RUHR
ECONOMIC PAPERS

Yuliya Shrub
Jonas Rieger
Henrik Müller
Carsten Jentsch

Text Data Rule – Don't They?

A Study on the (Additional) Information of Handelsblatt Data for Nowcasting German GDP in Comparison to Established Economic Indicators

technische universität
dortmund

#964

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Ludger Linnemann
Technische Universität Dortmund, Department of Business and Economics
Economics – Applied Economics
Phone: +49 (0) 231/7 55-3102, e-mail: : Ludger.Linnemann@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Manuel Frondel, Prof. Dr. Torsten Schmidt, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #964

Responsible Editor: Ludger Linnemann

All rights reserved. Essen, Germany, 2022

ISSN 1864-4872 (online) – ISBN 978-3-96973-128-4

The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors.

Ruhr Economic Papers #964

Yuliya Shrub, Jonas Rieger, Henrik Müller, and Carsten Jentsch

Text Data Rule - Don’t They?

A Study on the (Additional) Information of Handelsblatt Data for Nowcasting German GDP in Comparison to Established Economic Indicators

technische universität
dortmund

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973128
ISSN 1864-4872 (online)
ISBN 978-3-96973-128-4

Yuliya Shrub, Jonas Rieger, Henrik Müller, and Carsten Jentsch[1]

Text Data Rule – Don't They?

A Study on the (Additional) Information of Handelsblatt Data for Nowcasting German GDP in Comparison to Established Economic Indicators

Abstract

The prompt availability of information on the current state of the economy in real-time is required for prediction purposes and crucial for timely policy adjustment and economic decision-making. While important macroeconomic indicators are reported only quarterly and also published with substantial delay, other related data are available more frequently, that is monthly, weekly, daily or even more often. In this regard, the goal of nowcasting methods is to make use of such more frequently collected variables to update predictions of less often reported variables such as e.g. GDP growth. In this paper, we propose a mixed-frequency model to investigate the potential of using text data in form of newspaper articles for nowcasting German GDP growth. Newspaper text data appears to be very helpful in this regard as it directly explains economic and social progress influencing GDP growth and as it is updated frequently without any substantial delay. We compare several setups based on commonly used macro variables with and without additionally included information from text data (extracted in an unsupervised manner) as well as a setup only based on such text data. To deal with the high dimensionality of the considered data, we make use of principal component regression, penalization techniques and random forest. Comparing our results leads to the conclusion that there are certain benefits achievable when text data are included for nowcasting, but the unsupervised extraction of information from text data tends to still contain too much irrelevant noise hampering the performance of the resulting nowcasting approach.

JEL-Codes: C52, C53, C55, E37

Keywords: Topic model; latent Dirichlet allocation; text mining; econometrics; gross domestic product; prediction; forecast

August 2022

[1] All TU Dortmund. – All correspondence to: Jonas Rieger, Technical University Dortmund, Otto-Hahn-Str. 6, 44227, Dortmund, Germany, e-mail: rieger@statistik.tu-dortmund.de

1 Introduction

The Organisation for Economic Cooperation and Development (OECD) defines Gross Domestic Product (GDP) as the measure of the value added created through the production of goods and services in a country during a specific period (OECD, 2022a). GDP is one of the main measures of economic activity. In order to compare the GDP values of the most recent period to the previous period, the GDP growth measure is used. The analysis of GDP growth plays an important role in economic decision-making. Usually, GDP growth is published quarterly with a substantial lag, which impedes assessing the state of the economy in real-time. Therefore, the problem of nowcasting GDP growth has received significant attention in the scientific research (Ashwin et al., 2021, Gayer et al., 2014, Thorsrud, 2020).

Besides the common approaches, which use economic and survey data (Carriero et al., 2015, Gayer et al., 2014, McCracken et al., 2021), there is increasing interest in novel text-based methods (Gentzkow et al., 2019) for predicting GDP growth (Ashwin et al., 2021, Kalamara et al., 2020, Thorsrud, 2020). Textual information extracted from newspaper articles has several advantages compared to typically used economic and survey data. First, newspaper data directly explains the economic and social processes that can influence GDP growth. Second, the news data updates frequently without any substantial delay. In addition, the news influences a broader range of economic agents in comparison to the professional economic reports.

In this work, we construct a nowcasting approach for German GDP growth. The approach is based on financial, economic and survey indicators and on German-language newspaper data. We extract the textual information using the latent Dirichlet allocation (LDA) topic model (Blei, 2012, Ke et al., 2020). According to our experiments, text data can be competitive with other types of information for nowcasting German GDP growth.

The follwoing sections are structured as follows. In Section 1.1, we review the recent studies on nowcasting GDP growth and the use of text data. Section 1.2 contains the description of the problem and the goals of this work. Section 2 describes the statistical methods for nowcasting German GDP growth. Section 3 includes information about the used data sets. In Section 4, we present the empirical set-up for nowcasting German GDP growth. The results of the work are explained in Section 5. Finally, Section 6 summerizes the results and Section 7 gives an outlook for future work.

1.1 Literature Overview

Traditional approaches for GDP prediction use economic indicators and/or survey data (Bec and Mogliani, 2015, Carriero et al., 2015, Gayer et al., 2014, McCracken et al., 2021). However, recent studies show that text data from newspapers can be competitive with economic indicators and survey data for nowcasting or forecasting GDP growth and other macroeconomic indicators (Ashwin et al., 2021, Ellingsen et al., 2021, Kalamara et al., 2020, Thorsrud, 2020). Several approaches use text data in predictions. For example, a number of studies work with raw text data - words and/or sentences (Aguilar et al., 2021, Ashwin et al., 2021). In order to capture the structure of text, some studies use N-grams to turn text into time series (e.g., see Kalamara et al., 2020). Alternatively, topic-based approaches are exploited to aggregate text and extract structured information. For example, Ardia et al. (2019) and Aprigliano et al. (2022) use manually topic-labeled newspaper articles. In order to increase the efficiency of topic extraction, topic models can be used. The latent Dirichlet allocation (LDA) is a widely applied topic model, showing promising results for predicting macroeconomic indicators (Ellingsen et al., 2021, Thorsrud, 2020). Thorsrud (2020) shows that sentiment-adjusted topics, which the author adds in the mixed frequency dynamic factor model, produce competitive nowcasting results with forecast combination systems for Norwegian GDP growth. Text data are extracted with the LDA from the largest Norwegian business newspaper. Ellingsen et al. (2021) incorporate text data from US newspapers in the form of LDA topics, with financial and economic indicators from the FRED-MD data set into nowcasting and forecasting models of US GDP, consumption and investment growth.

In order to transform the pre-processed text data into time series, dictionary-based sentiment (Algaba et al., 2020) calculation is frequently used (Aprigliano et al., 2022, Ardia et al., 2019, Ashwin et al., 2021, Kalamara et al., 2020, Thorsrud, 2020). Many sentiment dictionaries are available in English. Therefore, some studies (Aprigliano et al., 2022, Ashwin et al., 2021) translate non-English articles into English for sentiment adjustment. In contrast, the sentiment dictionary can also be translated into a target language (Thorsrud, 2020). In addition, available non-English dictionaries can be used. For example, Ashwin et al. (2021) pre-process text data based on the German-language business dictionary, created by Bannier et al. (2019). Dictionaries can operate at the word (Correa et al., 2017) or sentence level (Hutto and Gilbert, 2014) and cover general-purpose or economic vocabularies.

When predicting GDP growth, relevant data are released with different frequencies. Hence, the mixed frequency problem occurs. As a solution approach, Baffigi et al. (2004) consider bridge models for forecasting euro area GDP. Carriero et al. (2015) use Bayesian mixed frequency models for nowcasting US GDP growth. Thorsrud (2020) and Andreini et al. (2020) estimate the mixed frequency dynamic factor models via the Kalman filter. Ellingsen et al. (2021) use unrestricted MIDAS models (Ghysels et al., 2004). Gayer et al. (2014) and Bec and Mogliani (2015) apply the so-called blocking approach (Chen et al., 2012), which shows good nowcasting and forecasting performance for high dimensional data with good interpretation and fast calculation properties.

Generally, the prediction of GDP growth is based on statistical and machine learning models. Ellingsen et al. (2021) combine principal component analysis (Stock and Watson, 1989), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996)) and random forest (Breiman, 2001) models. Kalamara et al. (2020) compare US GDP growth forecasts with different models. As a result, Ridge regression (Hoerl and Kennard, 1970), artificial neural networks (Rumelhart et al., 1986) and support vector machine (SVM) (Drucker et al., 1996) show the most promising performances. Ashwin et al. (2021) analyse various models and present that Ridge regression provides the best results in normal times and non-linear models during the periods of large shifts.

1.2 Problem Statement

The main objective of this work is to construct the nowcasting approach of German GDP growth with the use of text data from a German-language newspaper. This approach should produce *monthly* nowcasts of German GDP growth based on the different types of mixed frequency data including economic data, financial indicators, newspaper data and survey data. In the context of text data, we focus on the use of LDA topic models to extract important information from newspaper articles. The goal is to analyze the role of the number of topics in producing the nowcasts. Additionally, we aim to compare the nowcasts produced with the sentiment-unadjusted and sentiment-adjusted topics. Finally, we intend to compare the role of text data with the other types of mixed frequency data in the nowcasting of German GDP growth over the quarter.

2 Statistical Methods for Nowcasting German GDP Growth

This section describes the statistical methods applied for the nowcasting model construction. In particular, we explain the principle of model construction for the mixed frequency data case as well as statistical and machine learning models for GDP growth prediction, including methods for dimensionality reduction of correlated data. Moreover, we describe the theoretical background of text data aggregation with topic models. We also show time series data analysis and transformation, which are essential for constructing nowcasting models.

2.1 Nowcasting Model Construction for Mixed Frequency Data

Information about German GDP growth is released on a quarterly basis (see Section 3.1). In contrast, news text data are published on a daily basis, and other economic, financial and survey indicators are released with daily or monthly frequencies. Data published at different frequencies results in the mixed frequency problem, which has to be adequately addressed when modeling and analyzing the data. In addition, most of the data are published with a delay that leads to incompleteness of predictors' information during the nowcasting period. Thus, we face the so-called ragged edge problem.

In order to overcome these issues and produce monthly nowcasts, various statistical methods have been proposed in the literature. In this work, following Gayer et al. (2014); Carriero et al. (2015) and Bec and Mogliani (2015), we apply the so-called blocking approach. The blocking approach has several advantages over other methods when dealing with mixed frequency data. In comparison to bridge models, the blocking approach can directly use data that are available at any time, without the need to extrapolate intra-quarterly missing information (Gayer et al., 2014). Unlike MIDAS and dynamic factor models, the coefficients of the blocking approach can be estimated with the ordinary least squares (OLS) method (Bec and Mogliani, 2015). For MIDAS and dynamic factor models, it is computationally expensive to consider a large time series dimension leading to a large number of coefficients in the model. Finally, the blocking approach enables to directly interpret and evaluate the impact of predictor variables into the nowcasting model (Bec and Mogliani, 2015).

The idea of the blocking approach originates from the engineering literature (Chen et al., 2012). The blocking approach is based on splitting high-frequency information into multiple low-frequency time series (Gayer et al., 2014). In our case, a monthly time series is divided into three time series with a quarterly frequency. That means the first time series collects information of the first month of the quarter, the second time series the data published in the second month of the quarter, while the third time series gathers information regarding the third month of the quarter.

For a formal description, let L denote the set of indices of all predictor variables, where p is the lag order of a response quarterly variable and q is the maximal lag order of predictor variables. We consider the following model for nowcasting GDP growth:

$$y_t = a_0 + \sum_{r=1}^{p} a_r y_{t-r} + \sum_{l \in L} \sum_{h_{mon}=0}^{q} c_{l,h_{mon}} x_{l,t-h_{mon}} + \epsilon_t, \tag{1}$$

where y_t is the nowcast of GDP growth at quarter t, a_0 is a constant term, r is an index of response lags with the coefficients a_r, and $x_{l,t-h_{mon}}$ is a predictor variable with index l, a monthly lag h_{mon} and a model coefficient $c_{l,h_{mon}}$. The model error ϵ_t is a zero mean error (Carriero et al., 2015). From Equation (1) we can notice that the response lag has the index $r \in \{1, \ldots, p\}, r, p \in \mathbb{N}$. The lag of the monthly predictors has a non-negative rational number $h_{mon} \in \left\{0, \frac{1}{3}, \frac{2}{3}, \ldots, q\right\}$.

In this work, we focus on nowcasting GDP growth based on data that are available at the end of each month of the predicted quarter. In addition, we do not produce the forecasting or backcasting GDP growth, as German GDP growth is published with a one month delay (see Section 3.1). Therefore, we include into Equation (1) only the first lag of German GDP growth ($p = 1$) and the first three lags of predictor variables $\left(q = \frac{2}{3}\right)$. However, Equation (1) cannot be estimated without complete information of predictors in the predicted quarter. According to the blocking approach, we can divide the model into three parts considering the information, which is available at the end of each month. For notation simplicity, instead of real number lags for each predictor $x_{l,t-h_{mon}}$, we use $x_{l,t}^{(m)}$, where $m \in \{1, 2, 3\}$ detects the month of the quarter. Thus, each predictor variable consists of three blocked time series for quarter t: $x_{l,t-\frac{2}{3}} = x_{l,t}^{(1)}$, $x_{l,t-\frac{1}{3}} = x_{l,t}^{(2)}$ and $x_{l,t} = x_{l,t}^{(3)}$. For this purpose, let L_1, L_2 and L_3 be pairwise disjoint subsets of indices of predictor variables in the set L ($L = L_1 \cup L_2 \cup L_3$). Precisely, L_1 is a subset of indices with the first monthly release at the first month ($m = 1$) of the quarter t, L_2 has indices of predictor variables with the first monthly release at the second month ($m = 2$) of the quarter t, etc.

Following the newly introduced notation, we obtain three nowcasting models

$$y_t = a_{0,1} + a_{1,1}y_{t-1} + \sum_{l \in L_1} c_{l,1}^{(1)} x_{l,t}^{(1)} + \epsilon_{t,1}, \tag{2}$$

$$y_t = a_{0,2} + a_{1,2}y_{t-1} + \sum_{l \in L_1} \left(c_{l,2}^{(1)} x_{l,t}^{(1)} + c_{l,2}^{(2)} x_{l,t}^{(2)} \right) + \sum_{l \in L_2} c_{l,2}^{(1)} x_{l,t}^{(1)} + \epsilon_{t,2}, \tag{3}$$

$$\begin{aligned} y_t = a_{0,3} + a_{1,3}y_{t-1} + \sum_{l \in L_1} \left(c_{l,3}^{(1)} x_{l,t}^{(1)} + c_{l,3}^{(2)} x_{l,t}^{(2)} + c_{l,3}^{(3)} x_{l,t}^{(3)} \right) + \\ \sum_{l \in L_2} \left(c_{l,3}^{(1)} x_{l,t}^{(1)} + c_{l,3}^{(2)} x_{l,t}^{(2)} \right) + \sum_{l \in L_3} c_{l,3}^{(1)} x_{l,t}^{(1)} + \epsilon_{t,3}. \end{aligned} \tag{4}$$

Altogether, in general form, we get

$$y_t = a_{0,m} + a_{1,m}y_{t-1} + \sum_{i=1}^{m} \sum_{j=1}^{m-i+1} \sum_{l \in L_i} c_{l,m}^{(j)} x_{l,t}^{(j)} + \epsilon_{t,m}, \tag{5}$$

where $a_{0,m}$ and $a_{1,m}$ are coefficients of the nowcasting model at the end of the month $m \in \{1, 2, 3\}$, $c_{l,m}^{(j)}$ is a coefficient for the predictor with the index number l, for the model at the end of the month m with available predictor's information for the month (j). Thus, in Equation (5) we sum all information available at the end of the month m over all subsets of released indices of predictor variables. The error $\epsilon_{t,m}$ is a error with a zero mean.

According to Chen et al. (2012) and Zamani et al. (2011), the blocked system should be stationary. To achieve this, we have to transform some variables.

2.2 Time Series Stationarity and Data Transformation

Suppose we are dealing with a real-valued time series process $\{X_t, t \in \mathbb{Z}\}$ and observe a realization of length $n \in \mathbb{N}$ of the time series process denoted by $x_1, \ldots, x_n$. The classical decomposition of a time series process $\{X_t, t \in \mathbb{Z}\}$ is according to the following formula (Brockwell and Davis, 1991):

$$X_t = m_t + s_t + Z_t^{rand}, \tag{6}$$

where s_t is a seasonal component with a period d_{per}, m_t is a deterministic trend component, and Z_t^{rand} is a random (stationary) component.

To apply the blocking approach, we should ensure the stationarity of the system. Generally, this means that the distributional properties of the system are time-invariant (Zamani et al., 2011). Because of the time dependence structure of the monthly time series in the blocked system, stationarity of the system is in general questionable. Hence, in practice, we should remove possible non-stationary behavior of all the time series to be included in the nowcasting model (McCracken et al., 2021). In the ideal case, we would like to obtain random component Z_t^{rand} in Equation (6) that is stationary for each time series. A stationarity property allows applying statistical models for time series prediction and drawing statistical inferences.

Following Gayer et al. (2014), Carriero et al. (2015) and McCracken et al. (2021), we do not focus on finding the appropriate stationary model for each time series, but on the elimination of possible non-stationary behavior of the time series.

Non-stationarity of a time series can be caused by many factors, e.g., by the existence of a seasonal component or the presence of a deterministic or non-deterministic trend. Eliminating such factors can potentially provide a stationary time series. Hence, for the data analysis, our goal is to detect possible non-stationary time series y_t, transform them and check whether the non-stationary behavior is properly extracted and eliminated in Z_t^{rand}. There exist several approaches for defining non-stationary behavior of time series, e.g., based on the visual analysis of time series plots, evaluating sample autocorrelation functions and hypothesis testing (Cryer and Chan, 2008). We use a combination of these methods. The visual analysis allows detecting possible trending behavior and variance changes over time. The sample autocorrelation function $\hat{\rho}(h)$ at lag $h \geq 0$ computed from the sample $x_1, \ldots, x_n$ with sample mean $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is defined as (Brockwell and Davis, 1991):

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\frac{1}{n}\sum_{j=1}^{n-h}\left(x_{j+h}-\bar{x}\right)\left(x_j-\bar{x}\right)}{\frac{1}{n}\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)},$$

where $\hat{\gamma}(h)$ denotes the sample autocovariance function, which can be estimated only for a stationary time series in theory. However, following Brockwell and Davis (1991) and Cryer and Chan (2008), we can use it to detect possible non-stationarities in time series data. In particular, a slow decay of $|\hat{\rho}(h)|$ as h increases can be an indicator for non-stationarity e.g. in form of a (deterministic) trend in time series.

Another approach for detecting non-stationary behavior is hypothesis testing. In our work, to test also for stochastic trends, we apply the test for the presence of a unit root causing such a trend in the autoregressive polynomial, as suggested by Dickey and Fuller (1979). The idea consists in Dickey–Fuller reparametrization of

$$\Delta X_t = \phi_1^* X_{t-1} + \phi_2^* \Delta X_{t-1} + \cdots + \phi_p^* \Delta X_{t-p+1}, \forall t \in \mathbb{Z},$$

where $\Delta X_t = X_t - X_{t-1}$ is a difference operator, $\phi_1^* = \sum_{i=1}^{p} \phi_i - 1$, and $\phi_j^* = -\sum_{i=j}^{p} \phi_i, j \in \{2, \ldots, p\}$. In cases when the autoregressive polynomial $\phi(z)$ has a unit root at 1, then $0 = \phi(1) = -\phi_1^*$. Therefore, the null hypothesis of the augmented Dickey–Fuller test (ADF test) on the presence of a unit root can be formulated as

$$H_0 : \phi_1^* = 0 \text{ vs. } H_1 : \phi_1^* < 0.$$

The test statistic τ has the form

$$\tau := \frac{\hat{\phi}_1^*}{\widehat{SE}\left(\hat{\phi}_1^*\right)},$$

where $\hat{\phi}_1^*$ is the ordinary least squares estimator of ϕ_1^* and $\widehat{SE}\left(\hat{\phi}_1^*\right)$ is the estimated standard error of $\hat{\phi}_1^*$ (see Brockwell and Davis, 2016). If the null hypothesis can be rejected at the predetermined significance level α, a time series sample is assumed to be generated by a stationary time series process.

When we cannot reject the null hypothesis of the ADF test, then it is suggested to take first differences of the time series (Brockwell and Davis, 2016). In general, there are different types of time series data transformation and taking differences, which can remove non-stationary behavior. In our work, we follow Gayer et al. (2014) and perform two types of time series transformation depending on their form. We calculate percentage changes with respect to the previous quarter for all trending real activity and financial time series expressed in absolute or index values. That is, we consider the following transformation

$$\dot{x}_t^{(m)} = \frac{1}{3}\left[\frac{x_t^{(m)}}{\frac{1}{3}\left(x_{t-1}^{(1)} + x_{t-1}^{(2)} + x_{t-1}^{(3)}\right)} - 1\right], \tag{7}$$

where $m \in \{1, 2, 3\}$ is the mth month in quarter t, $x_t^{(m)}$ is the initial time series observation in month m and quarter t, and $\dot{x}_t^{(m)}$ is the transformed time series observation. The percentage change transformation provides trend elimination and, like a log-transformation, stabilizes a variance over time (Brockwell and Davis, 2016). Equation (7) ensures additivity of monthly terms for the quarter term definition in the form $\dot{x}_t = \dot{x}_t^{(1)} + \dot{x}_t^{(2)} + \dot{x}_t^{(3)}$.

For most financial indicators, which describe rates, yields, etc., we subtract the average value of the previous quarter from monthly values, leading to

$$\ddot{x}_t^{(m)} = \frac{1}{3}\left[x_t^{(m)} - \frac{1}{3}\left(x_{t-1}^{(1)} + x_{t-1}^{(2)} + x_{t-1}^{(3)}\right)\right], \tag{8}$$

where $x_t^{(m)}$ is the initial time series observation in month $m \in \{1, 2, 3\}$ and quarter t, and $\ddot{x}_t^{(m)}$ is a difference-transformed time series observation. As in cases with percentage changes, the transformation in Equation (8) provides values, in which monthly components are additive: $\ddot{x}_t = \ddot{x}_t^{(1)} + \ddot{x}_t^{(2)} + \ddot{x}_t^{(3)}$.

Although month-on-month differences and percentage changes are more common for monthly data, we calculate differences and percentage changes with respect to the previous quarter, as this shows more accurate forecasting results due to noise reduction and the smoothing of data irregularities (see Rünstler et al., 2009).

Before we perform any transformation, we seasonally adjust all time series. In this work, most of the observed data have already been seasonally adjusted with the X-13ARIMA-SEATS software (United States Census Bureau, 2022), which we also applied to the unadjusted time series.

After seasonal adjustment and time series transformation, we obtain a data set that we can use for constructing nowcasting models.

2.3 Latent Dirichlet Allocation for Topic Modeling

We make use of topic models to explore hidden semantic structures in texts. Topic models are able to discover main themes that cover a large and unstructured collection of documents (Blei, 2012). In our case, we consider news articles as documents. The topic is considered as a distribution of words over a fixed vocabulary.

We analyze the topic structure with the latent Dirichlet allocation (LDA) topic model. The advantage of the LDA is the high interpretability of extracted topics (Chang et al., 2009). The LDA is a method of generative probabilistic modeling. In the generative approach, we assume that the distribution over the data has a parametric form, and we focus on estimating the parameters (Shalev-Shwartz and Ben-David, 2014). The idea of the LDA is based on the document generating process. The LDA assumes that topics are specified before words and documents are generated:

1. For each document d in the collection we choose a topic distribution $\boldsymbol{\theta}_d$.

2. For generating the nth word $W_{d,n}$, $n \in \{1, 2, \dots, N_d\}$ in document d, we choose the topic assignment $Z_{d,n}$ for the nth word in the document d from the topic distribution $\boldsymbol{\theta}_d$, where $Z_{d,n} \mid \boldsymbol{\theta}_d \sim \text{Multinomial}(\boldsymbol{\theta}_d)$.

3. We choose the word $W_{d,n}$ from the distribution over the vocabulary of the assigned topic, where $W_{d,n} \mid Z_{d,n}, \boldsymbol{\psi}_k \sim \text{Multinomial}(\boldsymbol{\psi}_k)$.

In practice, the document's topic distribution $\boldsymbol{\theta}_d$ and topic assignments $Z_{d,n}$ for the nth word of document d as well as the topic's word distributions $\boldsymbol{\psi}_k$ are unknown. However, the documents and words are observed. This leads to the data generating process from the joint probability distribution over the observed and hidden random variables. By the observed random variables, we mean each word $W_{d,n}$ for the document d. The latent random variables are the topic assignment for each word $Z_{d,n}$, the topic distribution $\boldsymbol{\theta}_d$ for each document d and the distribution over the vocabulary $\boldsymbol{\psi}_k$ of each topic with the index k.

We assume that the topics and their allocation for each document have the prior Dirichlet distribution $\boldsymbol{\psi}_k \sim \mathit{Dirichlet}(\eta)$ and $\boldsymbol{\theta}_d \sim \mathit{Dirichlet}(\delta)$ (Blei et al., 2003). The Dirichlet prior distribution assumption is convenient to use, as the Dirichlet distribution is conjugate to the multinomial distribution. As in Thorsrud (2020) and Griffiths and Steyvers (2004), we use Gibbs sampling based on a Markov chain construction to compute the posterior distribution (Gilks et al., 1995). The parameters $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_K, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D$ can be estimated from the posterior distribution.

There are a few methodological limitations of LDA. Firstly, it does not consider the order of documents and their relevance over time. In addition, LDA is a bag of words procedure. Hence it does not take the order of words into account for the assignment of topics, but assumes independence between words. Moreover, it does not consider the correlation of the topics within each document. Finally, one of the main problems of the LDA relates to choosing the optimal number of topics. When we have a small number of topics, we get a mix of several topics, which may lead to misunderstanding and inaccurate interpretation of topics, whereas as large number of topics is difficult to interpret because the topics can be similar to each other. Nevertheless, LDA is characterized in particular by its low requirements, robustness and stability over more complex novel methods; and thus forms in particular a favorable starting point for fully automated application without human intervention.

2.4 Statistical and Machine Learning Models for High Dimensional Data and Evaluation Metrics

Considering that we have a model as described in Equation (5), we want to rewrite Equations (2)–(5) in a matrix form for further discussion of statistical and machine learning models. All our blocked models for $m \in \{1, 2, 3\}$ have the same logic structure. Hence, we can generalize the model structure for further definitions and concepts. We define $\mathbf{X} \in \mathbb{R}^{N_q \times M_m}$ as a matrix of all predictor variables, which are available at the end of the month m, and GDP growth values from the previous quarter. N_q is a number of quarters, M_m defines a number of all predictor variables, their lags and the GDP growth lag at the nowcasting month m. The vector $\mathbf{y} \in \mathbb{R}^{N_q}$ denotes quarter values of GDP growth. The matrix $\mathbf{X}^c \in \mathbb{R}^{N_q \times (M_m+1)}$ is the matrix $\mathbf{X}$ with with an additional vector of ones in its first column. Thus, we have:

$$\mathbf{y} = \mathbf{X}^c \boldsymbol{\beta}^c + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}^c = (\beta_0, \beta_1, \ldots, \beta_{M_m})^T \in \mathbb{R}^{M_m+1}$ is a vector of parameters with the constant term β_0. The vector $\boldsymbol{\epsilon} \in \mathbb{R}^{N_q}$ is the corresponding zero mean vector.

For defining the optimal vector of parameters of a linear model, we have to solve the equation

$$\hat{\boldsymbol{\beta}}^c = \arg\min_{\boldsymbol{b}} \left((\mathbf{y} - \mathbf{X}^c \boldsymbol{b})^T (\mathbf{y} - \mathbf{X}^c \boldsymbol{b}) \right). \tag{9}$$

In this case, the optimal parameter vector can be estimated with the ordinary least squares method (Hastie et al., 2009)

$$\hat{\boldsymbol{\beta}}^c = ((\mathbf{X}^c)^T \mathbf{X}^c)^{-1} (\mathbf{X}^c)^T \mathbf{y}, \tag{10}$$

where $\hat{\boldsymbol{\beta}}^c$ is the estimated parameter vector. Then, the fitted values $\hat{\mathbf{y}} \in \mathbb{R}^{N_q}$ of the model have the form

$$\hat{\mathbf{y}} = \mathbf{X}^c \hat{\boldsymbol{\beta}}^c.$$

We want to construct a nowcasting model that includes different financial, economic and survey indicators and aggregated text data. We extract the important information from text data by topic models. As a result, we face the problem of a large number of different predictor variables. Sometimes the number of predictor variables can be larger than the number of observations, i.e. quarters. In such a case, the OLS estimator in Equation 10 has an infinite number of solutions. Apart from many predictors, we can have a large correlation between some variables in the data set. This effect produces a multicollinearity problem when $\mathbf{X}^T\mathbf{X}$ or $(\mathbf{X}^c)^T\mathbf{X}^c$ becomes (approximately) singular. In addition, the model overfitting can be a large problem for high dimensional data sets (Hastie et al., 2009). To overcome these issues, there exist various approaches. In our work, we follow Gayer et al. (2014) and fit principal component regression as one of the dimensionality reduction methods. In addition, inspired by Hastie et al. (2009), we use shrinkage methods with help of the elastic net regularization (Zou and Hastie, 2005).

Finally, to consider possible non-linear effects, we fit a random forest (Breiman, 2001), which has built-in feature selection techniques. For all our models, we consider that the predictor matrix $\mathbf{X} \in \mathbb{R}^{N_q \times M_m}$ is standardized (centered and scaled), to make the estimated coefficients better interpretable.

2.4.1 Principal Component Regression

Firstly, we describe how we use principal component regression (PCR) (James et al., 2013). For this, we include the principal components into a linear regression model

$$y_t = c_0 + \sum_{j^{PC}=1}^{J_t} c_{j^{PC}} z_{t,j^{PC}} + \epsilon_t^{PCR}, \tag{11}$$

where $J_t \in \{1, \ldots, J_{max}\}$ is a number of added principal components in the quarter $t \in \{1, \ldots, N_q\}$ with $J_{max} \leqslant M_m$, where M_m is a possible maximum number of principal components in Equation (11), $z_{t,j^{PC}}$ is the corresponding principal component in quarter t with the index j^{PC}, and ϵ_t^{PCR} is the error of principal component regression. $c_0, \ldots, c_{J_t}$ are the corresponding coefficients (Gayer et al., 2014). To prevent overfitting, we should choose an appropriate maximal number of principal components J_{max}, which we can use in PCR. There exist different approaches for definition J_{max}. As we produce many nowcasts, we cannot use visual analysis methods like scree plots (Cattell, 1966). In this work, we choose J_{max} according to the Kaiser rule (Kaiser, 1960). The idea of the Kaiser rule is to choose the principal components with a variance larger than 1.

The important step is a proper choice of the number of the principal components J_t. In this work, we follow Gayer et al. (2014) and Caggiano et al. (2011) and calculate the Akaike information criterion (AIC) (Akaike, 1998). The lowest AIC value corresponds to the best J_t.

2.4.2 Shrinkage

Another class of methods for dealing with high-dimensional data are shrinkage methods. In this work, we choose the elastic net regularization (Friedman et al., 2010, Zou and Hastie, 2005). It can be considered as a combination and generalization of the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and Ridge regularization (Hoerl and Kennard, 1970). The idea of these methods consists of imposing a penalty on the regression coefficients. Hence, we can overcome the multicollinearity, $N_q < M_m$ and overfitting problems.

The penalty term of the elastic net regularisation (Friedman et al., 2010) has the form

$$\lambda \left(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_{\ell_2}^2 + \alpha||\boldsymbol{\beta}||_{\ell_1} \right) = \lambda \sum_{j=1}^{M_m} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right),$$

where $\lambda \geqslant 0$ is a tuning parameter, which indicates the strength of the coefficients' penalty. The value $\alpha \in [0, 1]$ is another tuning parameter that controls the type of regularization. For example, the Ridge penalty results with $\alpha = 0$ and the LASSO penalty with $\alpha = 1$. The norms $||\cdot||_{\ell_1}$ and $||\cdot||_{\ell_2}$ are ℓ_1- and ℓ_2-norms respectively. $\beta_1, \beta_2, \ldots, \beta_{M_m}$ are the corresponding coefficients and the elements of the vector $\boldsymbol{\beta}$. The penalty term has a division factor $\frac{1}{2}$ before the sum-of-squares penalisation for simplicity of function derivations. The elastic net regularization has the advantages of both Ridge and LASSO regularizations. Unlike LASSO, which can select only N_q variables when $N_q < M_m$, the elastic net contains the coefficients' results for all predictor variables in the regression model. Moreover, the elastic net overcomes the LASSO method problem when the LASSO regularization selects only one predictor variable from highly correlated predictor variables without caring which variable is reasonable to select. In comparison to the Ridge regularization, the elastic net has properties of the LASSO method, which usually outperforms the Ridge method for the case with correlated predictor variables and $N_q > M_m$ (Tibshirani, 1996). The parameters α and λ can be determined by cross-validation. We describe the details of cross-validation in Section 4.1.

2.4.3 Random Forest

The last model, which we use in our study, is the random forest (Breiman, 2001). Unlike linear regression models, the random forest can capture non-linear dependencies between a response variable and predictors.

The idea of the random forest lies in the construction of a large collection of de-correlated trees. The de-correlated trees are achieved through random selection with replacement of $mtry \leqslant M_m$ input variables. These input variables are possible candidates for the split, when constructing a tree. After growing all trees, the model averages the results over all constructed trees for the observation. Hence the model has the form

$$\hat{y}_t^{RF} = \frac{1}{B} \sum_{b=1}^{B} T(\mathbf{x}_t, \Upsilon_b),$$

where $\hat{y}_t^{RF}$ is a fitted value, B is a number of trees, T represents the tree itself and Υ_b the parameters of the bth tree. In general, parameters such as a number of trees or a number of selected input variables before each split can be tuned and determined by, e.g., cross-validation.

Despite random forests can capture non-linear dependencies, they can face several problems. For example, they can show a poor performance in extrapolation. Random forests assign prediction values based on the previously seen response values and cannot correctly predict the response value, which lies outside the response values in the training data set.

2.4.4 Performance Metrics

Following Gayer et al. (2014) and Ashwin et al. (2021), we evaluate and compare predictions of our models according to the root mean squared error (RMSE). As in Gayer et al. (2014), the best model is considered to be the one that minimizes the root mean squared error regardless of the significance of the improvement.

The RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{N_{test}} \sum_{t_{test}=1}^{N_{test}} (y_{t_{test}} - \hat{y}_{t_{test}})^2},$$

where N_{test} is a number of test quarters, $y_{t_{test}}$ is a response variable value for the t_{test}th quarter in the test set and $\hat{y}_{t_{test}}$ is a model prediction of a response variable for the t_{test}th quarter in the test set.

In addition, we want to evaluate our model in terms of overfitting and underfitting. Underfitting means that the model cannot correctly capture the structure of the training data. Overfitting means that the model adapts very closely to the training data and produces a large error on the test data, thus, has a poor ability to generalization (Hastie et al., 2009). To understand the nature of the mean squared error (MSE) of a model, we can apply bias-variance decomposition. Following Taieb and Atiya (2015), the bias-variance decomposition of the test MSE for nowcasting has the form

$$\begin{aligned} \text{MSE} = \sigma^2 + \mathrm{E}_{\mathbf{x}_{t_{test}}, y_{t_{test}}} \left[\mathrm{Var}_{\mathcal{D}_{train}}(\hat{y}_{t_{test}} \mid \mathbf{x}_{t_{test}}, y_{t_{test}})\right] + \\ \mathrm{E}_{\mathbf{x}_{t_{test}}, y_{t_{test}}} \left[\mathrm{E}^2_{\mathcal{D}_{train}}(y_{t_{test}} - \hat{y}_{t_{test}} \mid \mathbf{x}_{t_{test}}, y_{t_{test}})\right], \end{aligned} \tag{12}$$

where σ^2 is a variance of the data, $\hat{y}_{t_{test}}$ is a model prediction, and $y_{t_{test}}$ is a true response variable value for the prediction of the t_{test}th quarter in the test data set. The vector $\mathbf{x}_{t_{test}}$ defines predictor variable values for the t_{test}th quarter and $\mathcal{D}_{train}$ is a training data set, which is used for model estimation. The second term of Equation (12) is a variance of a model, while the third term defines a squared bias. In general, $y_{t_{test}}$, $\mathbf{x}_{t_{test}}$ and $\mathcal{D}_{train}$ come from unknown distribution. We estimate the bias and the variance by

$$\widehat{\text{Bias}}_{model} = \left| \frac{1}{N_{test}} \sum_{t_{test}=1}^{N_{test}} y_{t_{test}} - \frac{1}{N_{test}} \sum_{t_{test}=1}^{N_{test}} \hat{y}_{t_{test}} \right|,$$

and

$$\widehat{\text{Variance}}_{model} = \frac{1}{N_{test}-1} \sum_{t_{test}=1}^{N_{test}} \left(\hat{y}_{t_{test}} - \frac{1}{N_{test}} \sum_{t_{test}=1}^{N_{test}} \hat{y}_{t_{test}} \right)^2,$$

respectively, where $\widehat{\text{Bias}}_{model}$ detects the estimated model's bias, and $\widehat{\text{Variance}}_{model}$ is the estimated model's variance. The models that tend to overfitting may have a large variance, whereas the models that underfit the data may have a large bias.

3 Description of the Data Set

All data sets cover the time period from April 2005 to June 2021. This section contains information about GDP growth data and selected predictor data. Each predictor belongs to one of four data categories, according to its information. Specifically, this includes collected newspaper text data as well as real activity, financial indicators and survey data. In addition, we highlight the characteristics and sources of the used data.

3.1 Real Activity, Financial, Survey and German GDP Growth Data

For nowcasting German GDP growth, we use various established indicators such as real activity, financial indicators and survey data, which are available in the time period from April 2005 to June 2021. The choice of financial and real activity indicators is based on the work of Gayer et al. (2014). In comparison to Gayer et al. (2014), where the authors select indicators for nowcasting euro area GDP growth, we choose indicators published for the German economy. Thus, all selected real activity indicators and several financial indicators, e.g., a bond yield and monetary aggregates, relate to the German economy. Other financial indicators include information about global economic development, which can potentially influence German GDP growth.

Financial data contains the daily ten-year bond yield, euro exchange reference rates, interbank offered rates, gold and oil prices, stock market data and monetary aggregates. Due to early availability, financial indicators can be exploited at the early stages of nowcasting. Real activity data includes output in production sector indices, unemployment rate and industry turnover, consumer price index (CPI) and harmonized indices of consumer prices (HICP). The last two indices are treated as consumer price movements and inflation measures. We choose only indicators with a release date occurring before the end of the nowcasting quarter. The index values of real activity data are calculated by the Deutsche Bundesbank (2022) with the reference year 2015. In this work, we take values from the latest available release of the data sets because not all indicators have data vintages. Although the revised data can significantly overestimate the forecasting performance of the model (Diebold and Rudebusch, 1991), our focus is on comparing different models and the role of different types of data in nowcasting performance. The data are downloaded from several freely available sources, namely Deutsche Bundesbank (2022), European Central Bank (2022), ifo Institute (2022) and Federal Reserve Bank of St. Louis (2022).

In addition to financial and real activity sector data, we use also survey indicators. According to Gayer et al. (2014) and Andreini et al. (2020), survey data can improve nowcasts at the beginning of the quarter. Due to their earlier release dates compared to real activity indicators, survey indicators can complete some missing information. Survey data also include prognoses for the following months, which can help predict GDP growth values at the beginning of the quarter. In this work, we use a freely available source of the business survey, the ifo Business Climate Index for Germany (ifo Institute, 2022a). Apart from availability, the business survey is published at the end of each month and implies current monthly information about the German economy. Various studies (e.g., Lehmann, 2020 and Andreini et al., 2020) show a high forecasting quality of the ifo Business Climate Index for Germany business survey indicators. In our work, we choose two indices from the business survey: the ifo Business Situation Index for Germany and the ifo Business Expectations Index for Germany. The indices are calculated based on monthly survey responses of German firms concerning the current business situation and the business expectations for the next six months. Respondents have three ways to describe the business situation and expectations, as "good", "satisfactory", or "poor". After weighting and aggregating the survey answers, the balance value of the indicator is calculated. It corresponds to the difference in percentages of the survey answers "good" and "poor". Finally, the indices are evaluated according to the following formula (ifo Institute, 2022a):

$$Index\ value = \frac{balance\ value\ in\ the\ current\ month + 200}{average\ balance\ value\ in\ the\ base\ year + 200} \times 100,$$

where the base year is equal to 2015.

All predictors have different release dates and are published with different frequencies. The indicators are released on a daily (D) or monthly (M) basis. For the nowcasting approach with mixed frequency data, we construct models considering the monthly availability of data. If an indicator is published daily, then we average the indicator's values over a month. In this case, the release day of the last indicator's value in the selected month corresponds to the release date of the daily publishing indicator itself. Assume that we collect a set of indicators for a specific month. Depending on publication lag, we divide predictors into five classes

- DP – release on a current day of the observed month (only for daily frequency),
- MP1 – release before or on the last day of the observed month,
- MP2 – release in the first half (before the 15th day) of the next month,
- MP3 – release between the 15th day and the last day of the next month (both including),
- MP4 – release in the first half (before the 15th day) of the second month after the observed month.

Tables 1–3 contain all the selected financial, real activity sector and survey data indicators, including their description, source and publication characteristics.

German quarterly GDP growth data are downloaded from the OECD database (OECD, 2022b). The data show the percentage change in GDP compared with GDP from the previous quarter. The GDP growth data cover the time period from the second quarter of 2005 to the second quarter of 2021. Before July 2020, the GDP growth data were published with roughly a 45-day delay. Nowadays, the first release of German GDP growth appears in the second half of the next month after the end of the reference quarter (after around 30 days) (Investing.com, 2022). In this work, we assume the newest publication rules of GDP growth for the whole nowcasting time period.

Table 1: Overview of selected financial indicators.

Indicator	Publication frequency	Publication lag type	Unit	Description	Source
bond_10y_yield	D	MP2	%	The daily yield of the current ten-year federal bond	DBB, 2022
eurib_3m	M	MP2	%	Euro Interbank Offered Rate three-month funds	DBB, 2022
exr_usd	D	DP	USD	Euro United States dollar reference rates	ECB, 2022
exr_jpy	D	DP	JPY	Euro Japanese yen reference rates	ECB, 2022
exr_gbp	D	DP	GBP	Euro Pound sterling reference rates	ECB, 2022
gold_prices	M	MP2	EUR	The gold price per troy ounce from London Bullion Market Association	DBB, 2022
libor_3m_us	D	MP2	%	The three-month USD London Interbank Offered Rate	FRED, 2022
M1	M	MP3	EUR million	Money stock (narrow money), Germany	DBB, 2022
M2	M	MP3	EUR million	Money stock, Germany	DBB, 2022
M3	M	MP3	EUR million	Money stock (broad money), Germany	DBB, 2022
oil_price	D	DP	USD	The West Texas Intermediate crude oil price per Barrel	FRED, 2022
vix_us	D	DP	Index	Chicago Board Options Exchange Volatility Index	FRED, 2022

Table 2: Overview of selected real activity indicators.

Indicator	Publication frequency	Publication lag type	Unit	Description	Source
`hicp`	M	MP1	Index	Harmonised Index of Consumer Prices	DBB, 2022
`cpi`	M	MP1	Index	The national Consumer Price Index	DBB, 2022
`ip_capital_goods`	M	MP4	Index	Output in the production sector: capital goods	DBB, 2022
`ip_civil_engineering`	M	MP4	Index	Output in the production sector: civil engineering	DBB, 2022
`ip_consumer_goods`	M	MP4	Index	Output in the production sector: consumer goods	DBB, 2022
`ip_durable_consumer_goods`	M	MP4	Index	Output in the production sector: durable consumer goods	DBB, 2022
`ip_energy`	M	MP4	Index	Output in the production sector: energy	DBB, 2022
`ip_industry`	M	MP4	Index	Output in the production sector: industry	DBB, 2022
`ip_intermediate_goods`	M	MP4	Index	Output in the production sector: intermediate goods	DBB, 2022
`ip_main_construction_industry`	M	MP4	Index	Output in the production sector: main construction industry	DBB, 2022
`ip_non-durable_consumer_goods`	M	MP4	Index	Output in the production sector: non-durable consumer goods	DBB, 2022
`ip_structural_engineering`	M	MP4	Index	Output in the production sector: structural engineering	DBB, 2022
`turnover_industry`	M	MP4	Index	Turnover in industry	DBB, 2022
`ur_de`	M	MP4	%	Unemployment rate, Germany	DBB, 2022

Table 3: Overview of selected survey indicators.

Indicator	Publication frequency	Publication lag type	Unit	Description	Source
bus_sit_Index	M	MP1	Index	ifo Business Situation Index	ifo, 2022
bus_exp_Index	M	MP1	Index	ifo Business Expectation Index	ifo, 2022

3.2 Text Data

The problem of considerable publication lag for most real activity indicators constrains the use of these data at the early stages of nowcasting. In addition to survey and financial indicators, the inclusion of text data from newspapers into nowcasting models can be a possible source of early available information.

For the German GDP growth prediction, we employ the news data collected from the German-language business newspaper "Handelsblatt". The used data set covers the time period from 01/04/2005 to 30/06/2021 and is cleaned of the most obvious stop words: articles, most of the modal verbs, pronouns, prepositions, conjunctions and particles. For computation reasons (Strubell et al., 2019) and most likely without lack of model quality (Maier et al., 2020), we delete those words that occur less than 50 times in the corpus. In addition, we delete words with a length of just one character. The final data set consists of 417,504 articles. In Section 4.2, we describe further steps for text data preparation, sentiment adjustment and monthly aggregation, which are necessary for the nowcasting models.

4 Empirical Set-Up for Nowcasting German GDP Growth

First, we explain the construction of the nowcasting approach, the aspects of the models' fitting and the choice of parameters. After that, we describe the approach of text data aggregation, the sentiment and frequency adjustment. In addition, we describe the software tools used for nowcasting German GDP growth.

4.1 The Approach for Nowcasting German GDP Growth

As described in Sections 3.1 and 3.2, we have predictor variables with monthly or daily frequencies, which have different release dates. The first release of German GDP growth is published quarterly with a delay of around 30 days. We aggregate the predictor variables with daily frequencies on a monthly basis (see more details for non-text data in Section 3.1 and text data in Section 4.2). For the nowcasting model, following Gayer et al. (2014) and

Carriero et al. (2015), we apply the blocking approach to transform predictor variables with monthly frequencies into quarterly time series.

As already mentioned in Section 2.2, we want to extract and eliminate possible non-stationary behavior from all time series data. Firstly, we need to seasonally adjust all time series. Most of the time series data, which are published by Deutsche Bundesbank (2022), and all survey data are already seasonally adjusted with X-13ARIMA-SEATS software. If still necessary, we adjust unadjusted ones. We do not seasonally adjust text data as they do not have a significant seasonal component.

The next step is to eliminate non-stationary behavior. We analyze the visual representation of predictor variables over time and the sample autocorrelation function. Then, we transform the trending time series and the time series data with changing variance over time. For this purpose, we follow Gayer et al. (2014) and perform the difference (Equation (8)) and percentage change (Equation (7)) transformation for financial and real activity time series data, as described in Section 2.2. After transformation, we perform the ADF test to check the hypothesis on the existence of a unit root. Non-stationary behavior can produce worse results for the nowcasting models. Therefore, we additionally analyze the time series data for which the null hypothesis of the ADF test is not rejected using the sample ACF plots. In addition, if the nowcasts with these predictor variables are less accurate, we remove them from the data set. As proposed in Gayer et al. (2014), Giannone et al. (2008) and Andreini et al. (2020), survey indicators are treated as stationary in levels. Text data are also aggregated on a monthly basis and expressed in topic frequencies or sentiment scores for each topic over time (see Section 4.2). We treat text data as differenced according to Equation (8).

Based on the blocking approach, we construct three nowcasting models depending on each month of the nowcasting quarter. This means we produce the nowcast of German GDP growth at the end of each month. Table 4 presents the information about the possible predictor variables that we use in each model. "Month 1", "Month 2" and "Month 3" define the nowcasts at the end of the first, second and third month, respectively. The predictor variables are aggregated according to the publication lag type in Tables 1–3. The predictor variables that we use in our models are centered and scaled.

Regarding the nowcast models, the initial in-sample period covers the time period from July 2005 (Q3 2005) to September 2015 (Q3 2015). Hence, we have 41 quarters in the initial training data set. The prediction period runs from October 2015 (Q4 2015) to June 2021 (Q2 2021) and consists of 23 quarters. It is additionally divided into the "stable" period (Q4 2015 – Q4 2019) and the turbulent "COVID-19" period (Q1 2020 – Q2 2021). Starting from September 2015, we retrain our nowcasting models monthly to produce new nowcasts for the remaining out-of-time quarters. Thus, we use an expanding window of our data, starting not from April 2005 but from July 2005, as most of the data are transformed with respect to the previous quarter.

As described in Section 2.4, we choose the maximal number of principal components J_{max} according to the Kaiser rule for each PCR model. As the observed training data set changes each month, the J_{max} value should be chosen for each GDP growth prediction and each type of the nowcasting model ("Month 1", "Month 2" and "Month 3"). Then, we

Table 4: Possible predictors in the nowcasting models, which are published at the moment of producing the nowcast in the nowcasting quarter. "Month 1", "Month 2" and "Month 3" define models at the end of the corresponding month of the nowcasting quarter. Here, (m) with $m \in \{1, 2, 3\}$ indicates the month in the quarter for which data are available. Text data are defined as "text".

Month 1	Month 2	Month 3
GDP growth in the previous quarter	GDP growth in the previous quarter	GDP growth in the previous quarter
DP(1)	DP(1), DP(2)	DP(1), DP(2), DP(3)
MP1(1)	MP1(1), MP1(2)	MP1(1), MP1(2), MP1(3)
-	MP2(1)	MP2(1), MP2(2)
-	MP3(1)	MP3(1), MP3(2)
-	-	MP4(1)
text(1)	text(1), text(2)	text(1), text(2), text(3)

choose the best number of principal components J_t with the smallest AIC value for the training data set. Finally, we add only those principal components, which are significant at the 0.1 level according to the t-test.

For the elastic net models, we need to tune the parameters λ and α. The parameters are chosen according to the 5-fold cross-validation with initially assigned observations to each fold. We split the training data set into 5 folds, successively train the model on 4 of the 5 folds and then validate on the remaining fold. We evaluate the squared errors for each validation fold, accumulate and average them. Thus, the combination of the parameters with the lowest error is chosen. When the data have a time-dependent structure, there are possibilities using a time series variant of the cross-validation, which divides the training data set into, e.g., 5 folds and then sequentially trains the model on the first i folds ($i \in \{1, 2, 3, 4\}$) and validates on the $i + 1$th fold. As we do not have a large number of observations, we apply the ordinary 5-fold cross-validation. Nevertheless, we do not randomly assign the observations to the folds. We can hence partially save the structure of the time-dependent data. For tuning the α parameter, we use the sequence of values starting from 0 to 1 with the step size 0.1. The λ-sequence is obtained by an application of the cross-validation function `cv.glmnet` in the `R` package `glmnet` (Friedman et al., 2010).

For the random forest, most of the parameters are selected as default values from the `R` package `ranger` (Wright and Ziegler, 2017). Following Kalamara et al. (2020), we train 200 trees. The parameter *mtry* is chosen equal to the rounded down square root of the number of predictor variables, the default in the `ranger` package. For the random forest training, we use the whole training data set as for PCR.

Finally, we calculate RMSE, bias and variance after producing the nowcasts for the whole, "stable" and "COVID-19" out-of-sample period for each model. Based on these metrics, we compare our models. In addition, we compare the best models with the benchmark of an AR(1) model.

4.2 LDA Topics, Sentiment Adjustment and Frequency Aggregation of Text Data

As mentioned in Section 2.3, we want to analyze the impact of the number of LDA topics on the nowcasting performance. Following Ellingsen et al. (2021), we choose different numbers of topics (K) with the maximal number of topics equal to 120. Following the ideas from Thorsrud (2020) and Ellingsen et al. (2021), we use the sequence of a possible number of topics from 50 to 120 with the step size 10. As it might be a better choice for some of our aggregation techniques to have less topics, we run the LDA for the sequence from 3 to 10 topics with the step size 1 as well. We choose the parameters δ and η of the prior Dirichlet distribution for the distributions over topics and words equal to the default value $\frac{1}{K}$, proposed in the `ldaPrototype` package (Rieger, 2020).

For all nowcast scenarios, we run the LDA from scratch starting from April 2005 until the end of the given month and repeat this procedure for all selected numbers of topics. Thus, we can ensure the same topic structure over the training and test data sets for the nowcasting model. The motivation to re-estimate the LDA results monthly is to consider the change of the topic structure over time and to recognize new topics, which can appear in new months.

After applying the LDA, each word in each newspaper article is assigned to one of the possible topics. We aggregate the words from all articles according to the topic assignments. Then, we want to aggregate the words from each topic on a monthly basis, considering the time of the article's publication. The first type of aggregation is based on the frequencies with which the topic is monthly mentioned. In this case, we divide the number of words from each topic by the total number of words in each month.

The second type of aggregation is based on sentiment analysis of topics. In this work, we define sentiment analysis as positiveness or negativeness of each topic according to its sentiment score. There exist different types of sentiment score assignment to each word (Ashwin et al., 2021). In our work, the sentiment analysis is performed with the sentiment German-language SentiWS dictionary (Remus et al., 2010). This dictionary is convenient to use because we do not need to translate words into English. The general-purpose SentiWS dictionary contains more words than the business-purpose dictionary, created by Bannier et al. (2019), and has been successfully used, for instance, in the field of political communication (Haselmayer and Jenny, 2017). The dictionary contains 1,650 positive and 1,800 negative words in German. Additionally, around 16,000 positive and 18,000 negative forms of these words are included. Therefore, we do not need to perform stemming or lemmatization (Schütze et al., 2008). The sentiment scores of words are assigned in the interval $[-1, 1]$, defining their degree of positiveness or negativeness. Some words in the SentiWS dictionary have different degrees of positiveness/negativeness at the same time. For such words, we calculate the average sentiment score. Words that are not covered by SentiWS get zero scores. The monthly sentiment score for each topic is calculated as the sum of all sentiment scores of all word tokens assigned to this topic.

After aggregation, we calculate the differences of frequency and sentiment according to the previous quarter values (see Equation (8)).

5 Statistical Evaluation

We prepared our data sets, evaluated the nowcasting models, visualized and analyzed the results in `R` (R Core Team, 2021). The list of the used packages can be found in Table 10 in the Appendix.

In the following, we discuss the performance of established economic indicators such as financial, survey and real activity indicators for nowcasting German GDP growth. Furthermore, we add text data in the nowcasting models and discuss different changes in performance gained by this.

5.1 ADF Test

As we mentioned in Sections 3 and 4, all financial, survey and real activity indicators and text topics are aggregated on a monthly basis. All non-text indicators are seasonally adjusted. German GDP growth is published quarterly. The next step is to perform the transformation of predictor variables. Table 11 in the Appendix presents detailed information about the types of transformation selected. After transformation, we applied the ADF test to check for unit roots in the resulting time series. The p-values of the ADF test are shown in Table 12 in the Appendix.

According to Table 12, we can reject the null hypothesis of the existence of a unit root for almost all predictor variables at a significance level of 0.05. However, we cannot reject the null hypothesis for `hicp`, `cpi`, `M1`, `M2`, `M3`, `bus_sit_Index` and `bus_exp_Index`. Figure 4 in the Appendix shows the form of these indicators after the transformation. The survey indicators `bus_sit_Index` and `bus_exp_Index` are treated as stationary in levels (Section 4.1), and they provide better results without any transformation. Thus, we do not remove these indicators from the nowcasting models. For other predictor variables with potential non-stationary behavior, we analyze the structure using time series plots (Figure 4) and ACF plots.

For the predictor variables `hicp`, `cpi` we observe that after lag 3 we do not have any ACF values that differ significantly from zero. We assume that we have a finite-lag structure, so these predictor variables remain in our models. Moreover, the presence of these indicators improves the results of the nowcasts at the early stages. The ACF plots show significant lags of the higher order for the monetary aggregates `M1`, `M2`, `M3`. We keep the predictor variable `M1` in the nowcasting models and remove `M2` and `M3` from the models. This procedure is done due to the non-stationary behavior of monetary aggregates. However, we keep the predictor with the smallest p-value, which improves the nowcasting results. German GDP growth is also assumed to be stationary, and the ADF test rejects the null hypothesis of a unit root at level 0.05.

Table 5: The nowcasting results on a monthly basis of German GDP growth with the AR(1) model. "Var." is an estimated variance of each model.

Model	*Stable*			*COVID-19*			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	0.55	0.01	0.03	7.72	0.97	7.83	3.97	0.25	2.44
Month 2	0.55	0.01	0.03	7.72	0.97	7.83	3.97	0.25	2.44
Month 3	0.55	0.01	0.03	7.72	0.97	7.83	3.97	0.25	2.44

5.2 LDA Topics

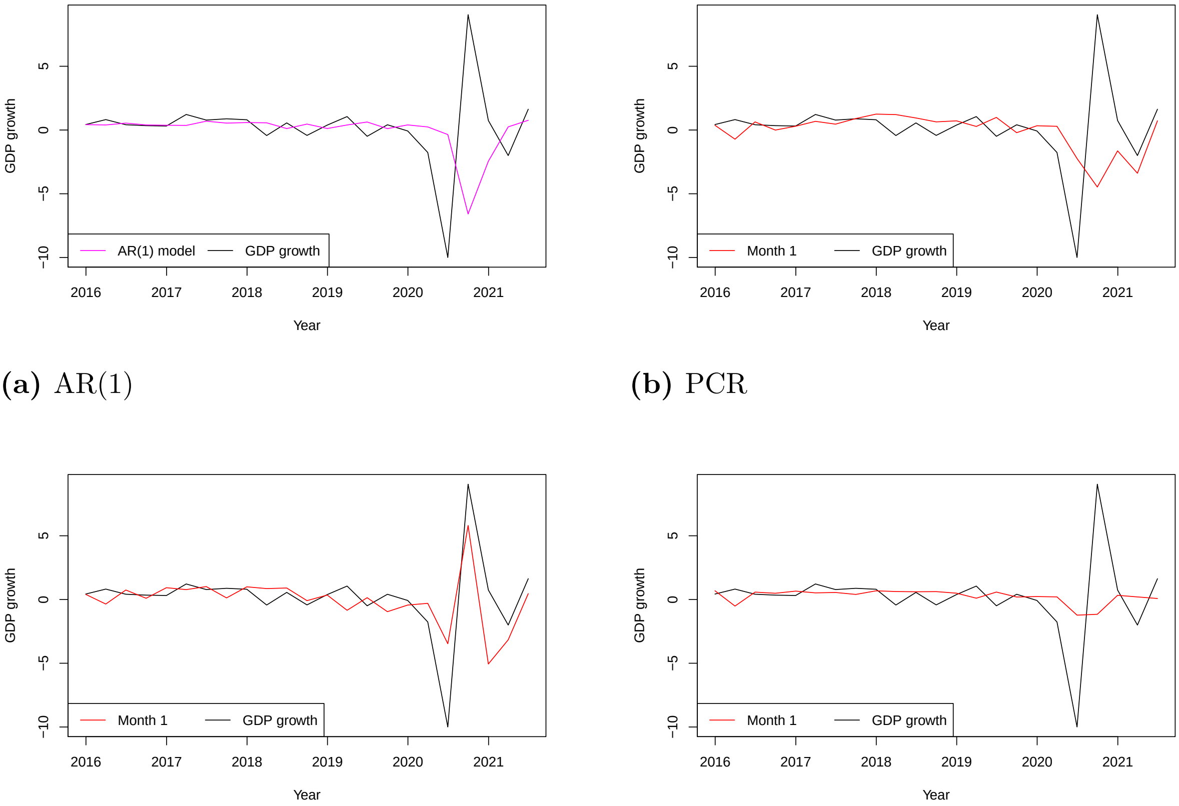
As described in Section 4.2, the number of the LDA topics varies from 3 to 120. The more topics we have, the more detailed information we can extract from topics. For example, the topic about the COVID-19 pandemic in April 2020 can be extracted as a separate one for a large number of topics (60 and more). However, with an increased number of topics, rather homogeneous topics tend to split-up as well leading to more, but similar topics. Moreover, the structure of topics change from month to month. With an increasing number of topics, we can clearly see the change in the structure of topics over time. As an example, we analyze the structure of 7 topics. As with 3 topics, most of the topics do not substantially change their structure. However, as an example, one topic shows a clear difference between top words from the time period April 2005 – October 2015 and from the time period April 2005 – June 2021. Table 13 in the Appendix presents this difference. These topical shifts may contain the major additional information of text data for nowcasting German GDP growth.

5.3 Results using Established Economic Indicators

We determine the statistical models that produce the best nowcasts using economic indicators but *no* text data. We train the nowcasting models and calculate RMSE, the estimated bias and variance of each model. We evaluate the models in the "stable" time period, before the COVID-19 pandemic, in the "COVID-19" and the whole time period, which is indicated as "all".

We compare our models with the AR(1) model. In Table 5, we present the results of the AR(1) model. The results do not change during the quarter because we observe the first release of German GDP growth at the end of the quarter's first month. In Figure 1a, we present the nowcasts of the AR(1) model with real German GDP growth for the whole out-of-sample period. We identify a (plausible) tendency of the AR(1) model to produce nowcasts which has a delay in comparison to the real GDP growth data.

We trained and produced the nowcasts in the out-of-sample period for PCR, the elastic net and the random forest models. The results of the models are shown in Table 6. In addition, we present the visual comparison of the AR(1), PCR, the elastic net and the random forest models in Figure 1 for the nowcasts at the end of the first month of the



(a) AR(1)

(b) PCR

(c) Elastic net

(d) Random forest

Figure 1: German GDP growth and its nowcasts using financial, economic and survey indicators at the end of the first month of the quarter with the AR(1), PCR, the elastic net and random forest models in the out-of-sample period.

nowcasting quarter. Figures 5 and 6 in the Appendix show the nowcasts at the end of the second and third month of the nowcasting quarter.

According to Table 6, the elastic net model shows the best results in the whole and the turbulent "COVID-19" time periods among all models, including the AR(1) model, while the random forest model outperforms all indicator-based models in the "stable" time period. The random forest model shows better results in the third month of the "stable" period in comparison to the AR(1) model. However, both, the random forest and the AR(1) model, realize estimates of the model’s variance close to zero. Such behavior of the random forest can be explained by its low ability to extrapolate. In contrast, the elastic net and PCR models show lower RMSE but substantially larger estimates of the model’s variance in the "all" and "COVID-19" time periods. In comparison to the random forest and AR(1) models, the estimated variance and bias of the PCR and elastic net models in the "stable" period are also larger. The fact of the larger estimated variance is a sign of potential overfitting. However, in the "COVID-19" and "all" time periods, we accept the larger variance to predict the turbulent period with extremely large volatility in German GDP growth. The extreme behavior of German GDP growth is not considered as non-significant outliers and, in general, ought to be predicted. In the "stable" period, we notice

Table 6: The nowcasting results on a monthly basis of German GDP growth with principal component regression (PCR), the elastic net and the random forest using financial, economic and survey indicators. "Var." is an estimated variance of each model. The italic, bold numbers indicate the best performing models for each month of the "stable", "COVID-19" and "all" periods.

PCR

Model	*Stable*			*COVID-19*			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	0.79	0.11	0.26	6.52	1.40	4.12	3.40	0.29	2.19
Month 2	0.73	0.10	0.30	5.56	0.94	1.78	2.91	0.32	1.17
Month 3	0.60	0.10	0.17	3.12	0.04	10.65	1.67	0.08	2.65

Elastic net

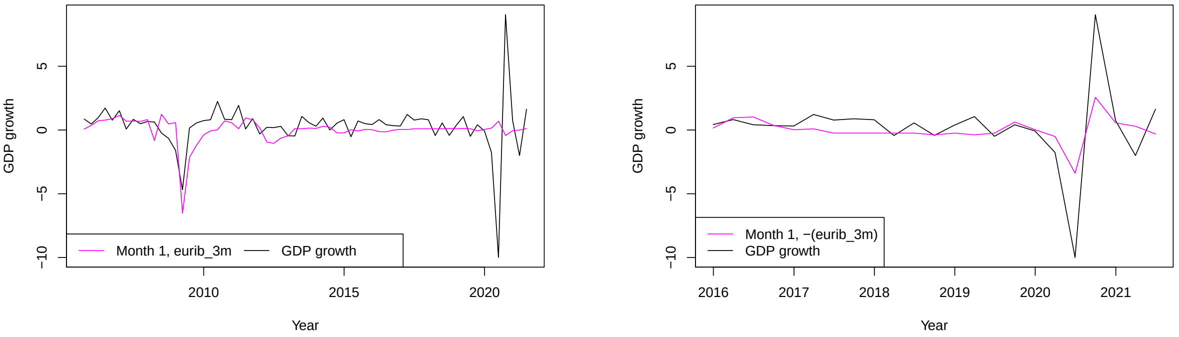
Model	*Stable*			*COVID-19*			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	0.79	0.14	0.41	***3.91***	0.56	15.22	***2.11***	0.25	4.06
Month 2	0.68	0.19	0.23	***4.04***	0.52	9.29	***2.15***	0.28	2.53
Month 3	0.55	0.13	0.14	***2.17***	0.58	20.92	***1.21***	0.06	4.86

Random forest

Model	*Stable*			*COVID-19*			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	***0.65***	0.03	0.09	5.66	0.13	0.53	2.95	0.06	0.29
Month 2	***0.59***	0.07	0.04	5.64	0.27	0.40	2.92	0.12	0.19
Month 3	***0.47***	0.07	0.05	5.27	0.50	0.54	2.72	0.08	0.17

that the elastic net model tends to overfit during the plateau time periods. Unlike the random forest model, the elastic net model poorly predicts the peaks in Q1 2019 and Q3 2019 but recognizes the decrease in Q3 2018, especially at the end of the first and the second month of the nowcasting quarter. In addition, it shows the best prediction results in the turbulent time period. The PCR model also produces unstable behavior when German GDP growth does not change.

Almost all models in Table 6 show the tendency to improve the nowcasts with an increasing number of published indicators. As an exception, we have an increase in RMSE at the end of the second month for the elastic net model in the "COVID-19" and "all" time periods. This unusual behavior is explained by the release of the `eurib_3m` indicator at the end of the second month. If we want to predict the decreases in Q1 2020, Q2 2020, Q4 2020 and Q1 2021, we should add the `eurib_3m` indicator in the model with the negative coefficient. However, the coefficient estimated by the elastic net model is positive. The possible problem is that the model cannot recognize the sign of relation from previous periods. For example, in Q1 2009 the `eurib_3m` indicator decreases with German GDP growth, while in 2016–2021 the movement of the `eurib_3m` indicator reflects the change of German GDP growth (see Figure 2).



(a) Q3 2005 – Q2 2021 data set

(b) Q4 2015 – Q2 2021 data set

Figure 2: German GDP growth and the scaled value of the transformed `eurib_3m` indicator, collected at the end of the first month. The first plot includes the whole in-sample and out-of-sample data sets, while the second plot includes only the out-of-sample data set. The `eurib_3m` is multiplied by -1 in the second plot.

According to the results we choose the elastic net for our further analysis. The elastic net shows the best results in the "COVID-19" time period and a good ability to capture the direction of German GDP growth when it increases or decreases. In comparison to PCR, this model shows better performance results. In addition, with the elastic model, we can analyze the impact of the predictor variables in GDP growth prediction.

5.4 Results of Models Only with Text Data

As a complementary analysis, we tried nowcasting using *only* the information from text data. We evaluate the nowcasting models with the following sequences of the number of topics: from 3 to 10 with the step size 1 topic and from 50 to 120 with the step size 10 topics. The results are evaluated for the topic frequencies and the sentiment scores. For the random forest and the elastic net separately, we analyzed the best number of topics for "Month 1", "Month 2" and "Month 3" in the "stable", "COVID-19" and "all" time periods. It turns out that most of the models mimic a constant model. Thus, we cannot consider such models suitable for predicting GDP growth. Therefore we only provide the main findings from this analysis.

For the elastic net model, the nowcasts produced by topics do not show large variability over time. This is caused by the frequent choice of the intercept as the only predictor by the elastic net model. The random forest shows a small estimated variance in the "stable" and "all" time periods. The small estimated variance in the "all" time period occurs because the random forest models cannot predict the turbulent time period. In general, we observe better performance results of the random forest models in the "stable" period and the elastic net models in the "all" period. This can be explained by better nowcasts of the elastic net in the "COVID-19" time period. This fact corresponds to the results in Section 5.3. The pronounced volatility of German GDP growth in the "COVID-

19" period influences the large difference between the nowcasts and real GDP growth. Therefore, sometimes when the model better predicts the peaks of German GDP growth, we have a larger model's variance in the "all" time period.

We notice that in many cases, the RMSE, the estimated bias and variance do not show a sequential decrease (or increase) with the new information added by the published articles at the end of the second or third month, which is an indicator for little information in the text data.

5.5 Results of Models Adding Text Data to the Established Economic Indicators

We now add text data to the elastic net nowcasting models with financial, survey and real activity indicators and analyze the role of text data in nowcasting German GDP growth. We evaluate the models with different numbers of topics, which are treated as frequency-based or sentiment-adjusted time series.

We extract the best configurations of text data according to the predicted time period ("stable", "COVID-19" and "all") and time of the nowcast's production ("Month 1", "Month 2", "Month 3"). All models based on frequencies show worse results in the "COVID-19" and "all" periods than the models without text. Therefore, for the frequencies case, we focus on the results in the "stable" period. For the sentiment scores, we also discuss the results of the best models in the "COVID-19" and "all" periods. The results of the best models are presented in Table 7.

In addition to elastic net, we also considered random forest models. However, as discussed in Section 5.3, these showed worse results, so we have omitted their discussion here.

We notice that the best models with sentiment scores usually tend to produce better nowcasts in comparison to the best models with topic frequencies. Sentiment scores are useful not only in the "stable" period but in the more turbulent periods (Table 7). We are unable to observe any stable dependence that a larger or smaller number of topics always produces better nowcasts. However, we detect that smaller numbers of topics produce better nowcasts at the end of the quarter. This effect arises from the fact that we have many variables with mutual dependencies at the end of the quarter. We do not observe any model that improves the nowcasts in the turbulent time period at the end of the first month. The improvement in the nowcasts by the elastic net models with 10 and 9 sentiment-adjusted topics, financial, survey and real activity indicators at the end of the second and third nowcasting months is explained by the improvement in the nowcast in Q4 2020.

In general, RMSE values become smaller with an increasing number of available indicators in the "stable" period. Exceptions are models with 110 and 90 sentiment-adjusted topics in the "stable" period. The problem occurs because the topics change their structure. Moreover, some topics coincide with the behavior of German GDP growth only in several months and do not influence GDP growth in the way that the model predicts.

Table 7: The best results of the elastic net models with text data, financial, economic and survey indicators. The topics are transformed into time series either by the frequencies ("freq."), or by the sentiment scores ("sent"). The number before "sent." or "freq." is a number of sentiment-adjusted or frequency-based topics. "Var." is an estimated variance of each model. The numbers in bold indicate the best performing models for each month in the "stable", "COVID-19" and "all" periods for the frequencies and sentiment scores separately. The best results have lower RMSE than the models without text data. The italic, bold numbers detect the best results of the models for each nowcasting month in the "stable", "COVID-19" and "all" periods among frequencies and sentiment scores.

The "stable" period: frequencies and non-text indicators

Model	110 freq.			120 freq.			6 freq.		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	**0.72**	0.09	0.16	0.79	0.06	0.29	0.79	0.18	0.42
Month 2	0.65	0.11	0.12	**0.60**	0.09	0.14	0.67	0.06	0.19
Month 3	0.55	0.10	0.09	0.57	0.10	0.12	**0.54**	0.10	0.16

The "stable" period: sentiment scores and non-text indicators

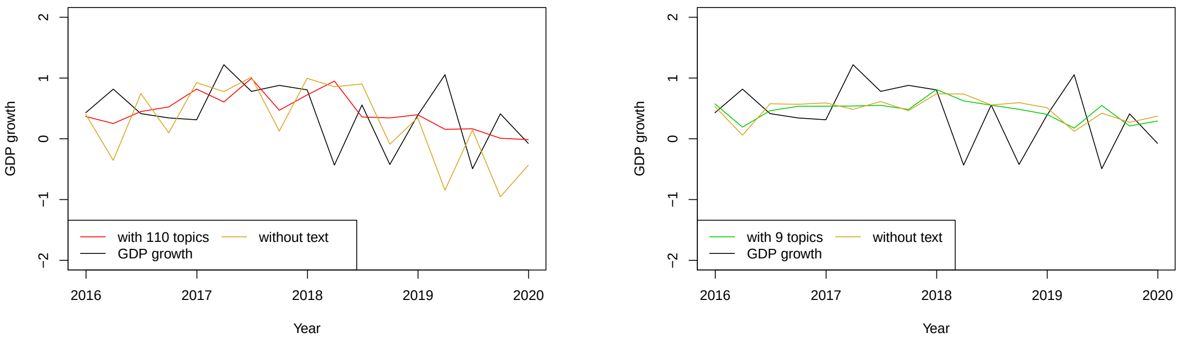
Model	70 sent.			110 sent.			90 sent.		
	RMSE	Bias	Var.	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	0.70	0.07	0.26	***0.55***	0.03	0.09	0.73	0.05	0.23
Month 2	0.68	0.01	0.17	***0.55***	0.14	0.17	0.74	0.07	0.17
Month 3	***0.52***	0.08	0.09	0.62	0.09	0.12	***0.52***	0.02	0.09

**The "COVID-19" and "all" periods:
sentiment scores and non-text indicators**

Model	10 sent.					
	COVID-19			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	4.06	0.31	12.27	2.17	0.21	3.23
Month 2	***3.89***	0.08	9.20	***2.13***	0.09	2.47
Month 3	2.07	0.73	20.69	1.22	0.09	4.83

**The "COVID-19" and "all" periods:
sentiment scores and non-text indicators
(continued)**

Model	9 sent.					
	COVID-19			*All*		
	RMSE	Bias	Var.	RMSE	Bias	Var.
Month 1	4.07	0.54	14.83	2.17	0.34	3.91
Month 2	4.19	0.05	8.98	2.23	0.06	2.36
Month 3	***2.02***	0.70	20.41	***1.18***	0.12	4.77



(a) The elastic net, the nowcasts at the end of the first month.

(b) The random forest, the nowcasts at the end of the second month.

Figure 3: German GDP growth and the nowcasts of the model with financial, survey and real activity indicators and with/without sentiment-adjusted topics in the "stable" period. The nowcasts are presented for the elastic net model with/without 110 topics at the end of the first month and for the random forest with/without 9 topics at the end of the second month.

At this point, the question arises in which time periods text data improve the nowcasts. Figure 3a shows the nowcasts with and without 110 sentiment-adjusted topics at the end of the first month of the nowcasting quarter in the "stable" period of the elastic net model. Figure 3b presents the nowcasts with and without 9 sentiment-adjusted topics at the end of the second month of the random forest model. Both plots show that the models with text data better predict German GDP growth in Q1 2016. The elastic net model with topical information shows better results in the Q4 2018 – Q4 2019 time period. For the random forest, we notice a small improvement in the nowcasts in the Q1 2016 – Q1 2017 and Q4 2017 – Q1 2019 time periods.

Apart from the results of the best models, described in Tables 7, most of the models with text data improve the nowcasts at the end of the first month in comparison to the elastic net and random forest models without text in the "stable" period. The results of these models ("Month 1") are presented in Table 8. However, all these models show a larger or equal RMSE to the AR(1) model in the "stable" period at the end of the first month (Table 5 on page 23).

Finally, we illustrate the influence of text data compared to financial, survey and real activity data and perform counterfactual analysis for the elastic net and the random forest models. The counterfactual analysis compares RMSE values of the models with or without text data, financial, survey and real activity indicators. Based on this, we can identify the data which improves the model performance. For this, we focus on the advantages that text data can bring in nowcasting German GDP growth. In this case, we choose 70 sentiment-adjusted topics for the elastic net model. This combination shows good results in the "stable" period, where the model does not show unusual behavior (e.g., the larger RMSE at the end of the second or third month compared to the first month). In the random forest case, we choose the model with 9 sentiment-adjusted topics. The

Table 8: Top 15 best results of the models at the end of the first month of the nowcasting quarter ("Month 1") with text data, financial, economic and survey indicators. The topics are transformed into time series either by the frequencies ("freq."), or by the sentiment scores ("sent"). The number before "sent." or "freq." is a number of sentiment-adjusted or frequency-based topics. "Var." is an estimated variance of each model.

The "stable" period: "Month 1" model

Model	RMSE	Bias	Var.	*Model*	RMSE	Bias	Var.
elastic net without text	**0.79**	**0.14**	**0.41**	**random forest without text**	**0.65**	**0.03**	**0.09**
4 sent.	0.77	0.17	0.38	6 freq.	0.63	0.01	0.05
8 sent.	0.76	0.19	0.40	7 sent.	0.63	0.02	0.05
70 freq.	0.76	0.07	0.25	5 freq.	0.61	0.02	0.06
9 sent.	0.74	0.27	0.42	8 freq.	0.60	0.02	0.05
10 sent.	0.74	0.17	0.37	8 sent.	0.60	0.05	0.06
3 freq.	0.74	0.15	0.34	9 sent.	0.60	0.05	0.05
9 freq.	0.74	0.22	0.30	10 sent.	0.60	0.04	0.04
10 freq.	0.74	0.14	0.28	70 sent.	0.60	0.03	0.03
50 sent.	0.73	0.16	0.29	7 freq.	0.59	0.03	0.05
90 sent.	0.73	0.05	0.23	9 freq.	0.59	0.02	0.05
8 freq.	0.73	0.20	0.34	50 freq.	0.58	0.01	0.03
110 freq.	0.72	0.09	0.16	50 sent.	0.58	0.00	0.03
80 sent.	0.71	0.07	0.38	120 sent.	0.58	0.04	0.03
70 sent.	0.70	0.07	0.26	10 freq.	0.57	0.05	0.04
110 sent.	0.55	0.03	0.09	80 sent.	0.56	0.00	0.03

counterfactual analysis is performed in the "stable", "COVID-19" and "all" periods. Table 9 presents the RMSE values of the elastic net model, Table 14 in the Appendix shows the results of the random forest in the "stable", "COVID-19" and "all" time periods.

We notice again that the RMSE increases for some cases with new information published during the nowcasting quarters, especially for the elastic net model without real activity indicators in the "COVID-19" period. The reasons are changes in the topic structure and the `bus_exp_Index` variable. In the "COVID-19" period, expectations about the business situation for the next six months substantially change over the quarter compared to the "stable" period and influence German GDP growth because we do not have any real activity indicators.

In all cases, we see that the models without real activity indicators have the worst results at the end of the third month. The reason is that the turnover and production output indices directly influence German GDP growth and are released in the third month of the quarter. Also, we see the tendency that the models without text data have worse or equal results in comparison to the models with text data and all other indicators in the "stable" period (Table 9). In addition, all models with text data have better results than the models without text data at the end of the first month in the "stable" period (Table 8). These observations reveal that newspaper data can improve nowcasts at the early stages of nowcasting in the "stable" period. However, the models with text do not produce the same effect in the turbulent period (see Tables 9 and 14 in the "COVID-19"

Table 9: The counterfactual analysis of the elastic net model with 70 sentiment-adjusted ("sent.") topics in the "stable", "COVID-19" and "all" time periods (reports RMSE). "Text+FSR" is a model with text and financial, survey and real activity indicators, "FR" includes financial and real activity indicators, "FS" financial and survey indicators, and "SR" survey and real activity variables. "FSR" is the model with all types of indicators without text data. The bold numbers indicate the best model in each month in the "stable", "COVID-19" and "all" time periods separately, the italic numbers the worst.

The "stable" period: 70 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	0.70	**0.65**	0.68	0.66	*0.79*
Month 2	**0.68**	0.71	0.71	*0.74*	**0.68**
Month 3	**0.52**	0.55	*0.56*	**0.52**	0.55

The "COVID-19" period: 70 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	4.49	*5.47*	4.55	4.47	**3.91**
Month 2	5.12	5.53	5.32	*5.54*	**4.04**
Month 3	2.68	2.52	*5.58*	2.73	**2.17**

The "all" period: 70 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	2.37	*2.85*	2.39	2.35	**2.11**
Month 2	2.68	2.89	2.78	*2.90*	**2.15**
Month 3	1.44	1.37	*2.89*	1.46	**1.21**

and "all" time periods). The proper aggregation of text data and their combination with indicators can also improve the nowcasts at the end of the second and third months (see Tables 9 and 14). Survey data mostly improve the nowcasts in the turbulent period, especially at the end of the first and second months (in Tables 9 and 14 the "Text+FR" models have larger RMSE values than the "Text+FSR" models at the end of the first and second months in "COVID-19" and "all" time periods). This reveals that survey data can replace the missing information about real activity indicators. Financial indicators tend to improve the nowcasts at the end of the second month in all time periods (in Tables 9 and 14 the "Text+SR" models have larger RMSE values than the "Text+FSR" models at the end of the second month). The reason is the publishing of monetary aggregates and interest rates.

6 Conclusion

This work presents an approach for nowcasting German GDP growth using text data from newspaper articles. We apply the blocking approach to overcome the mixed frequency problem. We train the PCR, random forest and elastic net models based on real activity, financial and survey data to produce the monthly nowcasts. The random forest model shows the best results in the stable time period when German GDP growth does not change substantially. The elastic net model outperforms other models in the turbulent COVID-19 pandemic period. In addition, we compare the results with the AR(1) model. The random forest shows better nowcasting performance than the AR(1) model in the COVID-19 pandemic and whole out-of-sample time period and it shows better nowcasting results than the AR(1) model at the end of the third month in the time period before the COVID-19 pandemic. The elastic net shows better results than the AR(1) model during the COVID-19 pandemic.

In order to aggregate unstructured text data, we apply the LDA topic model with a wide range of topics and two aggregation techniques based on frequency and sentiment. Our experiments reveal that the number of topics influences the quality of nowcasts. However, we do not find the relation that a larger or smaller number of topics always produce better results. The choice of models plays an important role in producing nowcasts. Thus, the elastic net models more often tend to have good results with a large number of topics, though the random forest tend to perform better with a small number of topics. The nowcasting results show that the sentiment-adjusted topics better predict the unstable COVID-19 time period compared to the frequencies of topics. Additionally, in combination with other indicators, sentiment scores usually produce better nowcasts than frequencies in the whole prediction period, especially before the COVID-19 pandemic.

The counterfactual analysis of the models with financial, survey and real activity indicators and text data show that text data can be useful at the early stages of the nowcasts in the time period before the "COVID-19" pandemic. Text data can complement the missing information of real activity and financial indicators, which is unavailable at the beginning of the nowcasting quarter.

7 Outlook

There are several directions for further analysis and potential improvements of the nowcasts. First, it may be reasonable to combine the models for producing the final nowcasts. This can be realized similar to the approach by Ellingsen et al. (2021). Thus, we can combine the results of the random forest models and the elastic net models to improve the nowcasts. Moreover, it may be useful to try more complex models such as (deep) artificial neural networks (Kalamara et al., 2020) or dynamic factor models (Thorsrud, 2020).

In the context of sentiment analysis, it may be promising to compare the SentiWS dictionary with other German-language sentiment dictionaries (e.g., the German-language dictionary created by Bannier et al., 2019). We can additionally investigate different sentiment aggregation techniques or exclude potentially irrelevant words using the term frequency–inverse document frequency (Salton and Buckley, 1988) transformation to enhance the expressiveness of topics.

Regarding topic selection, further analysis may cover the use of only the relevant topics. The relevance of the topics can be defined by their correlation to German GDP growth or by expert analysis. To keep the human effort within reasonable limits, the number of potential models need to be greatly reduced for this purpose. Therefore, it could be worth using the RollingLDA (Rieger et al., 2021) model, which makes repeated recalculation of the LDA models unnecessary due to the architecture of the method, as well as it makes the topics consistently interpretable over time.

Acknowledgments

The present study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA). In addition, the authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

References

Aguilar, Pablo, Corinna Ghirelli, Matías Pacce, and Alberto Urtasun (2021). "Can news help measure economic sentiment? An application in COVID-19 times". In: *Economics Letters* 199, p. 109730. DOI: 10.1016/j.econlet.2021.109730.

Akaike, Hirotogu (1998). "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of Hirotugu Akaike*. Springer, pp. 199–213. DOI: 10.1007/978-1-4612-1694-0_15.

Algaba, Andres, David Ardia, Keven Bluteau, Samuel Borms, and Kris Boudt (2020). "Econometrics meets sentiment: An overview of methodology and applications". In: *Journal of Economic Surveys* 34.3, pp. 512–547. DOI: https://doi.org/10.1111/joes.12370.

Andreini, Paolo, Charlotte Senftleben-König, Thomas Hasenzagl, Lucrezia Reichlin, and Till Strohsal (2020). *Nowcasting German GDP*. CEPR, Discussion Paper No. DP14323. URL: https://ssrn.com/abstract=3526048.

Aprigliano, Valentina, Simone Emiliozzi, Gabriele Guaitoli, Andrea Luciani, Juri Marcucci, and Libero Monteforte (2022). "The power of text-based indicators in forecasting Italian economic activity". In: *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2022.02.006.

Ardia, David, Keven Bluteau, and Kris Boudt (2019). "Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values". In: *International Journal of Forecasting* 35.4, pp. 1370–1386. DOI: 10.1016/j.ijforecast.2018.10.010.

Ashwin, Julian, Eleni Kalamara, and Lorena Saiz (2021). *Nowcasting Euro area GDP with news sentiment: A tale of two crises*. ECB, Working Paper No. 2616. DOI: 10.2139/ssrn.3971974.

Baffigi, Alberto, Roberto Golinelli, and Giuseppe Parigi (2004). "Bridge models to forecast the euro area GDP". In: *International Journal of Forecasting* 20.3, pp. 447–460. DOI: 10.1016/S0169-2070(03)00067-0.

Bannier, Christina, Thomas Pauls, and Andreas Walter (2019). "Content analysis of business communication: introducing a German dictionary". In: *Journal of Business Economics* 89.1, pp. 79–123. DOI: 10.1007/s11573-018-0914-8.

Bec, Frédérique and Matteo Mogliani (2015). "Nowcasting French GDP in real-time with surveys and "blocked" regressions: Combining forecasts or pooling information?" In: *International Journal of Forecasting* 31.4, pp. 1021–1042. DOI: 10.1016/j.ijforecast.2014.11.006.

Blei, David M. (2012). "Probabilistic Topic Models". In: *Communications of the ACM* 55.4, pp. 77–84. DOI: 10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Brockwell, Peter J and Richard A Davis (1991). *Time Series: Theory and Methods*. Springer Science & Business Media. DOI: 10.1007/978-1-4419-0320-4.

Brockwell, Peter J and Richard A Davis (2016). *Introduction to Time Series and Forecasting*. Springer. DOI: 10.1007/978-3-319-29854-2.

Caggiano, Giovanni, George Kapetanios, and Vincent Labhard (2011). “Are more data always better for factor analysis? Results for the euro area, the six largest euro area countries and the UK”. In: *Journal of Forecasting* 30.8, pp. 736–752. DOI: 10.1002/for.1208.

Carriero, Andrea, Todd E Clark, and Massimiliano Marcellino (2015). “Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility”. In: *Journal of the Royal Statistical Society. Series A,(Statistics in Society)* 178.4, p. 837. DOI: 10.1111/rssa.12092.

Cattell, Raymond B (1966). “The scree test for the number of factors”. In: *Multivariate Behavioral Research* 1.2, pp. 245–276. DOI: 10.1207/s15327906mbr0102_10.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei (2009). “Reading Tea Leaves: How Humans Interpret Topic Models”. In: *NIPS: Advances in Neural Information Processing Systems*. Vol. 22. Curran Associates Inc., pp. 288–296. URL: https://papers.nips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554
Abstract.html.

Chen, Weitian, Brian DO Anderson, Manfred Deistler, and Alexander Filler (2012). “Properties of blocked linear systems”. In: *Automatica* 48.10, pp. 2520–2525. DOI: 10.1016/j.automatica.2012.06.020.

Correa, Ricardo, Keshav Garud, Juan M Londono, Nathan Mislang, et al. (2017). *Constructing a dictionary for financial stability*. IFDP notes. Board of Governors of the Federal Reserve System, Washington. DOI: 10.17016/2573-2129.33.

Cryer, Jonathan D and Kung-Sik Chan (2008). *Time Series Analysis: With Applications in R*. Vol. 2. Springer. DOI: 10.1007/978-0-387-75959-3.

Deutsche Bundesbank (2022). *Time series databases*. URL: https://www.bundesbank.de/en/statistics/time-series-databases (visited on 01/10/2022).

Dickey, David A and Wayne A Fuller (1979). “Distribution of the estimators for autoregressive time series with a unit root”. In: *Journal of the American Statistical Association* 74.366a, pp. 427–431. DOI: 10.2307/2286348.

Diebold, Francis X and Glenn D Rudebusch (1991). “Forecasting output with the composite leading index: A real-time analysis”. In: *Journal of the American Statistical Association* 86.415, pp. 603–610. DOI: 10.2307/2290388.

Dowle, Matt and Arun Srinivasan (2020). *data.table: Extension of "data.frame"*. R package version 1.13.2. URL: https://CRAN.R-project.org/package=data.table.

Drucker, Harris, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik (1996). “Support vector regression machines”. In: *NIPS: Advances in neural information processing systems* 9. URL: https://papers.nips.cc/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html.

Eddelbuettel, Dirk (2013). *Seamless R and C++ Integration with Rcpp*. Springer. DOI: 10.1007/978-1-4614-6868-4.

Ellingsen, Jon, Vegard Larsen, and Leif Anders Thorsrud (2021). “News media versus FRED-MD for macroeconomic forecasting”. In: *Journal of Applied Econometrics*. DOI: 10.1002/jae.2859.

European Central Bank (2022). *Euro foreign exchange reference rates*. URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html (visited on 01/10/2022).

Federal Reserve Bank of St. Louis (2022). *FRED, Economic Data*. URL: https://fred.stlouisfed.org/categories (visited on 01/10/2022).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: https://www.jstatsoft.org/v33/i01/.

Gagolewski, Marek (2020). *R package stringi: Character string processing facilities.* URL: http://www.gagolewski.com/software/stringi/.

Gayer, Christian, Alessandro Girardi, and Andreas Reuter (2014). *The role of survey data in nowcasting euro area GDP growth.* European Commission, Directorate-General for Economic and Financial Affairs. See also 10.1002/for.2383. URL: https://ec.europa.eu/economy_finance/publications/economic_paper/2014/pdf/ecp538_en.pdf.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". In: *Journal of Economic Literature* 57.3, pp. 535–574. DOI: 10.1257/jel.20181020.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov (2004). *The MIDAS touch: Mixed data sampling regression models.* URL: https://escholarship.org/uc/item/9mf223rs.

Giannone, Domenico, Lucrezia Reichlin, and David Small (2008). "Nowcasting: The real-time informational content of macroeconomic data". In: *Journal of Monetary Economics* 55.4, pp. 665–676. DOI: 10.1016/j.jmoneco.2008.05.010.

Gilks, Walter R, Sylvia Richardson, and David Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice.* CRC press. DOI: 10.1201/b14835.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235. ISSN: 0027-8424. DOI: 10.1073/pnas.0307752101.

Grolemund, Garrett and Hadley Wickham (2011). "Dates and times made easy with lubridate". In: *Journal of Statistical Software* 40.3, pp. 1–25. URL: https://www.jstatsoft.org/v40/i03/.

Hamner, Ben and Michael Frasco (2018). *Metrics: Evaluation Metrics for Machine Learning.* R package version 0.1.4. URL: https://CRAN.R-project.org/package=Metrics.

Haselmayer, Martin and Marcelo Jenny (2017). "Sentiment analysis of political communication: combining a dictionary approach with crowdcoding". In: *Quality & Quantity* 51.6, pp. 2623–2646. DOI: 10.1007/s11135-016-0412-4.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media. DOI: 10.1007/978-0-387-84858-7.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Hutto, C. and Eric Gilbert (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

ifo Institute (2022a). *ifo Business Climate Index for Germany.* URL: https://www.ifo.de/en/survey/ifo-business-climate-index (visited on 01/10/2022).

ifo Institute (2022b). *ifo Time Series.* URL: https://www.ifo.de/en/umfragen/time-series (visited on 01/10/2022).

Investing.com (2022). *Germany Gross Domestic Product (GDP) QoQ.* URL: https://www.investing.com/economic-calendar/german-gdp-131 (visited on 01/10/2022).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning.* Vol. 112. Springer. DOI: 10.1007/978-1-4614-7138-7.

Kaiser, Henry F. (1960). “The application of electronic computers to factor analysis”. In: *Educational and Psychological Measurement* 20.1, pp. 141–151. DOI: 10.1177/001316446002000116.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2020). *Making text count: economic forecasting using newspaper text.* Bank of England, Working Paper No. 865. URL: https://www.bankofengland.co.uk/working-paper/2020/making-text-count-economic-forecasting-using-newspaper-text.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit (2020). *Robust Machine Learning Algorithms for Text Analysis.* URL: http://www.joseluismontielolea.com/LDA_2021.pdf (visited on 08/22/2022).

Koppers, Lars, Jonas Rieger, Karin Boczek, and Gerret von Nordheim (2021). *tosca: Tools for statistical content analysis.* R package version 0.3-1. DOI: 10.5281/zenodo.3591068. URL: https://github.com/Docma-TU/tosca.

Lehmann, Robert (2020). *The forecasting power of the ifo business survey.* CESifo, Working Paper No. 8291. URL: https://www.cesifo.org/en/publikationen/2020/working-paper/forecasting-power-ifo-business-survey.

Maier, Daniel, Andreas Niekler, Gregor Wiedemann, and Daniela Stoltenberg (2020). “How document sampling and vocabulary pruning affect the results of topic models”. In: *Computational Communication Research* 2.2. DOI: 10.31219/osf.io/2rh6g.

McCracken, Michael W, Michael T Owyang, and Tatevik Sekhposyan (2021). “Real-time forecasting and scenario analysis using a large mixed-frequency Bayesian VAR”. In: *International Journal of Central Banking.* URL: https://www.ijcb.org/journal/ijcb21q5a8.htm.

OECD (2022a). *Gross Domestic Product (GDP) (indicator).* DOI: 10.1787/dc2f7aec-en. (Visited on 01/10/2022).

OECD (2022b). *Quarterly GDP (indicator).* DOI: 10.1787/b86d1fc8-en. (Visited on 01/10/2022).

Ooms, Jeroen (2021). *writexl: Export Data Frames to Excel ’xlsx’ Format.* R package version 1.4.0. URL: https://CRAN.R-project.org/package=writexl.

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R.* Second Edition. Springer. DOI: 10.1007/978-0-387-75967-8.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Remus, R., U. Quasthoff, and G. Heyer (2010). “SentiWS – a publicly available German-language resource for sentiment analysis”. In: *Proceedings of the 7th International Language Resources and Evaluation (LREC’10).* URL: https://aclanthology.org/L10-1339/.

Rieger, Jonas (2020). “ldaPrototype: A method in R to get a prototype of multiple latent Dirichlet allocations”. In: *Journal of Open Source Software* 5.51, p. 2181. DOI: 10.21105/joss.02181.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2021). “RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data”. In: *Findings Proceedings of the 2021 EMNLP-Conference.* ACL, pp. 2337–2347. DOI: 10.18653/v1/2021.findings-emnlp.201.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). “Learning representations by back-propagating errors”. In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0.

Rünstler, Gerhard, Karim Barhoumi, Szilard Benk, Riccardo Cristadoro, Ard Den Reijer, Audrone Jakaitiene, Piotr Jelonek, António Rua, Karsten Ruth, and Christophe Van Nieuwenhuyze (2009). "Short-term forecasting of GDP using large datasets: a pseudo real-time forecast evaluation exercise". In: *Journal of Forecasting* 28.7, pp. 595–611. DOI: 10.1002/for.1105.

Salton, Gerard and Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval". In: *Information Processing & Management* 24.5, pp. 513–523. DOI: 10.1016/0306-4573(88)90021-0.

Sax, Christoph and Dirk Eddelbuettel (2018). "Seasonal Adjustment by X-13ARIMA-SEATS in R". In: *Journal of Statistical Software* 87.11, pp. 1–17. DOI: 10.18637/jss.v087.i11.

Schauberger, Philipp and Alexander Walker (2020). *openxlsx: Read, write and edit xlsx files.* R package version 4.2.3. URL: https://CRAN.R-project.org/package=openxlsx.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to Information Retrieval.* Cambridge University Press. URL: https://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press. URL: https://dl.acm.org/doi/10.5555/2621980.

Stock, James H and Mark W Watson (1989). "New indexes of coincident and leading economic indicators". In: *NBER Macroeconomics Annual* 4, pp. 351–394. URL: http://www.nber.org/chapters/c10968.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th ACL-Conference.* ACL, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://www.aclweb.org/anthology/P19-1355.

Taieb, Souhaib Ben and Amir F Atiya (2015). "A bias and variance analysis for multistep-ahead time series forecasting". In: *IEEE Transactions on Neural Networks and Learning Systems* 27.1, pp. 62–76. DOI: 10.1109/TNNLS.2015.2411629.

Thorsrud, Leif Anders (2020). "Words are the new numbers: A newsy coincident index of the business cycle". In: *Journal of Business & Economic Statistics* 38.2, pp. 393–409. DOI: 10.1080/07350015.2018.1506344.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

United States Census Bureau (2022). *X-13ARIMA-SEATS Seasonal Adjustment Program.* URL: https://www.census.gov/data/software/x13as.References.html (visited on 01/10/2022).

Weston, Steve (2020a). *doParallel: Foreach Parallel Adaptor for the "parallel" Package.* R package version 1.0.16. URL: https://CRAN.R-project.org/package=doParallel.

Weston, Steve (2020b). *foreach: Provides Foreach Looping Construct.* R package version 1.5.1. URL: https://CRAN.R-project.org/package=foreach.

Wright, Marvin N. and Andreas Ziegler (2017). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17. DOI: 10.18637/jss.v077.i01.

Zamani, Mohsen, Wetian Chen, Brian D.O. Anderson, Manfred Deistler, and Alexander Filler (2011). "On the zeros of blocked linear systems with single and mixed frequency data". In: *50th IEEE Conference on Decision and Control and European Control Conference*, pp. 4312–4317. DOI: 10.1109/CDC.2011.6160434.

Zeileis, Achim and Gabor Grothendieck (2005). "zoo: S3 infrastructure for regular and irregular time series". In: *Journal of Statistical Software* 14.6, pp. 1–27. DOI: 10.18637/jss.v014.i06.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendix

Table 10: List of packages for the nowcasting approach construction, evaluation and visualization.

Package	Citation	Used for
`openxlsx`	Schauberger and Walker, 2020	read data from "xlsx" format
`tosca`	Koppers et al., 2021	text data preparation and statistical analysis
`ldaPrototype`	Rieger, 2020	LDA algorithm
`foreach`	Weston, 2020b	parallel executing loop
`doParallel`	Weston, 2020a	parallel execution of "foreach"
`lubridate`	Grolemund and Wickham, 2011	date configuration
`zoo`	Zeileis and Grothendieck, 2005	time series data aggregation
`seasonal`	Sax and Eddelbuettel, 2018	seasonal adjustment with X-13ARIMA-SEATS
`urca`	Pfaff, 2008	ADF test
`data.table`	Dowle and Srinivasan, 2020	fast aggregation of the large data
`stringi`	Gagolewski, 2020	string's processing
`glmnet`	Friedman et al., 2010	elastic net regression
`ranger`	Wright and Ziegler, 2017	random forest regression
`Rcpp`	Eddelbuettel, 2013	R and C++ integration for `ranger`
`Metrics`	Hamner and Frasco, 2018	evaluation metrics
`writexl`	Ooms, 2021	export data frames to 'xlsx' format

Table 11: Overview of selected transformations for predictor variables. The transformation "1" is a difference transformation, "2" is a percentage changes transformation, and "0" defines no transformation. Type "f" defines financial indicators, "r" real activity indicators, "s" survey indicators, and "t" text data.

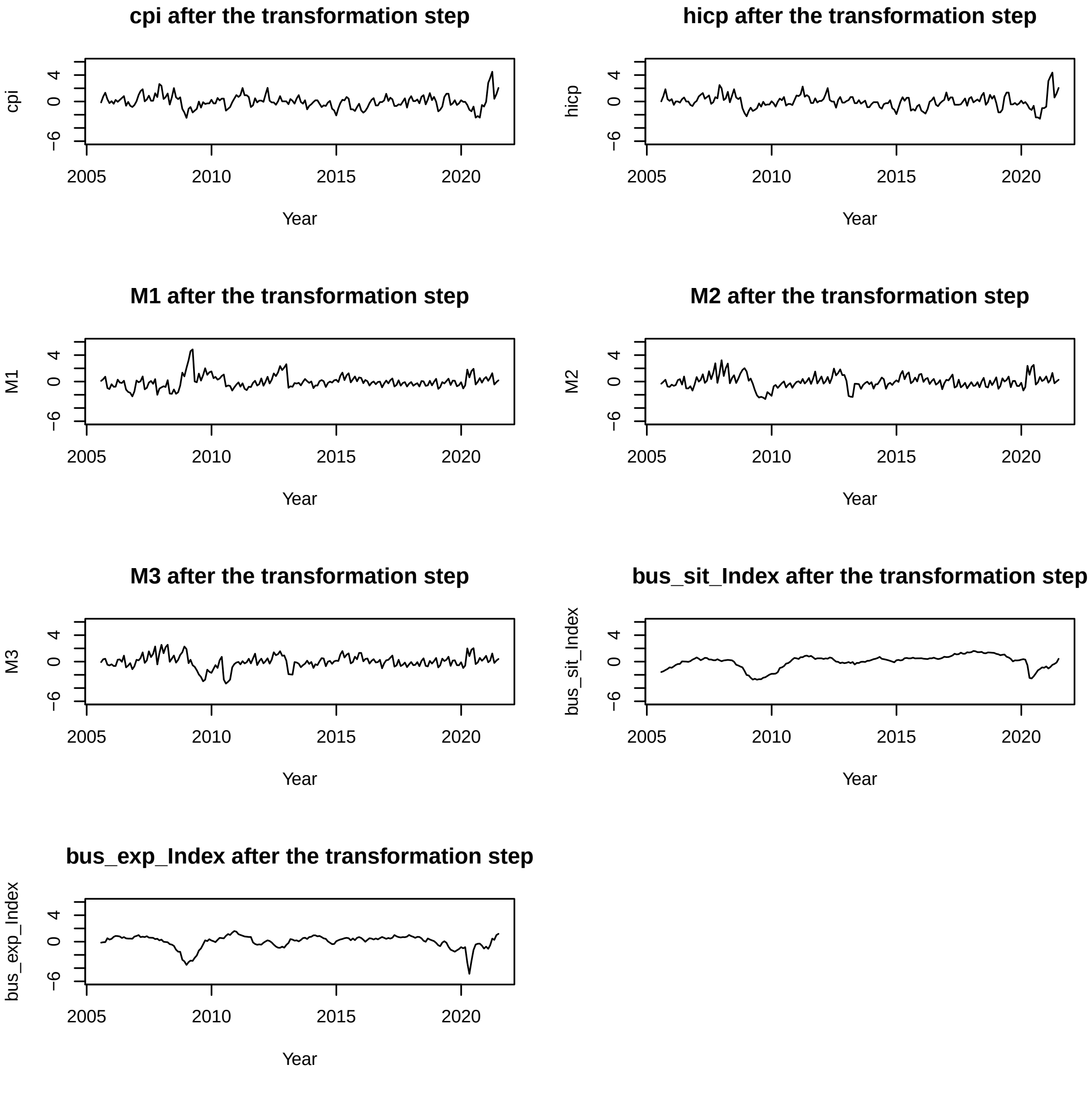
Indicator	Type	Transformation	Indicator	Type	Transformation
bond_10y_yield	f	1	hicp	r	2
eurib_3m	f	1	cpi	r	2
exr_usd	f	2	ip_capital_goods	r	2
exr_jpy	f	2	ip_civil_engineering	r	2
exr_gbp	f	2	ip_consumer_goods	r	2
gold_prices	f	2	ip_durable_consumer_goods	r	2
libor_3m_us	f	1	ip_energy	r	2
M1	f	2	ip_industry	r	2
M2	f	2	ip_intermediate_goods	r	2
M3	f	2	ip_main_construction_industry	r	2
oil_price	f	2	ip_non-durable_consumer_goods	r	2
vix_us	f	2	ip_structural_engineering	r	2
bus_sit_Index	s	0	turnover_industry	r	2
bus_exp_Index	s	0	ur_de	r	1
text (frequencies)	t	1	text (sentiments)	t	1

Table 12: The results of the ADF test without a trend and a drift term. The results show the p-values of the ADF test. When the p-values are equal to or smaller than 0.05, then the hypothesis about the existence of a unit root is rejected. The expression "< 0.01" means that the p-values are smaller than 0.01. For all text data, we check the ADF test on transformed topics for the last LDA update to consider the whole time period. The possible maximal lag in the ADF test equals 12 (the final results of the test rejection at level 0.05 also remain the same for 6 lags). For simplicity, the text data results for all topics are summarized in the last row.

Indicator	p-value	Indicator	p-value
bond_10y_yield	< 0.01	hicp	0.16
eurib_3m	< 0.01	cpi	0.22
exr_usd	< 0.01	ip_capital_goods	< 0.01
exr_jpy	< 0.01	ip_civil_engineering	< 0.01
exr_gbp	< 0.01	ip_consumer_goods	< 0.01
gold_prices	< 0.01	ip_durable_consumer_goods	< 0.01
libor_3m_us	0.01	ip_energy	< 0.01
M1	0.14	ip_industry	< 0.01
M2	0.22	ip_intermediate_goods	< 0.01
M3	0.20	ip_main_construction_industry	< 0.01
oil_price	< 0.01	ip_non-durable_consumer_goods	< 0.01
vix_us	< 0.01	ip_structural_engineering	< 0.01
bus_sit_Index	0.96	turnover_industry	< 0.01
bus_exp_Index	0.98	ur_de	< 0.01
text (frequencies)	< 0.01	text (sentiments)	< 0.01

Table 13: Example of a topic that changes its structure over time. The case is for 7 topics, extracted from the time periods April 2005 – October 2015 and April 2005 – June 2021.

Top 10 words (April 2005 – October 2015)	Top 10 words (April 2005 – June 2021)
china, russland, indien, eon, peking, prozent, strom, russischen, chinas, rwe	china, ezb, wirtschaft, regierung, europa, banken, griechenland, usa, land, eu



cpi after the transformation step

hicp after the transformation step

M1 after the transformation step

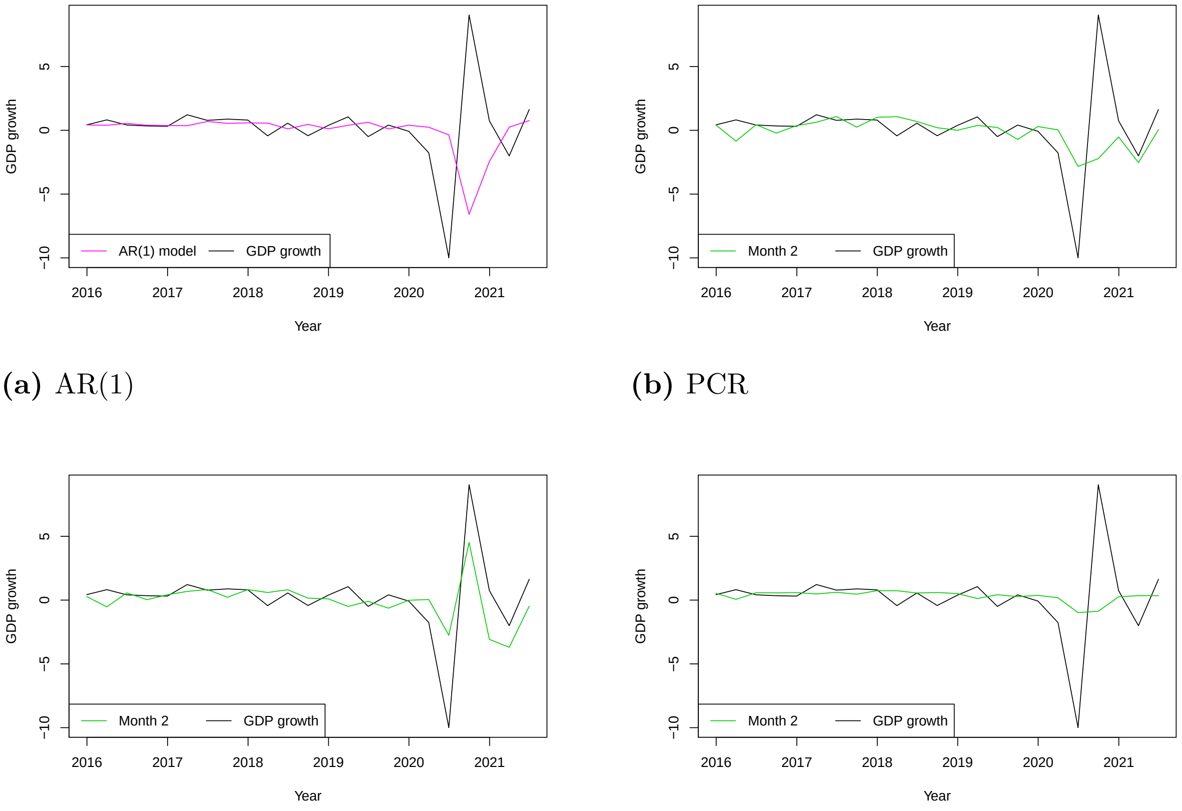
M2 after the transformation step

M3 after the transformation step

bus_sit_Index after the transformation step

bus_exp_Index after the transformation step

Figure 4: The predictor variables `cpi`, `hicp`, `M1`, `M2`, `M3` after the transformation step. All data are standardised. Each plot covers the time period from Q3 2005 to Q2 2021.



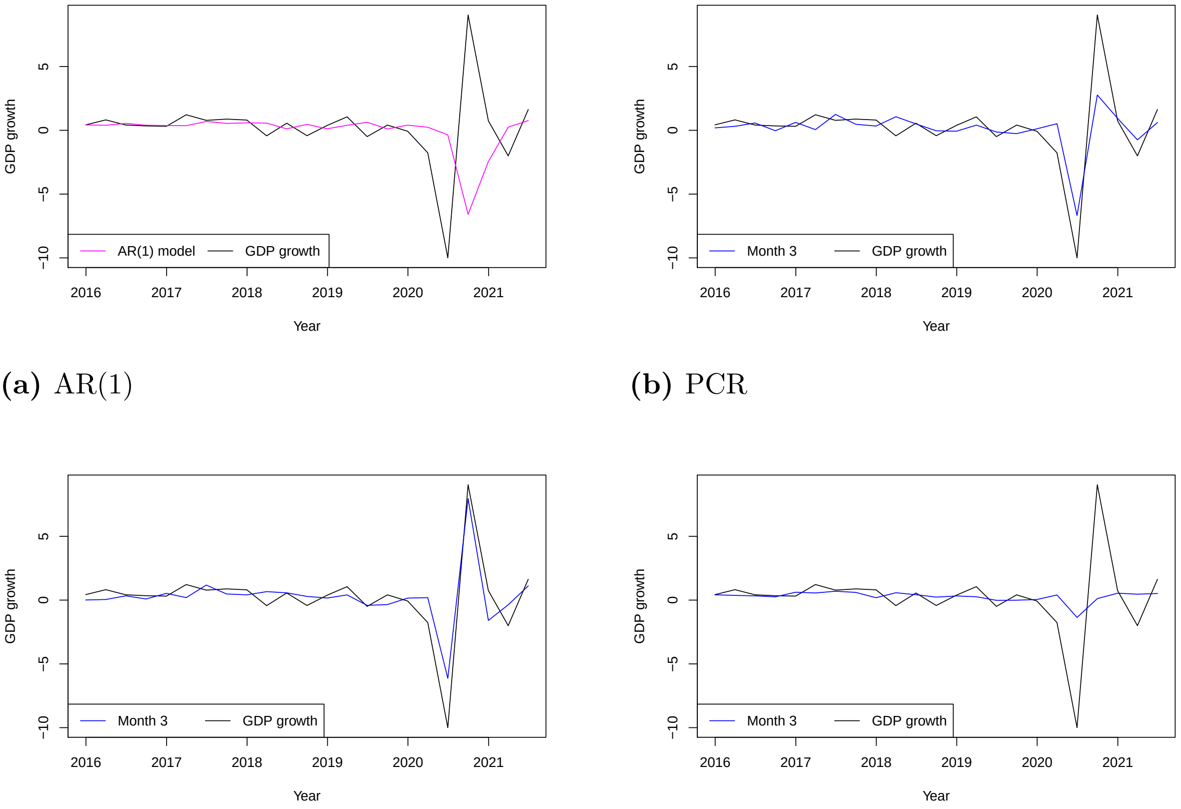
(a) AR(1)

(b) PCR

(c) Elastic net

(d) Random forest

Figure 5: German GDP growth and its nowcasts at the end of the second month of the quarter with the AR(1), PCR, the elastic net and random forest models in the out-of-sample period.



(a) AR(1)

(b) PCR

(c) Elastic net

(d) Random forest

Figure 6: German GDP growth and its nowcasts at the end of the third month of the quarter with the AR(1), PCR, the elastic net and random forest models in the out-of-sample period.

Table 14: The counterfactual analysis of the random forest model with 9 sentiment-adjusted ("sent.") topics in the "stable", "COVID-19" and "all" time periods (reports RMSE). "Text+FSR" is a model with text and financial, survey and real activity indicators, "FR" includes financial and real activity indicators, "FS" financial and survey indicators, and "SR" survey and real activity variables. "FSR" is the model with all types of indicators without text data. The bold numbers indicates the best model in each month in the "stable", "COVID-19" and "all" time periods separately, the italic numbers the worst.

The "stable" period: 9 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	0.60	0.60	**0.57**	0.61	*0.65*
Month 2	**0.55**	0.56	**0.55**	*0.62*	0.59
Month 3	0.47	0.47	*0.52*	**0.45**	0.47

The "COVID-19" period: 9 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	5.83	*5.98*	**5.64**	5.77	5.66
Month 2	**5.59**	*5.79*	5.65	5.71	5.64
Month 3	5.43	5.49	*5.59*	**5.20**	5.27

The "all" period: 9 sentiment-adjusted topics

Model	Text+FSR	Text+FR	Text+FS	Text+SR	FSR
Month 1	3.02	*3.09*	**2.92**	2.99	2.95
Month 2	**2.89**	*2.99*	2.92	2.96	2.92
Month 3	2.80	2.83	*2.89*	**2.68**	2.72