spending, and unemployment rate are closely followed by policymakers to assess the state of the economy. Often, economic surveys are used to gauge the respondents' sentiment about the future economy. For forecasting economic growth in Europe, the Economic Sentiment Indicator (ESI) is probably the best known metric. Its predictive power for forecasting economic growth has been studied by Mourougane and Roma (2003), Silgoner (2008), Gelper and Croux (2010), and Zanin (2010), among others. Even though the ESI can be seen as an early indicator of future economic activity, it still suffers from

[☆] We are grateful to Andreas Alfons, Nabil Bouamara, Samuel Borms, Dries Cornilly, Siem Jan Koopman, Wouter Torsin, as well as participants at the CFE 2017 conference for helpful comments. All analyses have been performed in the R statistical language (R Core Team, 2017). All errors and omissions are the sole responsibility of the authors.

[*] Corresponding author. University of Neuchâtel, Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland. Phone: +41 32 718 1365.

Email addresses: `david.ardia@unine.ch` (David Ardia), `keven.bluteau@unine.ch` (Keven Bluteau), `kris.boudt@vub.be` (Kris Boudt)

many drawbacks, such as release lags, cost of the survey, the changes in respondents over time, measurement errors, and, importantly, the deterministic nature of the survey design once the survey has been carried out (see, *e.g.*, Bound et al., 2001, for a review of measurement errors in economic surveys).

In this paper, we exploit the quantitative information in the sentiments expressed by authors of newspaper and magazine articles discussing a country's economic state to obtain timely forecasts of the country's economic growth. In contrast with the use of surveys, where the number of degrees–of–freedom is limited by the questions asked, textual sentiment analysis starts off with a rich (big) data environment of a virtually infinite number of texts. A significant advantage of text–based sentiment analysis is thus its flexibility. 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 want to only extract the relevant information from the text and create informative indices for predicting economic growth.

To address this challenge, we propose a methodology which first computes thousands of sentiment values capturing the tone expressed by the authors of newspaper and magazine articles discussing topics related to the country's economic growth.[1] It then maps the hordes of sentiment values in a single economic growth prediction by means of aggregation based on topic (*e.g.*, "real estate market" or "job creation"), time, and a data–driven calibration using penalized least squares regression forecasting techniques. We refer to the resulting optimized aggregate value of sentiment as a text–based sentiment index, and treat it as both an alternative and a complement to the traditional framework of measuring economic sentiment through surveys. The resulting 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 to gauge the contribution of the various textual sentiment indices to the prediction.

Finally, note that, besides being flexible, timely, and data–rich, the proposed methodology to construct optimized sentiment indices for forecasting economic growth has the advantage that its design can be backtested. We validate the approach by showing that, for an out–of–sample evaluation window ranging from April 2000 to April 2016, the text–based sentiment indices provide additional predictive power for the semi–annual and annual growth rate of the German industrial production index, compared to the standard use of a univariate time–series model and the inclusion of the ESI. This result is shown to be robust to various alternative choices of implementation.

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 to predict is the h–period logarithmic change in the variable Y_t, expressed in percentage points:

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

[1] In this study we used lexicon–based unsupervised methods to calculate sentiment, but our methodology could also be applied to supervised sentiment classification methods.

where $t = 1, 2, \dots, T$ is a daily time index. We require y_t^h to be covariance stationary. This is typically the case when Y_t represents a country's economic activity (*e.g.*, its gross domestic product or industrial production), its price level (*e.g.*, the consumer price index or the exchange rate), and similarly for corporate variables, like the firm's sales or stock price. In our application, y_t^h is the logarithmic growth in industrial production for Germany over horizons ranging from one to twelve months. Note that, due to the publication lag, it may be that Y_t is not 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, $\mathbb{E}(y_T^h \,|\, \mathcal{I}_T)$. This is a common problem in time–series forecasting, for which we expand the information set $\mathcal{I}_T$ by including not only past values of Y_t that are known at time T, but also various sentiment values extracted from a corpus of texts published up to date T. We describe below the methodology, as depicted in Figure 1.

[Insert Figure 1 about here.]

2.1. Data preparation

Step 1: Classify texts by topic and use expert opinion to choose a subset of topics 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 directly from the texts. In our application, we use the corpus of newspaper and magazine papers 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 (see Liu et al., 2016, for a recent review) or, if topic–labelled news are available for a training set, the use of a support vector machine classifier (see, *e.g.*, Tobback et al., 2016). Expert opinion is then used to exclude the topics that, beforehand, can be qualified as being irrelevant for forecasting the variable of interest y_T^h. The resulting topic–markers for our application on forecasting economic growth are reported in Table 1. The corpus that we analyze are the texts that discuss at least one of the selected topics. The corpus is organized in terms of publication date t, with $t = 1, \dots, T$, where N_t is the number of texts in the corpus of texts published at time t. We use n to index the text available at time t, with $n = 1, \dots, N_t$.

[Insert Table 1 about here.]

Step 2: Compute for each text n of corpus t the sentiment using L methods. For each text, we compute the underlying sentiment using L different textual sentiment computation methods. For a general review of available methods, we refer the reader to Ravi and Ravi (2015). In our application, we use the bag–of–words approach to compute the net sentiment using L different lexicons to classify the words as positive, negative or neutral. We henceforth obtain for each text document $n = 1, \dots, N_t$, published at time $t = 1, \dots, T$, L different sentiment values, which we denote by $s_{n,t,l}$, where $l = 1, \dots, L$.

2.2. Aggregating sentiment into a prediction

At this stage, we have for each day t and for each of the N_t texts, 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 $N_t \times 1$. The next steps aim at reducing the high dimensionality of the available texts (*i.e.*, total of texts is $N_1 + \ldots + N_T$). To that end, we first compute the daily sentiment per topic–markers by aggregating across the sentiment of texts published on a given day. We 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: For each corpus n and method l, obtain K topic–based sentiments. 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 for each day t the text–to–topic aggregation matrix $\mathbf{W}_t$ of dimension $K \times N_t$ 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. In the application, we discard news from smaller news outlets, which is equivalent to setting the weight of the news articles coming from those sources to zero.

Step 4: For each topic k and method l, obtain time–series aggregated values. Next, we aggregate through time. We take a maximum time–aggregation lag τ $(0 \leq \tau < T)$, and, for a given l, we 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 the $LK \times (\tau + 1)$ 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 vec$(\cdot)$ is the vectorization operator.[2]

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

We recommend to use a data–driven calibration of the aggregation matrix $\mathbf{B}$ in order to strike a balance between a strong decay in weights to obtain timeliness on the one hand, and, on the other hand, an equal–weighting approach to obtain efficiency when all time–lags are equally informative. To do so, we recommend using multiple Almon polynomials to populate the aggregation matrix $\mathbf{B}$. They form a parsimonious way to obtain time–series weights that are flexible enough to describe a variety of weights; see Andersen et al. (2000), Bollerslev et al. (2000), and Boudt et al. (2011) for related applications where the Almon polynomials are used. Under this approach, the $(\tau+1)\times B$ aggregation matrix is given by:

$$\mathbf{B} \equiv \begin{bmatrix} p_1(\tau) & & p_B(\tau) \\ \vdots & \dots & \vdots \\ p_1(0) & & p_B(0) \end{bmatrix} . \tag{5}$$

The b–th column of $\mathbf{B}$ corresponds to the b–th order Almon polynomial given by:

$$p_b(i) \equiv c_b \left(1 - \left(\frac{i}{\tau} \right)^b \right) \left(\frac{i}{\tau} \right)^{3-b} , \tag{6}$$

where $i = 0, \dots, \tau$, and c_b is the normalization factor such that the sum of each column equals one. Note that, because of the decay in the Almon weights, τ can be relatively large.

Step 5: Calibration to optimize 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:

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

where α is an intercept, $\mathbf{x}_t$ is a $M\times1$ vector of (non–textual sentiment) variables at time t, $\boldsymbol{\gamma}$ is the corresponding vector of parameters, $\boldsymbol{\beta} \equiv (\beta_1, \dots, \beta_P)'$ is a vector of parameters associated with the textual sentiment indices, and ε_t is an error term at time t. Typically, $\mathbf{x}_t$ includes y_s where y_s is the latest information available on the dependent variable up to time t, that is $s \leq t$. This point is of particular relevance in economics due to the release lag faced by economic indicators. A further explanatory variable of interest can be the sentiment obtained using questionnaires. We will include the latter in our application to study the value added of text–based sentiment analysis compared to the traditional approach using only questionnaires.

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

[3] In our application, all calibrations are performed with the R package **glmnet** (Friedman et al., 2010).

expressed as:

$$\min_{\widetilde{\alpha},\boldsymbol{\gamma},\widetilde{\boldsymbol{\beta}}} \left\{ \frac{1}{T} \left\| y_t^h - \left(\widetilde{\alpha} + \boldsymbol{\gamma}' \mathbf{x}_t + \widetilde{\boldsymbol{\beta}}' \widetilde{\mathbf{s}}_t \right) \right\|_2^2 + \lambda_1 \left[\lambda_2 \| \widetilde{\boldsymbol{\beta}} \|_1 + (1 - \lambda_2) \| \widetilde{\boldsymbol{\beta}} \|_2^2 \right] \right\} , \tag{8}$$

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 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 LASSO regularization (when $\lambda_2 = 1$) introduced by Tibshirani (1996). The variable $\widetilde{\mathbf{s}}_t$ is the standardized version of $\mathbf{s}_t$ with components $\widetilde{s}_{i,t} \equiv (s_{i,t} - \mathrm{av}_i)/\mathrm{std}_i$, where av_i and std_i are the sample mean and standard deviation of $\{s_{i,t}; t = 1, \ldots, T\}$, respectively. The standardization is crucial in penalized regressions as the penalty depends on the scale of the components of $\boldsymbol{\beta}$.

The implementation of the elastic–net in (8) requires the calibration of the penalty parameters λ_1 and λ_2. We follow Zou et al. (2007) and minimize the so–called BIC–like criterion, where BIC stands for Bayesian Information Criterion.[4] We define the vector $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ of size $T \times 1$ as the forecast of $\mathbf{y}^h$ obtained by fixing λ_1 and λ_2. The BIC–like criterion is defined as:

$$BIC_{\lambda_1,\lambda_2} \equiv \frac{\| \mathbf{y}^h - \widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2} \|_2^2}{T\sigma^2} + \frac{\ln T}{T} \, \widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}) , \tag{9}$$

where σ^2 is defined as the variance of the forecast error given by the largest $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$ possible using different combinations of λ_1 and λ_2. The parameter $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$ is an estimator of the degree–of–freedom of the elastic–net given $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ (see Tibshirani and 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.

Step 6: Forecasting. For forecasting and interpretation, $\widetilde{\boldsymbol{\beta}}$ is rescaled to give the corresponding optimal unstandardized vector $\widehat{\boldsymbol{\beta}}$. Recall that $\widehat{\beta}_i$ and $\widetilde{\beta}_i$ are the i–th components of $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, respectively, and that av_i and std_i are the location and scaling parameters used to standardize s_i. The unstandardized regression parameter can be recovered by rescaling each component of $\widetilde{\boldsymbol{\beta}}$ as follows:

$$\widehat{\beta}_i \equiv \frac{\widetilde{\beta}_i}{\mathrm{std}_i} . \tag{10}$$

An additional value must then be subtracted from the regression intercept to account for the centering of the series:

$$\widehat{\alpha} \equiv \widetilde{\alpha} - \sum_{i=1}^{P} \frac{\widetilde{\beta}_i}{\mathrm{std}_i} \mathrm{av}_i . \tag{11}$$

Our forecast at time T is then given by:

$$\widehat{y}^h_T \equiv \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}' \mathbf{x}_T + \widehat{\boldsymbol{\beta}}' \mathbf{s}_T . \tag{12}$$

[4] In our study, the low sample size and cross–correlation generated by the overlapping data when $h > 1$ make the cross–validation calibration methodology unstable, which is undesirable.

2.3. Forecast precision and attribution

Given the predicted values of y_T^h, it is important 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. For the evaluation of the forecasting performance, we use the Root Mean Squared Forecast Error (RMSFE), the Mean Absolute Forecast Error (MAFE), and the Directional InAccuracy (DIA). Let $e_{i,t}^h \equiv y_t^h - \widehat{y}_{i,t}^h$ be the error term for model i at time t for an horizon h where $\widehat{y}_{i,t}^h$ is the forecast. The RMSFE, MAFE, and DIA measures of model i at horizon h are defined by:

$$\mathrm{RMSFE}_i^h \equiv \sqrt{\frac{1}{T_F}\sum_{t=T+1}^{T+T_F}(e_{i,t}^h)^2} \tag{13}$$

$$\mathrm{MAFE}_i^h \equiv \frac{1}{T_F}\sum_{t=T+1}^{T+T_F}|e_{i,t}^h| \tag{14}$$

$$\mathrm{DIA}_i^h \equiv -\frac{1}{T_F}\sum_{t=T+1}^{T+T_F} I\{y_t^h\widehat{y}_{i,t}^h > 0\}\,, \tag{15}$$

where T is the size of the estimation sample, T_F is the number of out–of–sample observations, and $I\{\cdot\}$ is an indicator function.

We apply the Model Confidence Set (MCS) procedure of Hansen et al. (2011) in order to evaluate the statistical significance of the differences in RMSFE, MAFE, and DIA across prediction models. Specifically, the MCS procedure allows us to identify a subset of equivalent models that are deem statistically superior to any other subsets of models in terms of predictive capability for a given loss function. For our purposes, we apply the MCS on the squared error (*i.e.*, $\mathcal{L}_{i,t}^h = (e_{i,t}^h)^2$), the absolute error (*i.e.*, $\mathcal{L}_{i,t}^h = |e_{i,t}^h|$), and the directional inaccuracy (*i.e.*, $\mathcal{L}_{i,t}^h = -I\{y_t^h\widehat{y}_{i,t}^h > 0\}$). First, let $\mathcal{M}$ be the set of all competing models. Also, let the differential loss at time t between model i and j for the horizon h be $d_{i,j,t}^h \equiv \mathcal{L}_{i,t}^h - \mathcal{L}_{j,t}^h$, and let the average differential loss be $\bar{d}_{i,j}^h \equiv \frac{1}{T_F}\sum_{t=T+1}^{T+T_F} d_{i,j,t}^h$. The MCS procedure starts by computing a series of test–statistics:

$$t_{i,j} \equiv \frac{\bar{d}_{i,j}^h}{\sqrt{\widehat{\mathbb{V}}(\bar{d}_{i,j}^h)}} \quad \text{for} \quad i,j \in \mathcal{M}\,, \tag{16}$$

where $\widehat{\mathbb{V}}(\bar{d}_{i,j}^h)$ is a bootstrap estimate of the true variance $\mathbb{V}(\bar{d}_{i,j}^h)$. Then, the equal predictive ability (EPA) hypothesis for the set $\mathcal{M}$ (*i.e.*, $\bar{d}_{i,j}^h = 0$ for all $i,j \in \mathcal{M}$) is tested by first computing the test–statistic:

$$t_{\mathcal{M}} \equiv \max_{i,j\,\in\mathcal{M}} |t_{i,j}|\,, \tag{17}$$

and by evaluating it at a certain level $1-\alpha$ against the distribution under the null hypothesis

($i.e.$, $\bar{d}^h_{i,j} = 0$ for all $i, j \in \mathcal{M}$) obtained using a bootstrap procedure.[5] If the EPA is rejected, an elimination rule is then applied to remove the worst performing model from the set $\mathcal{M}$ resulting in a new set $\mathcal{M}^-_{1-\alpha}$. The procedure is then repeated on the set $\mathcal{M}^-_{1-\alpha}$. Doing this procedure sequentially until the EPA is accepted results in the "superior set of models" $\mathcal{M}^\star_{1-\alpha}$. In the application, we use the MCS implementation in Bernardi and Catania (2016).

Step 8: Attribution. Until now, our exposition has been a bottom–up story of aggregating the sentiment of individual texts through cross–sectional, time–series, and elastic–net weighting into a prediction of economic growth. Once this prediction is obtained, it is important to top–down attribute the obtained prediction to the individual texts 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 $s_{n,t,l}$:

$$\widehat{y}^h_T = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}'\mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}'\boldsymbol{e}_{l,k,b} \cdot W_{t,k,n} B_{T-t,b} \cdot s_{n,t,l} \, , \tag{18}$$

where $\boldsymbol{e}_{l,k,b}$ is basis vector of size $LKB \times 1$, which extracts the relevant regression parameter in $\widehat{\boldsymbol{\beta}}$ 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 matrix $\mathbf{B}$. It is easy to see from (18) that the weight $\omega_{n,t,l}$ attributed to the sentiment $s_{n,t,l}$ is equal to:

$$\omega_{n,t,l} = \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}'\boldsymbol{e}_{l,k,b} \cdot W_{t,k,n} B_{T-t,b} \, , \tag{19}$$

such that:

$$\widehat{y}^h_T = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}'\mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \omega_{n,t,l} \cdot s_{n,t,l} \, . \tag{20}$$

Clearly, it is infeasible to analyze all (n, t, l)-combinations. We thus proceed by grouping them by common attributes, like time or topic. For example, to obtain the attribution of topic g $(1 \leq g \leq K)$, we can fix $k = g$ and compute the attribution by integrating the other dimensions:

$$a_g \equiv \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}'\boldsymbol{e}_{l,g,b} \cdot W_{t,g,n} B_{T-t,b} \cdot s_{n,t,l} \, . \tag{21}$$

3. Application to forecasting economic growth

We illustrate the complete framework by analyzing sentiment in text documents available in *LexisNexis* to forecast economic growth in Germany. We choose a European country to benchmark our results against the use of the Economic Sentiment Indicator, as published by the European Commission every last business day of the month. Among all European countries, we pick Germany as it is the most important EU economy in terms of gross domestic product.

We first introduce the data and the models that we compare. We then present our main results and interpret the attribution that we obtain.

[5] Typical values of α in the literature are 5%, 10%, and 25%.

3.1. Data and descriptive statistics

3.1.1. Quantitative data

We aim at forecasting the log–growth in German 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.[6] As mentioned in the introduction, several authors find that sentiment measures can predict industrial production growth (see, *e.g.*, Silgoner, 2008; Gelper and Croux, 2010; Ulbricht et al., 2017). 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.

In terms of readily available sentiment information, we use the Economic Sentiment Indicator (ESI) of Germany, as published by the European Commission. The German ESI is based on sentiment surveys carried out by the European Union. It is a composite index of five confidence indicators: (i) industrial confidence indicator, (ii) services confidence indicator, (iii) consumer confidence indicator, (iv) construction confidence indicator, and (v) retail trade confidence indicator, with weights 40%, 30%, 20%, 5%, and 5%, respectively.

Figure 2 presents the German ESI from February 1996, to April 2016 along with the realized twelve–month log–growth in the German industrial production for a total of 242 observations.[7] We observe a strong co–movement between the two series. This also shows up in terms of a high contemporaneous correlation of approximately 0.71.

[Insert Figure 2 about here.]

3.1.2. Qualitative data – corpus

To compute textual sentiment indices for Germany, we retrieve the set of news consisting of all English articles from European sources in the *LexisNexis* database with reference to Germany. Dates range from September 31, 1994, to March 31, 2016. We apply the following filters:

- We use the geographic location such that we select only news relevant to Germany (relevance score greater or equal to 85 in *LexisNexis*).
- We remove articles with less than 500 words to exclude sentiment estimates with high standard errors due to the small number of words used to estimate the sentiment value.[8]
- We limit the number of sources by filtering out the media that have published less than 20 articles during our first training sample (*i.e.*, the period ranging from September 31, 1994, and March 31, 2000) as a way to remove smaller news outlets.
- We use the topic filter and filter out non–economic related topics.

[6] We retrieve data for the German industrial production from the Organisation for Economic Co–operation and Development (OECD) database: `https://data.oecd.org/industry/industrial-production.htm`.

[7] We retrieve data for the German ESI from `http://ec.europa.eu/economy_finance/db_indicators/surveys/time_series/index_en.htm`.

[8] As we rely on a word frequency model for sentiment analysis, the standard error of the sentiment estimates are directly related to the number of words in the text analyzed. Other studies relying on the *LexisNexis* database, such as Shapiro et al. (2017), use a lower limit of 200 words. In our application, we use the default option provided in *LexisNexis* of 500 words. Similar results are obtained when using 200 words.

Table 1 presents the topics selected as well as the number of documents associated with them. The final corpus is a combination of several types of media including newspapers, newswire, magazines, and newsletters that amount to a total of 26,139 articles and 47 topics. Note that a news article might refer to more than one topic. Indeed, we observe that the average number of topics by news article is 3.53 (the median is 3).

[Insert Table 1 about here.]

3.1.3. Qualitative data – sentiment calculation

We use standard lexicon–based sentiment analysis to measure the sentiment. The fundamental 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 (*HAR*; 2,550 positive words and 3,695 negative words).[9] 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. For the analysis of financial and economic discourses this implies the use of specialized financial dictionaries, such as those developed by Henry (2008) (*HEN*; 105 positive words and 85 negative words) and Loughran and McDonald (2011)[10] (*LM*; 354 positive words and 2,355 negative words). We use the three lexicons leading to $L = 3$ sentiment calculation methods.

Another aspect of sentiment analysis is valence shifting words (Polanyi and 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.[11]

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

$$s_{n,t,l} \equiv \frac{N^{+}_{n,t,l} - N^{-}_{n,t,l}}{N^{+}_{n,t,l} + N^{-}_{n,t,l} + N^{0}_{n,t,l}}\,, \tag{22}$$

where $N^{+}_{n,t,l}$ is the number of positive words in document n at 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. This use of the net sentiment measure, computed as the difference in the frequency of the positive words (positive sentiment) and the frequency of the negative words (negative sentiment) normalized by the total number of words, is widespread in the literature (see, *e.g.*, Arslan-Ayaydin et al., 2016, and the references therein).

Figure 3 presents the yearly average of the individual news article sentiments. First, while the average value is negative across all years, we see that the time–variation coincides with the

[9] Available at `http://www.wjh.harvard.edu/~inquirer/homecat.htm`.

[10] Available on the authors' website; see `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

[11] 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.*

economic cycle. In particular, we observe a large drop in the average yearly sentiment during the dot–com bubble burst of 2001 and the financial crisis of 2008. These events are preceded by a large, almost linear, increase in the yearly average. Following the financial crisis of 2008, we see that, while sentiment is recovering from its 2008 low, it does not follow the linear pattern seen before 2001 and 2008, indicating more uncertainty.

[Insert Figure 3 about here.]

3.1.4. Qualitative data – aggregation of sentiment

We build the aggregation matrices $\mathbf{W}_t$ $(t = 1, \ldots, T)$ in such a way that each of the 47 topics is summarized by a sentiment index. The time–series aggregation matrix $\mathbf{B}$ contains $p_1(\tau)$ and $p_2(\tau)$, that is $B = 2$, which corresponds to the polynomials that give higher weights to the observations that are near the forecasting date. We set the value $\tau = 180$ days. This gives a total of $P = LKB = 3 \times 47 \times 2 = 282$ sentiment indices when using all three lexicons.

Table 2 presents quantiles of the correlation between the realized German industrial production log–growth at several growth horizons and the 282 sentiment indices over the full sample. The sentiment indices are generally positively correlated with the German industrial production log–growth. We also observe a wide range of correlation values for all horizons. For example, correlation ranges from -0.19 to 0.35 for the one–month horizon and from -0.04 to 0.63 for the twelve–month horizon.

[Insert Table 2 about here.]

Figure 4 presents the yearly average of the 47 sentiment indices calculated with the Loughran & McDonald lexicon using the first order Almon polynomial with $\tau = 180$.[12] Similarly to the yearly average of the non–aggregated sentiment shown in Figure 3, we observe that there is a general decrease in all sentiment indices during the years 2001 and 2008. We also observe a large variability in the cross–section of yearly averages.

[Insert Figure 4 about here.]

3.2. Models

The forecasting models that we consider are nested in the linear framework (7). They use the ESI in level and in first differences as a predictor of economic growth, in addition to the last available information on economic growth. Our use of the ESI as an early indicator of future economic activity is consistent with the documented predictive power of ESI in forecasting economic growth by Mourougane and Roma (2003), Gelper and Croux (2010), Silgoner (2008), and Zanin (2010), among others.

[12] We observe the same pattern for other groups of sentiment indices such as those calculated with the Henry lexicon or with a different Almon polynomial order.

More precisely, we study the following four specifications:

$$\begin{aligned}
\mathcal{M}_1 : &\quad y_t^h = \alpha^h + \gamma_1^h y_{t-2}^1 + \varepsilon_t^h & (23)\\
\mathcal{M}_2 : &\quad y_t^h = \alpha^h + \gamma_1^h y_{t-2}^1 + \gamma_2^h ESI_t + \gamma_3^h \Delta ESI_t + \varepsilon_t^h & (24)\\
\mathcal{M}_3 : &\quad y_t^h = \alpha^h + \gamma_1^h y_{t-2}^1 + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h & (25)\\
\mathcal{M}_4 : &\quad y_t^h = \alpha^h + \gamma_1^h y_{t-2}^1 + \gamma_2^h ESI_t + \gamma_3^h \Delta ESI_t + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h\,, & (26)
\end{aligned}$$

for $t = 1, \ldots, T$ months, where y_{t-2}^1 is the latest available one–month log–growth of the German industrial production taking into consideration the six–week publication lag, ESI_t is the German Economic Sentiment Indicator at time t, and ΔESI_t is the log difference of the ESI at time t, $\Delta ESI_t \equiv \ln ESI_t - \ln ESI_{t-1}$. Note that we are now dealing with a monthly frequency as opposed to the daily frequency used in the construction of the sentiment indices.

We estimate models (23)–(24) by standard ordinary least squares, as the number of covariates is low. On the other hand, we estimate models (25)–(26) using the elastic–net regularization as a way to penalize the parameters related to the textual sentiment indices as presented in (8). Each model is estimated on a rolling window basis of 60 months.[13]

We then evaluate each model using the next period out–of–sample observations. That is, if the sample window is from months $t = 1$ to $t = 60$, we evaluate the out–of–sample forecast made with the observations at month $t = 61$. Out–of–sample forecasting performance is then evaluated using the RMSFE, MAFE, and DIA measures. The performance measures of each model are then compared using the MCS approach with a 75% confidence level following Ulbricht et al. (2017). Finally, to account for possible changes in out–of–sample forecasting performances over time, we analyze the full out–of–sample period and three subperiods: pre–crisis, crisis, and post–crisis. The full out–of–sample period ranges from April 2000 (March 2001 for $h = 12$) to April 2016 (193 observations for $h = 1$ and 182 observations for $h = 12$). The pre–crisis period ranges from April 2000 to November 2008 (104 observations for $h = 1$ and 93 observations for $h = 12$). The crisis period ranges from December 2008 to June 2010 (19 observations). Finally, the post–crisis period ranges from July 2010 to April 2016 (70 observations).[14]

3.3. Main results

3.3.1. Comparing predictors

Table 3 presents the RMSFE, MAFE, and Directional Accuracy (DA) (*i.e.*, negative DIA) measures for the four model specifications and the five forecasting horizons over the four time windows.

[Insert Table 3 about here.]

For the full sample, we see that textual sentiment–related specifications (*i.e.*, $\mathcal{M}_3$ and $\mathcal{M}_4$) exhibit the best performance for all horizons above (or equal to) three month ($h \geq 3$) and particularly at the nine– to twelve–month horizons. This gain in outperformance as the forecasting

[13] When a topic–based sentiment index has a missing value in the estimation window (*i.e.*, no articles are published on that topic), we set the corresponding regression parameter to zero.

[14] The crisis period is defined as in Ulbricht et al. (2017), that is as the period when considerably more models than usual suffer a forecast breakdown as defined in Giacomini and Rossi (2009).

horizon grows is also observed in Ulbricht et al. (2017) for news–derived economic sentiment indices. It is consistent with the "time–lag" effect in economics (see George et al., 1999). While financial markets can react (quasi) instantaneously to the sentiment expressed in the 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. This may explain why the sentiment becomes more predictive for economic growth over longer horizons. We also observe that combining the ESI and textual sentiment ($\mathcal{M}_4$) tends to improve forecasting precision when compared to the forecasting model using textual sentiment alone ($\mathcal{M}_3$). Hence, textual sentiment is complementary to the questionnaire–based ESI appraisal of economic sentiment.

Regarding the other subperiods, we can see that the relative performance between the textual sentiment–related models ($\mathcal{M}_3$ and $\mathcal{M}_4$) and the autoregressive model ($\mathcal{M}_1$) drops during the post–crisis period for the three– and six–month forecasting horizons. The same is observed for the ESI model ($\mathcal{M}_2$). This indicates that the German industrial production has become harder to forecast with sentiment–related model. However, at longer growth horizons (nine and twelve months), the textual sentiment model outperforms the ESI and autoregressive models. The period where textual sentiment indices seem to provide the most information about the future growth in industrial production is during the crisis period. Garcia (2013) finds a similar result for the stock market and attributes this to the behavioral finding that in a recession regime, decision makers have a greater sensitivity to news.

Figure 5 presents the twelve–month out–of–sample forecasts of the German industrial production for the ESI model ($\mathcal{M}_2$) and the textual sentiment model ($\mathcal{M}_3$). We clearly observe that the textual sentiment model follows more closely the German industrial production than the ESI model. This is even clearer when looking at the crisis period where the ESI model forecast shows a steep decrease only several months after the crisis period has begun.

[Insert Figure 5 about here.]

Overall the news–related sentiment indices seem to provide information on the future log–growth of the German economy. The added predictive information is significant given the large out–of–sample error reduction compared to more standard models.

3.3.2. Attribution

A common criticism for big data approaches to economic forecasting is that their results seem to come from a black box. In our setting, this criticism can be easily countered, since the attribution analysis described in Section 2.3 allows us to pinpoint the contribution of each sentiment value to the growth prediction. Given the large number of sentiment values, we recommend to analyze the attribution at the intermediate level of the grouping per cluster of topics. Table 4 presents six clusters of topics, which have been manually constructed by identifying economic concepts that are closely related, namely "GDP Output", "Job Market", "Prices & Interest Rate", "Real Estate", "Surveys", and "Others". The latter is composed of the remaining topics.

[Insert Table 4 and Figure 6 about here.]

Figure 6 presents the normalized attribution of these clusters, where we divide each of its elements by the L^2–norm of the attribution vector at that date. This procedure makes it easier to do comparison 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. We further find that almost all clusters show negative attribution at the beginning of the crisis period. In addition, this is preceded by a time of high positive sentiment, as almost none of the clusters show negative attribution from 2007 to mid–2008. This illustrates how reactive the sentiment indices are to the crisis period. The "Real Estate" cluster has contributed the most to the positive growth in industrial production after the crisis period. This is followed, after 2012, by a mostly negative attribution driven by the "Price & Interest Rate" and the "GDP Output" clusters. Finally, the contribution of the "Surveys", "Job Market", and "Others" clusters to the forecast is negligible when compared to the "GDP Output", "Price & Interest Rate", and "Real Estate" clusters.

3.4. Robustness analysis

We now proceed to analyze the impact of some of the modeling choices employed in this study. First, we analyze the impact of the sentiment calculation method. Specifically, we want to know if there is a single lexicon that produces overall better sentiment indices and therefore forecasting performance. Second, we want to assess the impact of alternative aggregation schemes for the sentiment indices construction. In particular, we evaluate the importance of the topic dimension. Also, we evaluate the performance of a simple moving average time–aggregation scheme over the more complex Almon polynomials.

3.4.1. Choice of lexicon

Table 5 presents the RMSFE and MAFE results for model $\mathcal{M}_3$ when a single lexicon is used to calculate the textual sentiment indices. We denote by *ALL* the benchmark results using all lexicons while *HEN*, *HAR*, and *LM* refer to model $\mathcal{M}_3$ calculated with the Henry (2008), Harvard General Inquirer, and Loughran and McDonald (2011) lexicons, respectively.

Consistent with our recommendation to consider various lexicons when optimizing the prediction of economic sentiment, we find that, in general, there is no lexicon that outperforms consistently. Overall, the *HAR* lexicon tends to perform worst. This is to be expected since the *HAR* lexicon has not been built specifically for the evaluation of financial and economic texts. We further find that, among the two financial–domain lexicons, *LM* tends to perform better than *HEN*. Finally, the use of all lexicons (*i.e.*, *ALL*), which relies on the elastic–net regression do the lexicon selection, provides superior forecasting performance when compared to the individual lexicons.

[Insert Table 5 about here.]

3.4.2. Alternative aggregation scheme

First, we evaluate the impact of removing the topic dimension from the indices construction by considering global economic sentiment indices. The global sentiment indices (*All Topics*) are constructed by letting $\mathbf{W}_t$ $(1, \ldots, T)$ be a $1 \times N_t$ (*i.e.*, $K = 1$) matrix and setting all elements of the matrix equal to $\frac{1}{N_t}$. We consider the use of the three lexicons and the first and second order Almon polynomials which leads to 6 global sentiment indices ($LKB = 3 \times 1 \times 2 = 6$).

Second, we evaluate the impact of using a simple moving average over the more complex Almon polynomial. The moving average indices (MA) are constructed by letting $\mathbf{B}$ be a matrix of size $(\tau + 1) \times 1$ (*i.e.,* $B = 1$) where each element of $\mathbf{B}$ is equal to $\frac{1}{1+\tau}$ and $\tau = 180$. In this case the topic dimension is not discarded resulting in 141 sentiment indices ($LKB = 3 \times 47 \times 1 = 141$).

Table 6 presents the results for RMSFE and MAFE for these alternative setups along with the benchmark results given by our original sentiment indices construction setup (see Section 3.1.4) which is denoted as *Almon*. First, we see that removing the topic dimension has a large negative impact on the forecasting performance. Aggregation of all topics into one general economic topic removes most of the explanatory power of the news sentiment. This is the case for all periods with exception of the pre–crisis period three–month German industrial production log–growth. The full sample results indicate that the use of Almon aggregation is superior to moving average aggregation. However, the sub–sample results give a more precise view of what is happening. We observe that the moving average aggregation yields, in general, a better forecasting precision than the Almon aggregation during the pre–crisis and the post–crisis periods. In these cases, the Almon polynomial does not seem to put enough weight on older news. Nevertheless, the Almon polynomial aggregation provides more precise industrial production growth forecasts during the crisis period.

[Insert Table 6 about here.]

4. Conclusion

Do textual sentiment indices provide any added value to the prediction accuracy of economic growth when compared to the use of economic surveys? To answer this question, one needs to first capture the relevant sentiment–based growth prediction from textual analysis of news releases. The latter is a big data problem, given the large number of texts published every day, the number of possible historical dates at 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 predicting economic growth using both topics–based aggregation, time–series aggregation, and predictive regressions using the elastic–net regularization.

We test the predictive power of textual sentiment by forecasting the growth in German industrial production using the news database *LexisNexis* over the period April 2000–2016. We find that the proposed optimized text–based sentiment analysis can significantly improve the forecasting performance, especially for the six–, nine–, and twelve–month forecasting horizons. Importantly, these forecasting accuracy gains are complementary to the use of the survey–based European Sentiment Index. Our main recommendation is thus to use both the questionnaire–based sentiment index and the proposed text–based sentiment analysis for forecasting economic growth.

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

References

Andersen, T.G., Bollerslev, T., Cai, J., 2000. Intraday and interday volatility in the Japanese stock market. Journal of International Financial Markets, Institutions and Money 10, 107–130. doi:10.1016/s1042-4431(99)00029-3.

Arslan-Ayaydin, Ö., Boudt, K., Thewissen, J., 2016. Managers set the tone: Equity incentives and the tone of earnings press releases. Journal of Banking & Finance 72, S132–S147. doi:10.1016/j.jbankfin.2015.10.007.

Bernardi, M., Catania, L., 2016. The model confidence set package for R. International Journal of Computational Economics and Econometrics URL: https://CRAN.R-project.org/package=MCS. in press.

Bollerslev, T., Cai, J., Song, F.M., 2000. Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the US Treasury bond market. Journal of Empirical Finance 7, 37–55. doi:10.1016/S0927-5398(00)00002-5.

Boudt, K., Croux, C., Laurent, S., 2011. Robust estimation of intraweek periodicity in volatility and jump detection. Journal of Empirical Finance 18, 353–367. doi:10.1016/j.jempfin.2010.11.005.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data, in: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics. volume 5. chapter 59, 3705–3843. doi:10.1016/S1573-4412(01)05012-7.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22. doi:10.18637/jss.v033.i01.

Garcia, D., 2013. Sentiment during recessions. Journal of Finance 68, 1267–1300. doi:10.1111/jofi.12027.

Gelper, S., Croux, C., 2010. On the construction of the European economic sentiment indicator. Oxford Bulletin of Economics and Statistics 72, 47–62. doi:10.1111/j.1468-0084.2009.00574.x.

George, E., King, M., Clementi, D., Budd, A., Buiter, W., Goodhart, C., Julius, D., Plenderleith, I., Vickers, J., 1999. The transmission mechanism of monetary policy. Bank of England.

Giacomini, R., Rossi, B., 2009. Detecting and predicting forecast breakdowns. Review of Economic Studies 76, 669–705. doi:10.1111/J.1467-937X.2009.00545.X.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453–497. doi:10.3982/ECTA5771.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? Journal of Business Communication 45, 363–407. doi:10.1177/0021943608319388.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67. doi:10.1080/00401706.1970.10488634.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5, 1–22. doi:10.1186/s40064-016-3252-8.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

Mourougane, A., Roma, M., 2003. Can confidence indicators be useful to predict short term real GDP growth? Applied Economics Letters 10, 519–522. doi:10.1080/1350485032000100305.

Polanyi, L., Zaenen, A., 2006. Computing Attitude and Affect in Text: Theory and Applications. Springer-Verlag. volume 20 of *The Information Retrieval*. chapter Contextual Valence Shifters. 1–10.

Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89, 14–46. doi:10.1016/j.knosys.2015.06.015.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Shapiro, A., Sudhof, M., Wilson, D., 2017. Measuring news sentiment. Working paper.

Silgoner, M.A., 2008. The economic sentiment indicator: Leading indicator properties in old and new EU member states. Journal of Business Cycle Measurement and Analysis 2007, 199–215. doi:10.1787/jbcma-v2007-art11-en.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B 58, 267–288. doi:10.1111/j.1467-9868.2011.00771.x.

Tibshirani, R.J., Taylor, J., 2012. Degrees of freedom in LASSO problems. Annals of Statistics 40, 1198–1232. doi:10.1214/12-AOS1003.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E.J., Martens, D., 2016. Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting, in press. doi:10.1016/j.ijforecast.2016.08.006.

Ulbricht, D., Kholodilin, K.A., Thomas, T., 2017. Do media data help to predict German industrial production? Journal of Forecasting 36, 483–496. doi:10.1002/for.2449.

Zanin, L., 2010. The relationship between changes in the economic sentiment indicator and real GDP growth: A time-varying coefficient approach. Economics Bulletin 30, 837–846.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the LASSO. Annals of Statistics 35, 2173–2192. doi:10.1214/009053607000000127.

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

This table presents the number of articles related to a given topic in the corpus. The list of topics is manually selected 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–economic related topics have been removed resulting in a corpus that focuses exclusively on the German economy. Documents with less than 500 words and small news outlets identified by the number of publications in the corpus between September 31, 1994, and March 31, 2000 are removed. Note that each article may be related to multiple topics. See Section 3.1 for details.

Topic	#	Topic	#
CURRENCIES	7,354	ECONOMIC DECLINE	1,266
ECONOMIC CONDITIONS	6,696	EMPLOYMENT GROWTH	1,250
ECONOMIC POLICY	6,103	ECONOMIC STIMULUS	1,197
INTEREST RATES	5,827	JOB CREATION	1,129
GERMAN CHANCELLORS	5,108	WAGES SALARIES	1,126
PRICES	4,185	OIL GAS INDUSTRY	1,062
RECESSION	3,844	IMPORT TRADE	960
ECONOMIC GROWTH	3,800	BUSINESS CONFIDENCE	863
GROSS DOMESTIC PRODUCT	3,594	SALES FIGURES	827
INFLATION	3,339	CONSUMER CONFIDENCE	710
UNEMPLOYMENT RATES	2,212	PRICE CHANGES	693
OUTPUT DEMAND	2,144	MANUFACTURING FACILITIES	684
EXPORT TRADE	2,101	RETAILERS	633
DEBT CRISIS	1,983	CONSUMPTION	596
BUDGET DEFICITS	1,977	HOUSING MARKET	590
COMPANY EARNINGS	1,956	CONSTRUCTION	563
COMPANY PROFITS	1,913	BUSINESS CLIMATE CONDITIONS	562
REAL ESTATE	1,834	EMPLOYMENT	462
OIL GAS PRICES	1,768	COMMODITIES PRICES	459
INTERNATIONAL TRADE	1,695	RETAIL SECTOR PERFORMANCE	303
MANUFACTURING OUTPUT	1,675	ECONOMIC SURVEYS	300
TRENDS	1,603	HOME PRICES	298
BOND MARKETS	1,475	UTILITIES INDUSTRY	278
PRICE INCREASES	1,374		
Number of topics			47
Number of articles			26,139
Number of topics per article			
Average			3.53
Minimum			1
25th percentile			1
50th percentile			3
75th percentile			5
Maximum			19

Table 2: Correlation between sentiment indices and the German industrial production
This table presents the minimum, 25th, 50th, 75th, and maximum percentiles of the contemporaneous correlation between the realized German industrial production at several growth horizons h and the 282 sentiment indices. The correlation is computed over the entire sample ranging from April 2000 to April 2016 (194 observations for $h = 1$ and 183 observations for $h = 12$). See Section 3.1 for details.

	Horizon (in months)				
Percentile	$h = 1$	3	6	9	12
Min	-0.19	-0.18	-0.07	-0.06	-0.04
25th	0.14	0.23	0.22	0.22	0.19
50th	0.19	0.31	0.33	0.32	0.30
75th	0.24	0.37	0.41	0.44	0.42
Max	0.35	0.56	0.68	0.68	0.63

Table 3: Forecasting results – RMSFE, MAFE and DA measures
This table presents the Root Mean Squared Forecast Error (RMSFE), Mean Absolute Forecast Error (MAFE) and the Directional Accuracy (DA) in percent for the four models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_2$, and $\mathcal{M}_4$ presented in (23)–(26). Lower RMSFE and MAFE values are preferred while larger DA values are preferred. We consider the one– ($h = 1$), three– ($h = 3$), six– ($h = 6$), nine– ($h = 9$), twelve–month ($h = 12$) log–growth in the German industrial production. The full out–of–sample period ranges from April 2000 (March 2001 for $h = 12$) to April 2016 (194 observations for $h = 1$ and 183 observations for $h = 12$). The out–of–sample pre–crisis period ranges from April 2000 to November 2008 (104 observations for $h = 1$ and 93 observations for $h = 12$). The out–of–sample crisis period ranges from December 2008 to June 2010 (19 observations). The out–of–sample post–crisis period ranges from July 2010 to April 2016 (70 observations). Gray cells indicate that the model belongs to the model confidence set at the 75% confidence level. See Section 3.3.1 for details.

		RMSFE				MAFE				DA			
Period	h	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
Full sample	1	1.61	1.56	1.61	1.58	1.22	1.20	1.22	1.19	55.4	60.1	55.4	58.0
	3	2.91	2.45	2.45	2.31	1.79	1.66	1.71	1.62	61.7	68.5	64.9	67.0
	6	4.81	4.23	2.96	2.94	2.96	2.72	2.18	2.17	63.8	72.9	75.5	75.0
	9	6.28	5.65	3.26	3.18	3.99	3.67	2.37	2.30	65.4	71.4	76.7	77.8
	12	7.53	6.72	3.67	3.62	4.91	4.50	2.65	2.61	67.0	71.4	80.7	81.3
Pre–crisis	1	1.45	1.49	1.45	1.45	1.17	1.21	1.17	1.17	54.8	57.7	54.8	54.8
	3	1.91	1.84	1.85	1.84	1.36	1.35	1.35	1.33	65.7	65.7	64.7	65.7
	6	2.63	2.52	2.09	2.10	2.06	1.96	1.69	1.67	70.7	71.7	77.7	75.7
	9	3.45	3.40	2.67	2.63	2.66	2.70	2.02	1.97	77.7	72.9	76.0	77.0
	12	3.93	3.95	2.99	2.92	3.19	3.25	2.22	2.18	76.3	75.2	78.4	79.5
Crisis	1	2.97	2.49	2.97	2.86	2.28	1.91	2.28	2.17	57.8	63.1	57.8	63.1
	3	7.51	5.38	5.23	4.90	5.87	3.90	4.19	3.93	52.6	84.2	73.2	73.2
	6	12.82	10.64	5.93	5.88	10.57	8.67	4.66	4.60	42.1	84.2	94.7	94.7
	9	16.32	14.11	6.23	6.16	14.39	11.88	5.50	5.39	26.3	63.2	94.7	94.7
	12	19.06	16.67	7.37	7.32	17.37	14.13	6.33	6.21	21.1	68.4	94.7	94.7
Post–crisis	1	1.26	1.31	1.26	1.22	1.00	1.00	1.00	0.95	55.7	62.8	55.7	64.3
	3	1.60	1.88	2.00	1.76	1.32	1.52	1.56	1.42	58.5	68.5	62.8	67.1
	6	2.78	2.88	2.79	2.76	2.17	2.18	2.20	2.21	60.0	71.4	67.1	68.5
	9	3.96	3.80	2.78	2.64	2.98	2.78	2.01	1.91	60.0	71.4	72.9	74.2
	12	5.32	4.60	2.90	2.87	3.82	3.55	2.23	2.21	67.1	67.1	80.0	81.4

Table 4: List of topic clusters
This table presents the six clusters of topics, which have been constructed manually. See Section 3.3.2 for details.

(1) GDP Output	(2) Job Market	(3) Prices & Interest Rate
BUDGET DEFICITS	EMPLOYMENT	BOND MARKETS
COMPANY EARNINGS	EMPLOYMENT GROWTH	COMMODITIES PRICES
COMPANY PROFITS	JOB CREATION	CURRENCIES
CONSUMPTION	UNEMPLOYMENT RATES	DEBT CRISIS
ECONOMIC CONDITIONS	WAGES SALARIES	ECONOMIC POLICY
ECONOMIC DECLINE		EXPORT TRADE
ECONOMIC GROWTH		IMPORT TRADE
ECONOMIC STIMULUS		INFLATION
GROSS DOMESTIC PRODUCT		INTERNATIONAL TRADE
MANUFACTURING OUTPUT		INTEREST RATE
OUTPUT DEMAND		OIL GAS PRICES
RECESSION		PRICES
SALES FIGURES		PRICE CHANGES
		PRICE INCREASES
(4) Real Estate	(5) Surveys	(6) Others
CONSTRUCTION	BUSINESS CLIMATE CONDITIONS	GERMAN CHANCELLORS
HOME PRICES	BUSINESS CONFIDENCE	MANUFACTURING FACILITIES
HOUSING MARKET	CONSUMER CONFIDENCE	OIL GAS INDUSTRY
REAL ESTATE	ECONOMIC SURVEYS	RETAILERS PERFORMANCE
	TRENDS	RETAIL SECTOR
		UTILITIES INDUSTRY

Table 5: Forecasting results – impact of the lexicon

This table presents the RMSFE and MAFE results for $\mathcal{M}_3$ when an alternative lexicon is used. *ALL* is the model that uses all lexicons to calculate the sentiment (282 sentiment indices), *HEN* stands for the Henry (2008) lexicon (94 sentiment indices), *HAR* stands for the Harvard General Inquirer lexicon (94 sentiment indices), and *LM* stands for the Loughran and McDonald (2011) lexicon (94 sentiment indices). See Table 3 for details.

		RMSFE				MAFE			
Period	h	*ALL*	*HEN*	*HAR*	*LM*	*ALL*	*HEN*	*HAR*	*LM*
Full sample	1	1.61	1.61	1.62	1.61	1.22	1.22	1.23	1.22
	3	2.45	2.39	2.35	2.50	1.71	1.63	1.70	1.69
	6	2.96	3.07	3.29	3.27	2.18	2.17	2.32	2.20
	9	3.26	3.65	3.58	3.41	2.37	2.60	2.62	2.52
	12	3.67	3.97	3.80	3.66	2.65	2.88	2.78	2.80
Pre–crisis	1	1.45	1.45	1.45	1.45	1.17	1.17	1.17	1.17
	3	1.85	1.74	1.94	1.83	1.35	1.26	1.39	1.32
	6	2.09	2.10	2.14	2.06	1.69	1.62	1.66	1.66
	9	2.67	2.07	2.71	2.66	2.02	2.02	2.07	2.08
	12	2.99	2.88	2.84	2.81	2.22	2.21	2.28	2.22
Crisis	1	2.97	2.97	3.02	3.00	2.28	2.28	2.33	2.29
	3	5.23	4.86	4.64	5.57	4.19	3.93	3.94	4.29
	6	5.93	6.20	6.59	6.96	4.66	4.98	5.29	5.43
	9	6.23	7.04	6.81	6.49	5.50	5.95	5.94	5.98
	12	7.37	7.72	7.90	6.29	6.33	6.49	6.96	5.47
Post–crisis	1	1.26	1.28	1.28	1.26	1.00	1.01	1.01	0.99
	3	2.00	2.19	1.93	1.95	1.56	1.54	1.55	1.52
	6	2.79	2.94	3.29	3.09	2.20	2.18	2.44	2.10
	9	2.78	3.43	3.37	3.10	2.01	2.50	2.48	2.17
	12	2.90	3.71	3.14	3.69	2.23	2.80	2.32	2.85

Table 6: Forecasting results – impact of the aggregation scheme

This table presents the RMSFE and MAFE results for $\mathcal{M}_3$ when an alternative aggregation schemes is used. We use *Almon* as the benchmark model. It uses the first and second order Almon polynomials to calculate the sentiment (282 sentiment indices) with $\tau = 180$. Robustness of the Almon polynomial aggregation is studied by considering a moving average aggregation with $\tau = 180$ (141 sentiment indices). Those results are shown in column *MA*. Robustness to the topic aggregation is studied by considering a global aggregation of all the topics while using the first and second order Almon polynomials (6 sentiment indices). Those results are shown in the column *All Topics*. All forecasts are obtained under the default approach of computing the sentiment values per text using the Henry (2008), Harvard General Inquirer, and Loughran and McDonald (2011) lexicons. See Table 3 for details.

		RMSFE			MAFE		
Period	*h*	*Almon*	*MA*	*All Topics*	*Almon*	*MA*	*All Topics*
	1	1.61	1.61	1.66	1.22	1.22	1.25
	3	2.45	2.50	2.80	1.71	1.70	1.86
Full sample	6	2.96	3.04	4.82	2.18	2.04	3.17
	9	3.26	3.54	6.55	2.37	2.34	4.32
	12	3.67	3.49	8.06	2.65	2.54	5.44
	1	1.45	1.45	1.45	1.17	1.17	1.17
	3	1.85	1.85	1.73	1.35	1.35	1.23
Pre–crisis	6	2.09	1.92	2.20	1.69	1.48	1.66
	9	2.67	2.54	2.95	2.02	1.87	2.14
	12	2.99	2.60	3.57	2.20	2.05	2.70
	1	2.97	2.97	3.08	2.28	2.26	2.41
	3	5.23	5.45	6.72	4.19	4.32	5.86
Crisis	6	5.93	6.83	11.55	4.66	5.37	10.25
	9	6.23	7.87	15.25	5.50	6.52	13.96
	12	7.37	7.00	18.53	6.33	5.92	16.93
	1	1.26	1.26	1.38	1.00	1.00	1.07
	3	2.00	2.00	2.19	1.56	1.50	1.70
Post–crisis	6	2.79	2.63	4.39	2.20	1.93	3.38
	9	2.78	2.72	6.19	2.01	1.84	4.70
	12	2.90	3.05	7.68	2.23	2.28	5.96

Figure 1: Methodology
This figure presents a scheme of the nine steps of the building blocks of the methodology. See Section 2 for details.

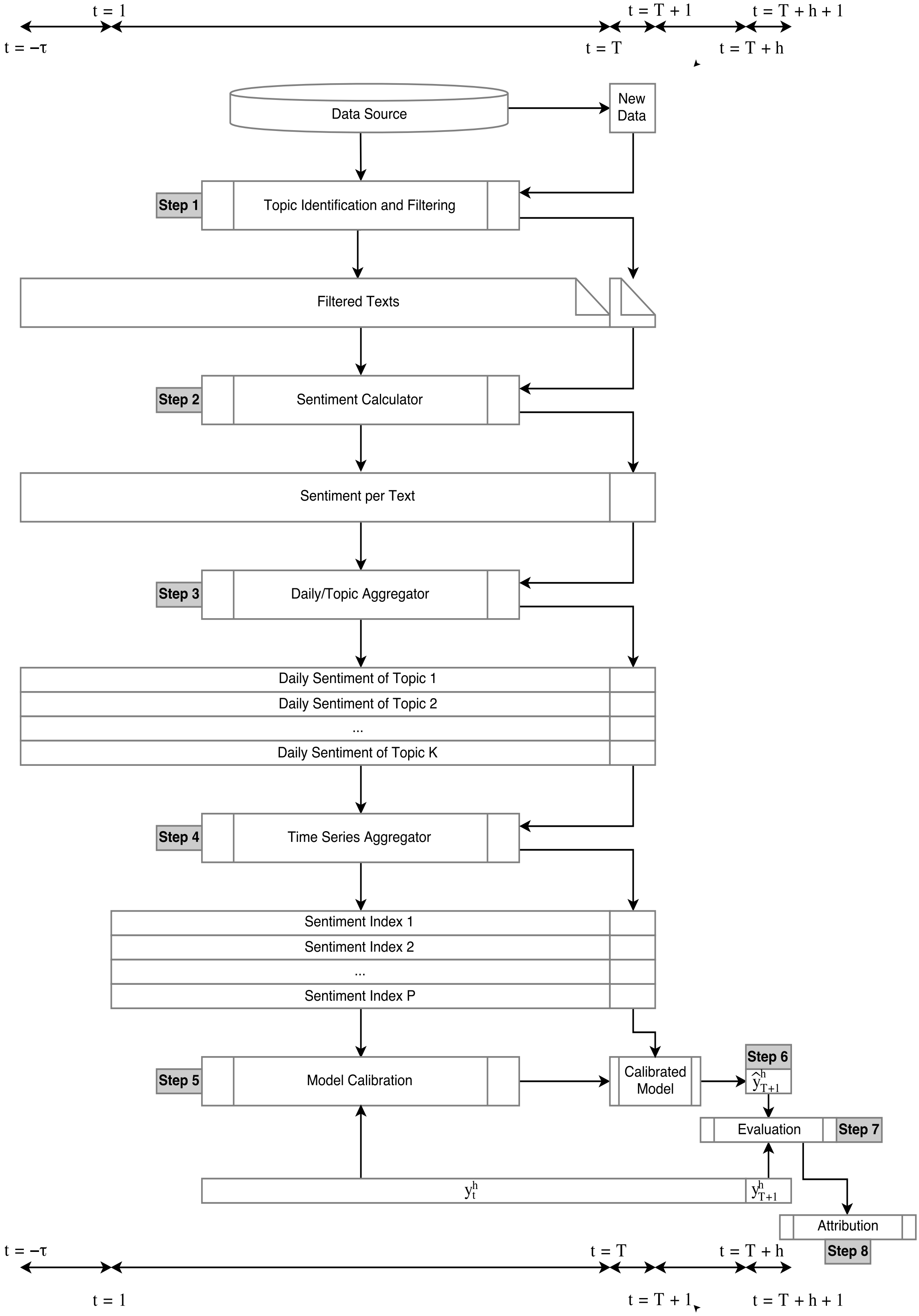


Figure 2: German ESI and realized industrial production growth
This figure presents the standardized monthly German Economic Sentiment Indicator along with the standardized realized rolling twelve–month log–growth of the German industrial production from February 1996 to April 2016 (242 observations). Standardization is done over the entire series by first subtracting the mean from the series and then dividing it by its standard deviation. The first vertical line from the left indicates the start of the out–of–sample periods, that is March 2001, for the twelve–month growth in German industrial production. The gray zone indicates the German financial crisis period, which, according to Ulbricht et al. (2017), spans from December 2008 to June 2010 (19 observations). See Section 3.1 for details.

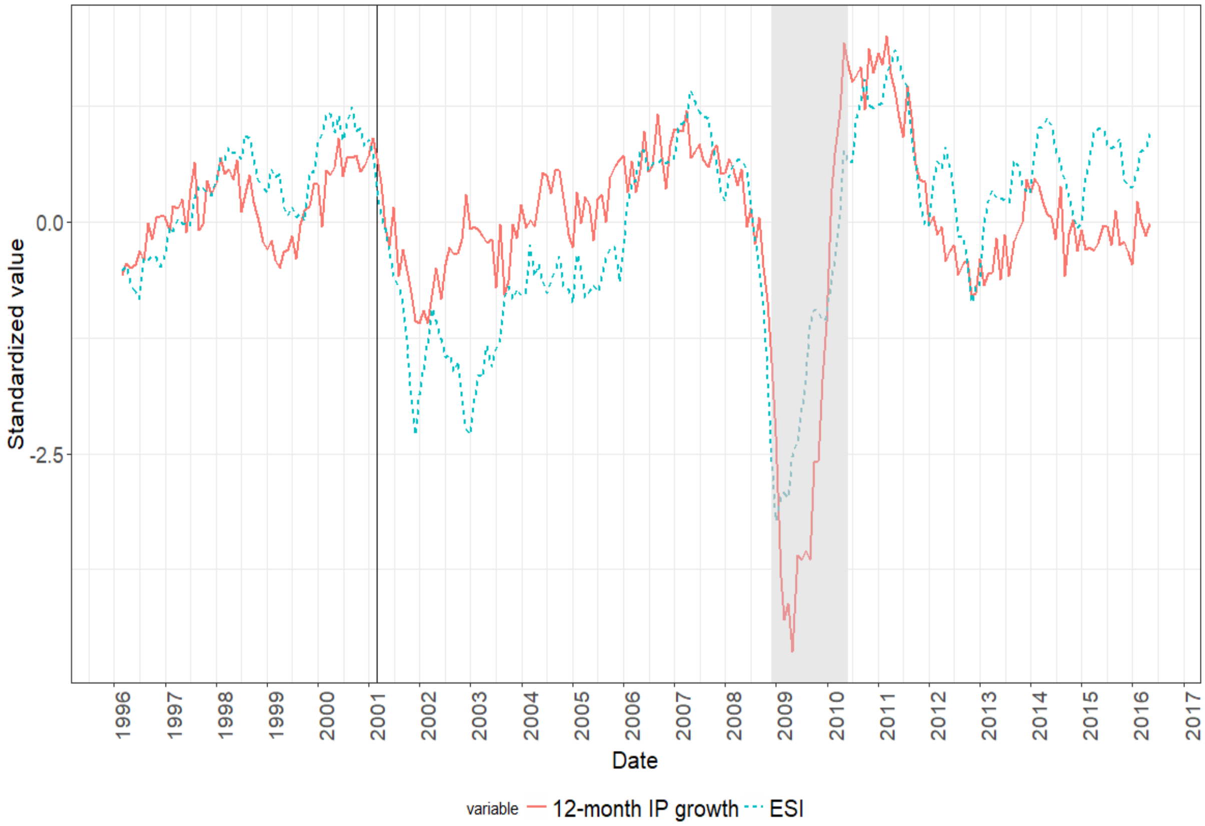


Figure 3: Yearly averages of the individual news articles' sentiments
This figure presents the yearly averages of the individual news articles sentiment (without aggregation) for the period ranging from 1994 to 2016. The gray zone indicates the December 2008-June 2010 German financial crisis period. See Section 3.1 for details.

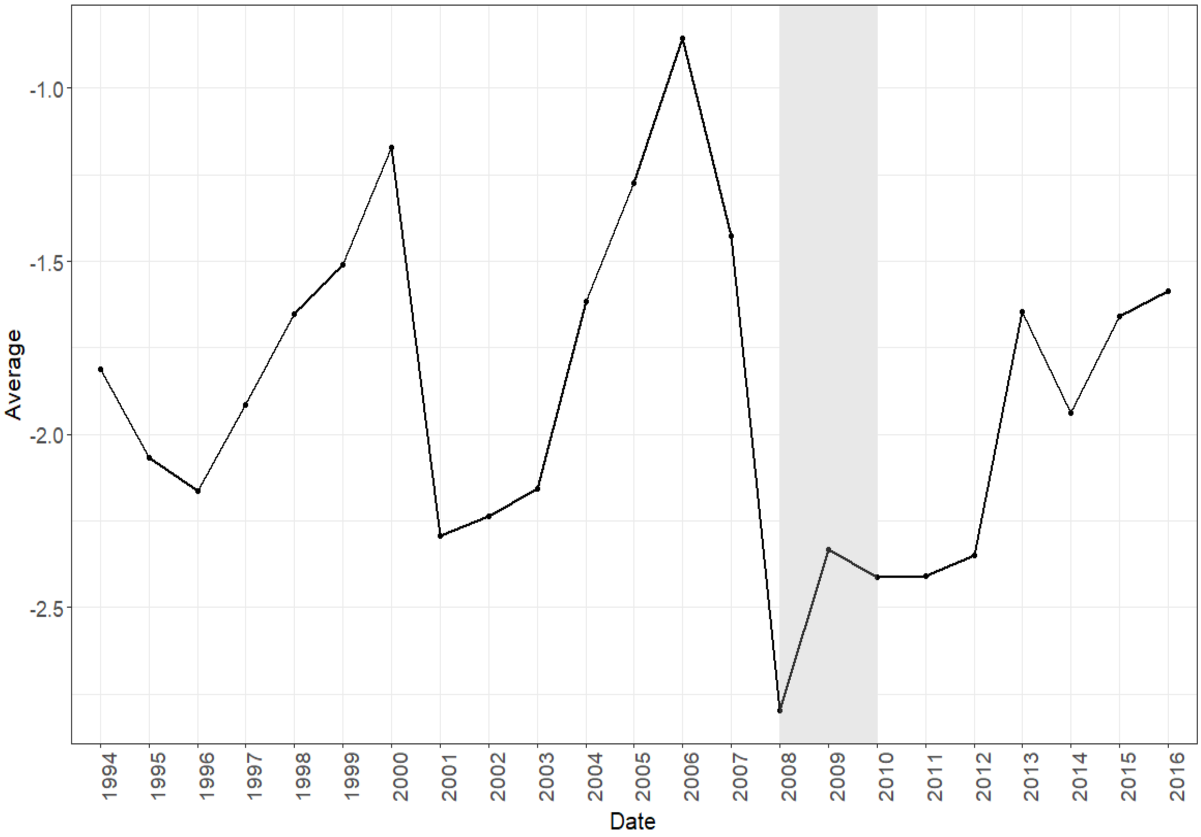


Figure 4: Yearly average of the 47 topic sentiment indices
This figure presents the yearly average of 47 sentiment indices for the period ranging from 1994 to 2016. Sentiment measures are constructed using the Loughran and McDonald (2011) lexicon and the first order Almon polynomial using $\tau = 180$. See Section 3.1 for details. Gray cells indicate that there were no text documents about that topic during that year.

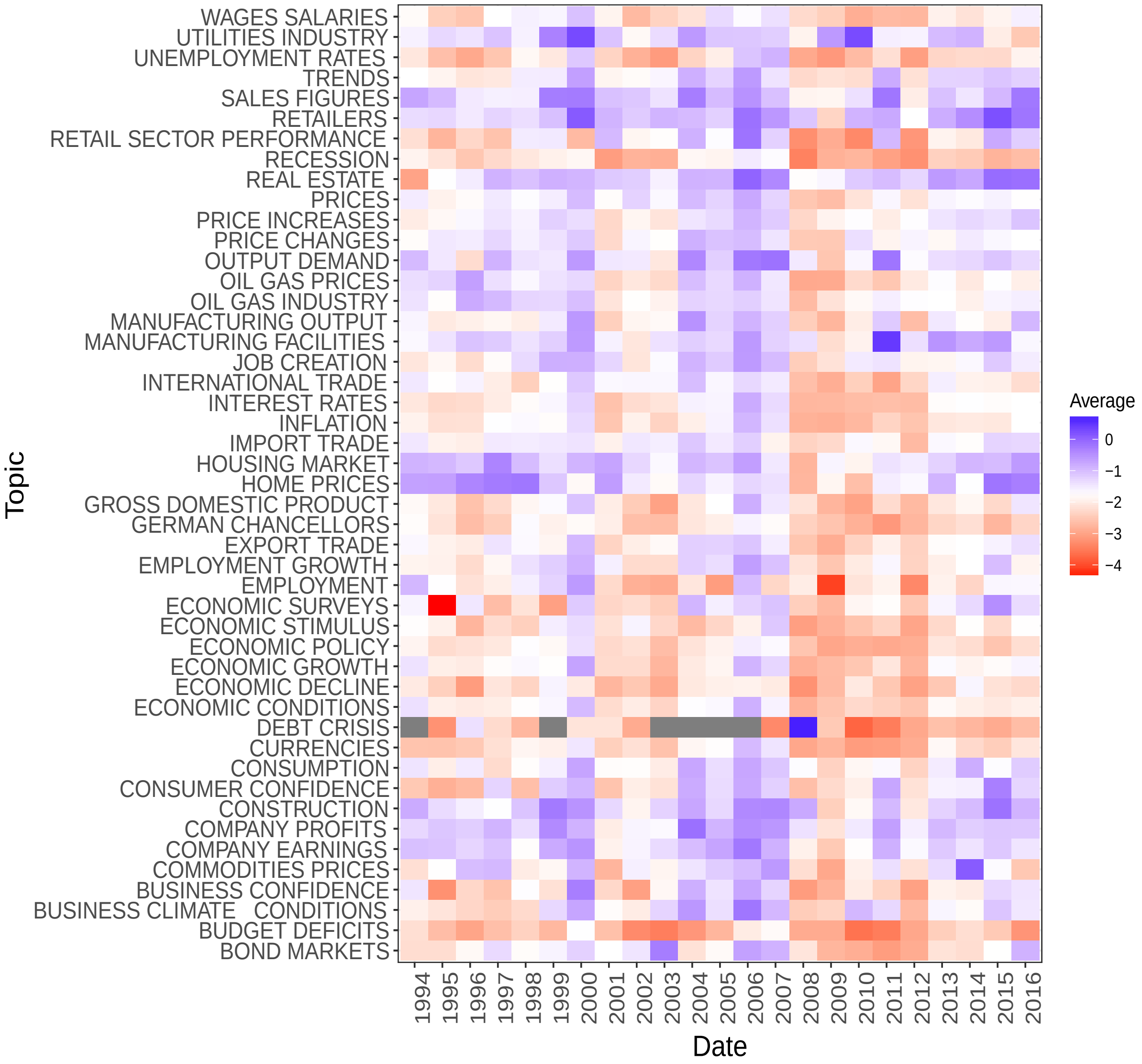


Figure 5: Model forecasts

This figure presents the forecasts provided by the ESI model ($\mathcal{M}_2$) and the text sentiment model ($\mathcal{M}_3$) together with the realized values of the twelve–month German industrial production log–growth. The out–of–sample window ranges from March 2001 to April 2016 (183 observations). The gray zone indicates the December 2008-June 2010 German financial crisis period. See Section 3.3.1 for details.

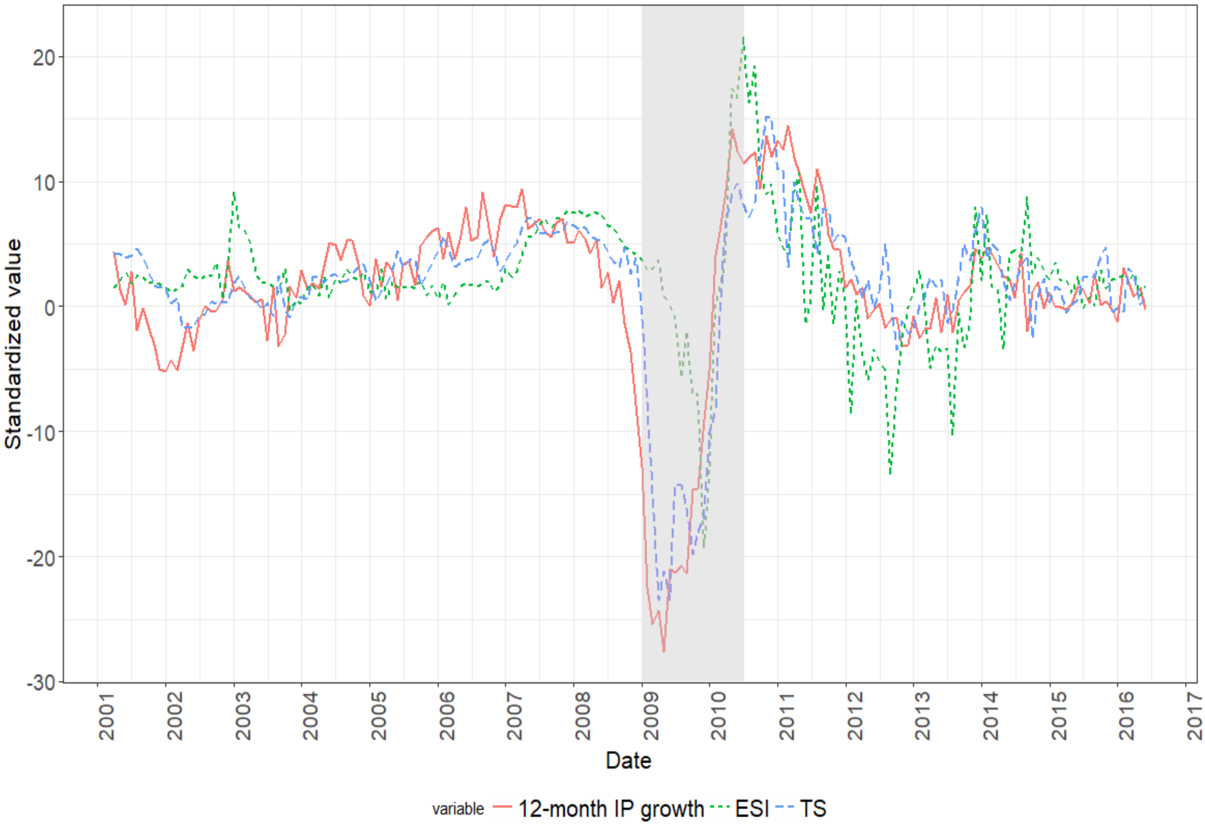
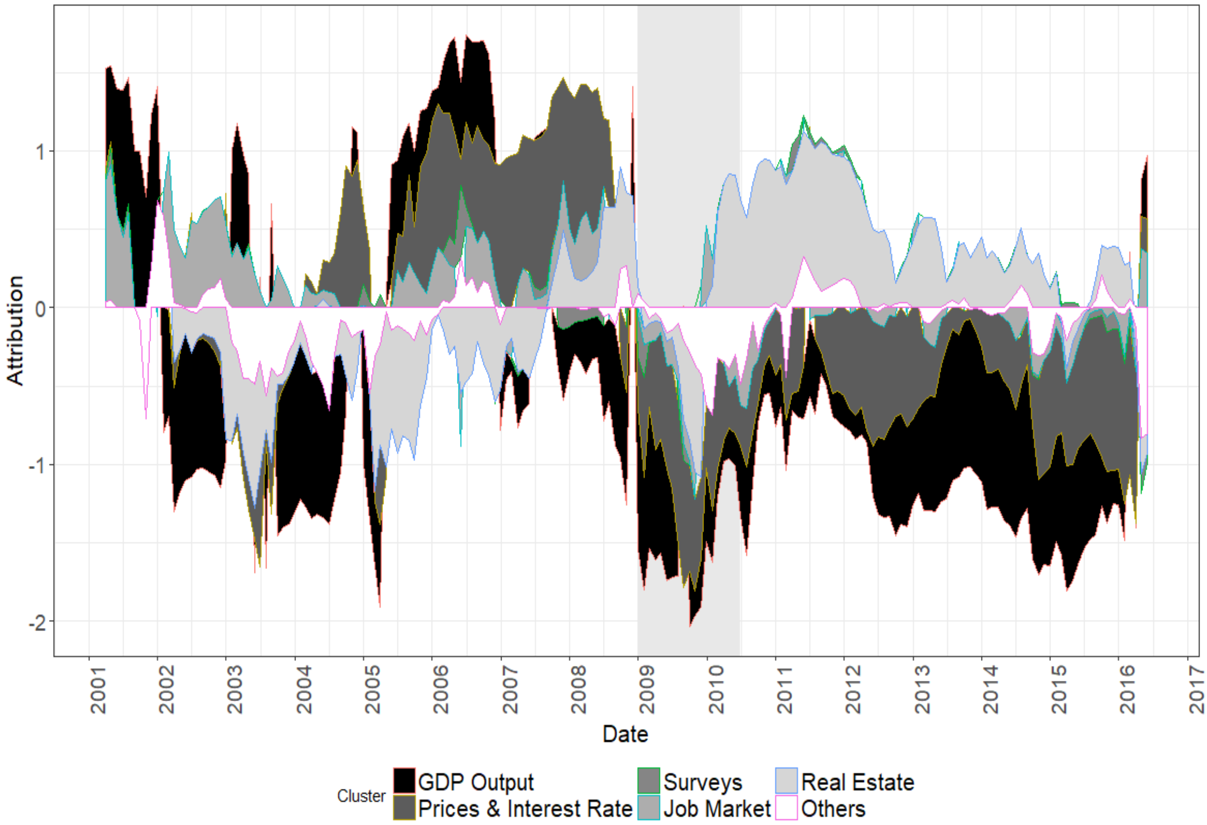


Figure 6: Forecast attribution
This figure presents the cluster attribution for the out–of–sample forecasts of the twelve–month German industrial production log–growth. The period ranges from March 2001 to April 2016 (183 observations). The attribution vector for a given date is scaled by dividing each element of the attribution vector by the L^2–norm of the attribution vector for that date. The gray zone indicates the December 2008-June 2010 German financial crisis period. Positive (negative) value indicates that the topic contributes positively (negatively) to the forecast and therefore increases (decreases) the forecast of the German industrial production log–growth. See Section 3.3.2 for details.


Attribution
Date
Cluster
GDP Output
Prices & Interest Rate
Surveys
Job Market
Real Estate
Others