Nowcasting GDP using tone-adjusted time-varying layered news topics: Evidence from the Financial press

Dorinth W. van Dijk*
d.w.van.dijk@dnb.nl, De Nederlandsche Bank
Department of Econometrics & Modelling, Amsterdam, Netherlands
and
Jasper M. de Winter
j.m.de.winter@dnb.nl, De Nederlandsche Bank
Department of Econometrics & Modelling, Amsterdam, Netherlands

February 16, 2024

Abstract

In this paper, we extract tone-adjusted, time-varying, and hierarchically ordered topics from a large corpus of Dutch financial news to investigate their usefulness for monitoring the business cycle and nowcasting GDP growth in the Netherlands. Our newspaper sentiment indicators exhibit a high concordance with the business cycle. Furthermore, we find that newspaper sentiment greatly increases the forecast accuracy of a standard dynamic factor model and outperforms a simple benchmark model by a large margin. We conclude that our tone-adjusted newspaper topics contain valuable information not embodied in monthly indicators from statistical offices. The layering and time-variation of the news topics allow for ample opportunities to determine the main drivers of the nowcast.

Keywords: dynamic factor model, latent dirichlet allocation, sentiment analysis.

* We gratefully acknowledge comments from Maurice Bun, Peter van Els, Olaf Sleijpen, and participants of the 2021 Joint Statistical Meeting, the 2023 Conference of the International Association for Applied Econometrics, the 2023 conference of the European Economic Association and research seminars at the Bank of England, the European Commission Joint Research Centre, the MIT Center for Real Estate, Atradius Netherlands, the Dutch Foreign Bankers' Association, Ortec Finance, the Autoriteit Financiële Markten, and De Nederlandsche Bank. We thank Edwin Fillekes, Mathijs Rotteveel, and Eugenie van Wiechen of Het Financieele Dagblad and Olaf van der Wiel of De Nederlandsche Bank for useful discussions and their help to unlock the digital archive of Het Fiancieele Dagblad. The views expressed herein are those of the authors and do not necessarily reflect the position of De Nederlandsche Bank. All remaining errors are our own.

1 Introduction

The use of big data sources for short-term forecasting have recently gained traction. A relatively novel data source is text, particularly newspaper articles (see e.g. Gentzkow et al., 2019, Thorsrud, 2020, Kalamara et al., 2022, Aprigliano et al. (2023) and Barbaglia et al., 2022). The advent of machine-learning techniques has facilitated the extraction of sentiment and topics from these texts.

This research primarily focuses on the improvement in forecast accuracy using textual data. We delve into two related aspects. Firstly, we examine if tone-adjusted news topics serve as good indicators for gauging the business cycle. Secondly, we investigate whether the tone-adjusted news topics increase the forecast accuracy of a state-of-the art nowcasting model, focusing on short-term forecast of the quarterly growth rate of gross domestic product (GDP).

We employ a novel newspaper corpus, covering the period from January 1^{st}, 1985, to January, 18^{th}, 2021, with articles published in the largest and only Dutch financial newspaper *Het Financieele Dagblad*. In total, we analyze approximately one million news articles. Because we observe the articles over a relatively long period of time we are able to analyze periods of crisis (Global Financial Crisis and the COVID-19–crisis) and more tranquil times separately. We construct a novel dictionary, stemming technique, and sentiment list tailored to Dutch financial news articles.

There is an expanding literature on Bayesian estimation of topic models and nowcasting. The most popular variant, Latent Dirichlet Allocation (LDA), was first introduced by Blei et al. (2003). This paper sparked a growing literature in the machine-learning field, see, Churchill and Singh (2022) for a recent survey. However, its use in economics has been quite limited until recently. With the increasing availability of large text databases and newspaper corpora and the development of machine-learning methods, economic applications are becoming more widespread. Hansen et al. (2018) is one of the first economic applications, examining the effect of transparency on the deliberation of the Federal Open Market Committee (FOMC) of the Board of Governors of the Federal Reserve System. More recently, the LDA model was applied to a corpus of financial newspaper articles by, amongst others, Thorsrud (2020). Recent contributions to the nowcasting literature combine sentiment and uncertainty extracted via pre-defined lists from newspaper texts (e.g. Shapiro et al., 2022 and Gentzkow et al., 2019) and incorporate these in nowcasting models (e.g. Barbaglia et al., 2022 and Rambaccussing and Kwiatkowski, 2020).

Our study contributes to the existing body of knowledge in several significant ways. To begin, we propose a new variant of the plain-vanilla LDA model, the tone-adjusted time-varying layered topic model. This model allows us to categorize news into time-varying hierarchical topics, each with its own sentiment. We then assess whether the sentiment indicator derived from this model accurately reflects the Dutch business cycle. Secondly, we develop a unique Dutch economic dictionary designed specifically to gauge the sentiment of economic news related to the Dutch economy. Our approach is further enhanced by the inclusion of Dutch-specific valence shifters. Thirdly, we present a streamlined Bayesian estimation method for easily estimating our topic model. This method uses the posterior distributions of previous time slices and layers as priors for the estimation of new time slices and deeper layers. Lastly, we contribute to the nowcasting literature by incorporating our novel tone-adjusted topics into a pseudo real-time out-of-sample forecast comparison. This comparison is conducted between a state-of-the-art nowcasting model with and without various versions of tone-adjusted newspaper topics. We also examine the impact of the modeling choices we made in relation to the topic model, such as layering and time variation, on the forecast accuracy of our nowcasting model.

Our main findings are twofold. First, we find that our measure for news sentiment accurately tracks the business cycle. More specifically, the correlation between aggregate newspaper sentiment and y-o-y GDP growth is high (0.79) and the downswings in sentiment correspond with recession indicators. The layered character of our model specifies the underlying dynamics of aggregate sentiment. We distinguish three layers totaling 64 topics. Sentiment swings within these topics make intuitive sense, e.g. at the onset of the Global Financial Crisis in 2008 financial markets sentiment declines most notably. The time-varying nature of the model shows significant shifts of important words within topics, e.g. words like ECB and Euro gained prominence within the financial markets topic.

Second, making use of newspaper sentiment topics significantly improves the forecast accuracy of GDP growth in the short-run. Compared to a simple benchmark model (i.e. prevailing mean), our dynamic factor model with newspaper topics results in MSFEs that are up to 42% lower. The newspaper topics also add significant value to a dynamic factor model without newspaper topics: the MSFE declines up to 37%. The forecast enhancement is significant across all considered horizons (i.e. backcasts, nowcasts, forecasts), but is strongest when nowcasting. Within the newspaper topics, we find that financial market sentiment has the highest contribution.

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the dataset used in our analysis, which include a corpus of newspaper articles and a set of macroeconomic indicators. The construction process of our tone-adjusted time-varying layered topic model is outlined in Section 3. Section 4 presents the nowcasting exercise. Section 5 presents the outcomes of our topic model, while the outcomes of the nowcasting exercise are discussed in Section 6. Section 7 concludes.

2 Data

This section describes the dataset we used for the estimation of our tone-adjusted time-varying topic model and our nowcasting exercise. Section 2.1 outlines the dataset of newspaper articles and the vocabulary employed in the estimation of the time-varying layered topic model. Section 2.2 details the dataset of monthly macro-economic indicators that we employ in the nowcasting exercise.

2.1 Corpus of newspaper articles

Our source of textual data is a comprehensive database of the only and largest financial newspaper of the Netherlands, *Het Financieele Dagblad* (FD). The database encompasses all articles published in the newspaper (both in print and online) for the period January 1^{st} 1985 up until January 18^{th} 2021. The raw database comprises $1,093,477$ articles. The data includes the complete text of each article, the article title, the publication URL, the publication date, the newspaper section in which the article was published, and one or more one-word tags describing the article content. Some of the articles consist of opinions by policymakers, plans by government and Parliament, and news on topics not directly related to the economy. We use these attributes to clean the database of articles that are not directly related to economic developments, and are therefore deemed non-relevant.

The cleaning process is executed in several steps: we remove irrelevant articles from the database, eliminate stopwords, correct for collocations, and stem all words and verbs to their root. These steps are elaborated in more detail in the Appendix.[1] After the cleaning process, we are left with $582,981$ articles, which is approximately a 47% reduction compared to the raw database. Below, we outline the steps taken to prepare the text of the remaining articles in the database for use in our topic model.

The cleaned corpus contains $1,287,851$ unique word tokens, which poses significant computational challenges. Furthermore, a topic model with such a large number of tokens is at risk of severe overfitting and may not generalize well to hold-out data. Following the approach of others, including Thorsrud (2020) and Barbaglia et al. (2022), we implement several steps to reduce the number of unique tokens.

Firstly, we establish a minimum and maximum number of articles in which a token should appear, based on the entire database of articles. We set the minimum number of documents a word appears in at least 0.1% thereby excluding words with very low frequency from the vocabulary. To remove very common words from the vocabulary, we only include words that occur in a maximum of 50% of the articles.[2] This so-called 'pruning', reduces the number of unique word-tokens from $1,287,851$ to $9,613$, a reduction of more that 99%.

Second, we examine all words and exclude verbs, adjectives, count-words (e.g. million, thousand), retaining only the nouns. The rationale behind this is to capture only the main topic of the sentence, with the noun being the most valuable word in this context. We check for synonyms and convert all words to their singular form. After this second 'pruning' step we are left with a final list of $2,135$ nouns in our vocabulary.[3]

2.2 Dataset of macro-economic indicators

In our nowcasting exercise, as described in Section 4, we merge the topics extracted from the corpus of newspaper articles with a monthly dataset of macro-economic indicators. We have constructed a pseudo real-time dataset of macro-economic indicators, which con-

[1] Our list of Dutch stopwords and transformation from conjugate verbs to their verbstem is publicly available, and can be downloaded here and here, respectively.

[2] An automated variant of this approach would be to use the term frequency–inverse document frequency (tf-idf). We experimented with this automated approach, but in our case the more labor intensive manual approach delivered a more meaningful vocabulary.

[3] Our vocabulary, including a English translation, is publicly available, and can be downloaded here.

sists of 58 monthly time series and quarterly GDP that were downloaded on February 1^{th} 2021. The selection of indicators is similar to those in the real-time database of the FRED-MD database. The statistical monthly information set reflects public knowledge at the beginning of the month and covers a wide range of information readily available to economic agents. The indicators are categorized into four groups. The first category includes hard, quantitative information on production and sales, such as industrial production in various sectors, retail trade turnover, household consumption, world trade and unemployment. The second category comprises financial variables, both quantities (money stock and credit volume) and prices (interest rates and stock prices). These determine the financing conditions for firms and consumers. Moreover, financial market prices partly reflect financial market expectations on output developments in the near future. The third category contains input and output prices, i.e. headline consumer and producer prices, and world market commodity prices. The fourth category contains monthly information on the economic development for the main trading partners of the Netherlands within the euro zone, i.e. Germany, France, Italy, Spain and Belgium. These indicators are potentially important for a small open economy such as the Netherlands. The Appendix provides detailed information on the sources, availability and transformation of the indicators.

3 Tone-adjusted time-varying layered topic model

The Latent Dirichlet Allocation (LDA) model has become a powerful tool for analyzing document collections in an unsupervised manner. We enhance the basic LDA model (Blei et al., 2003), by incorporating time variation in the topic content and introducing a hierarchy in the extracted topics. Firstly, we explain the operation of the basic topic model in Section 3.1. The subsequent sections extend the base model: time variation is introduced in Section 3.2, layering is discussed in Section 3.3, and tone adjustment is covered in Section 3.4. Section 3.5 outlines the Bayesian inference algorithm used to infer the model parameters.

3.1 Latent Dirichlet Allocation: Base model

The main idea of LDA is that each document is a part of a probability distribution over topics, and each topic is part of a probability distribution over words. The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. Importantly, LDA makes the implicit assumption that each word is generated from an underlying topic.

Define a document is a sequence of N words denoted by $\mathbf{d} = (w_{d1}, \ldots, w_{dn})$, where w_{dn} is the nth word in the sequence of document $\mathbf{d}$. A corpus is a collection of D documents denoted by $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$. Each document is composed of T topics. LDA assumes the following generative process for each document $\mathbf{d}$ in a corpus $\mathcal{D}$:

1. For each topic $t = 1, \ldots, T$,
   - Draw a distribution over words from a Dirichlet distribution with hyperparameter β, i.e: $\phi_t \sim \text{Dir}(\beta)$.
2. For each document, $\mathbf{d}$,
   - Draw a vector of topic proportions from a Dirichlet distribution with hyperparameter α, i.e: $\theta_d \sim \text{Dir}(\alpha)$.
   - For each word w_{dn}:
     (a) Draw a topic assignment x_{dn} for word $w_{dn} \sim \text{Mult}(\theta_d)$, $x_{dn} \in \{1, \ldots, T\}$ and Mult(.) is a multinomial distribution.
     (b) Draw a word w_{dn} from the $\sim \text{Mult}(\phi_t)$, where t is the drawn topic assignment in the previous step.

The generative probabilistic process with repeated sampling described above, can be conveniently illustrated using plate notation. This graphical notation, depicted in Figure 1, uses shaded variables to indicate observed variables, in our case, the words (w). Latent, or unobserved, variables are unshaded. Arrows represent conditional dependencies between variables, while plates (the boxes in Figure 1) denote repetitions of sampling steps, with the variables in the lower right corner indicating the number of samples. For instance, the inner plate over topic x and w illustrates the repeated sampling of topics and words until N words have been generated for document $\mathbf{d}$. The plate surrounding θ represents sampling over topics for each document $\mathbf{d}$, totalling D documents. The plate surrounding

ϕ_t demonstrates the repeated sampling of word distributions for each topic until T topics have been generated. Hyperparameters α and β determine the shape of the Dirichlet distribution. All subscripts for a variable on a plate carry over to all variables on that plate. For example, w in Figure 1 equals w_{dn} because it is located within the plates D and N.

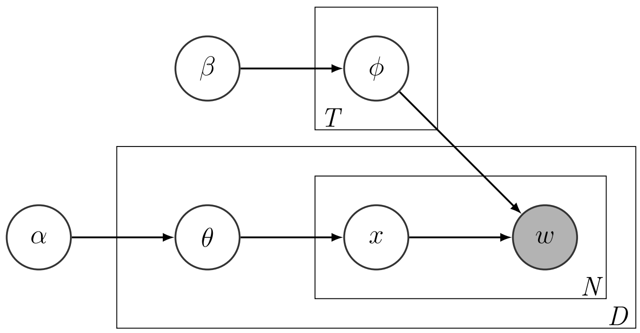


Figure 1: The graphical model for the topic model using plate notation

3.2 Extension 1: Time variation

The first extension we propose to the base model above is the incorporation of time variation. In our 'time-varying' topic model, we use the estimated topic-word distribution for a specific time slice as the starting point for estimating the topic-word distribution in the subsequent time slice. This approach is designed to address a limitation of standard topic models in real-time forecasting competitions, where topic models are typically estimated over the entire sample period. Many recent studies do not use a time-varying approach. By doing so, these studies implicitly assume that the documents originate from the same set of topics, which are time-invariant. Therefore, not considering time slices and estimating a single topic model over the entire time period can lead to a incorrectly specified topic model. For instance, with the word 'virus', this could result in an overestimation of its significance in the pre-COVID virus era and an underestimation during and after 2020. However, for our newspaper corpus, the sequence of the documents represents a changing set of topics. That is, the content of topics in 2022 will differ from those in 1985.

Figure 2 illustrates the generative probabilistic process with repeated sampling. Dashed arrows indicate that the variable is used in a subsequent time slice. Note, that the time slice subscript for a variable on a plate carries over to all variables on that plate, as shown in Figure 1. For instance, the variable w in the plate above ϕ_1 equals w_{d_1n} because it is situated within the plates D_1, indicating the first time slice, and N, indicating that the repeated sampling process is performed for each word in each document in time slice 1. Our topic model's dynamic modeling is related to the class of dynamic topic models (e.g. Blei and Lafferty, 2006 and Bittermann and Rieger, 2022). The primary distinction with dynamic topic models is that we do not make any explicit assumptions about the dynamics of the topic-word distributions and re-estimate the model for each time slice, rather than a single estimation with time-varying word distributions within topics. Our method is more flexible, and can readily accommodate shifts in vocabulary. Additionally, we estimate overlapping time slices instead of non-overlapping adjacent time slices, and we explicitly use the estimates of the Gibbs sampler in time slice $t-1$ as starting values in the Gibbs sampler in time slice t, resulting in more stable topics.

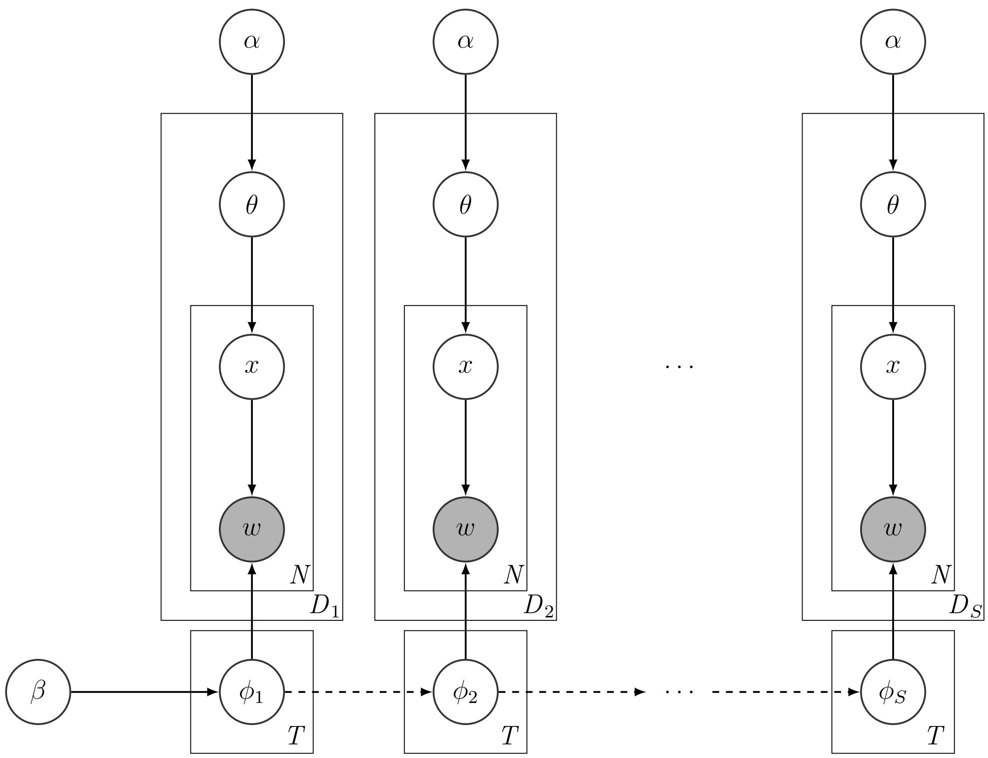


Figure 2: The graphical model for the time-varying topic model using plate notation

3.3 Extension 2: Hierarchy

In addition to time variation, our model also incorporates three layers. For clarity, we begin our explanation with the basic, non-time-varying LDA model. We then propose that the estimated topics can be further subdivided into more detailed topics.

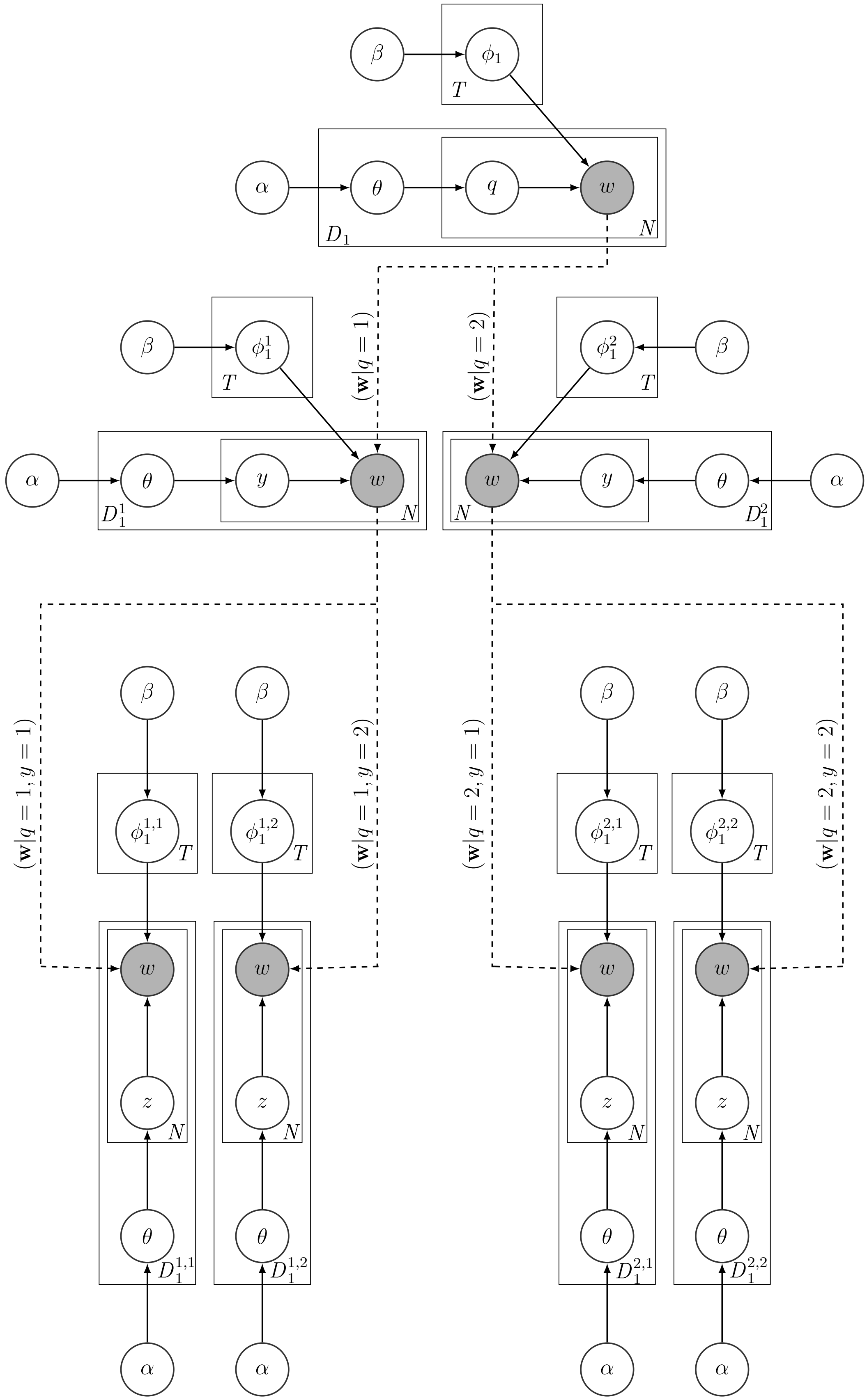


Figure 3: Plate notation for a stylized layered topic model with three layers

The layers in the topic model are denoted as follows: $D^{\text{layer1...,current layer}}$. For instance, $D^{1,2}$ represents all words in documents assigned to topic 1 in the first layer and topic 2 in the second layer. To keep our notation parsimonious, we denote the topic assignments in layer 1, 2 and 3 as q, y and z, respectively. Figure 3 provides a stylized illustration of our layered topic model with three layers, each extracting two topics. In the second layer, we estimate separate LDA models based on a subset of words in articles assigned to topic 1 and 2 in the first layer. This is denoted by the dashed arrows labeled with $(\mathbf{w}|q = 1)$ and $(\mathbf{w}|q = 2)$.

For illustrative purposes, we label the topic in the first layer as 'economics' and 'politics'. Our layered topic model aims to further refine these topics into two distinct subtopics within both 'economics' and 'politics'. We name these subtopics as 'indicators' $(\mathbf{w}|q = 1, y = 1)$ and 'Euro Area' $(\mathbf{w}|q = 1, y = 2)$ within economics, and 'Parliament' $(\mathbf{w}|q = 2, y = 1)$ and 'social partners', $(\mathbf{w}|q = 2, y = 2)$ within politics. The third and final layer of our layered topic model divides these four topics into 8 topics, i.e: $(\mathbf{w}|q = 1, y = 1, z = 1, 2)$, $(\mathbf{w}|q = 1, y = 2, z = 1, 2)$, $(\mathbf{w}|q = 2, y = 1, z = 1, 2)$ and $(\mathbf{w}|q = 2, y = 2, z = 1, 2)$. Note that the time slice subscript and the superscript for the layers for a variable on a plate are carried over to all variables on that plate, as shown in Figure 1 and Figure 2. The layering approach we propose differs from the canonical hierarchical topic model of Griffiths et al. (2003), where the hierarchy stems from the correlation between topics. In our model, we explicitly enforce hierarchy and estimate the model in separate layers.

3.4 Extension 3: Tone-adjustment

Using a measure of sentiment, we can 'tone-adjust' the topics in our model. This allows us to reallocate the headline sentiment of the newspaper for a specific day to the identified topics. As suggested by Thorsrud (2020), we start from article-level sentiment and aggregate it to headline sentiment. We proceed as follows. Initially, we compute the sentiment for each article by summing the sentiment scores of the words and dividing by the total number of words in the article. We explored several variations of weighting our newspaper sentiment score, such as weighting by the number of sentiment words per article or weighting by both the total number of words and sentiment words. Overall, the results were comparable to our chosen measure, but our measure resulted in a better fit with the year-on-year GDP growth, our measure for the business cycle. Results are available upon request from the authors. For a comprehensive treatment on the measurement of sentiment, see Algaba et al. (2020).

Next, we assign sentiment to topics according to the estimated θ (topic weight) per article for all third-layer topics. This yields a sentiment score for each third-layer topic per article and is calculated as article sentiment $\times$ topic weight. Next, these articles scores are aggregated across all articles per day to obtain a sentiment score per third-layer topic per day. Similarly, we can construct tone-adjusted topics in the second and first layer per day. From the topic-adjusted sentiment in the first layer, we can construct daily headline sentiment. Note that the calculated headline sentiment in this aggregation procedure is equivalent to calculating headline sentiment by using all articles without using topic proportions. This is the result of the topic weights in each layer summing to 1.

We employ a dictionary-based technique to construct our measure of newspaper sentiment, following the methodology in Tetlock (2007). The basis of our sentiment measure is the sentiment list introduced by Loughran and McDonald (2011). We translate this list to Dutch and supplemented it with words and collocations that are specific to the Dutch language. The augmented Loughran and McDonald dictionary contains $1,672$ words and collocations. Our list of Dutch sentiment words is publicly available, and can be downloaded here.

The integration of the techniques discussed in the previous sections, including time-variation, layering, and tone-adjustment, results in our tone-adjusted time-varying layered topic model. In other words, we extend the time-varying aspect to all layers of the model. Briefly, the estimation of successive slices is equivalent to the median of the estimated posterior topic-word distribution from the preceding time slice. In our base model, we estimate three layers, which is one more layer than the model depicted in the stylized example above and in Figure 3.

3.5 Bayesian inference

At its core, the topic model we propose remains a traditional LDA. The central concept of our approach is to re-estimate this LDA model for each layer and time slice separately, utilizing the results of the previous layer and time slice as a prior for the Gibbs sampler. Our inference procedure is as follows: For the first layer of the first time slice, spanning from January 1^{st} 1985 up until January 1^{st} 2000, we estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ using the collapsed Gibbs sampling procedure. This procedure is explained in more detail in the Appendix. We construct newspaper time slices with a monthly rolling approach as described in Section 3.2 and re-estimate the topic model each month. In total, we have 253 time slices. We experimented with smaller time windows, i.e. 5 and 10 year but found the shorter time windows entailed a costs in terms of less interpretable and more volatile topics.[4] For the first time slice, we use random initialization of the Gibbs sampler. For all subsequent time slices, we use the posterior estimate of $\boldsymbol{\phi}$ of the previous time slice to initialize the count matrix in the Gibbs sampler.

The aim of this 'chain' of count initialization is to stabilize the topics, i.e. In other words, the topics we extract are likely to maintain the same order in each time slice, resulting in a more stable topic-word distribution. To illustrate, let's say we extract two topics. After $4,000$ Gibbs iterations, the word 'inflation' has a count of 500 in topic 1 and 10 in topic 2. If we use this estimate as our final estimate of the topic-word distribution in slice t, and use the posterior estimate of $\boldsymbol{\phi}$ from the first time slice to initialize the count matrix in the Gibbs sampling algorithm in time slice $t+1$, the count of the word 'inflation' in topic 2 would need to increase significantly to become more prominent in topic 2 than in topic 1. This would only happen if the data strongly indicates that the word 'inflation' should be shifted to topic 2. Our approach differs from random initialization, where the word 'inflation' is randomly assigned to topic 1 and 2 in the first iteration, without considering the counts in the first time slice. We verify the stability of the topics over time by calculating the cosine distance of the topics. For more information on the calculation of this measure, see e.g. Aletras and Stevenson (2014). A score of 1 (0) indicates to topics are very similar (dissimilar). In our application, the calculated cosine distance for the same topic numbers in consecutive time slices consistently exceeds 0.95. Conversely, the cosine distances with other topics are always smaller.

After estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ for each time slice in the first layer of the topic model, we

[4] The results of the estimations with these shorter time windows are available upon request with the authors.

assign each word $\boldsymbol{w}$ to a specific topic assigned based on $\boldsymbol{\phi}$. Next, we assign every word that is assigned to topic 1 to the first 'branch' in the second layer (see Figure 3). For the third layer, we further divide into branches based on the 16 topic assignments in the second layer. We then estimate 4 topics per branch, following the same procedure as previously described. In total, we infer three topic layers, with 4 topics in the first layer, 16 topics in the second layer, and 64 topics in the third layer. The first layer indicates the general topic of the article, the second layer provides more details about the topic, and the third layer offers the most granular information.

We follow the same process for topics $2-4$, assigning the words belonging to these topics at an article level to 'branch' $2-4$. For each branch, we estimate four plain vanilla LDA's, one for the words of every topic in the first layer.

Finally, we attach sentiment to all articles and topics by calculating the sentiment score for each of the original article texts, as described in Section 3.4.

4 Nowcasting model and forecast design

To assess whether the sentiment indicators extracted from the newspaper contribute to short-term forecasts of GDP growth, we compare nowcasting models both with and without the inclusion of the extracted tone-adjusted topics. Our baseline model is a dynamic factor model (DFM), a commonly used forecasting model for many central banks and policymakers (e.g. Bańbura et al., 2013 and Jansen et al., 2016). The baseline model does not incorporate newspaper sentiment. The model is detailed in Section 4.2. Section 4.1 outlines the specifics of our forecast design to ascertain the added value of our newspaper indicators.

4.1 Forecast design

We estimate a DFM based on the specification in Bańbura et al. (2013), using a dataset of 58 monthly economic indicators. These indicators are categorized into four distinct groups: production & sales, surveys, financial indicators, prices, and indicators of the Netherlands' main trading partners. The data is further enriched with time series of topics derived from the tone-adjusted time-varying layered topic model described in Section 3. We transform the daily estimated topic models into a monthly format by extracting the sentiment per topic per day and averaging it over a month. This facilitates comparison with a model that only includes monthly economic indicators. Newspapers offer the advantage of higher frequency data, but this comes with increased volatility.

We construct a sequence of eight forecasts for GDP growth in a given quarter, obtained in consecutive months. Table 1 explains the timing of the forecasting exercise, using the forecast for the second quarter as an example. The first forecast is made on January 1^{st} using the monthly macro-economic series and the news topics available at that time. This forecast is referred to as the one-quarter-ahead forecast in month one. We then produce a monthly forecast for the next seven months, up to August. The final forecast is made on August 1^{st}, two weeks before the first release of GDP for the second quarter.[5] Following the conventional terminology, *forecasts* refer to (one-quarter) ahead forecasts, *nowcasts* refer to current quarter forecasts and *backcasts* refer to forecasts for the preceding quarter, as long as official GDP Figures have not be released.

Table 1: Timing of forecast exercise for a second quarter

Nr.	Forecast type	Month	Forecast made on the 1^{st} of
1	Forecast	1	January
2		2	February
3		3	March
4	Nowcast	1	April
5		2	May
6		3	June
7	Backcast	1	July
8		2	August

Our forecast design involves creating six sequential forecasts of the DFM for real GDP growth for each quarter in the period from 2003Q3–2020Q3. The estimation period starts in 1996M1, meaning that no monthly data prior to 1996M1 are used in the model's estimation. All monthly indicators were downloaded on February 1^{st} 2022. We begin evaluating the forecast errors of the model in 2003M9, using an expanding window for the model's estimation. This implies that the final backcast for 2020Q3 is on November 1^{st}. To clean the data for outliers, we set variables at the lower/upper bound of the distribution of the seasonally adjusted month-on-month growth rates for each variable. These bounds are

[5] Statistics Netherlands publishes the first estimate of GDP growth approximately 45 days after the end of a quarter.

defined as 3 $\times$ the interquartile range above/below the $1^{st}/3^{rd}$ quartile. If values fall below/above the lower/upper bound, we adjust these values to match the lower/upper bound growth rate.

4.2 Dynamic factor model

In practical terms, leveraging auxiliary information for short-term real GDP forecasting presents several challenges. The first is the 'curse of dimensionality', which arises from the large size of the information set. There are numerous potential variables for forecasting GDP growth. The data used in empirical literature can vary significantly in size, ranging from a few variables to over 300. Additionally, the limited length of the time series can lead to overparameterization. The second challenge is the 'ragged edge' problem, which arises from the fact that indicator variables are observed on a monthly basis, while GDP is observed on a quarterly basis. Furthermore, publication lags can vary due to different release dates.

DFMs address the 'curse of dimensionality' by summarizing the information from a potentially large dataset into a limited number of factors. The dynamic behavior is specified as a VAR process. Another key feature of the model is the use of the Kalman filter, which efficiently handles the unbalanced nature of the dataset and can manage differences in frequency. The Kalman filter replaces any missing monthly indicator observations with optimal predictions and also generates estimates of unobserved monthly real GDP, subject to a temporal aggregation constraint for the quarterly observation. Jansen et al. (2016) find in their comparative multi-country study that the DFM has the highest forecast accuracy on average, particularly for nowcasting and backcasting. In this paper, we use the DFM version proposed by Bańbura and Rünstler (2011) due to its relatively high forecast accuracy.

As demonstrated by Koopman and Harvey (2003), Bańbura and Rünstler (2011), and Rünstler (2016), it is relatively straightforward to derive the variable importance, or weights, for each variable in the dynamic factor model using the formulas in the state space representation of the dynamic factor model. In Section 6.1, we will report the average weights, or variable importance, of all variables in our dynamic factor model for each of the forecasting horizons we consider.

The Appendix provides detailed information on the DFM model equations, the state space representation of the DFM, and the impact on forecast accuracy of various modeling choices.

5 Outcome topic model

5.1 Topics and their interpretation

The primary topic model employed in our paper is the time-varying layered topic model, which is elaborated in Section 3. The determination of the optimal number of topics is a critical aspect of using topic models. Our main topic model consists of 64 topics. This is organized as follows: the first layer comprises 4 topics, each of which is further divided into 4 topics, leading to a total of 16 topics in the second layer. Each of these 16 topics is then subdivided into 4 topics, culminating in a third layer with 64 topics. Our decision to use 64 topics in our model is based on two considerations. Firstly, we conducted a series of statistical tests that are commonly used in standard topic models to determine the number of topics. The results of these tests are provided in the Appendix. While the tests did not provide definitive conclusions, it seemed that the use of 64 topics had a minimal impact on the explanatory power of the LDA. Additionally, our observations indicated that using more than 64 topics led to highly event-specific topics, suggesting a potential overfitting. Conversely, using significantly fewer than 64 topics resulted in overly broad topics. When applying layered and dynamic topic models, the specifics of topical layering, such as the number of layers, are not known beforehand. However, our decision to use 4 topics in the first layer was primarily influenced by the organizational structure of Het Financieele Dagblad. We were informed that the editorial team was divided into four groups, each focusing on a theme that coincidentally aligned significantly with the results of an time-varying layered topic model with 4 topics in the first layer. In conclusion, our choice of the number of topics and layers in our model is a combination of statistical criteria, interpretability, and insights into the organization of the newspaper's editorial staff. We leave a more detailed investigation of the optimal number of topics and layers or the 'depth' of the topic model for future research.

Table 2 shows how we label the extracted topics. In the first layer, we distinguish four topics, 'financial markets', 'firms', 'economics', and 'politics'. To illustrate further layering, consider the first-layer topic 'financial markets'. This topic can be broken down in four second-layer topics, i.e.: 'markets', 'financials', 'news', and 'financial indices'. These second-layer topics are each divided in four third-layer topics. For example, the 'financials' topic is divided in 'corporate finance', 'financials', 'banks', and 'insurance'.

Table 2: Names of topics in the three different layers of the topic model.

Layer 1	Financial Markets	Firms	Economics	Politics
Layer 2	Markets	Infrastructure	Elections	Parliament
Layer3	Raw materials	Chemical & pharma	Elections	Politics
	Exchanges	Indices	Easten Europe	Budgettary policy
	International	Mobility	Africa & Asia	Cabinets
	Monetary policy	Company results	United States	Ministries
Layer 2	Financials	Multinationals	Indicators	National
Layer3	Corporate finance	Telecom	International	Justice
	Financials (international)	Customers	Europe	Pensions & health care
	Banks (national)	Big tech	Trading partners	Supervision
	Insurance companies	Media	Fiscal policy	Education & research
Layer 2	News	Construction & Energy	Raw Materials	Lower Government
Layer3	Emissions	Construction	Asia	Housing
	Take-overs	Logistics	Oil & gas	Public-private
	Trade	Energy	Conflicts	Agriculture & fishery
	Insurers	Industry	Emerging economies	Transport
Layer 2	Fin. Indices	Demography	European Union	Social Partners
Layer3	Stock market	Retail	Germany	Wage negotiations
	Euronext	Bankruptcies	European Union	Labor market
	Analists	Listed	Italy & Spain	Entrepeneurs
	Results	International	France	Social security & pensions

In LDA the topics need to be labeled by the researcher. We employ the measure of a term's *relevance*, as introduced by Sievert and Shirley (2014), to interpret the topics. The relevance is calculated by weighing the most probable term in a topic, using the topic-word distribution $\boldsymbol{\phi}$ and the *lift*. The lift is defined as the ratio of a term's probability within a topic to its marginal probability across the corpus (Taddy, 2012).[6]

It is important to note that our approach uses a unified vocabulary across all time slices, incorporating significant words from various periods. The term 'COVID', for instance, was understandably prevalent in 2020 newspapers, but was seldom used in prior years. This does not hinder the inference of model coefficients in our framework, as words absent from a time slice are assigned an insignificant weight. This is evident when examining the collapsed

[6] We set the weight for $\boldsymbol{\phi}$ to 0.4 and the weight for the lift at $(1 - 0.6)$.

Gibbs sampling algorithm. Specifically, the weight assigned to a non-present word equals the smoothing parameter β, divided by the sum of the total number of words in each topic and the total number of words in all documents. Consequently, the collapsed Gibbs sampling algorithm will automatically adapt to the increased usage of the word 'COVID' if its count rises.

Figures 4 and 5 show the evolution of word rankings within topics over time, presenting the top 20 words for two of the four topics in the first layer of the topic model, namely financial markets and firms. These figures display the words of highest relevance in the initial time slice (panel A) and the final time slice (panel B), arranged in descending order of relevance. The light blue bar represents the overall term frequency, while the dark blue bar indicates the estimated term frequency within the topic.[7] The figures provide several interesting insights. Firstly, the topics can be relatively straightforwardly labeled based on the relevance rankings. For example, the topic of financial markets (Figure 4) is characterized by the high relevance of words such as 'stock', bank', 'stock exchange', and 'stock price'.

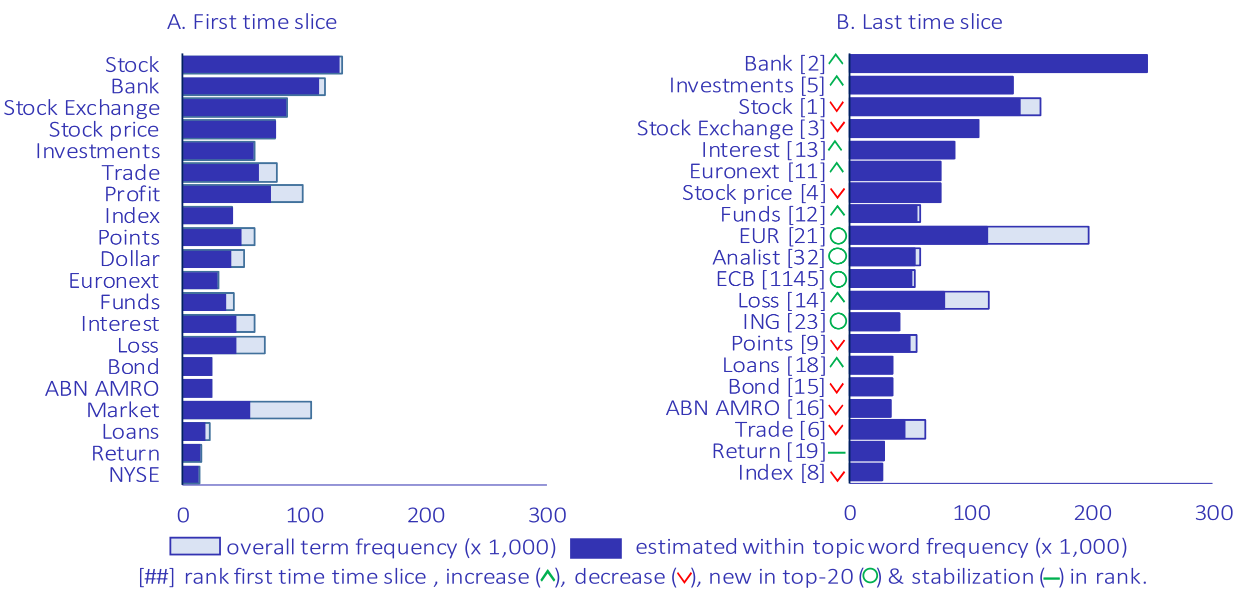


Figure 4: Top-20 words with highest relevance within financial markets topic, first and last time slice.

[7] The figures for the topics of economics and politics are included in the Appendix for the sake of brevity in the main text.

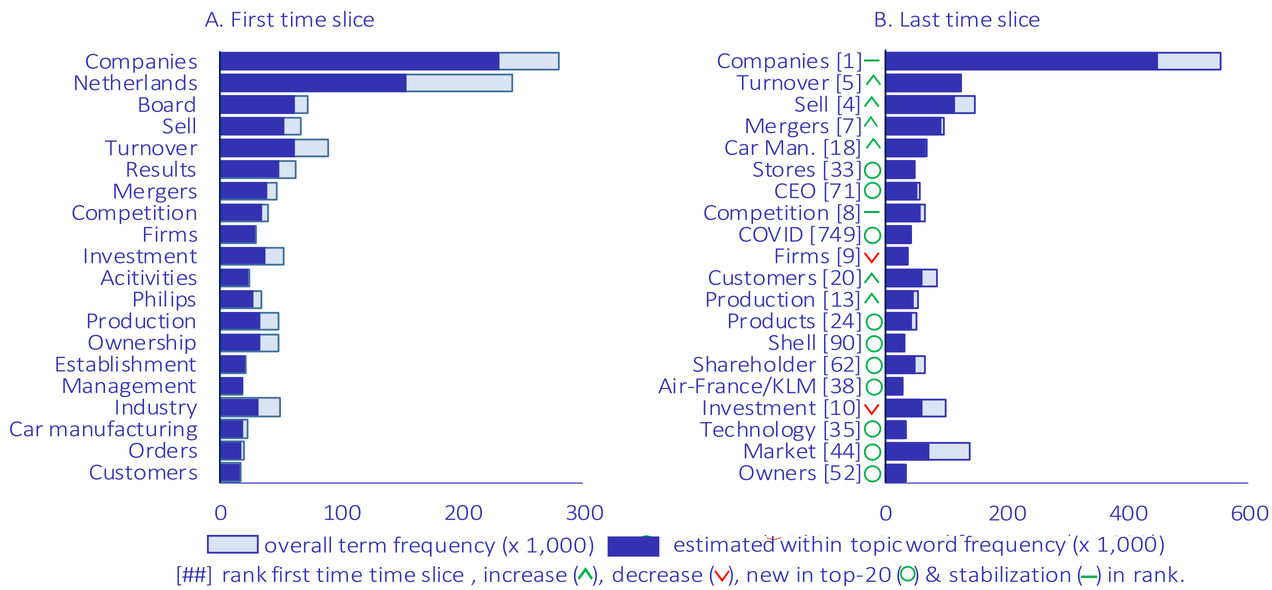


Figure 5: Top-20 words with highest relevance within firms topic, first and last time slice.

Secondly, there is a significant shift in the relevance of words within topics between the first and last time slices. This is evident from the red arrows, green arrows, and the green circles. These denote a decrease, increase, and new entry in the top 20 ranking, respectively. For example, within the financial markets topic, the terms 'ECB' and 'Euro' have gained prominence. This makes intuitive sense, given their increased media visibility following the introduction of the euro in 2002. Similarly, within the firms topic, the term 'COVID' has seen a significant rise in relevance, moving from position 749 to position 8.

5.2 Tone-adjusted topics

We assign sentiment scores to articles using dictionary techniques, as described in Section 3.4. The average sentiment across the entire sample is depicted in Figure 6.

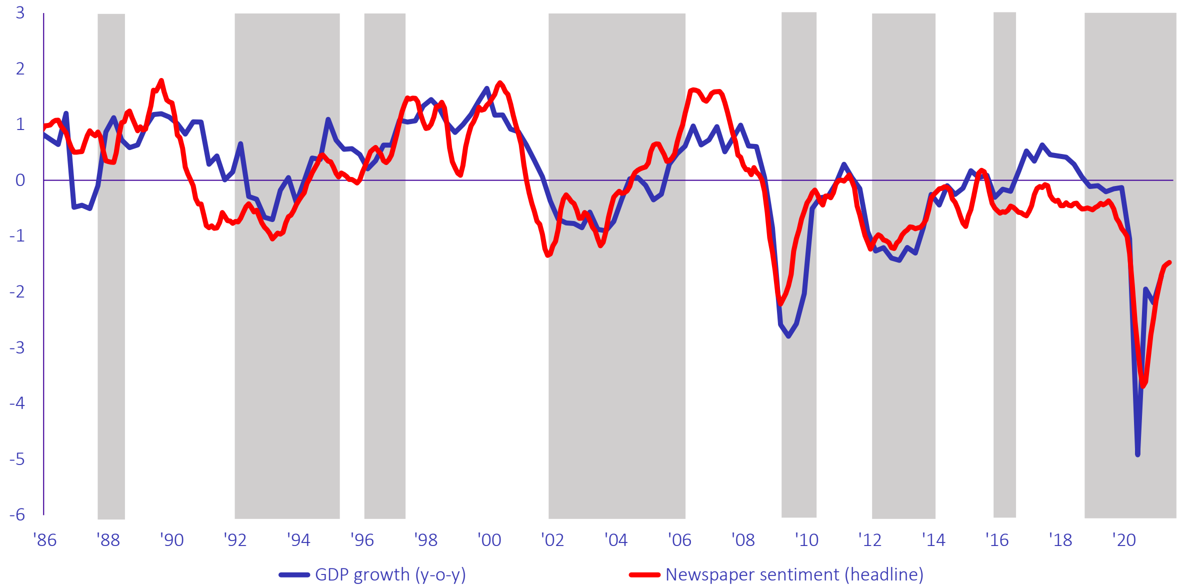


Figure 6: Normalized newspaper sentiment (headline, 6 month moving average) and normalized year-on-year GDP growth. Shaded areas indicate recessions.

Given the considerable noise in the daily and monthly sentiment series, we compute a moving average to discern the underlying trend in newspaper sentiment. Here, we present the trailing 6-month moving average. Figure 6 also incorporates our measure of the business cycle, which is the year-on-year growth of GDP. The shaded areas denote recessions, as defined by the reference turning points obtained from the OECD. The co-movement between the sentiment indicator and GDP growth is clearly visible. The correlation between the two series is high, i.e., 0.79. Moreover, sentiment becomes more negative when the economy is in a downturn. This suggests that sentiment derived from financial news could be a valuable indicator for tracking the business cycle, even without involved analysis.

The allocation of headline sentiment to topics can provide deeper insights into the reasons behind sentiment shifts, as illustrated in Figure 7. We divide the headline newspaper sentiment into the topics identified in the first layer, namely financial markets, firms, economics, and politics. Some intriguing patterns emerge. For instance, at the onset of the Global Financial Crisis in 2008, sentiment declines exclusively in the topic 'financial markets'. The declining sentiment manifests much later in the other topics, as the financial crisis intensifies. This contrasts sharply with the synchronous decline in all topics at the start of the COVID-19 crisis in 2020.

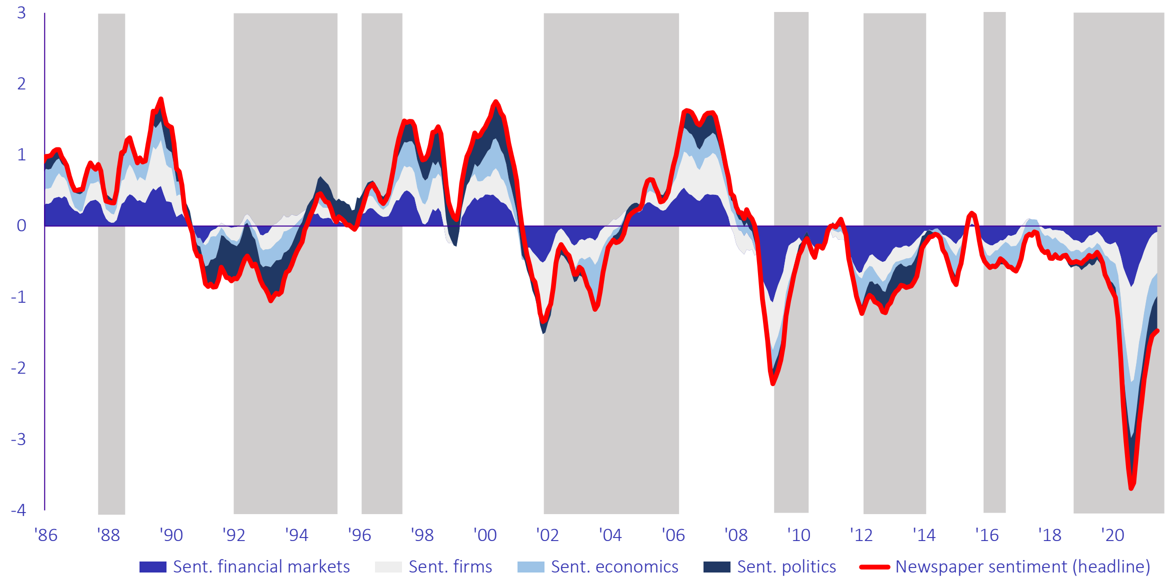


Figure 7: Normalized newspaper sentiment (headline, 6 month moving average) and by topic (first layer). Shaded areas indicate recessions.

6 Outcome nowcasting exercise

6.1 Forecasting accuracy news topics

Figure 8 shows the relative Cumulative Sum of Squared Forecast Errors (CSSFE) of the dynamic factor model, which is estimated on our dataset of monthly macroeconomic indicators and tone-adjusted time-varying news topics. When the relative CSSFE of the dynamic factor model falls below 1 (dotted blue horizontal line), it means that the CSSFE of the dynamic factor model is smaller than the CSSFE of the prevailing mean. An upward or downward trend in the relative CSSFE line implies a decaying or increasing forecasting advantage of the dynamic factor model, respectively. The relative CSSFE is computed using an expanding window, beginning with the first forecast error in 2003Q3. The figure presents the mid-quarter forecast accuracy on month 2 for three forecasting horizons, namely the backcast, nowcast, and forecasts. It also displays the relative CSSFE averaged over all forecasting horizons, i.e., from the one quarter ahead forecast up to the backcast in month 2. The relative CSSFE is plotted starting from 2006Q1 to smooth out the errors in the first three years of our forecast evaluation sample.

The relative average CSSFE of the dynamic factor model versus the prevailing mean over the total sample, represented by the red line in Figure 8, is 93%, i.e. the rightmost point of the red line. This finding emphasizes the significance of using monthly information when nowcasting GDP growth. The relative CSSFE is below one for all forecasting horizons. Figure 8 also supports previous findings in the literature (Bańbura et al., 2013 and Giannone et al., 2008) that conclude that the forecast accuracy of nowcasting models improves when more monthly information is available. As can be observed in Figure 8, the relative CSSFE decreases as more information becomes available, i.e., the relative CSFFE for the one quarter ahead forecast (blue solid line) is higher than the nowcast (blue dashed line) and the backcast (blue dotted line).

Interestingly, the average relative CSSFE measured is greatly influenced by the COVID period. In the pre-COVID period, the relative CSSFE averaged over all forecasting horizons is quite stable around 60%, but during the COVID period, the dynamic factor model is less accurate on average, as can be seen from the sharply increasing relative CSSFE. During the COVID period, the GDP growth rates in the Netherlands exhibited unprecedented volatility. The quarterly GDP growth rate in 2000Q1 was -8.9% quarter-on-quarter followed by a growth rate of 7.9% quarter on quarter in 2000Q1. For comparison, the average GDP growth rate over the period 2003Q1–2019Q4 was 0.3%. Apart from the COVID crisis the Global Financial Crisis in 2008/2009 also impacted the forecasting power of the dynamic factor model, but to a much lesser extent, as can be seen by the change in the forecasting accuracy in the first quarter of 2009Q1 when Dutch GDP growth declined by -3.7%. In contrast the COVID period, the forecast accuracy of the dynamic factor increased against the benchmark model for the mid-quarter backcast and nowcast, whilst deteriorating for the mid-quarter forecast.

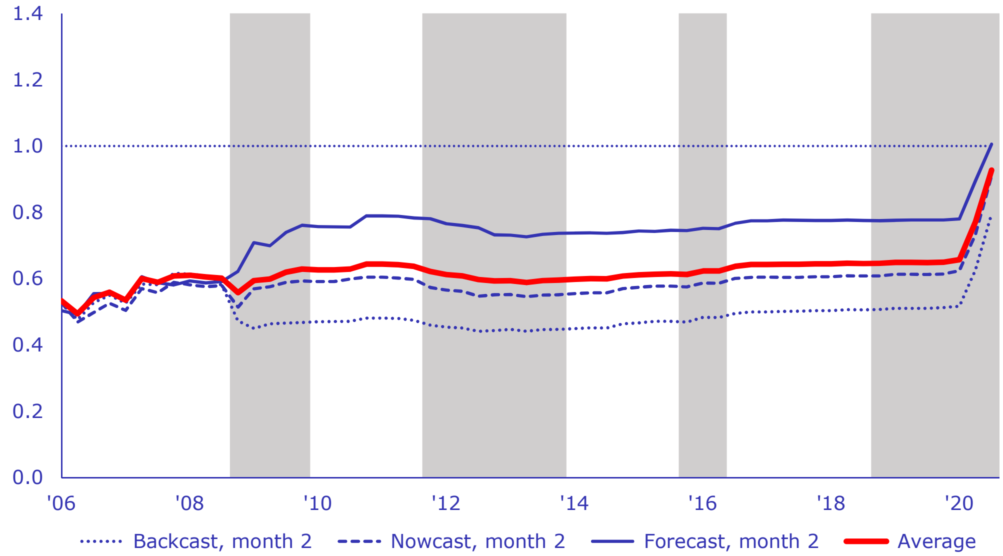


Figure 8: Relative cumulative sum of squared forecast errors: dynamic factor model versus prevailing mean, Shaded areas indicate recessions.

To formally ascertain the difference in forecast accuracy, Table 3 presents the Mean Squared Forecast Errors (MSFE) for each of the forecasting horizons we examined, ranging from the first quarter forecast in the first month to the backcast in the second month. To explore the substantial impact of excluding the extreme growth rates during the COVID-crisis and the Global Financial Crisis, we distinguish three periods: the total sample running from 2003Q3 to 2020Q3, the period excluding growth outliers in the COVID period, i.e., 2003Q3 to 2019Q4, and the period excluding the growth outliers and the Global Financial Crisis (2003Q3 to 2008Q4 and 2009Q2 to 2019Q4). These periods are indicated as total sample, excluding COVID outliers and excluding crisis outliers in Table 3. The MSFE of the benchmark model is shown in absolute terms, whilst the forecast accuracy of the dynamic factor model is expressed in terms of the MSFE of the benchmark model. Following, amongst others, Jansen et al. (2016) we present both a formal and informal measure to assess the observed differences in MSFEs. We roughly assess the economic importance of the gain by looking at the percentage difference in MSFE between two models. Bold-faced entries indicate that the MSFE of the DFM model including the tone-adjusted topics is at least 10% lower than the benchmark model. We conduct (one-sided) Diebold and Mariano (1995) (DM) tests as a formal test of statistical significance at the conventional levels (denoted by asterisks).[8] Non-starred, normal-type entries thus indicate models that

[8] The DM test broadly paints the same picture as the informal 10% improvement criterion, although

are equal in terms of forecast accuracy, both statistically and economically. We will follow the same two-way approach to statistical/economic significance in all tables that feature MSFEs in this paper.

Table 3: Forecast accuracy of dynamic factor model against benchmark model

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
(a) Prevailing mean (absolute MSFE)								
Total sample	2.47	2.47	2.47	2.50	2.50	2.50	2.51	2.51
No COVID outliers	0.56	0.56	0.56	0.58	0.58	0.58	0.59	0.59
No crisis outliers	0.30	0.30	0.30	0.31	0.31	0.31	0.32	0.32
(b) DFM: macro-economic indicators & news topics (relative MSFE)								
Total sample	**0.80**	**0.85**	**0.89**	0.91	0.96	1.03	1.01	0.98
No COVID outliers	**0.52****	**0.54****	**0.58****	**0.62****	**0.66****	**0.72****	**0.78***	**0.82***
No crisis outliers	**0.58*****	**0.60*****	**0.63*****	**0.65*****	**0.70****	**0.76****	**0.81***	**0.83***

Note: Prevailing mean (MSFE) refers to the MSFE of the prevailing mean of quarter-on-quarter GDP growth. DFM: macro-economic indicators & news topics (relative MSFE) refers to the relative (to the prevailing means) MSFE of a dynamic factor model estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics. The total sample spans from 2003Q1 to 2020Q3. The sample excluding COVID outliers ranges from 2003Q1 to 2019Q4, while the sample excluding crisis outliers covers the same period but excludes 2009Q1. Bold cells indicate that the MSFE is at least 10% lower than the baseline. Starred entries (*, **, ***) denote that the one-sided Diebold-Mariano test (where the alternative hypothesis is that the model is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

The results presented in Table 3 offer a more granular view than Figure 8. The table distinctly highlights the substantial impact of incorporating the GDP growth outliers during the COVID period. When measured over the total sample, the differences in forecast accuracy between the prevailing mean and the dynamic factor model are not statistically significant according to the DM-test, across all forecasting horizons. Our informal measure of economic significance suggests that all backcasts and the nowcasts in month 3 are, on average, more accurate than our benchmark model. However, when excluding the three quite extreme outliers during the COVID crisis, there is a much larger and statistically significant difference in forecast accuracy. The relative MSFE when nowcasting or backcasting is 0.30 percentage points lower. This reaffirms the uniqueness of the COVID period. The

the two do not always match. In some cases, large differences in accuracy are not statistically significant, while the reverse also happens. Jansen et al. (2016) note that the power of the DM test may be low due to the small number of observations. Moreover, the differences might signal that statistical significance and economic importance are different concepts.

impact of the Global Financial Crisis on forecast accuracy is considerably smaller. Consequently, we strongly believe that the weak forecast accuracy of the dynamic factor model during the unique and unexpected COVID period should not be the determining factor in our assessment of the forecasting quality of the dynamic factor model and tone-adjusted topics. This is especially true considering the dismal forecast accuracy of the benchmark model during this period.

6.2 Added value of forecasting with news topics

Overall, the outcomes in Figure 9 suggest that a dynamic factor model including news topics outperforms a dynamic factor model containing macro-economic indicators only. The average relative CSSFE (red line) indicates that the news topics derived contain unique information not captured by the hard economic monthly indicators, particularly when forecasting GDP growth in the current and next quarter, i.e. the red line is below 1 for almost the entire horizon. This outcome is in line with previous research (Ardia et al., 2019, Kalamara et al., 2022 and Ellingsen et al., 2022). In the period leading up to and including the Global Financial Crisis the relative forecast accuracy of the factor model including the news topics deteriorates (somewhat), and improves after the Global Financial Crisis after which the relative forecast accuracy stabilizes, before deteriorating during the COVID crisis.

Interestingly, the forecast accuracy of the mid-quarter backcast significantly deteriorates at the onset of the Global Financial Crisis. This result is entirely driven by the strong growth decline in 2009Q1 (-3.7%), where the model with news topics performed much worse than the model excluding the news topics. While the dynamic factor model with news topics performed worse during the quarter with the exceptionally large negative GDP growth during the strong growth decline in 2009Q1, the exact opposite is true for the large growth decline at the start of the COVID crisis. During this period, GDP growth declined by 8.9% quarter-on-quarter, and worse in the subsequent two quarters. This highlights the uniqueness of each of these crisis periods.

From these two crisis-episodes, it is challenging to conclude if news topics are beneficial for forecasting during these periods, as the outcomes contradict. Outside these crisis periods, both during recessions and periods of business cycle upturns, the forecast accuracy of the dynamic factor model with news topics is higher than the model without newspaper topics. This is a significant outcome, as it indicates that the information derived from newspapers consistently adds to the forecast accuracy of the forecasting model during normal times.

Furthermore, the lines indicate that the news topics increase the forecast accuracy when nowcasting and forecasting, but are less informative when there is more information on the forecasted quarter available when backcasting. In other words, the line showing the relative CSSFE of mid-quarter backcast lies above the lines showing the relative CSSFE for mid-quarter nowcasts and forecasts.

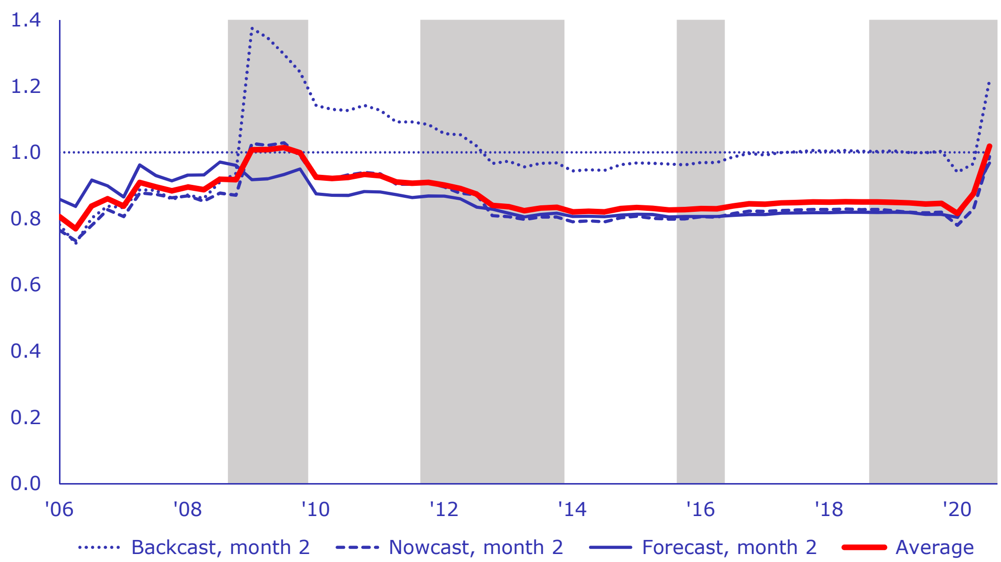


Figure 9: Relative cumulative sum of squared forecast errors: dynamic factor model monthly indicators versus dynamic factor model with monthly indicators and news topics, Shaded areas indicate recessions.

A more formal evaluation of the value added of the news topics is presented in Table 4. This evaluation compares the relative forecast accuracy of the dynamic factor model, which is estimated with hard monthly indicators both with and without the news topics, across all considered forecasting horizons. The results echo those from Figure 9: news topics enhance the forecast accuracy during nowcasting and forecasting, but provide less information during backcasting. Table 4 further indicates that, excluding the two unique crisis episodes, the forecasting advantage of the dynamic factor model with news topics is clearly visible, leading to an average decrease in the MSFE of 37% for nowcasting. This advantage is economically significant. Statistically, the DM-tests also indicate the difference is significant at at least the 5%-level.

Table 4: Forecast accuracy of dynamic factor model with and without tone-adjusted time-varying news topics

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
(a) DFM: macro-economic indicators (absolute MSFE)								
Total sample	1.61	1.83	2.13	2.31	2.47	2.65	2.61	2.55
No COVID outliers	0.31	0.35	0.41	0.46	0.50	0.53	0.57	0.57
No crisis outliers	0.25	0.27	0.3	0.32	0.33	0.34	0.35	0.33
(b) DFM: macro-economic indicators & news topics (relative MSFE)								
Total sample	1.22	1.14	1.03	0.99	0.97	0.97	0.97	0.97
No COVID outliers	0.94	**0.87**	**0.79****	**0.78****	**0.76*****	**0.78*****	**0.80*****	**0.86****
No crisis outliers	**0.72****	**0.68****	**0.64*****	**0.63*****	**0.65****	**0.70****	**0.74****	**0.80****

Note: DFM: macro-economic indicators (absolute MSFE) refers to the MSFE of the dynamic factor model estimated on a database of monthly macro-economic indicators. DFM: macro-economic indicators & news topics (relative MSFE) refers to the relative (to DFMs without news topics) MSFE of a dynamic factor model estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics. The total sample spans from 2003Q1 to 2020Q3. The sample excluding COVID outliers ranges from 2003Q1 to 2019Q4, while the sample excluding crisis outliers covers the same period but excludes 2009Q1. Bold cells indicate that the MSFE is at least 10% lower than the baseline. Starred entries (*, **, ***) denote that the one-sided Diebold-Mariano test (where the alternative hypothesis is that the model is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

To delve deeper into the contribution of news topics during nowcasting, Table 5 displays the (smoothed) average weights of the variables across different horizons, sorted by group. These weights are derived using the algorithm developed in Koopman and Harvey (2003), and applied to dynamic factor models in Bańbura and Rünstler (2011), Bańbura and Modugno (2014) and Rünstler (2016), among others. For a detailed explanation of the derivation of these weights, we refer to these papers. Table 5 presents the average weight per variable in the GDP forecasts over the total sample, excluding the outliers during the COVID period and the Global Financial Crisis. The average weights, which range from dark green to dark red, indicate the highest and lowest weight on that specific forecasting horizon, respectively.[9]

The table highlights several interesting findings. Firstly, news sentiment indicators

[9] Note that the weight of a variable differs from the average contribution to the forecast, which equals the average weight × average change of a variable. Table 5 presents the weights. For more details, see Bańbura and Rünstler (2011).

and financial market indicators have the highest weights overall. This can be seen from the (partly) green cells within these indicator groups. The weights of the indicators in the groups production & sales, prices and trade partners are relatively small when news sentiment and financial market indicators are included. Interestingly, and in line with intuition, macro-economic indicators that are traditionally seen as leading indicators (e.g. unemployment, bankruptcies, building permits, construction production), generally have a (much) lower weight for backcasting than for now- and forecasting.

Secondly, there is considerable variation in the weight of the variable groups within the group of financial market indicators. Within this group, the stock market indicators have the largest relative weight. The weights of the financial market indicators decline as the forecasting horizon shortens, with lower – but still relatively high compared to other indicators – weights of the stock market indices for nowcasting and forecasting. This reflects the fact that financial markets are forward looking.

Thirdly, within the news topics, financial markets and firms record the highest average weights. One of the advantages of our layered topic model is that we can assign the weights to detailed news topics. Within financial markets, we record the largest average weights for news on the development of financial indices and general news, and more specifically, stock market developments and news on takeovers. The high weight on financial market information aligns with the high weight on the financial market indices. The weights for news sentiment related to economics (e.g. elections and the European Union) and politics (e.g. parliamentary news and news on social partners) are relatively low on all forecasting horizons.

Fourthly, a minor yet discernible shift in the weight towards macroeconomic indicators on production & sales, prices, and the most important trading partners is observed when the time until the release of GDP decreases. The color of these indicators transitions from (dark) red during forecasting to (light) green during backcasting. This outcome aligns with previous findings on the weight of the relative importance of these so-called 'hard' indicators, as reported in Bańbura and Modugno, 2014 and Giannone et al., 2008.

Table 5: Description monthly database

Nr.	Variable name	Group	Backcast M2	Backcast M1	Nowcast M3	Nowcast M2	Nowcast M1	Forecast M3	Forecast M2	Forecast M1
1	Industrial production: total	Prod. & sales								
2	Industrial production: capital goods	Prod. & sales								
3	Industrial production: construction	Prod. & sales								
4	Industrial production: consumption goods	Prod. & sales								
5	Industrial production: durable consumption goods	Prod. & sales								
6	Industrial production: manufacturing	Prod. & sales								
7	Industrial production: non-durable consumption goods	Prod. & sales								
8	Household consumption: durable goods	Prod. & sales								
9	Household consumption: food, acohol and tobacco	Prod. & sales								
10	Household consumption: other goods	Prod. & sales								
11	Household consumption: service	Prod. & sales								
12	Building permits	Prod. & sales								
13	Imports	Prod. & sales								
14	Exports	Prod. & sales								
15	New commercial car registration	Prod. & sales								
16	New passenger car registration	Prod. & sales								
17	Retail trade turnover	Prod. & sales								
18	World Trade	Prod. & sales								
19	Unemployment rate	Prod. & sales								
20	Bankruptcies	Prod. & sales								
21	Loans to the private sector	Financial								
22	M1	Financial								
23	M3	Financial								
24	Interest rate: short-term	Financial								
25	Interest rate: long-term	Financial								
26	Interest rate: loans on mortgages, 5-10 year	Financial								
27	Amsterdam AEX-index	Financial								
28	Amsterdam AEX midkap-index	Financial								
29	Dow Jones Euro Stoxx 50-index	Financial								
30	Dow Jones Euro Stoxx basic materials-index	Financial								
31	Dow Jones Euro Stoxx financials-index	Financial								
32	Dow Jones Euro Stoxx technology-index	Financial								
33	Dow Jones Euro Stoxx healthcare-index	Financial								
34	Dow Jones Euro Stoxx industrials-index	Financial								
35	Dow Jones Euro Stoxx telecommunications-index	Financial								
36	Dow Jones Euro Stoxx utilities-index	Financial								
37	USD-EUR exchange rate	Financial								
38	Housing price index	Financial								
39	Consumerprice index: headline	Prices								
40	Consumerprice index: services	Prices								
41	World market commodity prices: total	Prices								
42	World market commodity prices: industrial materials	Prices								
43	World market commodity prices: agri/industrial prices	Prices								
44	World market commodity prices: metals	Prices								
45	World market commodity prices: energy	Prices								
46	Producer price: industry (domestic market)	Prices								
47	Producer price: industry (foreign market)	Prices								
48	Import prices	Prices								
49	Export prices	Prices								
50	Hourly wages	Prices								
51	Belgium: industrial production (excl. construction)	Trade partners								
52	France: industrial production (excl. construction)	Trade partners								
53	Germany: industrial production (excl. construction)	Trade partners								
54	Spain: industrial production (excl. construction)	Trade partners								
55	Italy: industrial production (excl. construction)	Trade partners								
56	Germany: retail trade	Trade partners								
57	Belgium: retail trade	Trade partners								
58	France: retail trade	Trade partners								
59	Markets: raw materials	News sentiment: fin. markets								
60	Markets: exchanges	News sentiment: fin. markets								
61	Markets: international	News sentiment: fin. markets								

continued on next page . . .

Table 5 – continued from previous page

Nr.	Variable name	Group	Backcast M2	Backcast M1	Nowcast M3	Nowcast M2	Nowcast M1	Forecast M3	Forecast M2	Forecast M1
62	Markets: monetary policy	News sentiment: fin. markets								
63	Financials: corporate finance	News sentiment: fin. markets								
64	Financials: financials	News sentiment: fin. markets								
65	Financials: banks	News sentiment: fin. markets								
66	Financials: insurance	News sentiment: fin. markets								
67	News: emissions	News sentiment: fin. markets								
68	News: takeovers	News sentiment: fin. markets								
69	News: trade	News sentiment: fin. markets								
70	News: insurers	News sentiment: fin. markets								
71	Financial indices: stock markets	News sentiment: fin. markets								
72	Financial indices: Euronext	News sentiment: fin. markets								
73	Financial indices: analists	News sentiment: fin. markets								
74	Financial indices: results	News sentiment: fin. markets								
75	Infrastructure: chemical & pharma	News sentiment: firms								
76	Infrastructure: indices	News sentiment: firms								
77	Infrastructure: mobility	News sentiment: firms								
78	Infrastructure: company results	News sentiment: firms								
79	Multinationals: telecom	News sentiment: firms								
80	Multinationals: customers	News sentiment: firms								
81	Multinationals: big-tech	News sentiment: firms								
82	Multinationals: media	News sentiment: firms								
83	Construction: construction	News sentiment: firms								
84	Construction: logistics	News sentiment: firms								
85	Construction: energy	News sentiment: firms								
86	Construction: industry	News sentiment: firms								
87	Demography: retail	News sentiment: firms								
88	Demography: bankruptcies	News sentiment: firms								
89	Demography: listed	News sentiment: firms								
90	Demography: international	News sentiment: firms								
91	Elections: elections	News sentiment: economics								
92	Elections: Eastern Europe	News sentiment: economics								
93	Elections: Africa & Asia	News sentiment: economics								
94	Elections: United States	News sentiment: economics								
95	Indicators: international	News sentiment: economics								
96	Indicators: Europe	News sentiment: economics								
97	Indicators: trade	News sentiment: economics								
98	Indicators: fiscal policy	News sentiment: economics								
99	Raw materials: Asia	News sentiment: economics								
100	Raw materials: oil & gas	News sentiment: economics								
101	Raw materials: conflicts	News sentiment: economics								
102	Raw materials: emerging markets	News sentiment: economics								
103	European union: Germany	News sentiment: economics								
104	European union: European Union	News sentiment: economics								
105	European union: Italy & Spain	News sentiment: economics								
106	European union: France	News sentiment: economics								
107	Parliament: politics	News sentiment: politics								
108	Parliament: budgetary policy	News sentiment: politics								
109	Parliament: cabinets	News sentiment: politics								
110	Parliament: ministries	News sentiment: politics								
111	National: justice	News sentiment: politics								
112	National: pensions & healthcare	News sentiment: politics								
113	National: supervision	News sentiment: politics								
114	National: education & research	News sentiment: politics								
115	Lower government: housing	News sentiment: politics								
116	Lower government: public-private	News sentiment: politics								
117	Lower government: agriculture & fishery	News sentiment: politics								
118	Lower government: transport	News sentiment: politics								
119	Social partners: wage negotiations	News sentiment: politics								
120	Social partners: labor market	News sentiment: politics								
121	Social partners: entrepeneurs	News sentiment: politics								
122	Social partners: social secutiry & pensions	News sentiment: politics								

Lowest weight — Highest weight

6.3 Robustness tests

The results presented so far are grounded in our base specification of the topic model, i.e. the sentiment adjusted news topics extracted from a time varying layered topic model. Table 6 offers a view of three robustness checks on the impact of: tone adjustment, the number of topics, and the structure of the topic model in terms of time variation and layering. Note that Table 6 shows the outcomes for the period excluding the outliers in the COVID period and the Global Financial Crisis. The results for the total sample and the sample excluding only the COVID outliers are available in the Appendix, as well as the relative CSSFE-lines.

In Panel (a), we compare the effect of using the proportion of news topics at each point in time, extracted from the topic model as in Bybee et al. (2020), versus using sentiment-adjusted news topics as in our base model, following the approach of Thorsrud (2020). We observe a significant enhancement in forecast accuracy when sentiment-adjusted topics are included instead of topic proportions. This improvement is particularly noticeable for horizons equal to or shorter than the 3rd month forecast, with the most substantial gain observed for the mid-quarter nowcast (M2), where the forecast accuracy sees an increase of 22%.

Panel (b) explores the influence on forecast accuracy when including different numbers of topics in our base model. We consider three scenarios: 64 topics, 16 topics, and 4 topics.[10]. We find that the number of topics included in the topic model plays a crucial role in the forecast accuracy. The bold entries across almost all forecasting horizons of the base model suggest that the inclusion of 64 topics from the 3^{rd} layer of the time-varying layered topic model yields more accurate forecasts on average than merely including the 4 topics from the 1^{st} layer. Including 64 topics also add to forecasting accuracy compared to 16 topics from the 2^{nd} layer. As expected, the added value is somewhat less than in the 64 vs 4 topic situation.

Finally, Panel (c) presents the outcomes of imposing layering and time-variation in the topic model. This panel shows that the imposition of layering and time variation in the

[10] The 64, 16 and 4 topics correspond to the topics in the third, second and first layer of our topic model, respectively.

Table 6: Forecast accuracy of several topic model variations, no crisis outliers

		Backcast		Nowcast			Forecast		
	MSFE	M2	M1	M3	M2	M1	M3	M2	M1
(a) Proportions versus sentiment adjusted									
Topic shares	absolute	0.20	0.22	0.24	0.26	0.27	0.28	0.28	0.28
Base model	relative	**0.87**	**0.82**	**0.78**	**0.78***	**0.81***	**0.86**	0.93	0.95
(b) Number of news topics									
4 topics	absolute	0.20	0.21	0.22	0.23	0.25	0.27	0.29	0.29
Base model	relative	**0.88**	**0.87***	**0.86***	**0.87***	**0.86***	**0.87***	**0.88***	0.92
16 topics	absolute	0.19	0.19	0.21	0.22	0.25	0.27	0.29	0.29
Base model	relative	0.95	0.94	0.92	0.91	**0.89***	**0.87****	**0.90***	0.92
(c) Time variation and layering									
TM	absolute	0.19	0.19	0.20	0.21	0.22	0.23	0.24	0.25
LTM	relative	0.98	0.98	0.98	0.98	0.99	0.98	0.98	1.00
TVTM	relative	0.99	0.99	0.99	1.00	1.01	1.00	0.99	0.99
TVLTM	relative	0.97	0.98	0.98	0.98	1.00	1.00	1.01	1.02

Note: Entries show the absolute and relative (to the absolute MSFEs in line above) MSFEs of dynamic factor models estimated on different datasets. Base model: dynamic factor model estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics, Topic shares: identical to base model but with topic shares instead of sentiment adjusted topics, 4 topics: identical to the base model but including only the 4 topics from the 1^{st} layer of the TVLTM, 16 topics: identical to the base model but including only the 16 topics from the 2^{nd} layer of the TVLTM, TM: identical to the base model but 64 topics from a plain vanilla topic model with fixed word-topic distribution after the first time slice, no layering in topics. LTM: identical to the TM but with three topic model layers. TVTM: identical to the base model, but 64 topics estimated without layering. Bold cells indicate the MSFE is at least 10% better than the baseline. Difference between small set of indicators and Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

topic model does not significantly enhance forecast accuracy.[11] This finding implies that for nowcasting purposes, estimating the more complicated layering and time-varying topic model may not be strictly necessary. The main added value of the layered and time-varying structure thus lies in the interpretation of the topics.

[11] In both the standard topic model (TM) and its layered variant (LTM), the initial time slice is used to estimate the topic-word distribution, which is subsequently held constant.

7 Conclusion

We investigate two related questions in this paper. The first question is whether newspaper sentiment can be used to assess the course of the business cycle. The second question is whether newspaper sentiment can be used to decrease the forecasting error of nowcasting models. To answer the first question, we combine a dictionary-based newspaper sentiment measure with a topic model that allows us to trace the sentiment by fine-grained time-varying and hierarchically ordered topics. Our tone-adjusted time-varying layered topic model is new in the literature. To examine the added value of including newspaper sentiment indicators in a formal nowcasting horse-race between a state-of-the-art nowcasting model using 58 monthly indicators versus a model using these indicators and 4 to 64, time-varying and layered, sentiment indicators derived from our topic model. Our main findings can be summarized in four points.

To begin, aggregate newspaper sentiment is a strong indicator of the business cycle. Sentiment correlates strongly with year-on-year GDP growth, and sentiment turns negative when the economy is in a downturn.

Second, combining sentiment with the topics from the topic model enables storytelling regarding sentiment movements. Moreover, the time-varying and layered character of the topics resulting from our time-varying layered topic model can provide policymakers and practitioners with valuable insights into the causes of sentiment swings.

Third, sentiment indicators derived from newspaper articles contain valuable information not embodied in traditional monthly indicators. The forecast accuracy of our DFM greatly improves when including the tone-adjusted topics derived from the newspaper articles, especially when now- and forecasting. The added value, however, of our layering and time-varying structure of the topic model is limited.

Fourth, we find that the number of extracted topics needs to be carefully analyzed before including the news topics in a nowcasting model. Specifically, we find that the forecast accuracy of a nowcasting model including 64 topics is higher than the accuracy of a nowcasting model including only 4 topics.

Our findings may be of interest to policymakers, financial analysts, and economic agents. We demonstrate that sentiment indicators derived from the financial press contain valuable additional information that can be extracted in real time. Future research could focus on developing selection criteria for the optimal number of layers and topics per layer in our topic model. Additionally, a deeper investigation of the real-time availability of the newspaper data could be an interesting avenue to explore. In our current analysis, we investigate the advantage of using newspaper data on one specific day in the month, i.e., the first day, using pseudo real-time vintages for the monthly economic indicators. Real-time vintages of the indicators could possibly better pinpoint when the newspaper data are most advantageous for nowcasting.

References

Aletras, N. and M. Stevenson (2014). Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 22–27. link.

Algaba, A., D. Ardia, K. Bluteau, S. Borms, and K. Boudt (2020). Econometrics meets sentiment: an overview of methodology and applications. *Journal of Economic Surveys 34*(3), 512–547. link.

Aprigliano, V., S. Emiliozzi, G. Guaitoli, A. Luciani, J. Marcucci, and L. Monteforte (2023). The power of text-based indicators in forecasting italian economic activity. *International Journal of Forecasting 39*(2), 791–808. link.

Ardia, D., K. Bluteau, and K. Boudt (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting 35*(4), 1370–1386. link.

Arun, R., V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi (Eds.), *Advances in knowledge discovery and data mining*, pp. 391–402. Springer Berlin Heidelberg. link.

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

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013). Now-casting and the real-time data flow. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part A, pp. 195–237. Elsevier. link.

Bańbura, M., D. Giannone, and L. Reichlin (2011). Nowcasting. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook on Economic Forecasting*, pp. 63–90. Oxford University Press. link.

Bańbura, M. and M. Modugno (2014). Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data. *Journal of Applied Econometrics 29*(1), 133–160. link.

Bańbura, M. and G. Rünstler (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting 27*(2), 333–346. link.

Barbaglia, L., S. Consolia, and S. Manzan (2022). Forecasting with economic news. *Journal of Business & Economic Statistics forthcoming*, 1–12. link.

Bittermann, A. and J. Rieger (2022). Finding scientific topics in continuously growing text corpora. In *Proceedings of the 3rd Workshop on Scholarly Document Processing*, pp. 7–18. link.

Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120. link.

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

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2020). The structure of economic news. NBER Working Papers 26648, National Bureau of Economic Research. link.

Cao, J., T. Xia, J. Lia, Y. Zhang, and ShengTanga (2009). A density-based method for adaptive LDA model selection. *Neurocomputing 72*(7), 1775–1781. link.

Churchill, R. and L. Singh (2022). The evolution of topic modeling. *ACM Computing Surveys 54*(10), 1–35. link.

Deveaud, R., E. Sanjuan, and P. Bellot (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique 17*(1), 61–84. link.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 134–144. link.

Ellingsen, J., V. H. Larsen, and L. A. Throrsrud (2022). News media vs. FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics 37*, 63–81. link.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74. link.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676. link.

Griffiths, T., M. Jordan, J. Tenenbaum, and D. Blei (2003). Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, Volume 16, pp. 17–24. MIT Press. link.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Science 101*, 5228–5235. link.

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

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal 42*(1), 177–196. link.

Jansen, W. J., X. Jin, and J. M. de Winter (2016). Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts. *International Journal of Forecasting 32*(2), 411–436. link.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics 37*(5), 896–919. link.

Koopman, S. J. and A. Harvey (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control 27*(7), 1317–1333. link.

Kuzin, V., M. Marcellino, and C. Schumacher (2013). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics 28*(3), 392–411. link.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10Ks. *Journal of Finance 66*(1), 35–65. link.

Porter, M. (1980). An algorithm for suffix stripping. *Program: Electronic library and information systems 14*(3), 130–137. link.

Rambaccussing, D. and A. Kwiatkowski (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting 36*(4), 1501–1506. link.

Resnik, P. and E. Hardisty (2009). Gibbs sampling for the uninitiated. mimeo, Institute for Advanced Computer Studies University of Maryland. link.

Rünstler, G. (2016). On the design of data sets for forecasting with dynamic factor models. In E. Hillebrand and S. J. Koopman (Eds.), *Dynamic Factor Models (Advances in Econetrics, Vol. 35)*, pp. 629–666. Emerald Group Publishing Limited. link.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2022). Measuring news sentiment. *Journal of Econometrics 228*(2), 221–243. link.

Sievert, C. and K. Shirley (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70. link.

Steyvers, M. and T. Griffiths (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*, pp. 424–440. Laurence Erlbaum. link.

Taddy, M. (2012). On estimation and selection for topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 1184–1193. link.

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

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics 38*(2), 393–409. link.

A Cleaning the database of newspaper articles

This section describes how we cleaned the raw database in more detail. Cleaning reduced the size of our article database from $1,093,477$ articles to $582,981$ articles, a reduction of approximately 47%. Section A.1 describes our procedure for removing irrelevant articles. Section A.2 describes how we remove stopwords, adjust for collocations and stem words and verbs.

A.1 Removal of irrelevant articles

First, we stripped the database of newspaper categories that were not relevant, such as fashion, radio and television pages, letters from readers, profiles of entrepreneurs, personal finance, advertorials, photo-pages, newspaper service pages, and announcements of events. This step reduces our database by $25,928$ articles to $1,067,549$.

Second, we used the publication URLs of the articles to eliminate irrelevant articles from the database. The URLs are composed of several parts, with the last part providing an abbreviated indication of the article title. After removing the last part of the URLs, we were left with $1,046$ unique URLs left. Each of these URL addresses contains information about the type of article. We manually checked and re-categorized all $1,046$ unique URLs into categories, provided the category was equal to or larger than 0.1% of all articles. Using this approach, we were able to categorize 72% of all articles. For the articles that had an informative URL, we grouped them into 11 categories, based on the tags provided by the FD, i.e., 1. company news, 2. economics & politics, 3. financial markets, 4. opinion-pages, 5. domestic news , 6. personal profiles, 7. foreign news, 8 human interest, 9. English pages and 10. short news, 11. archive. Category 9 might seem obscure for a Dutch financial newspaper, but until the mid 2000s, the FD contained an English back-page with the most important news of that day in English as a service to non-native readers. We deleted all articles in categories 4, 6, 8, 9 and 11, reducing our database by $438,550$ articles to $628,999$ articles.

Thirdly, we removed all articles with a title that clearly indicated the article was not relevant for our purpose. For example, articles that contained summaries of closing and opening prices for stock exchanges, articles containing agendas for upcoming events, company press releases, and overview summaries of newspaper content. This step reduced the database by 18,753 articles to 610,246 articles.

Fourthly, we eliminated articles that appeared more than once in the database. This can occur because articles are adjusted later on, or are first published on the website and later in the printed newspaper. We decided to remove these articles, and keep the article when it was first published (online). Deleting articles with the same title and the same publication date removed 23,080 articles from the database. Dropping articles with the same title and a one-day publication difference removed 1,978 articles from the database. A two-day publication delay is relatively rare, but we deleted these articles as well, removing another 538 articles. In total, this cleaning step reduced the number of articles in the database by 25,596 to 586,408.

Fifthly, we deleted all articles that had no content but did have a title. These articles can be broadly classified into the following groups: headlines appearing on the front-page of the FD, with the actual article in the database having a different article identifier, titles of infographics, one-line articles about changes in stock markets (e.g., ‘AHOLD -5.4%’), or one-liners indicating a recent release (e.g., ‘industrial production: $+2,8\%$’). This step reduced our database by 120 articles to 586,288.

Sixthly, we removed any remaining English articles. First, we identified articles that contained three of the most common English words, i.e., ‘the’, ‘and’ and ‘to’ . Next, we checked the language of these articles using the R-package `textcat` to verify the articles were in English. This procedure deleted another 3,307 articles from the database, and reduced the database to a total of 582,981 articles.

Finally, we converted all words to lowercase letters, stripped HTML codes from the text, and removed all punctuation and numbers from the texts. We retained the dot (‘.’) and the HTML-tag for sections ‘<\p>’ to identify the number of sections and sentences in an article.

A.2 Clean Content Articles

This Section outlines the process we used to clean the texts from the articles, which includes removing stopwords, checking for collocations, and stemming words and verbs.

A.2.1 Remove stopwords

We eliminated very common words, known as stopwords, such as 'the', 'a' and 'and', count words, (e.g., 'duizend', 'miljoen') and months and days of the week, which contribute little to understanding the meaning of an article. We compiled a list of these stopwords by combining the Dutch stopwords list in the R-package `snowballC` with stopwords that frequently appear in our corpus of articles. We do *not* delete stopwords that express sentiment, suck as 'nothing' or ' less', because these words are included in our sentiment list.

A.2.2 Collocations

A challenge with text analysis is that when two (or more) words naturally belong together, this is not automatically recognized by the topic model. To uncover the most relevant so-called *bi-grams*, we analyzed all single words and bi-grams with a minimum frequency of $4,000$ in the newspaper. In most cases, bi-grams relate to word combinations that do not have any special meaning like 'their customers', 'red car'. However, some bi- and tri-grams do have a special meaning, such as company names, like 'Royal Dutch Shell', 'London stock exchange", 'Thomas Cook', 'Standard & Poor's' and two or more words that have a different meaning when combined, e.g.: 'convertible stocks' , 'financial markets' , 'private equity' , 'Statistics Netherlands' , 'euro crisis', 'industrial production', 'current account', 'PMI index', 'interest rate' and 'Central Europe'.

A.2.3 Stem words

We convert all conjugate verbs and words to their stem. For example, the verbs 'is', 'be', 'are' are all reduced to the stem 'are'. This step reduces the number of unique words in the corpus, without loss of meaning of the words. Usually, this is done with a mechanical so-called Porter stemmer (Porter, 1980). We experimented with this stemming technique, but concluded that it yields unsatisfactory results in Dutch. Therefore, we follow another –more labor-intensive– process. We proceed as follows. Firstly, we stem all Dutch verbs using the verb list in the web-mining Python-module `Pattern` and augment the list with verbs that are specific to financial news. Our lists of conjugate and stemmed verbs contain

20,058 and 3,687 words, respectively. Secondly, we *manually* stem all nouns with a total frequency of 2,000 or higher. Furthermore, we check for synonyms, and replace words that are clearly synonyms, e.g. the Dutch language has two words for global: 'globaal' and 'mondiaal'. To keep our list of words parsimonious we combine these words into the word 'mondiaal'.

B Dataset of macro-economic indicators

Table A.1 provides details on the sources, availability, and transformations applied to the data series. The data can be categorized into four groups: production & sales, financial indicators, prices, and indicators of significant trading partners; see the headings in the table. Additionally, the table displays the transformation of the variables, variable names, data sources, link to the data series, as well as the start and end date for all variables in the dataset. Some of the data series are proprietary, and the link to the data series cannot be shared without prior permission from the data supplier. Table A.1 also indicates which indicators are included in the small database of monthly macro-economic indicators, which we use in Section D.2 to investigate the impact of using a large versus a small dataset in terms of forecast accuracy.

The available monthly data are typically already adjusted for seasonality (and calendar effects). Where necessary, raw data series are seasonally adjusted using the US Census X12 method. All monthly series are made stationary by differencing or log-differencing (in the case of trending data, such as industrial production, retail sales, and monetary aggregates). All variables are standardized by subtracting the mean and dividing by the standard deviation. This normalization is necessary to prevent overweighting series with large variances in the determination of common factors in our nowcasting model. The data transformations are consistent across all estimated models. In the nowcasting exercise, the 58 macro-economic indicators are combined with the appropriately transformed tone-adjusted topics extracted from the tone-adjusted time-varying layered topic model. The tone-adjusted topics are included in levels and normalized before inclusion in the nowcasting model.

Table A.1: Description monthly database

Nr.	Transform				Small	Variable name	Source	Code	Start	End
	log	dif.	fil.	sa.						
Production & sales										
1	1	1	3	1	1	Industrial production: total	CBS	link	Jan-65	Nov-20
2	1	1	3	1	0	Industrial production: capital goods	ECB	proprietary	Jan-70	Nov-20
3	1	1	3	0	1	Industrial production: construction	ECB	proprietary	Jan-85	Nov-20
4	1	1	3	1	0	Industrial production: consumption goods	ECB	proprietary	Jan-90	Nov-20
5	1	1	3	0	1	Industrial production: durable consumption goods	ECB	proprietary	Jan-90	Nov-20
6	1	1	3	0	0	Industrial production: manufacturing	ECB	proprietary	Dec-79	Nov-20
7	1	1	3	0	1	Industrial production: non-durable consumption goods	ECB	proprietary	Jan-90	Nov-20
8	1	1	3	1	1	Household consumption: durable goods	CBS	link	Jan-95	Nov-20
9	1	1	3	1	0	Household consumption: food, acohol and tobacco	CBS	link	Jan-95	Nov-20
10	1	1	3	1	1	Household consumption: other goods	CBS	link	Jan-95	Nov-20
11	1	1	3	1	1	Household consumption: service	CBS	link	Jan-95	Nov-20
12	1	1	3	1	1	Building permits	CBS	link	Jan-95	Nov-20
13	1	1	3	1	1	Imports	CBS	link	Jan-90	Nov-20
14	1	1	3	1	1	Exports	CBS	link	Jan-90	Nov-20
15	1	1	3	0	1	New commercial car registration	ECB	proprietary	Jan-90	Dec-20
16	1	1	3	1	1	New passenger car registration	ECB	proprietary	Jan-65	Dec-20
17	1	1	3	1	1	Retail trade turnover	ECB	proprietary	Jan-94	Dec-20
18	1	1	3	0	1	World Trade	CPB	link	Jan-91	Nov-20
19	0	1	3	0	1	Unemployment rate	CBS	link	Jan-83	Dec-20
20	1	1	3	1	1	Bankruptcies	CBS	link	Jan-65	Dec-20
Financial										
21	1	1	3	1	1	Loans to the private sector	ECB	proprietary	Jan-82	Dec-20
22	1	2	3	1	1	M1	ECB	proprietary	Jan-80	Dec-20
23	1	2	3	1	0	M3	ECB	proprietary	Jan-70	Dec-20
24	0	1	3	1	0	Interest rate: short-term	DNB	link	Nov-84	Jan-21
25	0	1	3	1	0	Interest rate: long-term	DS	NLGBD10	Jan-65	Jan-21
26	0	1	3	1	1	Interest rate: loans on mortgages, 5-10 year	ECB	proprietary	Jan-80	Nov-20
27	1	1	3	1	1	Amsterdam AEX-index	DS	AMSTEOE	Jan-83	Jan-21
28	1	1	3	1	1	Amsterdam AEX midkap-index	DS	AMSMKAP	Jan-83	Jan-21
29	1	1	3	1	1	Dow Jones Euro Stoxx 50-index	DS	DJES50I	Jan-87	Jan-21
30	1	1	3	1	1	Dow Jones Euro Stoxx basic materials-index	DS	S1ESBME	Jan-87	Jan-21
31	1	1	3	1	0	Dow Jones Euro Stoxx financials-index	DS	S1ESFNE	Jan-87	Jan-21

continued on next page ...

TableA.1 – Continued from previous page

Nr.	Transform				Small	Variable name	Source	Code	Start	End
	log	dif.	fil.	sa.						
32	1	1	3	1	1	Dow Jones Euro Stoxx technology-index	DS	S1ESG1E	Jan-87	Jan-21
33	1	1	3	1	0	Dow Jones Euro Stoxx healthcare-index	DS	S1ESH1E	Jan-87	Jan-21
34	1	1	3	1	1	Dow Jones Euro Stoxx industrials-index	DS	S1ESIDE	Jan-87	Jan-21
35	1	1	3	1	0	Dow Jones Euro Stoxx telecommunications-index	DS	S1EST1E	Jan-87	Jan-21
36	1	1	3	1	0	Dow Jones Euro Stoxx utilities-index	DS	S1ESU1E	Jan-87	Jan-21
37	1	1	3	1	0	USD-EUR exchange rate	DS	USECBSP	Dec-78	Jan-21
38	1	2	3	1	0	Housing price index	CBS	link	Jan-76	Dec-20
Prices										
39	1	2	3	1	1	Consumerprice index: headline	CBS	link	Jan-65	Dec-20
40	1	2	3	1	0	Consumerprice index: services	CBS	link	Jan-91	Dec-20
41	1	2	3	1	0	World market commodity prices: total	HWWI	proprietary	Sep-78	Dec-20
42	1	2	3	1	0	World market commodity prices: industrial materials	HWWI	proprietary	Sep-78	Dec-20
43	1	2	3	1	0	World market commodity prices: agri/industrial prices	HWWI	proprietary	Sep-78	Dec-20
44	1	2	3	1	1	World market commodity prices: metals	HWWI	proprietary	Sep-78	Dec-20
45	1	2	3	1	0	World market commodity prices: energy	HWWI	proprietary	Sep-78	Dec-20
46	1	2	3	1	0	Producer price: industry (domestic market)	CBS	link	Jan-81	Dec-20
47	1	2	3	1	1	Producer price: industry (foreign market)	CBS	link	Jan-81	Dec-20
48	1	2	3	1	0	Import prices	CBS	link	Jan-90	Nov-20
49	1	2	3	1	0	Export prices	CBS	link	Jan-90	Nov-20
50	1	2	3	1	0	Hourly wages	CBS	link	Jan-72	Dec-20
Trading Partners										
51	1	1	3	0	1	Belgium: industrial production (excl. construction)	ECB	proprietary	Jan-65	Nov-20
52	1	1	3	0	1	France: industrial production (excl. construction)	ECB	proprietary	Jan-65	Nov-20
53	1	1	3	0	1	Germany: industrial production (excl. construction)	ECB	proprietary	Jan-65	Nov-20
54	1	1	3	0	1	Spain: industrial production (excl. construction)	ECB	proprietary	Jan-65	Nov-20
55	1	1	3	0	1	Italy: industrial production (excl. construction)	ECB	proprietary	Jan-65	Nov-20
56	1	1	3	0	0	Germany: retail trade	ECB	proprietary	Jan-68	Dec-20
57	1	1	3	0	1	Belgium: retail trade	ECB	proprietary	Jan-70	Dec-20
58	1	1	3	0	0	France: retail trade	ECB	proprietary	Jan-70	Dec-20
Gross Domestic Product										
59	1	1	0	0	1	Gross Domestic Product	CBS	link	1977Q1	2020Q3

Transform: log: 0 = no logarithm, 1 = logarithm; dif.: degree of differencing 1 = first difference, 2 = second difference; filt.: 3 = change against the same month of the previous quarter. sa.: 0 = seasonal adjustment at source, 1= seasonal adjustment with X-13. **Source.**: CBS: Statistics Netherlands, CPB: CPB Netherlands Bureau for Economic Policy Analysis, DS: Datastream, DNB: De Nederlandsche Bank, ECB: European Central Bank, HWWI: Hamburgisches WeltWirtschaftsInstitut **Start**: First month of the time-series **End**: Last month of the time series. Codes for proprietary series available upon request.

C Details on Bayesian inference algorithm

C.1 Inference of the posterior distribution

The full posterior distribution of the latent variables ϕ, θ and $\boldsymbol{x}$, conditional on the observed corpus $\boldsymbol{w}$, and the priors α and β, can be inferred by using the definitions of conditional, marginal and joint distributions, as:

$$Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{w}, \alpha, \beta) = \frac{Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w}|\alpha, \beta)}{Pr(\boldsymbol{w}|\alpha, \beta)} \tag{A.1}$$

The joint distribution in the numerator can be written as:

$$Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w}|\alpha, \beta) = \underbrace{\prod_{t=1}^{T} Pr(\phi_t|\beta)}_{\text{topic-word}} \prod_{d=1}^{D^1} \left[Pr(\theta_d|\alpha) \prod_{n=1}^{N} Pr(x_{dn}|\theta_d) Pr(w_{dn}|\boldsymbol{\phi}, x_{dn}) \right] \tag{A.2}$$

The denominator, the *evidence* or *marginal likelihood*, can be obtained by marginalizing over the latent variables β, θ and $\boldsymbol{x}$, i.e.:

$$Pr(\boldsymbol{w}|\alpha, \beta) = \int \int \sum_{\boldsymbol{x}} \left(\prod_{t=1}^{T} Pr(\phi_t|\beta) \right) \left(\prod_{d=1}^{D^1} Pr(\theta_d|\alpha) \prod_{n=1}^{N} Pr(x_{dn}|\theta_d) Pr(w_{dn}|\boldsymbol{\phi}, x_{dn}) \right) d\theta d\phi \tag{A.3}$$

The numerator in equation (A.2) can be easily computed, but the evidence in equation (A.3) is intractable to compute as the latent variables β and θ are not separable in summing over all the possible values of the latent topic structure. This is because there is an exponentially large number of possible topic structures, making this sum intractable (see Blei et al., 2003 for a formal proof). Fortunately, there are several methods to approximate the inference, such as expectation-maximization (e.g., Hofmann, 2001), variational inference (e.g., Blei et al., 2003), or Gibbs sampling (e.g., Griffiths and Steyvers, 2004).

We used a variant of the Gibbs sampling algorithm for inference, i.e., the so-called *collapsed* Gibbs-sampling algorithm. The idea of collapsed Gibbs sampling is that it allows us to sample from a distribution that asymptotically follows the full joint distribution $Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w}|\alpha, \beta)$ without having to explicitly calculate any integrals. We present the main formulas of the collapsed Gibbs sampling algorithm without formal derivation, which are well described elsewhere (e.g. Resnik and Hardisty, 2009).

The collapsed Gibbs sampling procedure considers each word token in the text collection in turn and estimates the probability of assigning the current word token to each topic *conditional* on the topic assignments to all other word tokens. From this conditional distribution, a topic is sampled an stored as the new topic assignment for this word. We write this conditional distribution as $Pr(x_i = j|\mathbf{x}_{-i}, w_i, d_i, .)$, where $x_i = j$ represents the topic assignment of token i to topic j, $\mathbf{x}_{-i}$ refers to the topic assignments of all other word tokens, and '.' refers to all other known or observed information such as all other word and document indices and hyperparameters. Griffiths and Steyvers (2004) and Steyvers and Griffiths (2007) show this can be rather easily calculated, by a counting rule, i.e.:

$$Pr(x_i = j|\mathbf{x}_{-i}, w_i, d_i, .) \propto \frac{\mathbf{C}^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} \mathbf{C}^{WT}_{wj} + W\beta} \times \frac{\mathbf{C}^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} \mathbf{C}^{DT}_{w_i j} + T\alpha} \tag{A.4}$$

The first term on the right-side of the equal sign $\mathbf{C}^{WT}$ is the topic-word matrix, and $\sum_{w=1}^{W} \mathbf{C}^{WT}_{wj}$ is the total number of tokens(words) in each topic. In the second term $\mathbf{C}^{DT}$ is the document-topic matrix and $\sum_{t=1}^{T} \mathbf{C}^{DT}_{w_i j}$ indicates the total number of tokens (words) in document i. α and β are the hyperparameters of the Dirichlet distributions of the document-topic and topic-word distribution, respectively. W is the total number of words in the set of documents, and T is the number of topics. The first term is the probability of word w under topic j, whereas the second term is the probability that topic j has under the current topic distribution for document d. The intuition is that once many tokens of a word have been assigned to topic j, it will increase the probability of assigning any particular token of that word to topic j (the first term). At the same time, if topic j has been used multiple times in a document, it will increase the probability that any word from that document will be assigned to topic j (the second term). Therefore, words are assigned to topics depending on how likely the word is for a topic, *as well as* how dominant a topic is in a document.

The Gibbs sampling algorithm starts by assigning each word token to a random topic in $[1, \ldots, T]$. For each word token, the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are first decremented by one for the entries that correspond to the current optic assignment. Then, a new topic is sampled from the distribution in equation (A.4) and the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are incremented with the new topic assignments. Each Gibbs sample consists of the set of topic assignment tot all $N \times M$ word tokens in the corpus, achieved by a single pass through all documents. During the initial stage of the sampling process, the 'burnin period', the Gibbs samples have to be discarded because they are poor estimates of the posterior. After the burnin period, the successive Gibbs samples start to approximate the target distribution, i.e. the posterior distribution over topic assignments. At this point, to get a representative set of sampled from this distribution, a number of Gibbs samples are saved at regularly spaced intervals, to prevent correlations between samples ('skip sampling'). The sampling process is done sequentially and proceeds until the sampled values approximate the target distribution. We can then use the count-matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ to approximate the estimated posterior topic-word matrices $\boldsymbol{\phi}$ and the estimated posterior topic matrix per document $\boldsymbol{\theta}$, respectively. Following Griffiths and Steyvers (2004) these matrices can be build from calculated probabilities per topic-word and document-topic combinations, defined as:

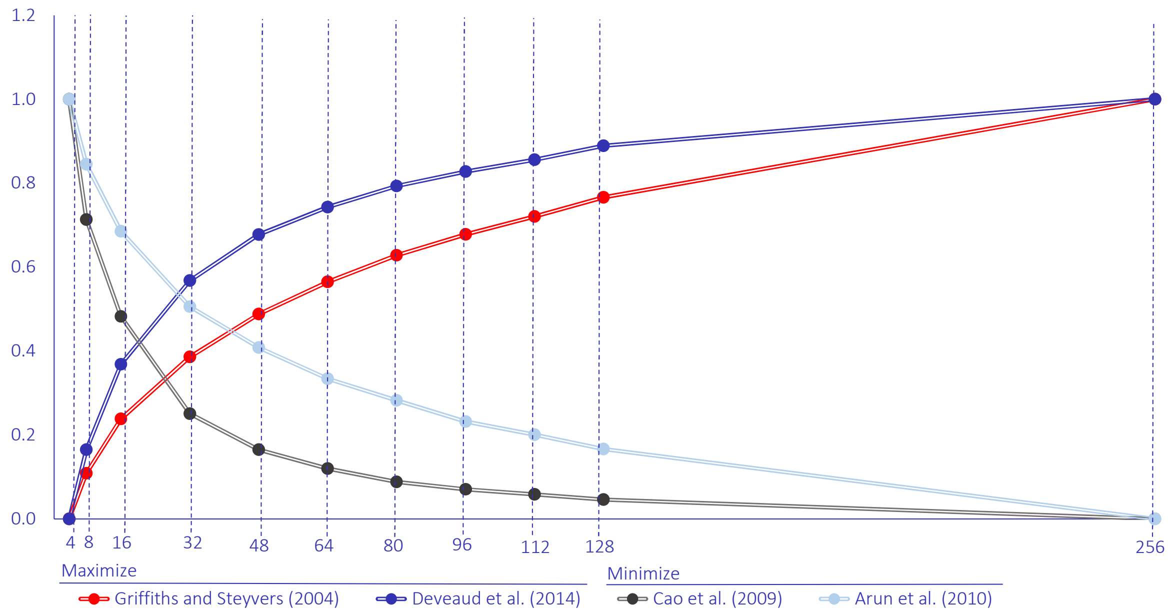
$$\hat{\phi}_{ij} = \frac{\mathbf{C}^{WT}_{ij} + \beta}{\sum_{k=1}^{W} \mathbf{C}^{WT}_{tj} + W\beta}, \qquad \hat{\theta}_{dj} = \frac{\mathbf{C}^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} \mathbf{C}^{DT}_{dt} + T\alpha} \tag{A.5}$$

C.2 Settings of Bayesian inference model

C.2.1 Number of topics

We conduct statistical tests to determine the optimal number of topics using the tests available in the R-package `ldatuning`, i.e. the test in Cao et al. (2009), Griffiths and Steyvers (2004), Arun et al. (2010) and Deveaud et al. (2014). For more information on the tests, we refer to these papers. We tested for the optimal number of topics in the first time slice 1, running from January 1^{st} 1985 to January 1^{st} 2000. The results are presented in Figure A.1. All tests are re-scaled with a min-max transformation to lie between 0 and 1. Generally the tests indicate an optimal number of topics that is higher than 64, but the improvement becomes relatively small with more than 64 topics.

Figure A.1: Selection of number of topics, indicators (re-scaled to 0 -1)



C.2.2 Iterations and hyperparameters

The Gibbs sampler also requires choices regarding the number of repeated samples (iterations) and the setting of the hyperparameters. In general, the choice of the hyperparameters α for the document-topic distribution (θ) and β for the topic-word distribution (ϕ) depends on the empirical application. With a higher α, documents are made up of more topics. Likewise, with a high β, topics are made up of most of the words in the corpus, and with a low β they consist of few words. Griffiths and Steyvers (2004) is an often used standard setting in the literature. in which α is to 50 divided by the number of topics ($50/T$) and β to 0.1. We set α to 0.1 and β to 0.01, after an informal grid-search, using several different hyperparameters in the range $0.01 - 0.1$. The chosen values for the hyperparameters resulted in the best interpretability of the topics.

We set the number of 'burn-in' draws for the Gibbs sampling algorithm in the first time slice to $1,000$. After the 'burn-in' period, we take another $2,000$ samples for the posterior distribution and save every 10^{th} iteration, resulting in 200 saved draws of the posterior distribution (skip-sampling). Next, we determine the draw with the highest posterior likelihood and take the posterior of the topic-word distribution and document-topic distribution as our estimate of ϕ and θ, respectively.

For the second until the last time slice, we do a maximum of $1,000$ iterations, and save every 10^th iteration, resulting in a maximum of 100 saved draws of the posterior distribution. We stop drawing from the posterior distribution if the increase in the likelihood for a draw is less than $1e^{-9}$ with respect to the likelihood ten draws earlier. This stopping algorithm leads to less than $1,000$ iterations for all time slices > 1. Again, we determine the draw with the highest posterior likelihood and take the posterior of the topic-word distribution and document-topic distribution as our estimate of ϕ and θ.

D Dynamic factor model

This section describes the model equations of the nowcasting model in the main text, i.e. the dynamic factor model. Section D.1 provides a detailed description of the model equations, while Section D.2 discusses the modeling choices we made and the impact of these choices on the forecast accuracy of the dynamic factor model.

D.1 Model equations

We use the dynamic factor model specification in Bańbura et al. (2011). The main equation are as follows.

$$\boldsymbol{x}_m = \Lambda \boldsymbol{f}_m + \boldsymbol{\xi}_m, \qquad \boldsymbol{\xi}_m \sim N(0, \Sigma_\xi) \tag{A.6}$$

which relates the n monthly indicators $\boldsymbol{x}_m = (\boldsymbol{x}_{1,m}, \ldots, \boldsymbol{x}_{n,m})'$ to r monthly static factors $\boldsymbol{f}_m = (\boldsymbol{f}_{1,m}, \ldots, \boldsymbol{f}_{r,m})'$ via an $n \times r$ matrix of factor loadings Λ and an idiosyncratic component $\boldsymbol{\xi}_m = (\boldsymbol{\xi}_{1,m}, \ldots, \boldsymbol{\xi}_{n,m})'$, where $r << n$. m is a monthly time index and the monthly indicators $\boldsymbol{x}_{i,m}$ are normalized three-month growth rates or differences. The DFM assumes that the idiosyncratic components are a multivariate white noise process, hence the covariance matrix Σ_ξ is diagonal. Furthermore, the DFM assumes that the factors follow a vector-autoregressive process of order p:

$$\boldsymbol{f}_m = \sum_{s=1}^{p} A_s \boldsymbol{f}_{m-s} + \boldsymbol{\zeta}_m, \qquad \boldsymbol{\zeta}_m \sim N(0, Q) \tag{A.7}$$

where A and Q are square $r \times r$ matrices. The final equation links the factors to mean-adjusted real GDP growth:

$$\boldsymbol{y}_m = \boldsymbol{\beta}' \boldsymbol{f}_m + \boldsymbol{\varepsilon}_m, \qquad \boldsymbol{\varepsilon}_m \sim N(0, \sigma_\varepsilon^2) \tag{A.8}$$

where $\boldsymbol{y}_m$ denotes the (unobserved) three-month growth rate of monthly real GDP. t is a quarterly time index. Quarterly real GDP growth in quarter t, $\boldsymbol{y}_t^Q$, is assigned to the third month of the quarter, i.e. month $3t$ on the monthly time scale. The relation between the quarterly GDP growth rate and quarter-on-quarter latent monthly GDP growth rates is given by

$$\boldsymbol{y}_t^Q = \frac{1}{3}(y_{3t} + y_{3t-1} + y_{3t-2}) \tag{A.9}$$

We derive the weights for all variables and forecasting horizons using the above state space equations. See Koopman and Harvey (2003) Bańbura and Rünstler (2011) and Rünstler (2016) for more details on the extraction of the variable weights. The dynamic factor model of Bańbura and Rünstler (2011) specified in equations (A.6)–(A.9) is estimated in three steps. In the first step we obtain the factors loadings Λ and initial estimates of the static factors $\hat{\boldsymbol{f}}_m$, applying a static principal components analysis to a balanced sub-sample of $\boldsymbol{x}_m$.[12] In the second step we estimate the coefficient matrices A_s in equations (A.7) using $\hat{\boldsymbol{f}}_m$, and $\boldsymbol{\beta}$ in equation (A.8) by using a quarterly version of equation (A.8).[13] In the third step, we cast the model in state space form and use the Kalman filter and smoother to re-estimate the estimated factors ($\hat{\boldsymbol{f}}_m$) and monthly GDP growth.

We calculate forecasts of quarterly GDP growth by applying equation (A.8) to forecasts of monthly factors generated by equation (A.7), and then aggregate to quarterly values. The state-space setup of our dynamic factor model is outlined in the next section. See Bańbura and Rünstler (2011) for a more detailed treatment of the dynamic factor model and the estimation procedures.

[12] The balanced sub-sample is obtained by discarding the rows in x_m that contain missing observations due to publication delays.

[13] As in Bańbura and Rünstler (2011), β is estimated via the regression $\boldsymbol{y}_t^Q = \boldsymbol{\beta} \boldsymbol{f}_t^Q + \boldsymbol{\varepsilon}_t^Q$, where $\boldsymbol{f}_t^Q$ are three-month averages of sample estimates of $\boldsymbol{f}_m$ using the aggregation rule in equation (A.9). σ_ε^2 is estimated as the sample variance of ε_t^Q divided by 3.

Equations (A.6)–(A.9) can be cast in state space form, as illustrated below for the case of $p = 1$. The aggregation rule is implemented in a recursive way in equation (A.11) by introducing a latent cumulator variable $\hat{\boldsymbol{y}}^{\boldsymbol{Q}}{}_m = \Xi_m \hat{y}^Q_{m-1} + \frac{1}{3} y_m$, where $\Xi_m = 0$ for m corresponding to the first month of the quarter and $\Xi_m = 1$ otherwise (see Bańbura and Rünstler, 2011). The monthly state space representation is given by the following observation equation:

$$\begin{bmatrix} \boldsymbol{x}_m \\ \boldsymbol{y^Q}{}_t \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_m \\ \boldsymbol{y}_m \\ \hat{\boldsymbol{y}}^Q_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_m \\ 0 \end{bmatrix} \tag{A.10}$$

and the transition equation:

$$\begin{bmatrix} I & 0 & 0 \\ -\boldsymbol{\beta}' & 1 & 0 \\ 0 & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_{m+1} \\ \boldsymbol{y}_{m+1} \\ \hat{\boldsymbol{y}}^Q_{m+1} \end{bmatrix} = \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Xi_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_m \\ \boldsymbol{y}_m \\ \hat{\boldsymbol{y}}^Q_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\zeta}_{m+1} \\ \boldsymbol{\varepsilon}_m \\ 0 \end{bmatrix} \tag{A.11}$$

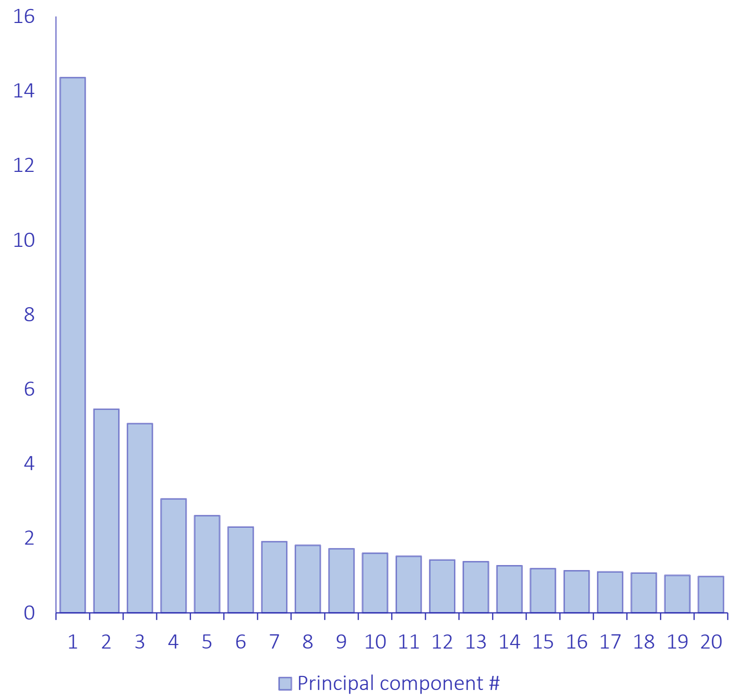
The application of the Kalman filter and smoother provides the minimum mean square linear estimates (MSLE) of the state vector $\boldsymbol{\alpha}_m = (\boldsymbol{f}_m, \boldsymbol{y}_m, \hat{\boldsymbol{y}}^Q_m)$ and enables the forecasting of quarterly GDP growth $\boldsymbol{y}^Q_t$ and dealing efficiently with an unbalanced dataset of missing observations at the beginning and at the end of the series by replacing the missing data with optimal predictions. Moreover, when compared with using principal components technique alone, the two-step estimator allows for dynamics of the common factors and cross-sectional heteroskedasticity of the idiosyncratic component.

D.2 Model specification

To estimate the model we need to specify the number of static common factors r and the number of lags p in the factor VAR process. See Figure A.2 for a 'scree plot' that indicates the explained variance for 1–20 factors. The plot clearly indicates that the model should include at least 3 factors.

We avoid selecting a specific combination of r and p to prevent potential potential misspecification and instability problems, as also noted by Kuzin et al. (2013) and Jansen

Figure A.2: Percentage of variance explained for 1–20 factors.



et al. (2016). Instead, we estimate a set of models for different combinations of r and p. We set the largest possible value of r at 3 and p at 6. The unweighted average of these forecasts form our baseline forecasts.[14] We take as our DFM forecast the (unweighted) average of the forecasts generated by all model specifications. This strategy helps avoid any hindsight bias.

Table A.2 through Table A.4 validate our selection of the number of factors (r) and lags in the VAR (p) for the total sample, the period without COVID outliers, and the period without crisis outliers. It is evident from our sample that the forecast accuracy of the backcasts, nowcasts, and forecasts is largely unaffected by the specification of the dynamic factor model. The simplest dynamic factor model specification, with just one factor (r) and one lag in the VAR part of the model (p), is not significantly inferior to a model with 6 factors and 6 lags in the VAR part of the model. After adjusting for outliers in all crisis periods, the forecasts on the 1 and 2 month horizon are statistically and economically

[14] Jansen et al. (2016) found that the forecasting power of the dynamic factor model increases if r increases (until at least six), while it hardly changes if p increases. A different approach is to choose the number of factors on the basis of in-sample criteria, as described in Bai and Ng (2002). Bańbura and Rünstler (2011) and Jansen et al. (2016) report that these procedures tend to result in more volatile and less accurate forecasts.

significantly better than our base model with 3 factors.

Table A.5 illustrates the effect on the forecast accuracy of our dynamic factor model when using the base dataset of monthly macroeconomic indicators, which includes 58 indicators in total, compared to using a smaller dataset of 34 indicators. The database description in Table A.1 indicates which selection of the large dataset is included in the small dataset in the 'Small' column. As shown in the table, the impact of using a smaller dataset of macroeconomic indicators on the forecast accuracy of the dynamic factor model is negligible, both statistically and economically.

Table A.2: Forecast accuracy of DFM with different number of factors and lags, total sample.

		Backcast		Nowcast			Forecast		
	MSFE	M2	M1	M3	M2	M1	M3	M2	M1
(a) Number of static factors (r)									
DFM: $\leq$ 3 factors (r)	absolute	1.97	2.09	2.19	2.28	2.41	2.58	2.53	2.46
DFM: $\leq$ 6 factors (r)	relative	0.96	0.99	1.00	1.01	1.02	1.01	1.00	1.01
DFM: $\leq$ 5 factors (r)	relative	0.97	0.99	1.00	1.01	1.01	1.01	1.00	1.00
DFM: $\leq$ 4 factors (r)	relative	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 2 factors (r)	relative	1.03	1.02	1.00	1.00	1.00	1.00	1.00	1.00
DFM: 1 factor (r)	relative	1.05	1.03	1.01	1.00	1.02	1.01	1.00	1.00
(b) Number of lags (p)									
DFM: $\leq$ 6 lags (p)	absolute	1.97	2.09	2.19	2.28	2.41	2.58	2.53	2.46
DFM: $\leq$ 5 lags (p)	relative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 4 lags (p)	relative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 3 lags (p)	relative	1.00	1.00	1.00	1.01	1.01	1.00	1.00	1.00
DFM: $\leq$ 2 lags (p)	relative	1.00	1.00	1.00	1.02	1.02	1.00	1.00	1.00
DFM: 1 lag (p)	relative	1.00	1.00	1.01	1.03	1.03	1.00	1.00	1.00

Note: DFM: $\leq$ X factors (r)refers to the MSFE of a dynamic factor model with X factors and 6 lags, estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics. DFM: $\leq$ X lags (r)refers to the MSFE of a dynamic factor model with X lags and 3 factors. Bold cells indicate that the MSFE is at least 10% lower than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) denote that the one-sided Diebold-Mariano test (where the alternative hypothesis is that the model is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

Table A.3: Forecast accuracy of DFM with different number of factors and lags, excluding COVID outliers.

	MSFE	Backcast M2	Backcast M1	Nowcast M3	Nowcast M2	Nowcast M1	Forecast M3	Forecast M2	Forecast M1
(a) Number of static factors (r)									
DFM: $\leq$ 3 factors (r)	absolute	0.29	0.31	0.33	0.36	0.38	0.42	0.46	0.49
DFM: $\leq$ 6 factors (r)	relative	1.01	1.00	1.00	1.01	1.03	1.03	1.02	1.03
DFM: $\leq$ 5 factors (r)	relative	1.01	1.00	0.99	1.00	1.01	1.01	1.01	1.02
DFM: $\leq$ 4 factors (r)	relative	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 2 factors (r)	relative	1.02	1.02	1.01	1.01	1.01	1.01	1.01	1.01
DFM: 1 factor (r)	relative	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99
(b) Number of lags (p)									
DFM: $\leq$ 6 lags (p)	absolute	0.29	0.31	0.33	0.36	0.38	0.42	0.46	0.49
DFM: $\leq$ 5 lags (p)	relative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 4 lags (p)	relative	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99
DFM: $\leq$ 3 lags (p)	relative	1.00	1.00	1.00	0.99	0.98*	0.98	0.98	0.99
DFM: $\leq$ 2 lags (p)	relative	1.00	1.00	1.00	0.99	0.98	0.98	0.99	0.99
DFM: 1 lag (p)	relative	1.00	1.00	1.00	0.99	1.00	1.02	1.01	1.02

Note: DFM: $\leq$ X factors (r)refers to the MSFE of a dynamic factor model with X factors and 6 lags, estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics. DFM: $\leq$ X lags (r)refers to the MSFE of a dynamic factor model with X lags and 3 factors. Bold cells indicate that the MSFE is at least 10% lower than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) denote that the one-sided Diebold-Mariano test (where the alternative hypothesis is that the model is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

Table A.4: Forecast accuracy of DFM with different number of factors and lags, excluding crisis outliers.

		Backcast		Nowcast			Forecast		
	MSFE	M2	M1	M3	M2	M1	M3	M2	M1
(a) Number of static factors (r)									
DFM: $\leq$ 3 factors (r)	absolute	0.18	0.18	0.19	0.2	0.22	0.24	0.26	0.27
DFM: $\leq$ 6 factors (r)	relative	1.02	1.02	1.02	1.02	1.04	1.04	1.04	1.04
DFM: $\leq$ 5 actors (r)	relative	1.00	1.00	0.99	0.99	1.01	1.02	1.02	1.03
DFM: $\leq$ 4 factors (r)	relative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DFM: $\leq$ 2 factors (r)	relative	1.00	1.00	1.01	1.00	1.00	1.00	1.00	1.00
DFM: 1 factor (r)	relative	0.99	0.98	0.98	0.97	0.96	0.96	**0.94***	**0.94**
(b) Number of lags (p)									
DFM: $\leq$ 6 lags (p)	absolute	0.18	0.18	0.19	0.20	0.22	0.24	0.26	0.27
DFM: $\leq$ 5 ags (p)	relative	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99
DFM: $\leq$ 4 lags (p)	relative	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98
DFM: $\leq$ 3 lags (p)	relative	1.00	1.00	1.00	1.00	0.98	0.97	0.97	0.97
DFM: $\leq$ 2 lags (p)	relative	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.97
DFM: 1 lag (p)	relative	1.00	1.00	1.01	1.01	1.02	1.05	1.04	1.06

Note: DFM: $\leq$ X factors (r)refers to the MSFE of a dynamic factor model with X factors and 6 lags, estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics. DFM: $\leq$ X lags (r)refers to the MSFE of a dynamic factor model with X lags and 3 factors. Bold cells indicate that the MSFE is at least 10% lower than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) denote that the one-sided Diebold-Mariano test (where the alternative hypothesis is that the model is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

Table A.5: Forecast accuracy of DFM with large and small dataset, no crisis outliers.

		Backcast		Nowcast			Forecast		
	MSFE	M2	M1	M3	M2	M1	M3	M2	M1
(a) Total sample									
Small dataset	absolute	2.01	2.12	2.02	2.28	2.43	2.58	2.54	2.47
Large dataset (base)	relative	0.98*	0.99	1.00	1.00	0.99	1.00	0.99	1.00
(b) Excluding COVID outliers									
Small dataset	absolute	0.30	0.31	0.33	0.36	0.38	0.41	0.45	0.48
Large dataset (base)	relative	0.98	0.98	0.99	1.01	1.01	1.02	1.02	1.00
(b) Excluding crisis outliers									
Small dataset	absolute	0.18	0.18	0.19	0.20	0.21	0.23	0.25	0.26
Large dataset (base)	relative	0.99	1.00	1.00	1.01	1.02	1.02	1.03	1.01

Note: Small dataset with 34 variables, Large dataset with 58 variables (baseline model). See Table A.1 for more information. Bold cells indicate the MSFE is at least 10% better than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

E Additional Figures and Tables

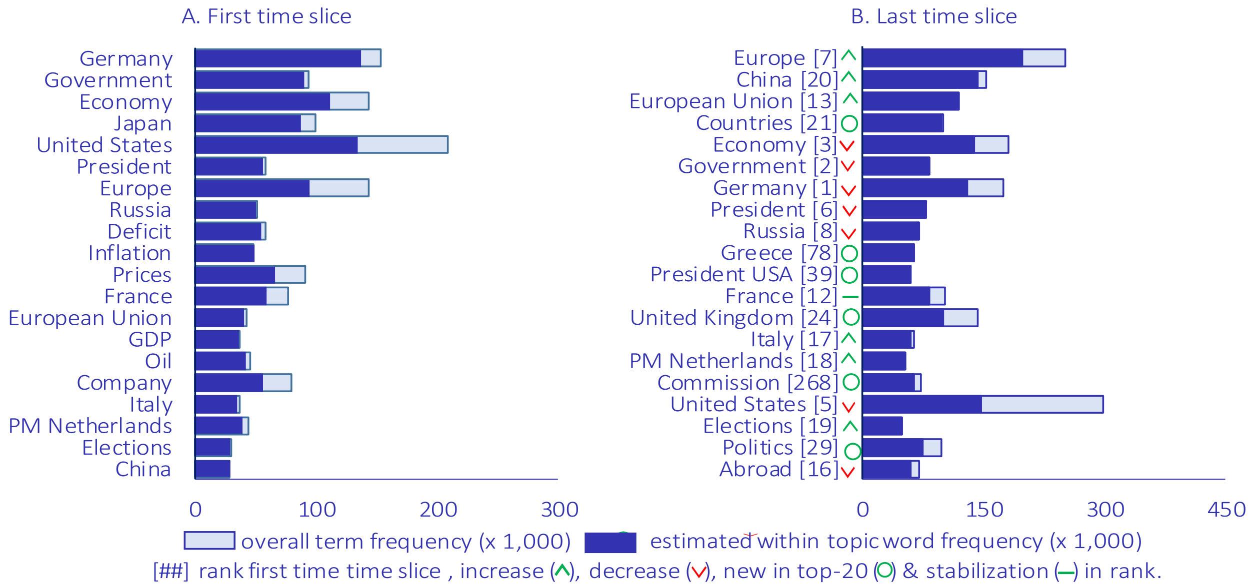


Figure A.3: Top-20 words with highest relevance within economics topic, first and last time slice.

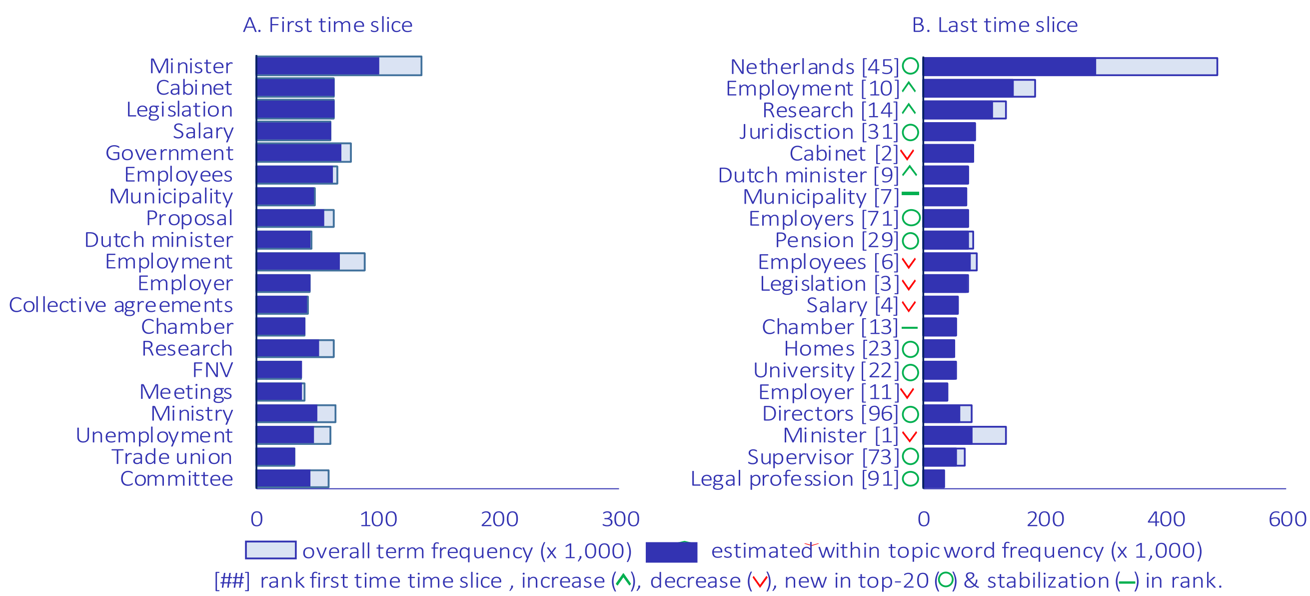


Figure A.4: Top-20 words with highest relevance within politics topic, first and last time slice.

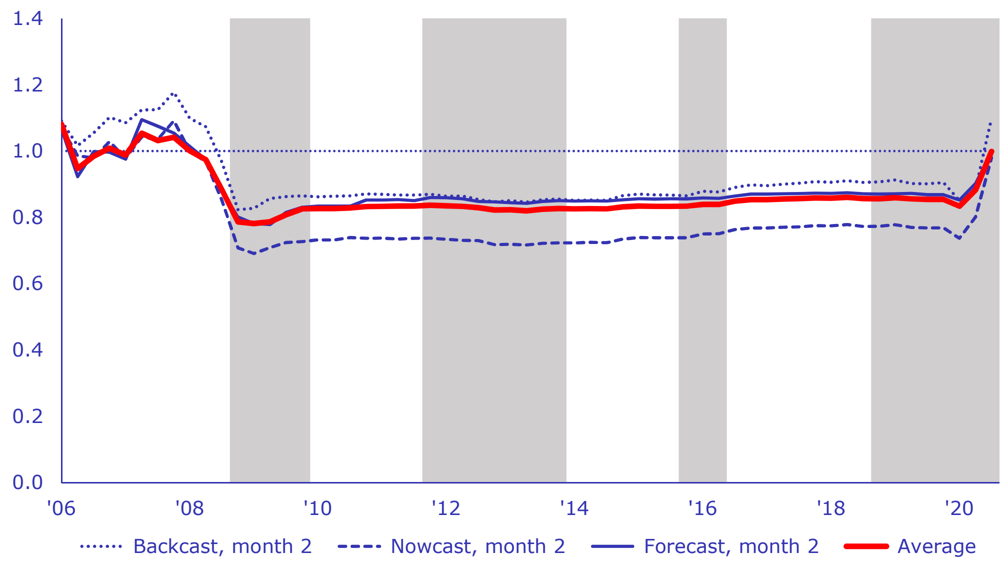


Figure A.5: Relative cumulative sum of squared forecast errors: dynamic factor model with monthly indicators and sentiment adjusted news topics versus dynamic factor model with monthly indicators and news topic proportions, Shaded areas indicate recessions.

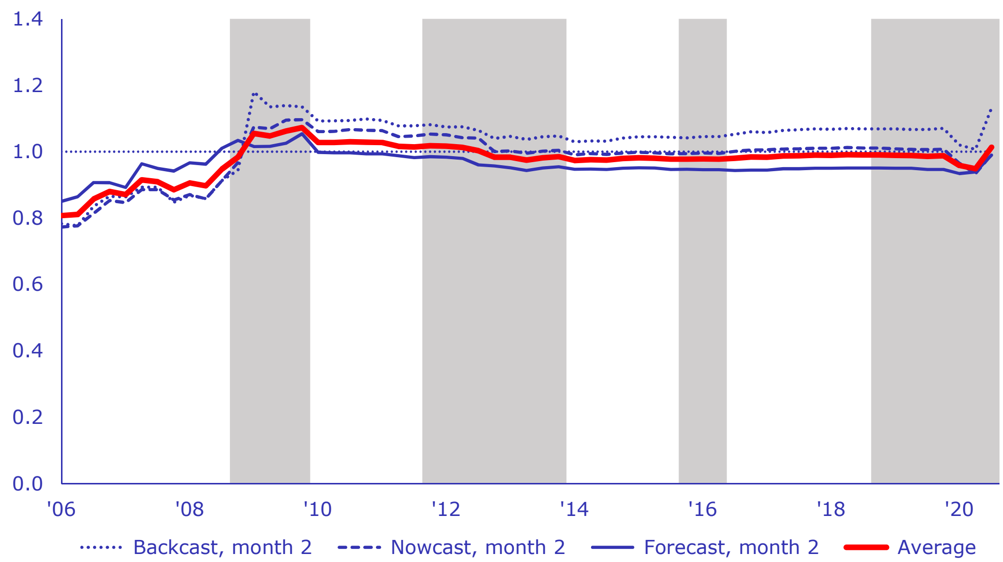


Figure A.6: Relative cumulative sum of squared forecast errors: dynamic factor model with monthly indicators and 64 sentiment adjusted news topics versus dynamic factor model with monthly indicators and 4 sentiment adjusted news topics, Shaded areas indicate recessions.

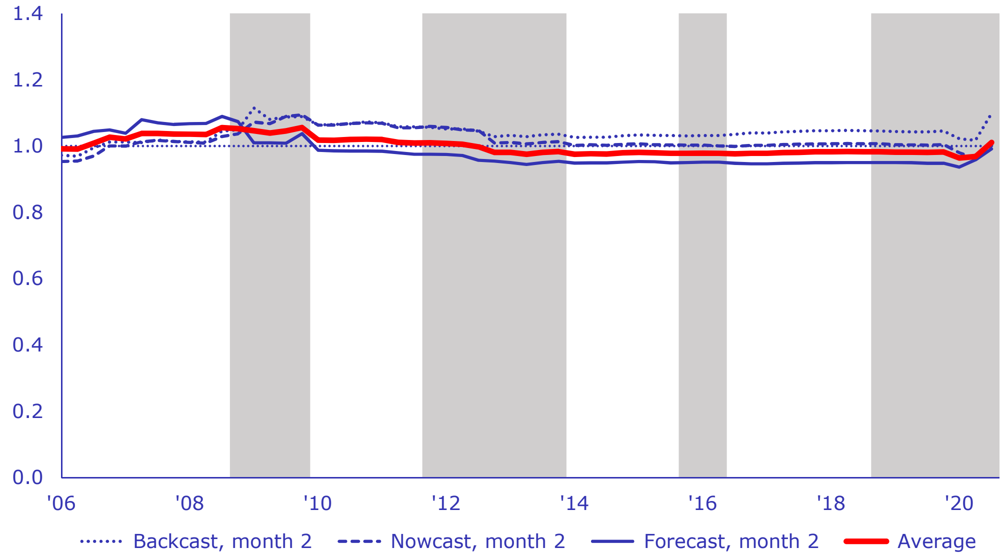


Figure A.7: Relative cumulative sum of squared forecast errors: dynamic factor model with monthly indicators and 64 sentiment adjusted news topics versus dynamic factor model with monthly indicators and 16 sentiment adjusted news topics, Shaded areas indicate recessions.

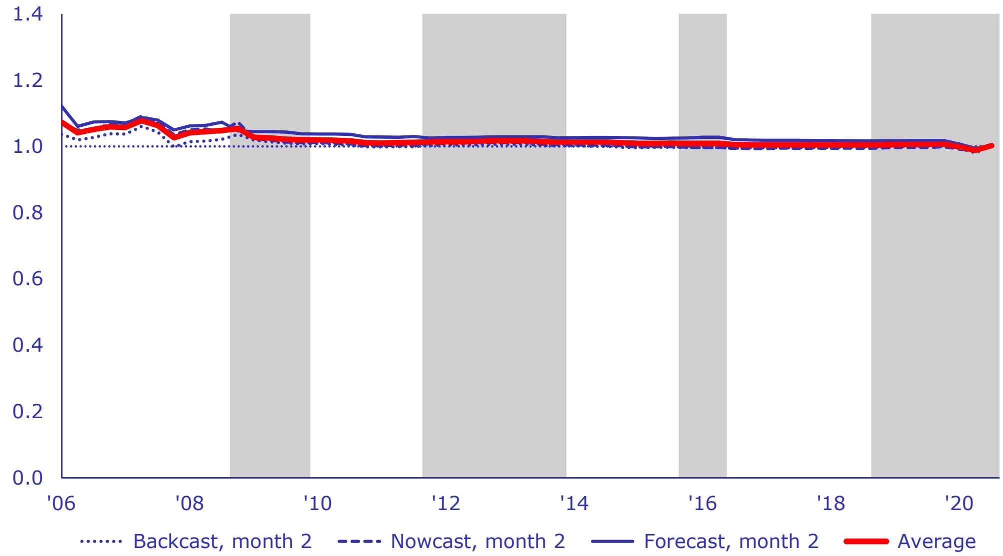


Figure A.8: Relative cumulative sum of squared forecast errors: dynamic factor model with 64 sentiment adjusted news topics from LTM versus dynamic factor model with 64 monthly indicators and sentiment adjusted news topics from TM, Shaded areas indicate recessions.

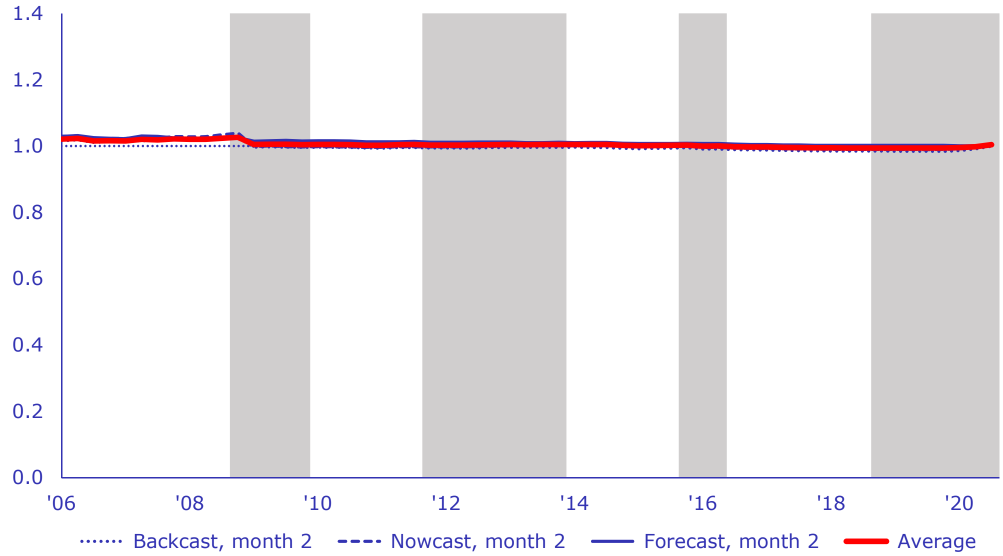


Figure A.9: Relative cumulative sum of squared forecast errors: dynamic factor model with 64 sentiment adjusted news topics from TVTM versus dynamic factor model with 64 monthly indicators and sentiment adjusted news topics from TM, Shaded areas indicate recessions.

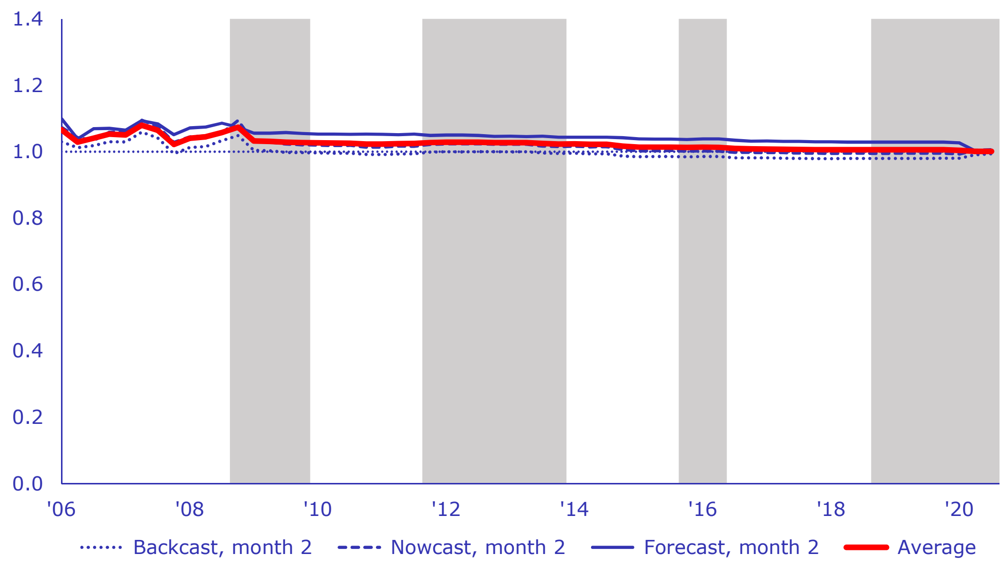


Figure A.10: Relative cumulative sum of squared forecast errors: dynamic factor model with 64 sentiment adjusted news topics from TVLTM versus dynamic factor model with 64 monthly indicators and sentiment adjusted news topics from TM, Shaded areas indicate recessions.

Table A.6: Forecast accuracy of several topic model variations, excluding COVID outliers

	MSFE	Backcast M2	Backcast M1	Nowcast M3	Nowcast M2	Nowcast M1	Forecast M3	Forecast M2	Forecast M1
(a) Proportions versus sentiment adjusted									
Only topics	absolute	0.34	0.40	0.45	0.49	0.51	0.53	0.54	0.55
Base model	relative	**0.85***	**0.77***	**0.72***	**0.74***	**0.75***	**0.79***	**0.85**	**0.89**
(b) Number of news topics									
4 topics	absolute	0.29	0.31	0.34	0.38	0.40	0.44	0.49	0.51
Base model	relative	1.02	0.99	0.96	0.97	0.94	0.94	0.93*	0.96
16 topics	absolute	0.28	0.30	0.33	0.37	0.40	0.45	0.49	0.51
Base model	relative	1.02	1.01	0.99	0.98	0.95	0.92**	0.94*	0.95
(c) Time variation and layering									
TM	absolute	0.34	0.34	0.35	0.37	0.38	0.41	0.44	0.47
TVTM	relative	0.99*	0.99*	0.99*	0.99	1.00	1.00	1.00	1.00
LTM	relative	0.99	0.99	0.99	0.99	1.00	1.00	1.01	1.01
TVLTM	relative	0.98	0.99	0.99	0.99	1.00	1.01	1.03	1.03

Note: Entries show the MSFE of a dynamic factor model estimated on different datasets. i.e. Base model: dynamic factor model estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics, Topic shares: identical to base model but with topic shares instead of sentiment adjusted topics, 4 topics: identical to the base model but including only the 4 topics from the 1^{st} layer of the TVLTM, 16 topics: identical to the base model but including only the 16 topics from the 2^{nd} layer of the TVLTM, TM: identical to the base model but 64 topics from a plain vanilla topic model with fixed word-topic distribution after the first time slice, no layering in topics. LTM: identical to the TM but with three topic model layers. TVTM: identical to the base model, but 64 topics estimated without layering. Bold cells indicate the MSFE is at least 10% better than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

Table A.7: Forecast accuracy of several topic model variations, total sample

		Backcast		Nowcast			Forecast		
	MSFE	M2	M1	M3	M2	M1	M3	M2	M1
(a) Proportions versus sentiment adjusted									
Topic shares	absolute	1.79	2.03	2.18	2.33	2.47	2.60	2.54	2.49
Base model	relative	1.10	1.03	1.01	0.98	0.98	0.99	1.00	0.99
(b) Number of news topics									
4 topics	absolute	1.74	1.96	2.16	2.30	2.46	2.61	2.55	2.49
Base model	relative	1.13	1.07	1.01	0.99	0.98	0.99	0.99	0.99
16 topics	relative	1.79	2.03	2.15	2.27	2.42	2.60	2.55	2.49
Base model	relative	1.10	1.03	1.02	1.00	0.99	0.99	0.99	0.99
(c) Time variation and layering									
TM	aboslute	2.25	2.25	2.26	2.32	2.45	2.56	2.52	2.44
TVTM	relative	1.00	1.00	1.00	1.01	1.01	1.01	1.00	1.00
LTM	relative	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TVLTM	relative	0.99	1.00	1.00	1.01	1.01	1.00	1.00	1.00

Note: Entries show the MSFE of a dynamic factor model estimated on different datasets. i.e. Base model: dynamic factor model estimated on a database of monthly macro-economic indicators & 64 monthly time-varying layered tone-adjusted news topics, Topic shares: identical to base model but with topic shares instead of sentiment adjusted topics, 4 topics: identical to the base model but including only the 4 topics from the 1^{st} layer of the TVLTM, 16 topics: identical to the base model but including only the 16 topics from the 2^{nd} layer of the TVLTM, TM: identical to the base model but 64 topics from a plain vanilla topic model with fixed word-topic distribution after the first time slice, no layering in topics. LTM: identical to the TM but with three topic model layers. TVTM: identical to the base model, but 64 topics estimated without layering. Bold cells indicate the MSFE is at least 10% better than the baseline. Relative MSFEs to the absolute MSFEs in the first line of each panel. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.