Okuneva, Mariia; Hauber, Philipp; Carstensen, Kai; Bär, Jasper

Working Paper

Nowcasting German GDP with Text Data

CESifo Working Paper, No. 11587

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Okuneva, Mariia; Hauber, Philipp; Carstensen, Kai; Bär, Jasper (2024) : Nowcasting German GDP with Text Data, CESifo Working Paper, No. 11587, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/312097

CESifo WORKING PAPERS

11587
2024

December 2024

Nowcasting German GDP with Text Data

Mariia Okuneva, Philipp Hauber, Kai Carstensen, Jasper Bär

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11587*

Nowcasting German GDP with Text Data

Abstract

This paper investigates the impact of news media information on improving short-term GDP growth forecasts by analyzing a large and unique corpus of 12.4 million news articles spanning from 1991 to 2018. We extract business cycle-related sentiment from each article using an annotated dataset from Media Tenor International and a Long Short-Term Memory neural network. This sentiment is then applied to adjust the sign of daily topic distributions estimated through the Latent Dirichlet Allocation algorithm. For the forecasting experiment, we select 10 sign-adjusted topics that show strong correlations with GDP growth, are highly interpretable, and economically relevant. An encompassing test reveals that these topics provide valuable information beyond professional forecasts. In an out-of-sample forecasting experiment, we also find that combining Dynamic Factor Model (DFM) forecasts—derived separately from hard data and text information—consistently outperforms the DFM model relying solely on hard data across all forecasting horizons, with the greatest improvements seen in nowcasts. These results underscore the effectiveness of integrating news media information into economic forecasting, in line with existing literature.

JEL-Codes: C530, C550, E370.

Keywords: textual analysis, topic modelling, sentiment analysis, macroeconomic news, machine learning, forecasting.

*Mariia Okuneva**
Kiel University / Germany
mokuneva@stat-econ.uni-kiel.de

Philipp Hauber
The Economist, London / United Kingdom
philipp.hauber@posteo.de

Kai Carstensen
Kiel University / Germany
carstensen@stat-econ.uni-kiel.de

Jasper Bär
Kiel University / Germany
j.baer@stat-econ.uni-kiel.de

*corresponding author

We are grateful to Rouven Lindenau and Britta Jensen for research assistance. We thank Nils Jannsen, Joscha Beckmann, as well as participants at internal research seminars at the University of Kiel and the Kiel Institute for the World Economy for helpful comments. Furthermore, we would like to express our gratitude to Media Tenor International for providing the annotated dataset, which was a crucial part of this project. Financial support from the Deutsche Bundesbank and the Fritz Thyssen Foundation is gratefully acknowledged. The code supporting the findings of this study is available in our GitHub repositories for data pre-processing https://github.com/MashenkaOkuneva/newspaper_data_processing/tree/master and for analysis https://github.com/MashenkaOkuneva/newspaper_analysis/tree/main.

1 Introduction

In Germany, during the sample period considered in our study, the first official estimate of gross domestic product (GDP) was released approximately six weeks after the end of the reference quarter. This lag in reporting creates a window where GDP growth can be predicted using more timely daily, weekly, and monthly data that become available before the official release. The main goal of this paper is to investigate whether incorporating information from news media can improve the accuracy of short-term forecasts, specifically for predicting the GDP growth rate.

An emerging body of research has focused on using newspaper articles for macroeconomic forecasting, which can be broadly divided into three categories: studies emphasizing article sentiment, such as Shapiro et al. (2022), Rambaccussing and Kwiatkowski (2020), and Kalamara et al. (2022); studies analyzing the topics discussed, as explored by Bybee et al. (2024) and Ellingsen et al. (2022); and hybrid approaches that incorporate both sentiment and topics, as in van Dijk and de Winter (2023), Aprigliano et al. (2023), and Thorsrud (2016). Our research falls into the latter category, which has demonstrated that integrating text data can significantly improve GDP growth forecasts. The success of such text-based measures can be attributed to two main factors. First, text data is available almost immediately, while many economic indicators are released with a considerable lag. Second, news-derived indices tend to be forward-looking, capturing the expectations of various economic agents, including firms, households, and governments.

In this study, we analyze an extensive corpus of 12.4 million news articles spanning from 1991 to 2018, sourced from three leading German newspapers—Süddeutsche Zeitung (SZ), Handelsblatt, and Welt—as well as Germany's largest news agency, dpa (Deutsche Presse-Agentur). The inclusion of a news agency alongside the newspapers provides daily coverage, ensuring that information on significant events occurring on weekends or public holidays is captured in real time. Recognizing the importance of thorough pre-processing for unsupervised learning, as highlighted by Denny and Spirling (2018), we carefully prepared our dataset. This involved three main categories of steps: excluding irrelevant content (e.g., regional news), applying common filtering techniques from the text mining literature (e.g., removing short articles and duplicates), and homogenizing the text (e.g., normalizing umlauts). Together, these steps improve the dataset's quality by eliminating irregularities and outliers while focusing on information relevant to economic forecasting. After pre-processing, the corpus was reduced to 3.3 million articles.

As noted earlier, this study combines two commonly used types of news information: sentiment and topics. To extract the tone from news articles, we apply an aspect-based sentiment analysis approach, similar to Barbaglia et al. (2023). Instead of evaluating the overall sentiment of an article, we focus on specific text segments that are semantically linked to the aspect, or concept, of interest. In this case, we focus on sentiment towards business cycle conditions—an aspect that has the potential to capture overall economic dynamics and thus be valuable for forecasting GDP growth.

For this purpose, we use a dataset provided by Media Tenor International (MTI), a research institute specializing in professional, aspect-based sentiment annotation of news articles. The dataset comprises 3,286 articles from six sources, including daily newspapers (e.g., BILD) and weekly or monthly journals (e.g., Spiegel), which are distinct from our main corpus of Handelsblatt, SZ, Welt, and dpa. Sentiment annotations were conducted by professional coders with a focus on business cycle conditions—our aspect of interest. Importantly, 18% of the articles were annotated by more than one coder, ensuring high-quality and reliable annotations.

We use this dataset to train a Long Short-Term Memory (LSTM) neural network, which is later applied to predict the sentiment of individual articles in the main corpus. The LSTM model is particularly well-suited for our task as it is designed to process long sequences of text. Since our goal is to extract sentiment specifically related to business cycle conditions, we train the model only on sentences containing terms associated with this aspect. These terms are identified through a word-embedding approach.

The MTI dataset contains annotations for articles published between 2011 and 2020. While it is possible to construct monthly sentiment indices for different sentiment aspects using these annotations, the limited time frame makes the resulting series insufficient for our out-of-sample forecasting exercise. Additionally, relying solely on MTI's annotations does not permit sentiment prediction for individual articles within our main corpus. Therefore, in this study, we explore an alternative use for the MTI dataset by employing it as a training set for our supervised machine learning approach. This allows us to extend the analysis to a much larger dataset, resulting in more stable indices that are available at a daily frequency.

Concerns about potential look-ahead bias may arise, as the training dataset includes articles from a time period that overlaps with the evaluation period of our out-of-sample forecasting experiment. However, we argue that this is not a significant issue, as our model is trained only

on sentences related to business cycle conditions and uses words that were prevalent in the main corpus before the out-of-sample forecasting period. The model concentrates on general, ongoing economic discussions about how events affect the business cycle, rather than the specific events themselves. The language used in these text segments remains relatively consistent over time, minimizing the risk of bias. Additionally, the fact that we train the model on articles from one set of German news media and apply it to articles from four other German news sources—achieving both a sentiment index and sign-adjusted topics that respond consistently to major economic events and exhibit strong correlations with GDP growth—further suggests that our methodology effectively captures business content that is covered reliably over time and across different sources.

To analyze semantic topics discussed in the news media over time, we apply the Latent Dirichlet Allocation (LDA) algorithm, introduced by Blei et al. (2003). LDA has become a popular method in economic forecasting (see, e.g., Ellingsen et al., 2022) due to its unsupervised nature and interpretable output. Specifically, it identifies the share of an article allocated to specific topics. From a subset of 887,300 articles that provide sufficient information on the aspect of interest, we extract 200 topics. We then adjust the sentiment of these topics using our business cycle-related sentiment and select the 10 sign-adjusted topics that show the strongest correlation with GDP growth.

To evaluate whether these topics provide valuable information beyond professional forecasts, we conduct encompassing tests. For this analysis, we rely on the Reuters Poll of German GDP forecasts, which aggregates predictions from around 20 experts, including representatives from private firms and research institutes.

Finally, we assess the role of text data in forecasting GDP growth through an out-of-sample real-time forecasting experiment. The main goal of this experiment is to evaluate whether incorporating text-based series can improve GDP growth forecasts compared to a model that relies solely on traditional economic and financial data. To achieve this, we estimate separate models for text data and hard data, a model that integrates both sources, and a combined forecast using predictions from the text-only and hard-data-only models.

The out-of-sample forecasting period spans from 2010 to 2018, during which we produce backcasts, nowcasts, as well as one-step-ahead and two-step-ahead forecasts at 30, 60, and 90 days into the quarter. Our primary model is the Dynamic Factor Model (DFM, Bańbura et al., 2011), which allows us to incorporate daily, monthly, and quarterly data while efficiently han-

dling missing observations at the start and end of the sample. As a benchmark, we use the Mixed Data Sampling (MIDAS) model (Foroni et al., 2015), a method valued for its simplicity and strong empirical performance. The MIDAS model allows us to examine whether using a different approach at a different frequency—where daily variables are aggregated to a monthly frequency—alters the answer to our central question: does text data improve forecasts based on hard data alone? To handle the high-dimensional setting in the MIDAS model, we apply several techniques, including LASSO, Ridge regression, Random Forests, and Principal Component Analysis (PCA), similar to Ellingsen et al., 2022.

Our research contributes to the existing literature in several ways. Firstly, our main contribution is addressing a gap highlighted by Thorsrud (2016), where many studies rely on a lexicon-based approach to adjust the sentiment of topics, which can be relatively inflexible. Furthermore, the sentiment dictionaries used in these studies are often tailored to other domains. In contrast, we extract sentiment specifically related to business cycle conditions, which is directly linked to economic dynamics and therefore potentially relevant for forecasting. This sentiment is derived using a supervised approach, known to be more precise, based on a high-quality training set. Secondly, our dataset of news articles is both large and unique—it includes not only newspapers but also a news agency, ensuring no missing observations in the extracted daily sign-adjusted topics. Thirdly, we have thoroughly prepared the dataset, reducing noise and improving its quality, with all steps carefully documented. Finally, for forecasting, we use the DFM model, which can handle daily data directly—an important feature when evaluating the value of new daily text series, as they might lose their timeliness advantage if transformed to a monthly frequency.

The remainder of the paper is structured as follows. Section 2 provides an overview of the dataset and discusses the pre-processing steps applied. In Section 3, we outline the sentiment analysis methodology, including details on the training set. Section 4 focuses on the extraction and sentiment adjustment of news topics, as well as the selection of topics for forecasting. Section 5 presents the results of the encompassing test. Section 6 describes the out-of-sample forecasting experiment. Finally, Section 7 concludes the paper.

2 Text Data

Our dataset is an extensive collection of German-language news, comprising articles from three leading newspapers with nationwide audience, Welt, Süddeutsche Zeitung (SZ), and Handelsblatt, and from Germany's largest news agency, dpa, which provides information and articles to almost all German daily newspapers. All four are known for quality journalism which is why we expect their articles to reflect current developments in a timely and reliable manner. As indicated by Table 1, our selected newspapers have high daily circulation figures in the German market, ensuring broad exposure. Articles from dpa are frequently reused by virtually all major news outlets in Germany, extending its reach far beyond direct subscribers. This widespread dissemination suggests that the articles used in our analysis affect the belief formation, and eventually the decisions, of the German public, thereby increasing its value for economic forecasting.

We sourced the Welt articles from the LexisNexis database, focusing specifically on the Economy and Finance sections, with these articles published from Monday to Saturday between March 1999 and January 2018, representing 2% of our total dataset. Articles from SZ, accounting for 29% of the dataset, were acquired from Genios[1] and span from Monday to Saturday between January 1994 and November 2018. The Handelsblatt corpus, also from Genios, includes publications from Monday through Friday over the same period and contributes 8% to the dataset. The largest portion originates from dpa and comprises 61% of the dataset with articles published daily from January 1991 to December 2018. See Table 1 for a detailed quantitative breakdown of these contributions. Overall, our dataset aggregates to approximately 12.4 million articles, making it, to the best of our knowledge, the largest collection of German-language news analyzed to date in the economic literature. Moreover, the dataset includes news coverage for every day of the week, which is particularly beneficial for short-term economic forecasting when using a model that directly handles daily data. This allows us to capture information about significant events occurring on weekends or public holidays in real time, without any publication lags.

The preparation of the dataset is thoroughly detailed in Appendix A and documented in the accompanying code repository[2]. It consists of three main categories of pre-processing steps which we briefly sketch in the following. The first category involves excluding information that is clearly irrelevant or may bias our results. From the SZ, we removed 1,611,327 articles (13%

[1]GBI-Genios (https://www.gbi-genios.de) is a leading provider of business databases in Germany, offering extensive resources for economic and financial information.

[2]https://github.com/MashenkaOkuneva/newspaper_data_processing

Table 1: Summary of the Original Dataset

Source	Period Covered	Days Published	Circulation	Articles	Share
Welt	Mar 1999 - Jan 2018	Mon to Sat	72,215	197,565	2%
SZ	Jan 1994 - Nov 2018	Mon to Sat	304,769	3,646,295	29%
Handelsblatt	Jan 1994 - Nov 2018	Mon to Fri	140,612	980,516	8%
dpa	Jan 1991 - Dec 2018	Mon to Sun	-	7,539,874	61%

Note: Articles from Welt are provided by LexisNexis, SZ and Handelsblatt are from Genois, and dpa articles are directly provided by dpa. The "Days Published" column indicates the weekdays on which articles from each source are published. The "Circulation" column reports the number of daily copies circulated in the first quarter of 2021, as reported by the *Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern* (German Audit Bureau of Circulation). The "Articles" column includes the total number of articles from each source, and the "Share" column denotes the percentage each source contributes to the total number of articles in the dataset.

of the dataset) with only regional news that arguably have limited relevance for macroeconomic forecasting. From dpa, we excluded all 1,403,690 articles (11% of the dataset) belonging to the *dpa-AFX Wirtschaftsnachrichten* (business news), a product tailored specifically to the needs of investors. Again, this information may be too granular to be fruitfully used to forecast aggregate developments. Finally, we removed 355 articles from Welt to account for data gaps during specific periods of insufficient coverage in LexisNexis.

The second category includes filtering steps that are commonly employed in the text mining literature. The most substantial effect came from discarding 3,056,317 articles (25% of the dataset) that included less than 100 words. This exclusion is necessary as our preferred topic modeling algorithm, LDA, struggles with short texts due to the lack of sufficient word co-occurrence information (see, for example, Cheng et al., 2014, Qiang et al., 2017, and Y. Bai et al., 2022). Additional filtering eliminated 1,870,368 articles (15% of the dataset) by identifying content not relevant for macroeconomic forecasting. We based the filtering on metadata markers, such as section types, as well as titles and specific text strings. For example, we excluded non-narrative or historically focused articles and we mostly refrained from including articles from sections not related to the economy, finance, and politics. Duplicate removal was also critical, deleting 1,035,860 articles (8.6% of the dataset) including exact and fuzzy duplicates, along with dpa-specific duplicates like news corrections. Another step involved removing irrelevant text segments, such as physical and e-mail addresses, editorial notes, and other uninformative content, and checking for minimal article length. This process resulted in the additional reduction of 61,039 articles. Splitting aggregated articles in Welt increased the number of articles by 35,004,

Table 2: Descriptive Statistics of the Final Dataset

Source	Articles per Day	Articles per Month	Articles total	Share
Welt	32	818	166,155	5%
SZ	73	1,844	551,453	17%
Handelsblatt	93	1,934	578,306	17%
dpa	200	6,073	2,040,385	61%
Total	326	9,929	3,336,299	100%

Note: The "Articles per Day" and "Articles per Month" columns report the average number of articles per day and per month, respectively, in our final data set. The "Articles total" column represents the total number of articles and the "Share" column shows the share of each source.

whereas for dpa, splitting followed by size filtering reduced the count by 16,768 articles.

The third category focuses on homogenizing the text to improve the quality of the input to our statistical models. It included normalizing umlauts in 186,933 articles, separating erroneously merged words and numbers in 144,035 articles, and correcting casing in 77,185 dpa articles. While we have highlighted the steps that had the largest impact on the corpus, many other processes from the second and third categories were also performed. These include language-based filtering, removal of number-heavy articles, exclusion of tables, and merging of article continuations, among others. Together, these additional steps further improve the quality of the data.

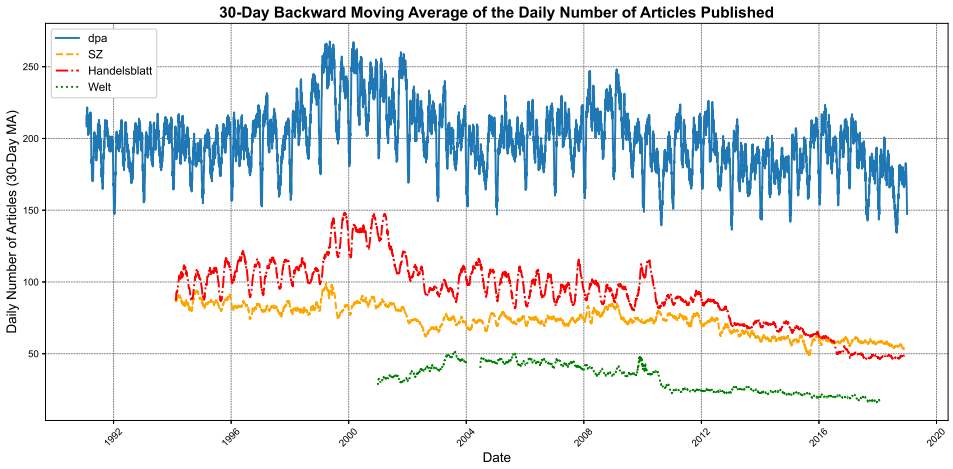


Figure 1: This graph displays the 30-day backward moving average of the daily number of articles published by dpa (blue), SZ (orange), Handelsblatt (red), and Welt (green) in the final dataset after pre-processing. The Y-axis shows this moving average, and the X-axis corresponds to specific days.

As a result of pre-processing, we reduce the total number of articles from 12.4 million in our

original dataset to 3.3 million in its pre-processed version, as reported in Table 2. The average number of articles published daily decreased from 1,199 to 326, and the monthly average fell from 36,396 to 9,929. While the share of dpa articles remained constant at 61%, the proportion of SZ articles dropped from 29% to 17%, Handelsblatt's share increased from 8% to 17%, and Welt's share rose from 2% to 5%. Pre-processing helps us remove irrelevant information, homogenize the texts, and eliminate irregularities and outliers. Additionally, it ensures we focus on content that is consistently covered over time. This standardization is important for estimating topic models, where the objective is to capture genuine spikes in discourse driven by newsworthy events rather than structural changes in reporting. We further discuss this in Appendix B.

To summarize our findings, Figure 1 shows the daily article numbers of the pre-processed datasets for all four media sources over time. Publication rates appear reasonably constant for most of the sample with a slight downward trend, particularly for Handelsblatt and Welt, that may reflect our focus on print editions: some content likely appeared exclusively online towards the end of the sample period. Furthermore, the reduced size of the dataset is beneficial for the computationally demanding task of topic model estimation.

3 Sentiment Analysis

Following the pre-processing of our dataset, we proceed to extract sentiment from the articles. In the literature, sentiment analysis is primarily approached in two ways: using lexicon-based and machine learning-based methods (Algaba et al., 2020). Lexicon-based approaches, which rely on fixed lists of words each with an assigned sentiment score, are more commonly used in the macroeconomic forecasting literature. A notable example is the Loughran and McDonald lexicon (Loughran & McDonald, 2011), particularly favored in this field due to its specific relevance to the economics and finance domains (see e.g., Fraiberger, 2016, Thorsrud, 2016, and van Dijk and de Winter, 2023). In contrast, machine learning approaches, especially supervised ones, often provide more precise sentiment identification by integrating complex models with expert domain knowledge (Ash and Hansen, 2023). A prominent example of this technique is Shapiro et al. (2022) who manually rate the negativity of 800 news articles. The relatively rare use of these methods is mainly due to the high costs and significant time investments required to develop annotated training sets. A promising solution to this challenge is to collaborate with

organizations that specialize in creating such annotations.

In this article, we apply supervised machine learning based on a dataset provided by Media Tenor International, a Swiss-based institute known for analyzing content from major German media outlets. Details about the training set are discussed in Subsection 3.1, our methodology for sentiment extraction in Subsection 3.2, and the resulting daily sentiment index in Subsection 3.3.

3.1 Training Set

Media Tenor International (hereafter referred to as MTI) employs professional coders to annotate news articles. These annotations include the country mentioned, the timing of the events described (past, present, or future), and several other characteristics, among which sentiment (categorized as negative, no clear tone, or positive) is particularly relevant to this article. An attractive feature of MTI's methodology is the focus on aspect-based sentiment (towards the business cycle, labor market, or monetary policy) rather than a general sentiment. This approach has the potential to provide a deeper understanding of the news content and produce sentiment indices that are closely correlated with important economic variables. Barbaglia et al. (2023) also emphasize the importance of aspect-based sentiment, though employing a lexicon-based approach, in contrast to the supervised method used here. Previous research by Ulbricht et al. (2017) demonstrated that sentiment indices derived from MTI's annotations can improve forecasts of industrial production over relatively long forecasting horizons.

The dataset we received from MTI contains details on each article's publication date, source (newspaper or journal), title, sentiment aspect (e.g., business cycle conditions), and the number of annotators who evaluated an article as negative, having no clear tone, or positive towards this particular aspect. Originally, the dataset included 16,874 annotations. It's important to note that annotations differ from articles as the same article may be annotated multiple times for different aspects. In fact, 1,916 articles received annotations at least twice. We excluded 295 annotations that had empty titles and 748 entries that lacked consensus among annotators (e.g., one annotator considered an article negative, while another rated it as having no clear tone). Annotations covered 58 aspects, but only a few were well-represented, with 4,108 sentiment annotations focused on business cycle conditions, 2,641 on fiscal policy, and 2,390 on the labor market.

The articles originated from eight different sources, including the major German daily tabloid newspaper—BILD—and its Sunday edition BILD am Sonntag (BamS), along with Welt am Sonntag (WamS) and Frankfurter Allgemeine Sonntagszeitung (FAS), which are the Sunday editions of the daily newspapers Welt and Frankfurter Allgemeine Zeitung, respectively. Additionally, four weekly and monthly magazines—Spiegel, Focus, Capital, and Manager Magazin—were also included in the analysis.

The annotations from this dataset can be directly used to construct monthly sentiment indices for different aspects by calculating the difference between the proportions of positive and negative articles for each month and aspect, as done by Ulbricht et al. (2017). We present the time series of business cycle sentiment and labor market sentiment, constructed using this methodology and smoothed with a 6-month backward rolling mean, in Figure 2. They clearly highlight the importance of aspect-based sentiment: while the indices for business cycle sentiment and labor market sentiment are positively correlated (correlation coefficient of 0.39), they also exhibit periods of marked divergence. For example, in July 2016, sentiment towards business cycle conditions drops sharply, whereas labor market sentiment does not show a similar decline.

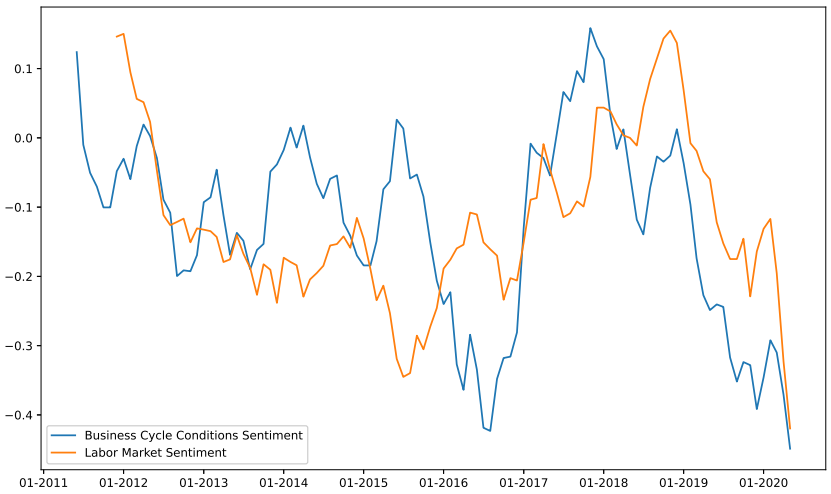


Figure 2: Business cycle sentiment (blue line) and labor market sentiment (orange line), constructed using Media Tenor data. The vertical axis represents the 6-month rolling mean of the difference between the proportion of positive an negative articles per month, while the horizontal axis indicates the corresponding month and year.

As the MTI dataset does not include the full texts of the annotated articles, we needed to access them manually. Although it was possible to download most of these articles from LexisNexis and Dow Jones Factiva (henceforth Factiva), manually retrieving 15,831 annotated

articles would have been excessively time-consuming. Therefore, we decided to concentrate only on articles that received sentiment annotations towards business cycle conditions. This focus is justified for two reasons. First, this aspect has the largest share in the dataset. Second, while this sentiment is specific to the economic domain, it is more general than sentiment towards fiscal policy or the labor market which are also sensible candidates. This broader scope provides a more accurate reflection of the overall economic situation, making it particularly useful for forecasting a variety of economic variables, with an emphasis on GDP growth in this research.

Out of the 4,108 articles annotated towards business cycle conditions, we successfully accessed 3,286 articles. Table 3 presents the number of articles by source. We did not download articles from Manager Magazin, as there were only 16 annotations from this source. Additionally, 392 annotated articles from FAS were unavailable in both LexisNexis and Factiva. We downloaded 2,216 articles from Factiva and 342 articles from LexisNexis based on the date of publication, source, and title provided in the MTI dataset. A further 505 articles from Spiegel and 223 from Focus were obtained from their online archives[3]. For details, please refer to our repository.[4]

Table 3: Articles matching MTI annotation

Source	Articles	Share of Total
Spiegel	1,020	31%
Focus	719	22%
BILD	571	17%
WamS	468	14%
Capital	362	11%
BamS	146	5%
Total	3,286	100%

We then merged the full texts of the downloaded articles with their sentiment annotations from the MTI dataset (see Appendix C.1 for details) and determined the sentiment of each article using a majority vote approach. For instance, if most annotators classified an article as positive, we labeled it with a +1 sentiment. The number of annotations per article ranged from 1 to 26. While the majority of the articles (2,681) were annotated by one person, 605 articles (representing 18% of the total) were reviewed by several coders. Among these, 256 articles were annotated by two people, 163 by three, and 75 by four, with smaller numbers annotated by five or more individuals. This approach ensures high-quality annotations and minimizes classification

[3]See the Spiegel archive at https://www.spiegel.de/spiegel/print/index-2024.html and the Focus archive at https://www.focus.de/magazin/archiv/.

[4]https://github.com/MashenkaOkuneva/newspaper_analysis/tree/main/MediaTenor_processing

error, especially in cases where the sentiment was unclear.

The final training set, following the general pre-processing steps outlined in Appendix C.2, covers the period from January 2011 until May 2020, with an average of 29 annotations per month. From Table 3, we see that around 50% of the articles come from Spiegel and Focus. Regarding the sentiment distribution, 49% of the articles (1,604 articles) express a negative sentiment towards the business cycle, 30% (992 articles) have positive sentiment, and only 21% (690 articles) were classified as having no clear tone.

3.2 Sentiment Extraction Methodology

In this subsection, we use the prepared MTI dataset to train a machine learning model specifically designed for predicting the sentiment of individual articles. Since our goal is to extract sentiment towards business cycle conditions, we train the model solely on sentences related to this aspect rather than on the entire text of each article. We will first describe the procedure for selecting relevant sentences and then proceed with the training and evaluation of the model.

3.2.1 Sentence Selection for Business Cycle Sentiment

Our main motivation for concentrating on sentences related to the aspect of interest is to mimic the annotation process used by professional coders at MTI. In their workflow, coders were instructed to read an article and annotate its sentiment towards business cycle conditions, naturally paying more attention to sentences containing relevant information while disregarding the rest. Although this approach can lead to some loss of information by omitting sentences that are indirectly relevant or provide additional context, it helps to focus on the most important content. This trade-off between relevance and completeness is a common challenge in aspect identification problems (see e.g., Liu, 2012).

To effectively isolate sentences that relate to business cycle conditions within the articles, we employ a lexicon-based approach. Specifically, a sentence is retained if it contains key terms directly associated with our aspect of interest, such as 'business cycle conditions' or 'economy'. We further expand this selection by including words that are either syntactically or semantically linked to these seed terms.

For the identification of such related words, we use the word2vec model, developed by Mikolov et al. (2013a), which has been particularly popular in the economic literature. It generates

so-called word embeddings—numerical vectors that capture the semantic properties of words, grouping similar words closer together in the vector space. This model has been successfully applied to define risk exposure categories (Davis et al., 2020), measure economic uncertainty (Soto, 2021), and assess climate change transition risks (Kapfhammer et al., 2020).

The objective function of the word2vec model, particularly in its skip-gram architecture, maximizes the probability of observing context words given a target word. Mathematically, this is expressed as maximizing the following log-likelihood function over the set of all words in the corpus:

$$L = \frac{1}{P} \sum_{p=1}^{P} \sum_{-C \leq j \leq C, j \neq 0} \log p(w_{p+j} | w_p) \,, \tag{1}$$

where P is the total number of words in the corpus, w_p is the target word at position p, C determines how many words before and after the target word are considered, and w_{p+j} represents a context word.

In line with neural-network language models, the conditional probability $p(w_{p+j}|w_p)$ that a context word w_{p+j} will appear given a target word w_p is modeled using a softmax function:

$$p(w_{p+j} | w_p) = \frac{\exp(v_{w_{p+j}}^\top u_{w_p})}{\sum_{i=1}^{V} \exp(v_i^\top u_{w_p})} \,, \tag{2}$$

where u_{w_p} and $v_{w_{p+j}}$ are the vector representations (embeddings) of the target word w_p and the context word w_{p+j}, respectively. The denominator serves as a normalization term that sums the exponentiated dot products of the target word vector with every other word vector in the vocabulary, V, of our corpus. In econometrics, this softmax function is equivalent to the multinomial logit model.

The output probabilities indicate how likely each vocabulary word is to appear near our target word. For instance, in a well-trained model with 'business cycle conditions' ('Konjunktur' in German) as the target word, we would expect significantly higher probabilities for contextually related words such as 'GDP' or 'consumption'. Conversely, unrelated words like 'dog' or 'weather' would correspondingly receive much lower probabilities.

As indicated by (2), the model's parameters consist of vector embeddings for target words u_{w_p}, combined into a matrix $\mathbf{U}$, and embeddings for context words $v_{w_{p+j}}$, which together form a matrix $\mathbf{V}$. The goal during training is to optimize these parameters by maximizing the log-likelihood as

formulated in (1). This optimization is achieved by iteratively updating the entries in matrices **U** and **V** via gradient descent, thereby refining the embeddings to more accurately capture semantic relationships between words. The final embeddings used in applications are drawn from matrix **U**. For detailed derivations of the parameter update equations, see Rong (2016). A more thorough explanation of word2vec in the context of economic literature is provided by Soto (2021).

We train our word2vec model on articles from dpa, Handelsblatt, SZ, and Welt, covering the period from 1991 to 2009. This timeframe ensures that our out-of-sample GDP growth forecasts are based solely on information that was historically accessible, thereby preventing any information leakage. Before estimation, we perform standard pre-processing steps, which are explained in detail in Appendix D.1. The primary goal of these steps is to standardize the input data and to reduce noise, for instance, by removing words that rarely appear in the dataset.

After pre-processing, the corpus of articles includes 757,990 unique words. With the size of embedding vectors set to 256, the matrices **U** and **V** would contain approximately 194 million parameters each. To handle this computational complexity, we apply three sampling techniques introduced by Mikolov et al. (2013b) and explained in Appendix D.2: subsampling of frequent words, shrinking the context window by random amounts, and negative sampling.

These methods, combined with carefully chosen hyperparameters, significantly contribute to the quality of the trained embeddings. For our model, we set the embedding dimension to 256 and the context window size C to 10—both commonly used in the literature. We experimented with context window sizes of 5 and 10, finding that a window size of 10 yielded more meaningful terms related to business cycle conditions, making it the better choice for our analysis. Other estimation details are summarized in Appendix D.3.

During training, we monitored the model's progress by periodically printing out and evaluating 16 words—8 from the 300 most frequently occurring in the corpus and 8 less common words—along with their related terms based on cosine similarity.[5] For example, during the first epoch (one complete pass through all the words in the corpus), the pronoun 'his' was incorrectly linked to unrelated terms like 'child allowance', 'aged', and 'newly built'. However, by the second

[5]Cosine similarity measures how similar two vectors are by calculating the cosine of the angle between them. Specifically, we measure the cosine similarity between the embedding of the selected word u_{w_i} and the embeddings of all words in the vocabulary u_{w_j} (for $j = 1, 2, \ldots, V$). The formula is given by $\cos\theta = \frac{u_{w_i} \cdot u_{w_j}}{\|u_{w_i}\| \|u_{w_j}\|}$. The value ranges from -1 (completely dissimilar) to 1 (identical), with "related" terms being those with the highest cosine similarity to the selected word.

epoch, the associations had refined to 'he' and various forms of 'his'. Similarly, the less common word 'Schröder' initially related to irrelevant terms like 'spheres of interest' and 'restaurant', but by the fifth epoch, it correctly associated with 'Gerhard', 'chancellor', and 'federal chancellor'.[6] We stopped training after 10 epochs, as the related terms for both common and uncommon words had become meaningful.

We further assessed the quality of our final word embeddings by evaluating their ability to identify terms related to the concept of interest. To do this, we visualized the embeddings of the 1,000 words most similar to 'business cycle conditions' based on cosine similarity, using the t-SNE technique. The main goal of t-SNE is to reduce the 256-dimensional embeddings to a 2-dimensional space, enabling a visual analysis of the relationships between words. The resulting visualization suggests that our embeddings effectively capture meaningful semantic relationships and can identify terms closely linked to the concept of interest. For a detailed explanation of the t-SNE algorithm, its application to this visualization, and an in-depth discussion of the results, please refer to Appendix D.4.

After confirming that words most similar to the estimated embeddings of 'business cycle conditions' and 'economy' are indeed relevant to the intended aspect, the next step was to determine the appropriate number of related terms to include. One straightforward approach is to manually select the top N words most similar to each target term. However, this method introduces an element of subjectivity. To address this, we explored an alternative approach applied by Soto (2021), which involves using K-means clustering to group the 1,000 word embeddings most cosine-similar to either 'business cycle conditions' or 'economy'. Words that fall within the same cluster as the target term are then considered related. Further details on K-means clustering and its application in our analysis can be found in Appendix D.5.

For 'business cycle conditions', the clustering approach proved effective, identifying 279 related terms, which are listed in Appendix D.6.1. However, for 'economy', this method yielded 653 related terms, many of which were overly generic, such as 'this', 'despite', and 'also'. This difference likely stems from the fact that 'economy' is a more frequently used term that appears in a wider variety of contexts. Therefore, for 'economy', we prefer to focus on the 100 most cosine-similar words listed in Appendix D.6.2.

Although economy and business cycle conditions are distinct concepts, there are two reasons

[6]Gerhard Schröder was German chancellor from 1998 until 2005.

why we include terms related to both. First, these concepts are closely interconnected and often co-occur in news articles. This is evidenced by the fact that 'economy' is a related term to 'business cycle conditions', and vice versa (see Appendix D.6). Additionally, after identifying terms associated with each concept, we noted that 44 terms appear in both lists, including 'economic upswing', 'global economy', and 'labor market'. Second, a review of articles from the MTI dataset indicates that coders considered sentences related to both economy and business cycle conditions when assessing sentiment.

Table 4: Sentences Retained for Sentiment Analysis

Sentiment	**Retained Sentences**
Negative	**Economy**: France has recorded hardly any growth for ten years, in addition to an enormous **foreign trade deficit** and high **national debt** (97% of **economic output**). **Unemployment**: At 10 percent, the **unemployment rate** is almost twice as high as in Germany, and youth **unemployment** is dramatic (currently 23.7 %, more than in Romania). Domestic destruction: In France, fear of **unemployment** is rampant.
No clear tone	It states that **economic growth** could be up to three percent higher if environmental and human rights groups did not lobby against coal mining and nuclear power.
Positive	Berlin - The German **economy** is growing! This is what the **economic experts** predict in their forecast for 2014. According to this forecast, **gross domestic product** is likely to increase by 1.9% in 2014, which is **stronger** than previously assumed. In addition to rising private **consumption**, the **economy** is also being boosted by steadily increasing corporate investment in new plant and machinery.

Note: These examples are drawn from the MTI dataset and display only the sentences that were retained. A sentence is included if it contains at least one term related to business cycle conditions; these terms are highlighted in bold for clarity. The articles were translated from German to English using DeepL. The original German versions of these articles are available in Appendix D.7.

Finally, we standardized the articles from the MTI dataset by retaining only those sentences that include at least one term related to either 'business cycle conditions' or 'economy'. Table 4 provides three representative examples: one article with a negative sentiment towards business cycle conditions, one with no clear tone, and one with a positive sentiment. The key terms that justified the inclusion of each sentence are highlighted in bold. As demonstrated, the retained content is directly relevant to business cycle conditions, and the sentiment in each article is clearly discernible. These filtered articles are then used in the supervised training for our sentiment analysis.

While there are alternative approaches for aspect identification, such as manually created dictionaries, topic model outputs, and supervised models trained on annotated corpora (see, e.g., Jangid et al., 2018, Ash and Hansen, 2023), our method offers several key advantages. It is

fully automated, computationally efficient, and, most importantly, proves to be highly successful in identifying terms relevant to the concept of interest.

3.2.2 Long Short-Term Memory Neural Network

For sentiment analysis, we selected the Long Short-Term Memory (LSTM) neural network, which we train on the filtered articles from the MTI dataset. This model is then used to predict sentiment towards business cycle conditions in the main corpus. The LSTM model, introduced by Hochreiter and Schmidhuber (1997), is a type of recurrent neural network (RNN) designed to process sequences of data—a critical feature given the sequential nature of language. Unlike standard RNNs, LSTMs effectively address the vanishing and exploding gradient issues, which improves their ability to learn dependencies over long sequences (for more details, see Hochreiter et al., 2001). LSTM networks have been successfully applied in various financial and economic contexts, including predicting stock closing prices (Jin et al., 2020), assessing financial system instability (Kanzari et al., 2023), forecasting inflation (Almosova & Andresen, 2023), and analyzing cryptocurrency-specific sentiment (Nasekin & Chen, 2020).

The architecture of our LSTM model is presented in Figure 3. In this model, each input word (e.g., 'decline') is first transformed into a 256-dimensional word embedding x^t using the same embedding matrix pre-trained with the word2vec algorithm that was previously used to identify terms related to business cycle conditions. These embeddings are then passed through two layers of LSTM cells (although for simplicity, only one layer is shown in the figure), where the cell states b_c^t and hidden states b_h^t are updated at each time step t. The hidden state b_h^t is used as the output of the LSTM cell. The cell state b_c^t functions as the model's memory, preserving essential information from previous time steps.

The LSTM cell updates this memory with the help of specialized neural network layers called "gates", which regulate the flow of information. At each time step, the *"forget gate"* decides which parts of the previous cell state should be discarded, allowing the model to remove irrelevant information. Simultaneously, a *candidate cell state* is generated, representing potential new information. The *"input gate"* regulates how much of this new information should be incorporated into the current cell state. Together, these updates result in an adjusted cell state that balances the retained information with the newly added content. Finally, the *"output gate"* controls how the updated cell state contributes to the hidden state b_h^t. This coordinated mechanism allows the

model to preserve relevant information over long periods of time. As a result, LSTM models are particularly well-suited for sentiment analysis, as they can capture and retain information from earlier words in a sequence, allowing these words to influence the final sentiment prediction.

The model processes the input sequentially, receiving one word at a time. The final layer of the network is an output layer with a sigmoid activation function, which is used for binary sentiment classification. As each word is processed, the sentiment prediction (based on the sigmoid activations) is updated, but the weights of the network remain constant throughout the sequence. The final sentiment prediction for the entire sequence is considered the sentiment of the article. The mathematical details of the LSTM model and its gates can be found in Appendix D.8.

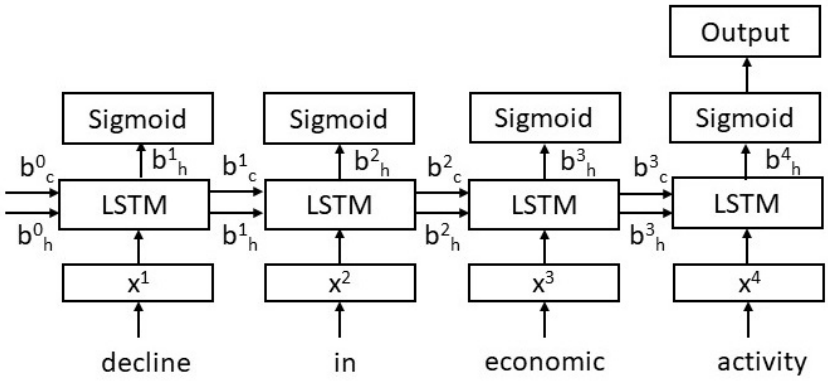


Figure 3: Architecture of the LSTM model used for sentiment analysis.

We opted for a two-class sentiment analysis rather than the original three-class approach (negative, no clear tone, positive) to focus on distinguishing clearly negative news from all other types. While most articles from the MTI dataset classified as positive or negative displayed a clear sentiment regarding business cycle conditions, the 'no clear tone' category presented significant challenges. This class was mixed: some articles genuinely reflected neutral sentiment about business cycle conditions, while others exhibited mixed sentiment or barely addressed the aspect. To illustrate the complexity of this category, Table 5 provides examples of articles annotated as 'no clear tone'. For brevity, we present only the sentences retained for sentiment analysis. The first example clearly addresses business cycle conditions and has a neutral tone, emphasizing statistics without expressing any sentiment. The second example shows mixed sentiment, beginning with optimism about expected economic growth but later highlighting the risks posed

by high national deficits. In the third example, terms related to business cycle conditions, like 'economy' and 'industry', are mentioned but are secondary to the article's primary focus on defense and security. This category's mixed nature, combined with its relatively small number of training examples, led to lower model accuracy when using the three-class approach. By simplifying the task to a binary classification, we improved the model's performance. Additionally, we want to highlight that the prevalence of negative sentiment in news articles, and the relatively small share of articles with no clear tone, reflects the inherent nature of news reporting rather than a limitation of our dataset. As documented by Soroka et al. (2019), news tends to exhibit a negative bias, and thus it is expected that the proportion of positive and no clear tone articles will always be lower than that of negative news.

Table 5: Examples of Articles Annotated as 'No Clear Tone'

Type	**Retained Sentences**
Neutral sentiment	This is how the **economy** has developed since then: **Unemployment**: before the vote, 1.64 million were unemployed; today (as of Oct. 2016) there are 16,000 fewer. Trade: **Exports** increased, the trade deficit fell from 11.4 billion (as of June) to 11 billion pounds (as of December). **Gross domestic product**: rose by 0.5 % from July to September.
Mixed sentiment	Late this afternoon, Chancellor Angela Merkel (59, CDU) received the heads of the world's most powerful **economic associations**, including Christine Lagarde (IMF) and Angel Gurria (**OECD**). The good news: the **global economy** will grow by 3.6% this year and by 3.9% in 2015. The high **national deficits** of many countries could jeopardize the **upturn**, the **economic experts** agreed.
Limited information on the aspect	As the most important European **economy**, we must live up to our global role. If **industry** is to maintain capacities for supplying the armed forces, this requires a clear commitment from politicians: for **sustainable** financial planning.

Note: This table provides examples of articles classified under the 'No clear tone' category, illustrating different cases such as neutral sentiment, mixed sentiment, and articles that barely discuss business cycle conditions. These examples are drawn from the MTI dataset and display only the sentences that were retained. A sentence is included if it contains at least one term related to business cycle conditions; these terms are highlighted in bold for clarity. The articles were translated from German to English using DeepL. The original German versions of these articles are available in Appendix D.9.

Before estimating the LSTM model on the filtered MTI articles, we performed model-specific pre-processing to prepare the data (see Appendix D.10). While most steps, such as lowercasing text and removing punctuation, are standard and aimed at focusing on essential information, two particular steps deserve attention. First, we excluded words without corresponding embeddings in our pre-trained word2vec model. These excluded words generally fall into three categories: rare terms like 'tweet' or 'video call' that seldom appeared in the economic and political news articles used to train word2vec; misspellings, which are absent from the main corpus; and words

that entered common usage after 2010, like 'Brexit' and 'COVID', which are missing because the embeddings were trained on articles published before 2010.

Excluding words from the first two categories helps reduce noise, while removing words from the third group minimizes the risk of data leakage by ensuring that the model focuses on vocabulary prevalent before 2010. Although this approach may leave some information in the texts unanalyzed, we recognize that our sentiment model is trained on articles from the out-of-sample forecasting period. Ideally, we would use only articles published before 2010, but such annotated corpora are rare. Thus, our solution, while not perfect, helps maintain a focus on historically relevant language without introducing new terms into the LSTM.

The second key step in our pre-processing involved excluding articles that contained 20 or fewer words. This exclusion led to the removal of 853 articles from the MTI dataset: 391 from the negative class, 171 from the no clear tone class, and 291 from the positive sentiment class. While this step does reduce the dataset size, it ensures that only articles with sufficient content related to business cycle conditions are included. By focusing on the aspect of interest, this approach, like the previously discussed step, also helps reduce the potential for data leakage by prioritizing consistently discussed economic topic over transient events like the COVID-19 pandemic.

Moreover, the original MTI articles varied significantly in length depending on the source. As shown in Table 6, Spiegel, a weekly journal, had much longer articles on average (1640 words) compared to the daily newspaper BILD, which had an average of 204 words. This substantial difference in length made it difficult to analyze these sources together. However, after filtering to retain only sentences with terms related to business cycle conditions and removing shorter articles, the average word count became much more comparable, ranging from 47 words in BILD to 186 words in Spiegel. This standardization of article length might be important for the performance of neural networks, which benefit from a simpler, more localized relationship between the inputs and the sentiment target (Graves, 2012b). Additionally, it facilitates the application of a model trained on one set of sources to other corpora containing different newspapers. This is because, by using inputs with a more consistent structure and a clear focus on the aspect of interest, the model is less likely to learn patterns specific to any particular source, such as variations in writing style, article formats, and length.

One could argue that 20 words is a rather low threshold and that with some generic terms in our list of words related to business cycle conditions—such as the verb 'create', which appears

Table 6: Article Length Statistics by Source

Source	Original			Filtered and Pre-processed		
	Mean	25th Perc.	75th Perc.	Mean	25th Perc.	75th Perc.
Spiegel	1640	726	2099	186	57	237
Focus	635	174	922	98	38	119
BILD	204	74	211	47	25	51
WamS	1001	468	1410	141	52	186
Capital	1512	511	2156	183	73	250
BamS	503	151	776	78	35	104

Note: This table provides statistics on the length of articles by source, including the mean, 25th percentile, and 75th percentile of the word count distribution for both the original MTI articles and their filtered and pre-processed versions.

720 times in the MTI dataset—we might end up keeping some articles that are only loosely or not at all related to the topic. However, we believe this is not a significant issue for two reasons. First, these generic terms, when used in articles related to the aspect of interest, provide valuable context (e.g., "*create* purchasing power", "*create* jobs and income"). Second, in cases where articles not closely tied to the aspect are retained primarily due to generic terms and the low threshold, our approach mitigates this issue by combining sentiment analysis with topic modeling. We discuss this further in the next section.

After pre-processing, our dataset included 2,433 articles: 1,213 categorized as negative (50%), 519 as having no clear tone (21%), and 701 as positive (29%). These were divided into three sets: the training set, comprising 1,920 articles (approximately 80% of the total), the validation set with 256 articles (about 10%), and the test set consisted of another 256 articles (also about 10%). The training set was used to develop the model, the validation set helped determine the optimal stopping point during training, and the test set was reserved for evaluating the model's final performance.

Our selected model consists of two LSTM layers, each with 32 hidden units. Further details about the model configuration, training process, and optimization settings are available in Appendix D.11. During development, we experimented with various pre-processing techniques, such as retaining words without pre-trained embeddings and allowing the model to learn their embeddings during training, lemmatizing words, and removing stopwords. We also tested several hyperparameters, including the number of hidden units, layers, maximum sequence length, and the number of epochs. The final model, which incorporated the pre-processing steps outlined in

Appendix D.10 and the architecture detailed in Appendix D.11, achieved the highest accuracy on the test set and delivered meaningful sentiment predictions.

To assess the performance of the model, we relied on common metrics such as accuracy and the F1 score. Accuracy measures the percentage of correct predictions, while the F1 score balances precision and recall to account for both false positives and false negatives. As shown in Table 7, our LSTM model achieved 62% accuracy for the negative class and 71% for the positive/no clear tone class. These results indicate that the model effectively distinguishes between sentiment categories, with slightly better performance in identifying articles with positive or no clear tone sentiments.

Table 7: LSTM Performance

	Accuracy	F1
negative tone	62%	0.64
postive/no clear tone	71%	0.69
Total (weighted average)	66.8%	0.67

We compared the LSTM model's performance with two alternative sentiment approaches: a Linear Support Vector Machine (LSVM) and a lexicon-based method. Although the LSTM only slightly outperformed the LSVM, it achieved a notably higher accuracy than the lexicon approach, demonstrating the advantages of our methodology. Full details of these comparisons are available in Appendix D.12.

3.3 Daily Sentiment Index

After successfully training the LSTM model on the MTI dataset and confirming its strong performance on unseen data, we applied it to predict sentiment for individual articles in the main corpus, which includes dpa, Handelsblatt, SZ, and Welt. This process begins by focusing exclusively on sentences containing at least one term related to business cycle conditions, followed by applying the same LSTM-specific pre-processing steps as described in Appendix D.10. One of these steps excludes articles with 20 or fewer words, meaning that the subsequent analysis is performed on a subset of the full corpus.

Table 8 shows that 887,300 articles, representing 27% of the full dataset, were retained for sentiment analysis. This selection allows us to focus on content that is highly informative and

more directly related to business cycle conditions. Notably, the descriptive statistics reveal an interesting shift: the average number of daily publications for Handelsblatt and dpa is now nearly identical, with Handelsblatt publishing 40 articles per day and dpa publishing 41. Furthermore, the share of dpa articles in the dataset has decreased from 61% (see Table 2) to 47%, while Handelsblatt's share has risen from 17% to 28%.

Table 8: Descriptive Statistics After LSTM-Specific Pre-Processing

Source	per Day	per Month	Total Articles	Share of Total
Welt	12	293	59,468 (36%)	7%
SZ	21	538	160,963 (29%)	18%
Handelsblatt	40	831	248,320 (28%)	28%
dpa	41	1,246	418,549 (21%)	47%
Total	87	2,641	887,300 (27%)	100%

Note: The "per Day" and "per Month" columns correspond to the average number of articles published per day and per month, respectively. The "Total Articles" column represents the total number of articles after LSTM-specific pre-processing has been completed. The percentages in parentheses indicate the proportion of articles that remained following this pre-processing step.

This shift is further illustrated in Figure 4, which highlights Handelsblatt's dominance in the dataset from 1994 to 2001. Given Handelsblatt's focus on business news compared to dpa's broader coverage, including political topics, this change reflects our success in narrowing the dataset to articles more relevant to the business cycle. Moreover, Figure 4 confirms that no unusual spikes in daily article publications occurred, ensuring consistent coverage over time. The noticeable rise in daily publications around the Great Recession further indicates that we are effectively capturing content directly related to the aspect of interest.

While the next phase of our analysis will focus on sentiment predictions for individual articles, we also constructed a daily Overall Business Sentiment Index (OBSI) for each media source and for the entire dataset. Calculated by subtracting the proportion of negative articles from the proportion of no clear tone/positive articles for each day, this index is based solely on sentences related to business cycle and economy. The OBSI combines topic-related sentiments with time-varying weights that capture shifts in topic relevance over time, providing an additional measure to validate our sentiment extraction methodology.

Figure 5 presents the standardized 180-day backward-looking rolling mean of the daily OBSI for each media source and the entire corpus. The sentiment indices for different sources are notably correlated and align with significant economic downturns in Germany, including the

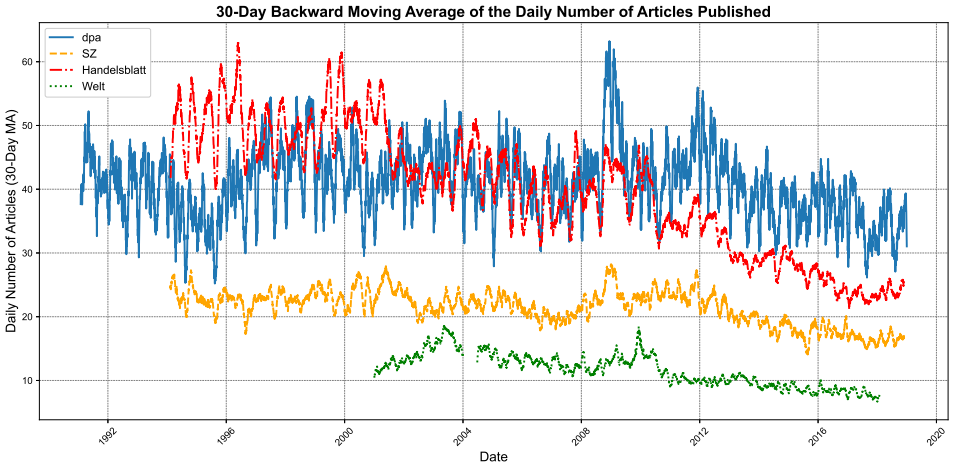


Figure 4: This graph displays the 30-day backward moving average of the daily number of articles published by dpa (blue), SZ (orange), Handelsblatt (red), and Welt (green) in the final dataset, after LSTM-specific pre-processing steps were applied. The Y-axis shows this moving average, and the X-axis corresponds to specific days.

post-reunification recession (1992-1993), the dot-com bubble (2001), the Great Recession (2008-2009), and the European sovereign debt crisis (2011-2013). This further supports the reliability of the sentiment predictions produced by our LSTM model, as the indices track key economic events and reflect overall sentiment trends in the media.

An additional insight from Figure 5 is that the sentiment index for Handelsblatt (in green) shows a stronger reaction to the Great Recession than the other media sources. Since sentiment is averaged across all topics covered by each source, and Handelsblatt contains a higher proportion of content related specifically to business cycle conditions, its more pronounced response is expected. This observation highlights the need to go beyond sentiment analysis alone. To ensure that sentiment is consistently calculated for articles covering relevant topics, we will integrate topic modeling into our analysis, as discussed in the next section.

4 Sign-Adjusted Topics

In this section, we begin by outlining our methodology for extracting news topics from the 887,300 articles retained for analysis and how these topics were integrated with sentiment related to business cycle conditions. We then explain the process of selecting sentiment-adjusted topics for the out-of-sample forecasting exercise. Finally, we discuss the resulting series to illustrate when and why combining topics with sentiment is particularly important.

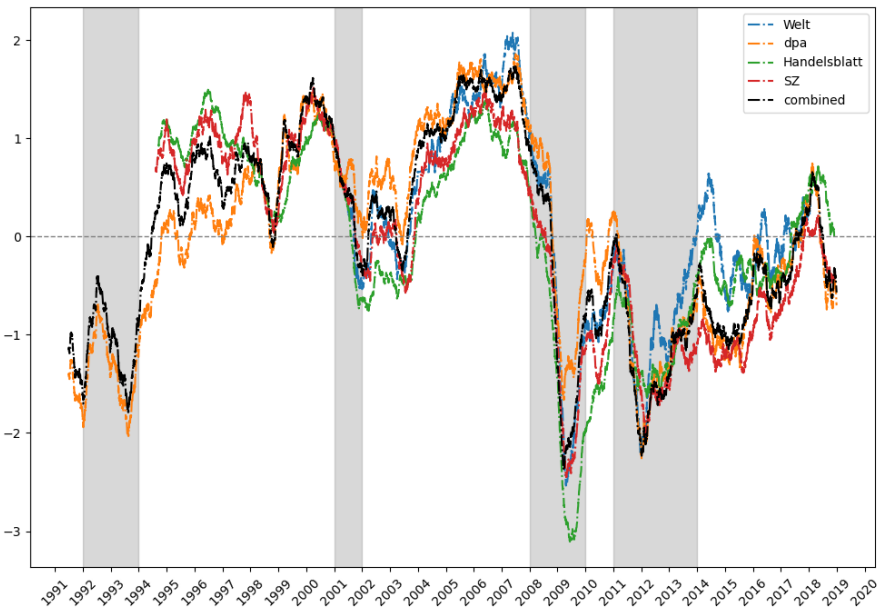


Figure 5: Standardized 180-day backward rolling mean of the daily OBSI for SZ (red), Handelsblatt (green), Welt (blue), dpa (orange), and the whole corpus combined (black). The Y-axis represents the 180-day backward moving average of the daily difference between the proportion of positive/no clear tone articles and negative articles. The X-axis shows the specific day. Shaded areas indicate periods of severe recessions experienced by Germany.

4.1 Latent Dirichlet Allocation

To identify topics within the news articles, we apply the Latent Dirichlet Allocation (LDA) model, originally introduced by Blei et al. (2003). LDA has become increasingly popular in economic forecasting (see e.g., Ellingsen et al., 2022, Bybee et al., 2024) due to its unsupervised nature and highly interpretable output. Specifically, the model estimates the proportion of an article's content dedicated to different topics, which also makes it possible to quantify the share of attention each topic receives on a daily basis. Given the widespread use of LDA in recent economic research, we provide a brief overview of the model in this subsection, focusing on the key aspects relevant to our implementation. For a more detailed discussion, we recommend Hansen et al. (2018), and for a deeper technical explanation, see Blei et al. (2003).

The LDA model can be understood through its generative process, which describes how the observed data—words in documents—are assumed to be generated. Consider a corpus consisting of D documents. Each document is modeled as a mixture of K topics, and each topic is char-

acterized by a distribution over a fixed vocabulary of V unique words. The generative process begins by drawing a distribution β_k over the vocabulary for each topic $k = 1, \ldots, K$, where $\beta_k \sim \text{Dirichlet}(\eta)$, with η being the hyperparameter. This distribution represents the likelihood of each word appearing in a given topic.

Next, for each document d, a distribution over topics θ_d is drawn from a Dirichlet distribution with hyperparameter α, such that $\theta_d \sim \text{Dirichlet}(\alpha)$. This document-specific topic distribution reflects the proportion of attention devoted to each topic within the document.

To generate the words $w_{n,d}$ (where $n = 1, \ldots, N_d$, and N_d is the total number of words in document d), the model first selects a topic for each word by sampling a topic assignment $z_{n,d}$ from a multinomial distribution parameterized by the document's topic distribution θ_d, i.e., $z_{n,d} \sim \text{Multinomial}(\theta_d)$, where $z_{n,d} \in \{1, \ldots, K\}$. After selecting a topic, the word $w_{n,d}$ is drawn from the vocabulary distribution $\beta_{z_{n,d}}$ associated with the assigned topic, i.e., $w_{n,d} \sim \text{Multinomial}(\beta_{z_{n,d}})$. This process is repeated for each word in the document, resulting in the document being represented as a mixture of topics.

In practice, we only observe the words $w_{n,d}$, while the topic assignments $z_{n,d}$, the document-specific topic distributions θ_d, and the topic-specific vocabulary distributions β_k must be inferred from the data. To estimate these latent variables, we used the collapsed Gibbs sampling algorithm as described by Griffiths and Steyvers (2004).

Before applying the LDA model to the 887,300 articles retained for sentiment analysis, we performed several pre-processing steps specific to topic modeling. These include adding collocations, which are combinations of two or three words with specific meanings, into the vocabulary. For example, the former German chancellor's name, Angela Merkel, is treated as a single token rather than two separate words. We also standardized the article texts by converting all words to lowercase, removing stopwords, applying stemming, and excluding terms with low tf-idf scores. For a more detailed explanation of these and other pre-processing steps, refer to Appendix D.13. Unlike in sentiment analysis, where only sentences containing terms related to business cycle conditions were analyzed, LDA was applied to the full text of each article, as the entire article is important for understanding what was discussed on a particular day.

The collapsed Gibbs sampling algorithm starts by randomly initializing the topic assignments $z_{n,d}$, drawing from a uniform distribution. Then, for each word in each document, the topic assignment $z_{n,d}$ is sequentially updated through multinomial sampling based on the following

conditional probability:

$$Pr(z_{n,d} = k|z_{-(n,d)}, w, \alpha, \eta) \propto \frac{m^k_{v,-(n,d)} + \eta}{\sum_v m^k_{v,-(n,d)} + V\eta} \times (m^d_{k,-n} + \alpha), \tag{3}$$

where $z_{-(n,d)}$ refers to all topic assignments except the current one for word $w_{n,d}$, and w represents all the words in the corpus. In this expression, $m^k_{v,-(n,d)}$ counts how many times word $w_{n,d}$ with the token index v has been assigned to topic k across the corpus, excluding the current assignment. Similarly, $m^d_{k,-n}$ denotes how many other words in document d have been assigned to topic k, again excluding the current word.

Intuitively, the first term in the Equation (3) measures how likely word $w_{n,d}$ is to belong to topic k, based on how frequently it has been assigned to that topic across the corpus. The second term indicates how prevalent topic k is within document d, increasing when more words in the document are linked to the same topic.

We estimated the LDA model using the training portion of the corpus (1991 to 2009) to avoid look-ahead bias. The process of updating topic assignments for all words in the training set was repeated 4,500 times, with the last 10 samples saved at a thinning interval of 50. To ensure that the Markov chain had converged, we used perplexity as a standard performance measure in the literature (introduced in Appendix D.14). The hyperparameters α and η were set to $\alpha = 50/K$ and $\eta = 200/V$, as recommended by Griffiths and Steyvers (2004). With these values, only a few topics received high probabilities in a document, while the remaining topics had near-zero probabilities, resulting in a sparse topic distribution.

To determine the optimal number of topics, 10-fold cross-validation was applied. We tested different values of K ranging from 10 to 250 (specifically, 10, 50, 100, 150, 200, and 250). The results suggest that perplexity averaged over 10 folds decreased as the number of topics increased, indicating an improved model fit. However, after reaching 200 topics, the gains became marginal, while the computational demands grew significantly due to the large dataset. For details, see Appendix D.14. Moreover, when K exceeded 200, the topics became too specific and harder to interpret, reducing their usefulness for analysis. Hence, our final LDA model was estimated with 200 topics.

For each of the 10 stored samples, we estimated the document-specific topic proportions using the following equation:

$$\hat{\theta}_d^k = \frac{m_k^d + \alpha}{\sum_{k=1}^{K}(m_k^d + \alpha)}, \tag{4}$$

where m_k^d represents the number of words in document d that were assigned to topic k. The final estimates of $\hat{\theta}_d^k$ were obtained by averaging these values across the 10 samples.

Similarly, the topic-specific vocabulary distributions were computed as:

$$\hat{\beta}_k^v = \frac{m_v^k + \eta}{\sum_{v=1}^{V}(m_v^k + \eta)}, \tag{5}$$

where m_v^k denotes how many times word $w_{n,d}$ was assigned to topic k across the entire training set.

All 200 estimated topic distributions, denoted as $\hat{\beta}_k$, were subjectively labeled using the final sample. To achieve this, we examined the most probable words associated with each distribution as well as the articles with the highest proportion of each topic. For example, the first topic (ID: T0), corresponding to the distribution $\hat{\beta}_1$, was labeled "Automotive Industry" because its most probable words include 'cars', 'vehicle', 'manufacturer', 'passenger car', 'car industry', and 'motor vehicle'. Additionally, articles with the highest proportion of this topic ($\hat{\theta}_d^1$) clearly centered on this theme. In Appendix E, we provide examples of the first 10 estimated topics, including their labels and the most probable words under each distribution, with both the original German stems and their English translations. The full table with labels and most probable words for all 200 topics, as well as the original German stems with their probabilities under each topic distribution, are available in our online repository.[7]

After estimating document-specific topic distributions for the training set, we queried topic assignments $\tilde{z}_{n,\tilde{d}}$ for each document $\tilde{d}$ in the test set (2010 to 2018), similar to Thorsrud (2016). The following distribution was used for re-sampling:

$$Pr\left(\tilde{z}_{n,\tilde{d}} = k \mid \tilde{z}_{-(n,\tilde{d})}, \tilde{w}, \alpha, \eta\right) \propto \hat{\beta}_k^v \left(m_{k,-n}^{\tilde{d}} + \alpha\right), \tag{6}$$

where $\tilde{w}$ represents the words in the test set.

In this step, we used only 100 iterations of the Gibbs sampler because $\hat{\beta}_k^v$ did not need to be re-estimated, as it corresponds to the topic-specific vocabulary distributions from the training

[7]See https://github.com/MashenkaOkuneva/newspaper_analysis/blob/main/topics/Topic_labels.pdf for the labels and most probable words, and https://github.com/MashenkaOkuneva/newspaper_analysis/blob/main/topics/topic_description.csv for the German stems and probabilities.

set. Using these topic assignments, we re-calculated the document-specific topic distributions for each of the 10 samples and then averaged them to obtain the final estimates for each article in the test set.

Our final goal was to generate time series that represent the proportion of attention devoted to each topic on a daily basis. To achieve this, we combined all articles from each day into a single document, then queried topic assignments and re-estimated document-specific topic distributions for the daily documents using the same approach as for the test set.[8] In the next subsection, we explain how these daily topics were combined with sentiment towards business cycle conditions and which topics were selected for the out-of-sample forecasting experiment.

4.2 Selected Sign-Adjusted Topics

To combine the estimated daily topic series with sentiment towards business cycle conditions, we applied a sign-adjustment process, similar to Thorsrud (2016). For each day, we identified the 11 articles with the highest proportion of each topic and assessed their sentiment. If the majority of articles were positive or had no clear tone, we assigned a value of +1 to the topic for that day; if they were negative, we assigned a value of -1. This value was then multiplied by the daily topic proportion to obtain the final series.

While all 200 sign-adjusted topic time series could be included in the out-of-sample forecasting experiment, we found that focusing on a smaller set of topics with stronger correlations to GDP growth improved both model performance and result interpretability. To achieve this, we calculated the correlation between each sign-adjusted topic and the annualized quarterly GDP growth in 34 real-time vintages, spanning from 2010 to 2018. We then selected the 10 topics that most frequently ranked among the top 10 most correlated topics. For example, Topic 27 appeared in the top 10 in all 34 vintages (see Appendix F.1).

To avoid look-ahead bias, we also experimented with the 10 topics that showed the strongest correlations with GDP growth only in the first vintage. While the out-of-sample forecasting results were qualitatively similar to those reported in the paper, we chose to focus on the 10 topics with consistent correlations across all 34 vintages for two reasons. First, these topics also demonstrated high correlations with GDP growth in the first vintage, indicating their importance

[8]The resulting time series for each topic are available in our repository: https://github.com/MashenkaOkuneva/newspaper_analysis/tree/main/topics/topics_plots.

as predictors even at the time of the initial forecast. In fact, the correlations of selected topics with GDP growth in the first vintage and average correlations across all 34 vintages are very similar to each other. For details, refer to Appendix F.1.

More importantly, one of the main strengths of our approach is that each topic comes with a clear narrative, allowing us to combine statistical evidence with economic reasoning when selecting the most relevant topics. For instance, T179, "Financial Crises and Market Regulation", had a strong correlation in the first vintage but was excluded from the final analysis because T27, "Economic Crises and Recessions", already captured a broader range of crises, making T179 redundant. Similarly, we excluded topics like T7, "Culture, Arts, and Literature", even before calculating correlations, as they lacked direct relevance to economic forecasting. In contrast, the selected topics (see Appendix F.1) not only demonstrate strong correlations with GDP growth but are also economically meaningful, discussed over a substantial portion of the sample period, and not overly focused on specific events or concepts.

We now turn to a few of the selected topics to illustrate when the sign-adjustment of daily topic series is particularly important. The first is T27, which has the highest correlation with GDP growth. We labeled it "Economic Crises and Recessions" because the most probable words under this topic are 'crisis', 'recession', and 'economic crisis' (see Appendix F.1). Figure 6 shows the 180-day backward rolling mean of the topic and its sign-adjusted version. The original topic series (in black) rises during all major recessions in Germany (shaded in gray), as well as during the Asian financial crisis of 1998-1999. Reassuringly, the topic reacts especially strongly to the 2008-2009 financial crisis, the most severe in recent history, which supports the validity of the series. The sign-adjusted version of this topic (in blue) mostly mirrors the original, but with the sign reversed, except during the post-reunification crisis, where sentiment was mixed. This pattern is expected for a topic centered on economic crises, as most articles discussing it tend to be negative. While our sign-adjustment process produces meaningful results for this topic, its added value is less clear, as the topic itself already conveys a strong sentiment.

In contrast, topics like T52, "German Automobile Industry and Major Manufacturers" (see Figure 7), highlight the importance of sentiment-adjustment. While the original topic series (in black) consistently accounts for about 1% of all news from 2006 onwards, it is the sign-adjusted series (in blue) that drops sharply during the financial crisis. Without this adjustment, it would not be possible to see how severely the crisis impacted one of Germany's most important

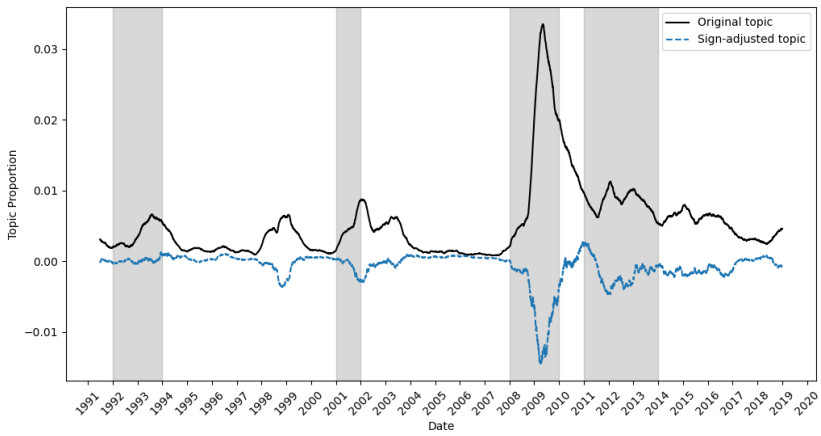


Figure 6: 180-day backward rolling mean of daily topic series (black) and sign-adjusted topic series (blue) for Topic 27 ("Economic Crises and Recessions"). The Y-axis represents the 180-day backward moving average of the daily topic proportion or the sentiment-adjusted topic proportion. The X-axis shows the specific day. Shaded areas indicate periods of severe recessions in Germany.

industries.

Other selected topics cover banking, mergers and acquisitions, job cuts, investments, financial and economic performance, as well as market reactions to news. The most probable words associated with each topic, along with their labels, are presented in Appendix F.1, while Appendix F.2 provides visualizations of the original and sign-adjusted series. Overall, the behavior of these topics and their sign-adjusted counterparts further supports the validity of our methodology, suggesting that they are meaningful predictors for the out-of-sample forecasting experiment.

We also conducted several robustness checks to ensure the reliability of our results. First, we examined the effect of adjusting the sign using 9 and 7 articles instead of 11. Second, we explored an alternative method for adjusting topic proportions. Instead of using a majority vote for sign adjustment, we calculated the average sentiment across the 11 articles. Our methodology proves robust to the considered modifications (see Appendix F.3 for details) and produces time series that are strongly correlated with the variable we intend to forecast.

5 Forecast Encompassing Analysis

Before proceeding with the out-of-sample forecasting experiment, we conducted a simple encompassing test similar to the approach used by Rambaccussing and Kwiatkowski (2020). The

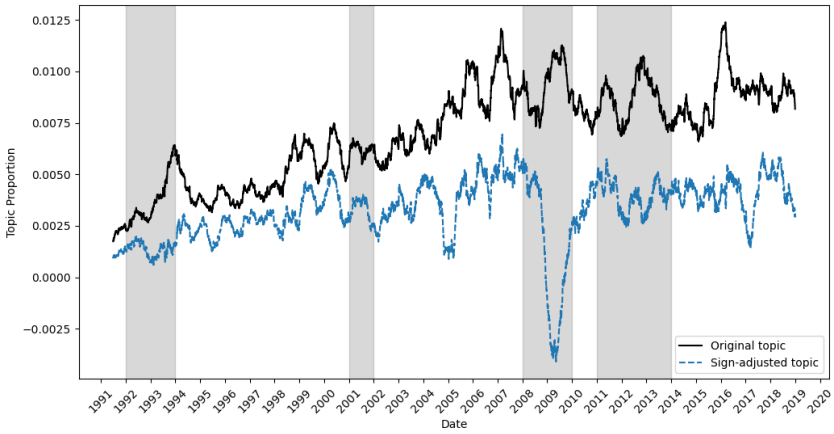


Figure 7: 180-day backward rolling mean of daily topic series (black) and sign-adjusted topic series (blue) for Topic 52 ("German Automobile Industry and Major Manufacturers"). The Y-axis represents the 180-day backward moving average of the daily topic proportion or the sentiment-adjusted topic proportion. The X-axis shows the specific day. Shaded areas indicate periods of severe recessions in Germany.

goal of this analysis is to determine whether the selected sign-adjusted topics contain valuable information that complements professional forecasts. For this purpose, we use the Reuters Poll of German GDP forecasts, which collects projections from approximately 20 professionals representing private firms and research institutes. Respondents in the survey are asked to provide a backcast for the previous quarter, a nowcast for the current quarter, and forecasts for quarters $t+1$ through $t+6$. The survey is conducted four times a year during the first month of each quarter and covers the period from 2006 to 2018. In our analysis, we focus on short-term forecasting and therefore exclude forecasts for $t+3$ and beyond.

From 2006 through the first half of 2014, forecasts were gathered during the first week of the quarter. Afterward, the collection period shifted to the end of the second week. The forecasts are expressed as quarter-on-quarter (q/q) GDP growth rates, and we use the median forecast from each survey round in our analysis.

The encompassing test is based on the following regression model:

$$\epsilon_{t+h|t} = \alpha + \beta f_{t+h|t} + \gamma s_t + u_{t+h}, \tag{7}$$

where $\epsilon_{t+h|t}$ is the conditional forecast error, defined as the difference between the actual q/q GDP growth rate at time $t+h$ and the median professional forecast $f_{t+h|t}$ made at time t. The

variable s_t represents the sign-adjusted topic value. A significant γ would indicate that the topic contains unique information that improves the point forecast.

For GDP, we use the first release data from the Deutsche Bundesbank's real-time database. The analysis is conducted for four different forecast horizons: backcasts ($h = -1$), nowcasts ($h = 0$), 1-step-ahead forecasts ($h = 1$), and 2-step-ahead forecasts ($h = 2$). The sample includes 51 quarters for $h = -1, 1$, and 2, and 52 quarters for $h = 0$. This difference arises because the survey was not conducted in the fourth quarter of 2005, resulting in a missing 1-step-ahead forecast for Q1 2006 and a 2-step-ahead forecast for Q2 2006. Additionally, the backcast for Q4 2018 is missing because the last available survey was conducted in that quarter.

Before performing the test, we pre-processed our daily sign-adjusted topics. First, we applied a 30-day backward-looking moving average filter, as in Thorsrud (2020), to reduce the noise inherent in the high-frequency data. Next, we used a biweight filter with a bandwidth of 1,200 days to eliminate very low-frequency variation. This filter identifies a long-run trend and provides greater flexibility than a linear filter (Stock and Watson, 2016). We then aggregated the daily series into quarterly values by calculating the mean and standardized both the topics and GDP growth. When estimating Equation (7), we use Newey-West heteroscedasticity and autocorrelation robust (HAC) standard errors, with the lag selected automatically as in Newey and West (1994).

ID	**Correlation**	**Label**
T11	0.55	Mergers and Acquisitions
T27	0.67	Economic Crises and Recessions
T52	0.62	German Automobile Industry and Major Manufacturers
T74	0.52	Concerns about Economic Bubbles and Recessions
T77	0.49	Private Investment
T81	0.71	Corporate Restructuring and Job Cuts in Germany
T100	0.59	Market Reactions to News
T127	0.60	Major Banks and Investment Banking
T131	0.42	German Investments in Emerging Markets
T138	0.73	Financial and Economic Performance

Table 9: Contemporaneous correlations between the first release of GDP growth and selected sign-adjusted topics for the period 2006–2018.

The first interesting result of our analysis is that, as seen in Table 9, the contemporaneous correlations between the sign-adjusted topics and GDP growth, calculated over the 2006–2018 period, are even stronger than those calculated for the shorter period in the previous section. For example, the correlation for T81 reaches 0.71, compared to the previously reported 0.54,

highlighting the predictive potential of these text-based indicators.

		T11	T27	T52	T74	T77	T81	T100	T127	T131	T138
h=-1	Intercept	-0.12***	-0.09***	-0.12***	-0.12***	-0.11***	-0.06**	-0.13*	-0.12**	-0.11***	-0.09**
	SPF	0.43***	0.33***	0.45***	0.44***	0.41***	0.24***	0.46***	0.43***	0.4***	0.33**
	Topic	-0.04	0.04	-0.04	-0.04	-0.02	0.11***	-0.06	-0.03	-0.01	0.04
h=0	Intercept	-0.26*	-0.14*	-0.2**	-0.27*	-0.3**	-0.09	-0.23*	-0.22	-0.35*	-0.06**
	SPF	0.63*	0.26	0.45*	0.67*	0.75**	0.13	0.53*	0.5	0.89*	0.04
	Topic	0.23**	0.32***	0.26***	0.21***	0.16**	0.39***	0.27***	0.28**	0.09	0.41***
h=1	Intercept	0.26	0.36***	0.34***	0.17	0.24	0.48***	0.1	0.17	0.23	0.37**
	SPF	-0.89**	-1.2***	-1.13***	-0.65	-0.87**	-1.52***	-0.46**	-0.65	-0.83**	-1.22***
	Topic	0.47**	0.57***	0.53***	0.42**	0.39*	0.63***	0.48***	0.5**	0.34*	0.62***
h=2	Intercept	0.34	0.32**	0.41***	0.17	0.4*	0.54***	0.06	0.22	0.46**	0.25
	SPF	-1.12**	-1.11***	-1.35***	-0.7	-1.31***	-1.64***	-0.43*	-0.84	-1.45***	-0.9***
	Topic	0.45**	0.55***	0.52***	0.44**	0.4**	0.6***	0.5***	0.5**	0.35*	0.6***

Table 10: Encompassing regressions for each forecast horizon, evaluating whether selected sign-adjusted topics provide additional information beyond the professional forecasts from the Reuters Poll. Significance levels are denoted as follows: *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Moreover, the results of the encompassing test, presented in Table 10, are encouraging. We find that γ is significant for all topics at horizons $h = 1$ and $h = 2$, for nearly all topics at the nowcast horizon (except for T131, "Investments in Emerging Markets"), and only for T81, "Job Cuts", at the backcast horizon. This suggests that most of the selected topics individually capture valuable information not already reflected in the professional forecasts, although improving backcasts is particularly challenging.

Overall, our sign-adjusted topics are closely linked to the business cycle and carry important information beyond professional forecasts. Next, we investigate whether all the topics together can improve GDP growth predictions in the out-of-sample exercise.

6 Out-of-Sample Forecasting Experiment

In the final section, we use the selected sign-adjusted topics, which were found to be economically meaningful and highly correlated with GDP growth, in an out-of-sample forecasting experiment. We begin by describing our main experimental setup using a Dynamic Factor Model (DFM henceforth). Following this, we outline the implementation of the unrestricted MIDAS (Mixed Data Sampling) model, which serves as a benchmark. Finally, we present and discuss the empirical results for both models to assess whether our text-based indicators can improve the accuracy of GDP growth forecasts.

6.1 Dynamic Factor Model

The main model we use to forecast GDP growth is the DFM proposed by Bańbura et al. (2011). We selected this model for its ability to handle data of different frequencies, such as daily, monthly, and quarterly, without requiring us to aggregate daily sign-adjusted topics or financial series. Daily data, which is valuable for capturing timely shifts in market expectations, might lose some of its informational content when transformed to a lower frequency.

The authors formulate the model for daily data as follows. Let $y_t^D = [y_{1,t}, y_{2,t}, \ldots, y_{n_D,t}]'$ denote a stationary n_D-dimensional vector process standardized to have a mean of 0 and unit variance. The vector y_t^D adheres to the following factor model representation:

$$y_t^D = \Lambda_D f_t + \varepsilon_t^D, \quad \varepsilon_t^D \sim \text{i.i.d.} N(0, \text{diag}(\sigma_1^2, \ldots, \sigma_{n_D}^2)), \tag{8}$$

where f_t represents an $r \times 1$ vector of unobserved common factors, and ε_t^D denotes the vector of idiosyncratic components. The matrix Λ_D, which is of size $n_D \times r$, contains the factor loadings for the daily series. In this context, $\Lambda_{D,i} f_t$ is referred to as the common component of $y_{i,t}$. The underlying premise is that the time series included in the model exhibit significant co-movement, enabling their behavior to be captured by a few common factors. Even though the errors are assumed to be both serially and cross-sectionally uncorrelated, maximum likelihood estimates of the model, when applied to a large number of daily series, remain robust to mild forms of misspecification (Doz et al., 2012).

Additionally, it is assumed that the factors f_t follow a VAR process of order p:

$$f_t = A_1 f_{t-1} + \ldots + A_p f_{t-p} + u_t, \quad u_t \sim \text{i.i.d.} N(0, Q), \tag{9}$$

where $A_1, \ldots, A_p$ are $r \times r$ matrices of autoregressive coefficients. Exploiting the dynamics of the factors can be particularly important when dealing with a substantial amount of missing observations.

To save space, we present the mixed-frequency structure of the DFM in Appendix G. It is also worth noting that the entire model is estimated within a state-space framework using the Kalman filter and Maximum Likelihood (ML). This estimation is facilitated by the EM (Expectation Maximization) algorithm. The Kalman filter is effective at addressing issues common to a nowcasting dataset: the non-synchronicity of data releases causing missing data at the end of

the sample (known as the "ragged" edge problem) and data coming in at different frequencies. A detailed explanation of the EM algorithm can be found in Bańbura and Modugno (2014).

The main goal of our forecasting experiment is to assess whether incorporating text-based series can improve GDP growth forecasts compared to a model using only traditional economic and financial data. To achieve this, we estimated separate models for text data and hard data, a model that integrates both sources, and a combined forecast using predictions from the text-only and hard-data-only models.

Our real-time dataset consists of 10 daily sign-adjusted topics, 5 daily financial indicators, 12 monthly economic series, and the target variable—annualized quarterly GDP growth. The financial indicators include a stock index, exchange rates, and government bond yields. The monthly series cover various aspects of the economy: hours worked for the labor market, the consumer price index and producer price index for prices, and industrial production, new orders, and turnover for real activity. Detailed information on the hard data can be found in Appendix H. Most of these series are sourced from the Deutsche Bundesbank's real-time database.

Before estimating the DFM, we pre-processed the data to ensure that all input series were suitable for forecasting. For the daily sign-adjusted topics, we used the same procedure as for the encompassing test: applying a 30-day backward-looking moving average filter and detrending with a biweight filter, with the difference that here all transformations are performed in real time. For the economic and financial series, we transformed the data to ensure stationarity. Specifically, we took the first difference for government bond yields and the first difference of the logarithm for the other series. Finally, all the series were standardized.

Regarding the design of the forecasting experiment, we fixed the timing of when forecasts were made, starting in the first quarter of 2010, and produced forecasts 30, 60, and 90 days after the beginning of each quarter. For example, the initial forecast was made on January 30, 2010, resulting in an estimation period from January 1, 1992, to January 30, 2010. Since the actual GDP figure for the previous quarter (Q4 2009) was not available on this date, we generated a backcast ($h = -1$) for December 31, 2009, a nowcast ($h = 0$) for March 31, 2010, as well as 1-step-ahead ($h = 1$) and 2-step-ahead ($h = 2$) forecasts. When producing forecasts 60 and 90 days into the quarter, backcasts were not generated because the actual GDP data had been released by then. This approach resulted in 34 short-term forecasts, with backcasts covering 2009 Q4 to 2018 Q1 and nowcasts spanning from 2010 Q1 to 2018 Q2.

For each series, we used the real-time vintage reflecting the data available to forecasters at the time the forecast was made, and all transformations to the series were performed in real time as well. The sample period began on January 1, 1992, providing enough observations for the 30-day moving average filter and aligning with the start of a new year. The model was re-estimated using an expanding estimation window as new data became available.

We experimented with different values for the number of extracted daily factors r and the lag order p of the VAR process in Equation 9. The final results are presented for models with 1 factor for text-only and hard-data-only specifications, and 2 factors for the model that combines both data sources. The lag order was set to $p = 10$. These choices were guided by better out-of-sample forecasting performance, likely attributable to the lower variance of parsimonious models. Additionally, we excluded the 5-year federal notes yield series from the final specifications, as its inclusion resulted in less accurate forecasts.

6.2 Unrestricted MIDAS

In this paper, we treat the DFM as our primary model, due to its appealing features, such as the ability to incorporate daily data, handle mixed-frequency series, and accommodate various patterns of missing observations. This method has also been successfully applied by Ashwin et al. (2024) to nowcast Euro area GDP using sentiment indices. However, another common approach in this literature is to estimate forecasting models with daily text series aggregated to a monthly frequency (see, e.g., Ellingsen et al., 2022 and van Dijk and de Winter, 2023). To explore this alternative, we selected the unrestricted MIDAS model (Foroni et al., 2015), a technique recognized for its simplicity and strong empirical performance.

The MIDAS model serves as our benchmark for several reasons: it enables us to evaluate the robustness and competitiveness of forecasts produced by the daily DFM, facilitates direct comparison with the existing literature, and, most importantly, allows us to examine whether using a different model at a different frequency alters the answer to the central question—does text data improve forecasts based on hard data alone? To maintain simplicity, we restricted the MIDAS analysis to variables consistently available over the entire sample period, avoiding issues related to missing data but potentially omitting some relevant information.

The MIDAS model we consider is represented by the following equation for $h = -1$:

$$y_{t+h} = \alpha_{h+1} + \sum_{i=0}^{K-1} \beta_{i,h+1} \cdot x_{t+h-i/m} + \sum_{j=1}^{K} \theta_{j-1,h+1} \cdot z_{t+h-j/m} + \epsilon_{t+h}, \tag{10}$$

where y_{t+h} is the low-frequency GDP growth rate being forecasted, x corresponds to the monthly text series, z represents a monthly economic or financial indicator, and K is the number of most recent monthly values included in the model. The index t denotes the quarter in which the forecast is made, and $m = 3$, as there are three months in each quarter.

For instance, if the backcast is generated on January 30, 2010, y_{t+h} corresponds to the GDP growth for Q4 2009. To construct this backcast, the most recent K-months of observations for both the text and hard data series are used. If $K = 1$, the model incorporates the December 2009 value for the text series ($x_{t+h-\frac{0}{3}}$) and the November 2009 value for the hard series ($z_{t+h-\frac{1}{3}}$), as the final monthly data point for some hard indicators in Q4 2009 is not yet released.

For other forecasting horizons, the model depends on the point within the quarter when the forecast is made: 30 days into the quarter ($v = 1/3$), 60 days ($v = 2/3$), or 90 days ($v = 1$). The model is given by:

$$y^v_{t+h} = \alpha_{h+1} + \sum_{i=0}^{K-1} \beta_{i,h+1} \cdot x_{t-(1-v)-i/m} + \sum_{j=2}^{K+1} \theta_{j-2,h+1} \cdot z_{t-(1-v)-j/m} + \epsilon^v_{t+h}, \tag{11}$$

where y^v_{t+h} represents GDP growth for quarter $t+h$, forecasted after observing v-th share of a quarter. For example, if $K = 1$ and a nowcast ($h = 0$) is made 30 days into Q1 2010 ($v = 1/3$), the model uses the January 2010 value for the text series ($x_{t-(1-1/3)-\frac{0}{3}}$) and the November 2009 value for the hard data ($z_{t-(1-1/3)-\frac{2}{3}}$), since not all hard series are yet available for December 2009 and January 2010.

As discussed above, in our MIDAS model, we only used series that were available throughout the entire sample period. Moreover, for the hard data, we treated all variables uniformly by using the most recent observation in each vintage where all economic and financial indicators were simultaneously reported. This simplified setup was chosen because the MIDAS model serves solely as a benchmark in this paper, whereas the main DFM model can handle variables with missing observations, both at the beginning and end of the sample. Consequently, out of the five daily financial indicators, we included three: the stock index and the 5-year and 10-year government bond yields, as exchange rates were available only from 1993. Similarly, economic indicators were excluded from the real-time vintages if they had gaps in their historical

coverage. For example, in the first vintage, the hours worked in manufacturing series was omitted because its most recent observation was from December 2008, whereas other series extended up to November 2009.

Additionally, smoothed and de-trended daily sign-adjusted topics were converted to a monthly frequency by averaging. For the DAX index, monthly log returns were calculated, and daily bond yields were averaged to produce their monthly counterparts. In contrast, the DFM model works directly with high-frequency daily data.

We estimated six different specifications, varying K from 1 to 6. Given the large number of regressors introduced by including multiple lags, we applied several techniques designed to handle high-dimensional settings and capture potential non-linear relationships between the predictors and the dependent variable. Specifically, we used LASSO (Tibshirani, 1996), Ridge regression (Hoerl & Kennard, 1970), Random Forests (RF) (Breiman, 2001), and PCA with factors estimated via the EM algorithm (Stock & Watson, 2002). To save space, explanations of each algorithm are provided in Appendix I, while only the main details are discussed here.

Before performing LASSO, Ridge, and PCA, we standardized the data, as these methods are sensitive to the scale of the input variables. For LASSO and Ridge, the tuning parameters controlling the degree of shrinkage were selected via cross-validation. For the RF models, we generated 500 bootstrap samples, and at each tree split, a random subset of one-third of the predictors was used to identify the optimal split, following standard practice. In the case of PCA, we experimented with different strategies to determine the number of factors, including the information criterion proposed by J. Bai and Ng (2002) and manually setting the number of factors to 1 or 2. Consistent with the DFM, parsimonious models performed best, with one factor for the text-only and hard-data-only specifications and two factors for the model integrating both sources. The resulting factors replaced the original variables in the MIDAS regressions, which were then estimated using OLS. For models that combined hard and text data, the factors were extracted jointly from both sources.

The design of our forecasting experiment mirrors that of the DFM model. For each forecasting horizon, we estimated 24 MIDAS specifications based on hard data, 24 using text data only, and 24 that integrated both sources. However, our main results focus on Ridge regression with $K = 3$, for several reasons.

First, in our analysis, Ridge regression consistently demonstrated robust and strong per-

formance across different data types and forecasting horizons. Second, this finding aligns with Ashwin et al. (2024), who report that Ridge regression achieves the most significant improvements in forecast accuracy when using text data, particularly during stable periods. Third, Eickmeier and Ng (2011) and Richardson et al. (2021) show that Ridge regression, among other shrinkage methods, can outperform factor-based methods in forecasting GDP growth, making it a strong alternative to the DFM.

Forecasts were generated for four horizons at 30 days into the quarter, and for three horizons at 60 and 90 days, beginning from Q1 2010. The entire exercise used real-time data, with the sample starting in 1992. All forecasts were produced through a direct forecasting approach and re-estimated with an expanding window.

6.3 Empirical results

In the final subsection, we present the results of our out-of-sample forecasting experiment, with the main goal of evaluating whether the selected sign-adjusted topics can improve forecasts based solely on hard data across four different horizons. To this end, Tables 11 and 12 report the forecasting performance of DFM and MIDAS models estimated with text data only, hard data only, and models that integrate both sources. We also include the results of forecast combinations derived from these text-only and hard-only models. Forecast accuracy is measured using root mean square forecast error (RMSFE), and the tables show relative RMSFEs compared to two standard benchmarks in the forecasting literature: the AR(1) model and the Survey of Professional Forecasters (SPF), represented here by the Reuters Poll.

Similar to van Dijk and de Winter (2023), we employ both economic and statistical approaches to compare RMSFEs. Economic importance is assessed by calculating the percentage difference in RMSFEs between two models, with entries in bold indicating at least a 5% improvement relative to the baseline. Statistical significance is determined using the one-sided Diebold-Mariano (DM; Diebold and Mariano, 1995) test with Newey-West standard errors, and significance levels are marked by asterisks.

The AR(1) benchmark is estimated using a direct multistep approach with an expanding window. The MIDAS models reported in the tables are based on Ridge regression with $K = 3$. As demonstrated in Appendix J, the results remain qualitatively similar when using the best-performing specifications. MIDAS with monthly data serves as a benchmark to assess whether

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Only text										
DFM	**0.83**	**0.91**	**0.85**	**0.84**	**0.90**	1.02	**0.82***	0.99	0.96*	0.97
MIDAS	**0.87**	**0.82***	**0.86**	**0.86**	1.11	0.98	1.05	1.00	1.08	1.06
Only hard data										
DFM	**0.65**	1.13	0.97	**0.85**	1.00	0.96	0.96	1.00	0.99	0.99
MIDAS	**0.74**	0.96	**0.78**	**0.73**	1.01	1.06	0.97	0.99	1.17	1.09
Text and hard data										
DFM	**0.67**	**0.89**	**0.83**	**0.78**	**0.88**	**0.92**	**0.70***	0.98	**0.95***	0.97
MIDAS	**0.74**	**0.85**	**0.87**	1.02	1.52	1.00	**0.91**	1.03	1.17	1.42
Forecast combination (optimal weights)										
DFM	**0.59***	**0.82***	**0.76**	**0.70**	**0.89**	**0.94**	**0.76****	0.99	0.96*	0.96
MIDAS	**0.67**	**0.80***	**0.70**	**0.71***	1.01	0.97	0.96	0.96	1.05	1.03
Forecast combination (equal weights)										
DFM	**0.61***	**0.84***	**0.76**	**0.70**	**0.91***	**0.94**	**0.78****	0.99	0.97**	0.97
MIDAS	**0.68**	**0.82***	**0.70***	**0.72***	1.03	0.97	0.97	0.96	1.06	1.03

Table 11: Relative RMSFE Scores: DFM and MIDAS Models vs AR(1). This table presents relative RMSFEs for DFM and MIDAS models estimated using text data only, hard data only, and models integrating both sources, with MIDAS results based on Ridge regression with $K = 3$. Relative RMSFEs for forecast combinations of hard-only and text-only models using optimal and equal weights are also included. All values are expressed relative to the RMSFE of the AR(1) benchmark. Bold entries indicate RMSFEs at least 5% lower than that of the AR(1). Asterisks denote statistical significance based on one-sided DM test (* 10%, ** 5%, *** 1%).

the success of text data in our forecasting experiment depends on the model choice and its frequency.

For the forecast combinations, we calculate optimal weights that minimize the mean squared error (MSE) of the combined forecast over the entire evaluation period (34 quarters) for each horizon. The weights are constrained to be non-negative and sum to unity, providing a reference for the ex-post contribution of the text-based data. Since this approach introduces a look-ahead bias, we also report results for combinations with equal weights for comparison.

Based on Table 11, the DFM model using only text data outperforms the AR(1) benchmark for backcasts, nowcasts, and 1-step-ahead forecasts when generated 30 and 90 days into the quarter, as well as for 2-step-ahead forecasts at 60 days. However, these differences are statistically significant only for 1-step-ahead forecasts made 90 days into the quarter and for 2-step-ahead forecasts. Given that the improvements for backcasts and nowcasts reach up to 17%, the lack of statistical significance may be due to the limited sample size used for evaluation. When com-

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Only text										
DFM	1.32	1.02	1.02	1.00	**0.89**	0.97	**0.78***	**0.94***	**0.92****	**0.93**
MIDAS	1.39	**0.92**	1.03	1.03	1.11	**0.94**	1.00	0.96	1.03	1.02
Only hard data										
DFM	1.04	1.26	1.16	1.01	1.00	**0.91**	**0.91**	0.96	**0.95***	**0.95***
MIDAS	1.19	1.08	**0.94**	**0.88**	1.00	1.01	**0.92**	**0.95**	1.12	1.05
Text and hard data										
DFM	1.07	1.00	1.00	**0.93**	**0.88**	**0.87**	**0.67****	**0.94***	**0.92****	**0.93**
MIDAS	1.18	**0.95**	1.04	1.22	1.51	**0.95**	**0.87**	0.99	1.12	1.36
Forecast combination (optimal weights)										
DFM	**0.95**	**0.92**	**0.91**	**0.83**	**0.89***	**0.89****	**0.73****	**0.94***	**0.92****	**0.93***
MIDAS	1.08	**0.90**	**0.83**	**0.84**	1.00	**0.92**	**0.91***	**0.92****	1.01	0.98
Forecast combination (equal weights)										
DFM	0.98	**0.94**	**0.91**	**0.83**	**0.91****	**0.89****	**0.74****	**0.95***	**0.93****	**0.93****
MIDAS	1.09	**0.92**	**0.84**	**0.86**	1.02	**0.93**	**0.93***	**0.92****	1.02	0.99

Table 12: Relative RMSFE Scores: DFM and MIDAS Models vs SPF. This table presents relative RMSFEs for DFM and MIDAS models estimated using text data only, hard data only, and models integrating both sources, with MIDAS results based on Ridge regression with $K = 3$. Relative RMSFEs for forecast combinations of hard-only and text-only models using optimal and equal weights are also included. All values are expressed relative to the RMSFE of the SPF benchmark (Reuters Poll). Bold entries indicate RMSFEs at least 5% lower than that of the SPF. Asterisks denote statistical significance based on one-sided DM test (* 10%, ** 5%, *** 1%).

pared to the SPF (Table 12), the DFM with text data shows higher errors for backcasts, performs similarly for nowcasts, and achieves lower forecasting errors for 1-step-ahead and 2-step-ahead forecasts, with the latter differences being statistically significant in three cases. A comparison of DFM and MIDAS models using text data alone suggests that the DFM generally performs better for backcasts, 1-step-ahead, and 2-step-ahead forecasts. However, the results for nowcasts are mixed, making it difficult to conclude that a model using daily sign-adjusted topics directly always outperforms an approach that aggregates them to a monthly frequency.

The DFM model estimated with hard data alone achieves lower errors than the AR(1) benchmark and is comparable to the SPF for backcasts, while also showing a clear improvement over the text-based DFM for this horizon. However, for nowcasts, we observe the opposite pattern: the hard-data DFM underperforms relative to the text-based DFM and has higher forecast errors compared to the MIDAS model with hard data, which only includes series consistently available over the entire sample period. This suggests that the DFM, which relies on the daily factor extracted from financial data, may be less effective for nowcasting than a model that aggre-

gates daily series into a monthly frequency. A potential direction for future research would be to conduct a more direct comparison between DFM models formulated at daily and monthly frequencies to better understand the implications of these choices for forecast accuracy.

While text-only and hard-data-only models provide valuable insights on their own, the more relevant question for this study is how well the models perform when both daily sign-adjusted topics and traditional economic and financial indicators are integrated. The DFM, which combines both sources, is estimated with two factors. For backcasts, it achieves an RMSFE that is 33% lower than that of the AR(1) benchmark. The model also consistently outperforms the AR(1) across all horizons, with statistically significant improvements for 1-step-ahead forecasts at 90 days and 2-step-ahead forecasts at 60 days. Compared to the SPF, the integrated DFM has higher errors for backcasts, performs similarly for nowcasts at 30 and 60 days, but shows better accuracy for nowcasts at 90 days and significantly lower errors for 1-step-ahead and 2-step-ahead forecasts.

It is important to note that the Reuters Poll is conducted at the beginning of the first month of the quarter, which means professional forecasters had access to less information than what is used in our models, especially when forecasting at 60 and 90 days into the quarter. Nevertheless, the results are promising: overall, the integrated model provides a clear improvement over the AR(1) across all horizons and surpasses the SPF for 1-step-ahead and 2-step-ahead forecasts. Furthermore, it consistently outperforms both the text-only and hard-data-only DFM models. When compared to the MIDAS model that incorporates both data sources, the integrated DFM demonstrates overall superior performance. This indicates that the DFM's ability to directly handle daily data and efficiently manage missing observations offers a distinct advantage.

The second approach to combining our data sources is through linear forecast combinations based on the text-only and hard-data-only models. We explore two types of combinations: optimal weights, determined ex-post, and equal weights, which are known ex-ante. For the DFM combinations, we observe improvements over the AR(1) benchmark across all horizons, with many of these differences being statistically significant. The gains are particularly notable for backcasts, where the optimal-weighted combination achieves a 41% reduction in forecast error. Furthermore, consistent with Ellingsen et al. (2022) and van Dijk and de Winter (2023), we see that forecast accuracy improves as more hard data becomes available. For example, the forecast error for nowcasts produced 30 days into the quarter is higher than for those generated at 90

days. This is further reflected in the declining optimal weight assigned to the text-based model in the DFM combinations, which drops from 66% for the nowcast at 30 days to 51% at 90 days (see Appendix J).

In addition, the combined DFM forecasts achieve lower errors than the SPF across all horizons, though these differences are statistically significant only for 1-step-ahead and 2-step-ahead forecasts. When compared to the MIDAS combinations using text-only and hard-data-only models, the DFM forecast combinations perform better for backcasts, 1-step-ahead, and 2-step-ahead forecasts while delivering similar performance for nowcasts. This suggests that, as with earlier findings, the relative advantage of the DFM, which relies on daily factors, depends on the specific forecasting horizon.

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Optimal Weights										
DFM	**0.91***	**0.73*****	**0.78****	**0.82****	**0.89***	0.97	**0.79****	0.99	0.97	0.97
MIDAS	**0.91**	**0.83***	**0.89**	0.96	1.00	**0.91**	0.99	0.97*	**0.90**	**0.94**
Equal Weights										
DFM	**0.94**	**0.75*****	**0.79****	**0.82****	**0.91****	0.98	**0.81*****	0.99	0.98*	0.98
MIDAS	**0.92**	**0.85****	**0.89**	0.98	1.02	**0.91**	1.00	0.97	**0.91**	**0.94***

Table 13: Relative RMSFE Scores: Forecast Combinations vs Hard-Data Models. This table presents relative RMSFEs for forecast combinations of DFM and MIDAS hard-only and text-only models using both optimal and equal weights, with MIDAS results based on Ridge regression with $K = 3$. All values are reported relative to the RMSFEs of the models estimated using hard data only. Bold entries indicate RMSFEs that are at least 5% lower than those of the hard-only models. Asterisks denote statistical significance based on one-sided DM tests (* 10%, ** 5%, *** 1%).

Finally, we address our main research question: can text data improve forecasts that rely solely on hard data? To answer this, we formally compare the forecast combinations to models that use only hard data (Table 13). We focus on combinations rather than models that directly integrate both sources, as they show better performance for backcasts and nowcasts. Our results indicate that combining DFM forecasts based on hard data and text data significantly improves upon the DFM using only hard data across all horizons, with the most notable gains observed for nowcasts. This aligns with findings in the existing literature, which suggest that text data can provide valuable information, especially when recent hard data is not yet available. A similar pattern is observed for the MIDAS model, though the benefits are more limited for nowcasts and 1-step-ahead forecasts.

Overall, our findings suggest that text data, represented here by sign-adjusted topics, can indeed enhance short-term economic forecasts—a conclusion further supported by our encompassing tests. Moreover, these text-based indicators offer a unique advantage: they are highly interpretable and provide a contextual narrative that complements traditional hard data.

7 Conclusion

This paper explores whether information extracted from German news media can improve short-term economic forecasts, specifically focusing on GDP growth. To achieve this, we analyzed a large dataset of news articles spanning from 1991 to 2018. We thoroughly pre-processed the corpus to retain only economically relevant content and to ensure it was in a format suitable for the machine learning methods applied in this study. The inclusion of the dpa news agency, which provides daily coverage including weekends, ensured that our dataset had no missing observations, a feature not typically available for standard economic and financial indicators.

To derive economically meaningful information from high-dimensional text data, we applied a combination of sentiment analysis and topic modeling. For sentiment extraction, we used a supervised learning approach based on a training set from Media Tenor International, a research institute that relies on professional coders to perform aspect-based sentiment annotation of news articles. We concentrated specifically on articles labeled with respect to business cycle conditions, as this type of sentiment is more likely to be directly relevant for economic forecasting. When training our LSTM model and applying it to the main corpus, we targeted sentences containing business cycle-related terms, which were identified through a word embeddings approach. This aspect-based sentiment extraction is one of the main contributions of our study, as it goes beyond general sentiment measures and offers a more focused indicator of economic dynamics. Importantly, our approach outperformed a lexicon-based method based on the Loughran-McDonald dictionary, which is often applied in macroeconomic forecasting. Furthermore, we constructed a daily sentiment index that showed clear responses to all major recessions in Germany, further validating our methodology.

For topic modeling, we applied the LDA algorithm and adjusted the sign of the extracted topics using our business cycle sentiment. From the initial 200 sign-adjusted topics, we selected the 10 most strongly correlated with GDP growth, allowing us to focus on topics that were both

economically relevant and easily interpretable. One notable result was the clear improvement provided by sentiment adjustment for certain topics, such as the one related to the automotive industry. While raw topic proportions remained relatively stable over time, the sign-adjusted series clearly captured the sharp downturn during the financial crisis, highlighting the critical role of incorporating sentiment. This result suggests that sentiment adjustment, particularly using business cycle-related sentiment, can be important for transforming raw topics into meaningful economic indicators.

Before applying sign-adjusted topics in the forecasting exercise, we conducted encompassing tests, which confirmed that our text-based series provide valuable information beyond professional forecasts, represented by the Reuters Poll. This result was especially pronounced for nowcasts and 1-step- and 2-step-ahead forecasts.

Finally, in the out-of-sample forecasting experiment, we used a Dynamic Factor Model as our primary model due to its ability to directly incorporate daily data and efficiently handle missing observations. The MIDAS model served as a benchmark to assess the robustness of our findings. Our results show that combining DFM forecasts based on text data and hard data consistently outperformed the DFM relying solely on hard data across all forecasting horizons, with the strongest gains seen for nowcasts. This suggests that text data, which is immediately available and highly interpretable, can effectively complement traditional hard data that captures more structural economic trends. Moreover, we found that the weight of text data decreases as additional hard data becomes available, especially for nowcasting, further highlighting the complementary nature of these two data sources.

Potential avenues for future research include refining the way topics are estimated over time. While our study focused on improving sentiment adjustment, we estimated the LDA model only once and used it to predict topic distributions throughout the evaluation period. A more dynamic approach would be to re-estimate topics on a monthly or quarterly basis, as done by van Dijk and de Winter (2023), allowing the model to better capture shifts in news content and changing economic narratives. Another potential improvement would be to include survey-based indicators, such as the ifo Business Climate Index, in the forecasting experiment to assess whether news data can further enhance forecasts when combined not only with traditional hard data but also with other soft indicators commonly used in macroeconomic forecasting.

References

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

Almosova, A., & Andresen, N. (2023). Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, *42*(2), 240–259.

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

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*, 659–688.

Ashwin, J., Kalamara, E., & Saiz, L. (2024). Nowcasting Euro area GDP with news sentiment: A tale of two crises. *Journal of Applied Econometrics*, *39*(5), 887–905.

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

Bai, Y., Li, X., Yu, H., & Jia, S. (2022). Crude oil price forecasting incorporating news text. *International Journal of Forecasting*, *38*(1), 367–383.

Bańbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting with daily data. *European Central Bank, Working Paper*.

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

Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication: Introducing a German dictionary. *Journal of Business Economics*, *89*(1), 79–123.

Barbaglia, L., Consoli, S., & Manzan, S. (2023). Forecasting with economic news. *Journal of Business & Economic Statistics*, *41*(3), 708–719.

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

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, *2*, 597–620.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*. https://doi.org/https://doi.org/10.1111/jofi.13377

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941.

Davis, S. J., Hansen, S., & Seminario-Amez, C. (2020). Firm-level risk exposures and stock returns in the wake of COVID-19. *Working Paper No. 27867. National Bureau of Economic Research*.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

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

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi—maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics*, *94*(4), 1014–1024.

Eickmeier, S., & Ng, T. (2011). Forecasting national activity using lots of international predictors: An application to New Zealand. *International Journal of Forecasting*, *27*(2), 496–511.

Ellingsen, J., Larsen, V. H., & Thorsrud, L. A. (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, *37*(1), 63–81.

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted Mixed Data Sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(1), 57–82.

Fraiberger, S. P. (2016). News sentiment and cross-country fluctuations. *Available at SSRN*.

Graves, A. (2012a). Long Short-Term Memory. *Supervised sequence labelling with recurrent neural networks. Studies in computational intelligence*, *385*, 37–45.

Graves, A. (2012b). Neural Networks. *Supervised sequence labelling with recurrent neural networks. Studies in computational intelligence*, *385*, 15–35.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: data mining, inference and prediction. *Springer New York*.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks. IEEE Pres*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Jangid, H., Singhal, S., Shah, R. R., & Zimmermann, R. (2018). Aspect-based financial sentiment analysis using deep learning. *Companion Proceedings of the The Web Conference*, 1961–1966.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, *32*, 9713–9729.

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

Kanzari, D., Nakhli, M. S., Gaies, B., & Sahut, J.-M. (2023). Predicting macro-financial instability – How relevant is sentiment? Evidence from long short-term memory networks. *Research in International Business and Finance*, *65*, 101912.

Kapfhammer, F., Larsen, V. H., & Thorsrud, L. A. (2020). Climate risk and commodity currencies. *CESifo Working Paper No. 8788. Available at SSRN.*

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147.

Lang, C., Schneider, R., & Suchowolec, K. (2018). Extracting specialized terminology from linguistic corpora. In E. Fuß, M. Konopka, B. Trawinski, & U. H. Waßner (Eds.), *Grammar and corpora 2016*. Heidelberg: University Publishing.

Lita, L. V., Ittycheriah, A., Roukos, S., & Kambhatla, N. (2003). tRuEcasIng. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 152–159.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

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

Mariano, R. S., & Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, *18*(4), 427–443.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Nasekin, S., & Chen, C. Y.-H. (2020). Deep learning-based cryptocurrency sentiment construction. *Digital Finance*, *2*, 39–67.

Newey, W. K., & West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, *61*(4), 631–653.

Qiang, J., Chen, P., Wang, T., & Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. *Pacific-Asia conference on knowledge discovery and data mining. PAKDD 2017. Lecture Notes in Computer Science*, *10235*, 363–374.

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting*, *36*(4), 1501–1516.

Rehurek, R., & Sojka, P. (2011). Gensim—statistical semantics in Python. *NLP Centre, Faculty of Informatics, Masaryk University*.

Reimers, N. (2016). Language independent truecaser in Python.

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941–948.

Rong, X. (2016). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

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

Shuyo, N. (2010). Language detection library for java.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychological reactions to news. *Proceedings of the National Academy of Sciences, USA*, *116*, 18888–18892.

Soto, P. E. (2021). Breaking the word bank: measurement and effects of bank level uncertainty. *Journal of Financial Services Research*, (59), 1–45.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*, 147–162.

Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.

Stock, J. H., & Watson, M. W. (2016). Why has GDP growth been so slow to recover? *Paper prepared for the Federal Reserve Bank of Boston 60th Economic Conference "The elusive 'great' recovery: Causes and implications for future business cycle dynamics".* https://www.bostonfed.org/-/media/documents/economic/conf/great-recovery-2016/james-h-stock.pdf

Thorsrud, L. A. (2016). Nowcasting using news topics. Big data versus big bank. *Norges Bank, Working Paper*.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.

Ulbricht, D., Kholodilin, K. A., & Thomas, T. (2017). Do media data help to predict German industrial production? *Journal of Forecasting*, *36*(5), 483–496.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

van Dijk, D., & de Winter, J. (2023). Nowcasting GDP using tone-adjusted time varying news topics: Evidence from the financial press. *De Nederlandsche Bank, Working Paper*.

Appendix A Data Preparation

Our dataset preparation followed a carefully structured series of steps, organized into three distinct categories. The first focuses on excluding irrelevant information, the second on filtering steps, and the third on text homogenization. For the latter two categories, the steps are further divided into those applied universally across all or most media sources and those specific to individual sources.

This section focuses on the technical details of each step—including those that impact only a small fraction of articles—with the primary goal of ensuring reproducibility and providing complete transparency about our data preparation choices. To further support this, we have made the code for all pre-processing steps publicly available[9].

Category 1: Exclusion of Irrelevant Information

The first category of pre-processing steps focuses on removing content that is either irrelevant to macroeconomic forecasting or could introduce bias into the analysis. While both the first and second categories involve filtering, the second category excludes articles based on general text mining practices, whereas the first is tailored specifically to the characteristics of our datasets and the requirements of our application.

1. The original SZ archive is divided into three distinct parts: historical articles (published before 2005), newer articles (published between 2005 and 2018), and regional news. For our analysis, we excluded regional news due to its limited relevance for economic forecasting. This decision reduced the dataset by 1,611,327 articles.

2. As the next step, we distinguished between two parts of the *dpa* dataset: *dpa-Basisdienst* (Basic Service) and *dpa-AFX Wirtschaftsnachrichten* (business news). The articles from *dpa-Basisdienst* closely resembled the news found in general newspapers, covering a range of major political and economic events. In contrast, *dpa-AFX Wirtschaftsnachrichten*, comprising 1,403,690 articles, catered more to a niche audience of investors and traders, focusing on stock market trends, financial results of companies, and the state of key industries.

[9] https://github.com/MashenkaOkuneva/newspaper_data_processing

While most of the *dpa-AFX* content appeared distinct from *dpa-Basisdienst* in both format and focus, we considered the potential relevance of 169,414 *dpa-AFX* articles tagged with the 'Konjunktur' (business cycle conditions) keyword for economic forecasting. However, a detailed analysis revealed that the proportion of these 'Konjunktur'-tagged articles in the *dpa-AFX* dataset increased significantly over time, from 3% in 2000 to 17% in 2018. This sharp rise suggested a structural change in the database's content rather than an accurate reflection of actual economic developments. Consequently, including all *dpa* and *dpa-AFX* articles with 'Konjunktur' in their keywords would likely distort the representation of this topic in our analysis, attributing the increase to the database's expansion rather than genuine economic trends.

Given the stark differences in content and target audience between *dpa* and *dpa-AFX*, as well as the risk of misrepresenting economic topics due to structural database changes, we decided to exclude *dpa-AFX Wirtschaftsnachrichten* from our analysis. While the *dpa-AFX* dataset undoubtedly contains valuable information, a separate topic model would be more appropriate for this data subset.

3. In the Welt archive, insufficient data availability in LexisNexis led to the removal of all articles published during two specific periods: March 1999–October 2000 and January 2004–April 2004. These periods had fewer than 100 articles per month, with no articles at all available from January to October 2000. Consequently, 355 articles were excluded.

Category 2: Filtering Steps

As discussed above, the second category involves filtering steps that follow general text mining practices. These steps aim to remove articles that are unsuitable for topic or sentiment modeling due to insufficient length or inappropriate formats, such as tables in text form, articles consisting mainly of numbers or names, or those written in languages other than German. To maintain unique content in the dataset, we exclude duplicate texts. Additionally, we filter out irrelevant articles using metadata, specific text strings, or article titles.

Steps Applied Universally or to the Majority of Sources:

1. *Short Article Removal*: We counted the number of words in each article, excluding numbers and treating hyphenated multi-word nouns (e.g., 'Experten-Gruppe', meaning 'expert

group') as single words. Articles with fewer than 100 words were removed, as topic models typically perform better with longer texts, which offer richer semantic features. This step resulted in the exclusion of 2,175,240 articles from dpa, 306,804 from Handelsblatt, 518,191 from SZ, and 56,082 from Welt.

2. *Removal of Exact Duplicates*: Articles with identical text content were identified as exact duplicates and removed. These duplicates often arose from minor variations in metadata. For instance, in the SZ archive, duplicates occurred when the same article was published on different pages for various regional editions. Similarly, if an article appeared twice in the corpus with different publication dates (e.g., 10.01.1991 and 11.01.1991), only the first entry was retained to avoid redundancy. For dpa publications, when identical articles were released by both dpa and dpa-AFX, we kept the version published by dpa. Consequently, this process led to the removal of 822,318 articles from dpa, 2,751 articles from Handelsblatt, 17,770 articles from SZ, and 1,005 articles from Welt.

3. *Extensive Filtering*: An important part of our dataset preparation involved performing extensive filtering based on text strings, article titles, and various types of metadata, including sections, subsections, keywords, genres, series, and sources.

   Section-based filtering was particularly critical for SZ, a newspaper covering a wide range of topics. After sampling articles from each section, we retained those most relevant to economic and political developments, namely "Politics", "Stock Market and Finance", "Money", "Opinion", "News", "Page Four", and "Economy". For Handelsblatt, a business-focused newspaper, the process was more straightforward, requiring the exclusion of only a few irrelevant sections. For Welt, we specifically targeted articles from the Economy and Finance sections right from the start, and for dpa, our focus was on sections related to Politics, Economy, and Finance, eliminating the need for further section-based filtering.

   Overall, filtering based on different kinds of metadata, titles, and text strings was designed to remove particular types of articles. These include:

   (a) *Articles with no narrative*: This category encompasses a variety of content that lacks a narrative structure or detailed discussion, including lists of upcoming events, names, donation accounts, emergency contacts, and company names. It also covers articles focused solely on prices, investment funds' values, stock prices, expected

earnings, exchange rate adjustments, and interest rates. Tables presented in text form, contact information for news media and organizations, press reviews consisting only of headlines from various outlets, financial news articles summarizing the ratings and target prices made by various investment banks, and content focusing exclusively on electoral statistics were also excluded. In addition, we removed graph legends, tables of contents summarizing what is included in current issues, and announcements from the media to their readers.

(b) *Internal communication articles not intended as news content for the general public*: This type includes schedules of planned coverage for specific events or topics, announcements about organizational changes within the media, and test messages used to verify the technical distribution systems.

(c) *General news articles unrelated to Economy or Politics*: This category covers a wide array of subjects such as natural disasters, criminal justice, legal news, environmental concerns, scientific research, societal issues, education and career guidance, public commentary including letters to the editor, IT and digitalization, automobiles, editorials, profile pieces, and marketing. It also includes highly specialized real estate articles aimed at potential homebuyers rather than those seeking broader economic insights, as well as pieces on health and medicine and topics tailored for younger readers.

(d) *Entertainment news*: Unlike general news articles, entertainment news concentrates on cultural and leisure activities. It includes a wide array of topics such as the latest developments in sports, arts, culture, the film and television industries, and updates on celebrities. It also spans literature, travel, style, fashion, leisure, hobbies, music, puzzles, language, and theater.

(e) *Regulatory disclosure articles*: They primarily consist of formal communications required by law or regulations. In our dataset, these articles are represented by directors' dealings and ad-hoc notifications.

(f) *Articles with low economic relevance*: Although these articles originate from sections related to Politics and Economy, they hold limited value for economic forecasting. They include local or regional news without broader economic impact and overviews of scheduled political events that lack examination of their possible effects on eco-

nomic trends. Background articles provide context and details, while analytical and commentary pieces offer valuable insights into political dynamics and societal issues; however, both types do not center around current events. Specialized informational pieces explaining specific concepts, rules, or regulations also fall into this category, as they provide general knowledge rather than current, event-driven economic analysis.

(g) *Articles with a format difficult for LDA*: These articles comprise collections of news briefs, each shorter than 100 words, providing minimal context for effective topic modeling. Additionally, some articles feature fragmented sentences and lists rather than continuous prose.

(h) *Articles with a historical focus*: These articles explore past events, presenting historical chronologies or detailed discussions on specific historical events. Although they provide a valuable understanding of the past, their lack of focus on recent developments makes them less useful for economic forecasting.

(i) *Retracted articles*: These are publications that have been formally withdrawn by the publisher due to containing incorrect, misleading, or outdated information.

(j) *Articles from sections and subsections covered within a limited time period*: To ensure a consistent thematic focus over time and concentrate on topics driven by external events rather than internal editorial decisions, we excluded articles from sections and subsections that were only active during specific periods. From Handelsblatt, for example, we removed sections like "Career", "Weekend Journal", and "Panorama". Similarly, from Welt, we omitted subsections such as "Hamburg Economy" and "Berlin Economy".

(k) *Advertisements*: We excluded all forms of advertisements as they are promotional content with no value for economic forecasting.

(l) *Right of reply pieces*: These legally mandated articles provide individuals or organizations with the opportunity to respond directly to published information they consider inaccurate.

The exact filtering criteria are detailed in our repository. Following this extensive filtering process, a total of 829,852 articles were removed from the dpa archive, 90,859 from Handelsblatt, 941,910 from SZ, and 7,747 from Welt.

4. *Language-Based Filtering*: An important component of our pre-processing strategy involved filtering out articles not written in German. To achieve this, we utilized the 'langdetect' library [10] by Shuyo (2010), a Python-based language detection tool that employs a probabilistic algorithm (Naive Bayes with character n-gram). To improve the accuracy of language detection, we determined the language for each article three times and calculated the average probability of the article being in German. We classified an article as written in German if this average probability exceeded a threshold of 90%. While a significant portion of the non-German articles in our dataset was in English, we also encountered articles in French and even a German dialect, Low German. Additionally, certain articles, technically in German, were misclassified as non-German due to the prevalence of English names, the use of informal language, or the presence of tables without narrative context. Consequently, we removed 94 articles identified as written in languages other than German from our dataset.

5. *Number-Heavy Article Removal*: We removed articles predominantly consisting of numbers. Employing regular expressions, we counted the occurrences of numbers in each text and classified an article as number-heavy if the ratio of numbers to words exceeded 0.5, a threshold determined through visual inspection. Once stripped of numerical data, these articles provided limited text for sentiment or topic analysis and often covered subjects irrelevant to our research question, such as detailed budget plans or car registration statistics. Through this approach, we successfully removed 273 number-heavy articles from dpa, 14 from Handelsblatt, 92 from SZ, and 7 from Welt.

6. *Table Exclusion*: We decided to exclude tables from our article analysis, as they typically lack sentiment or narrative elements and thereby add noise. For the dpa archive, our first step was to identify articles that contain at least one paragraph that is predominantly numerical and at least 10 words long, as such paragraphs are likely to be tables. We consider a paragraph predominantly numerical if the ratio of numbers to words exceeds 70%, a threshold set based on visual exploration. For the Handelsblatt, SZ, and Welt archives, we adopted a slightly modified approach due to a notably lower prevalence of tables. Here, we selected articles where the ratio of numbers to words was at least 20%.

[10] https://pypi.org/project/langdetect/

Once we identified articles with numeric paragraphs or high numerical density, we manually reviewed them to detect common strings that often precede tables. Subsequently, we used regular expressions to systematically remove tables identified by these strings. Through this process, 1,084 tables were eliminated from dpa, along with 308 from Handelsblatt, 292 from SZ, and 34 from Welt. Moreover, we excluded any articles that, after table removal, were shortened to fewer than 100 words. As a result, we removed 647 texts from the dataset.

7. *Article Continuation Merging*: In our dataset, articles split into multiple parts, known as 'chained articles', were merged to form complete units. This splitting often occurs when individual sections are lengthy, produced at different times throughout the day, or printed on separate pages of a newspaper. Merging these segments allowed us to accurately capture the overall sentiment of the news item, avoiding the repetition of sentiment analysis for its individual parts.

   Our merging criteria for dpa articles included two scenarios. Firstly, if an article's concluding sentence is referenced in the title of the next part, we merge these parts. Secondly, articles that share the same main headline are also combined. Importantly, we only merged parts published on the same day. Through this approach, we successfully combined 4,399 dpa articles.

   For Handelsblatt and SZ, we focused on articles containing phrases such as 'Continuation from page 10' and searched for their corresponding initial parts, typically indicated by strings like 'Continuation page 11'. In instances where several articles were potential candidates for the first part, we applied specific criteria to accurately determine the correct initial segment. This approach enabled us to merge 206 articles in the Handelsblatt archive. In contrast, chained articles were less of a concern in SZ, resulting in only 12 articles being merged. We also removed articles that remained under 100 words after merging. Eliminating these articles and those representing individual parts reduced our dataset by 5,298 articles.

8. *Fuzzy Duplicates Removal*: In our efforts to maintain a focus on unique content, we identified and removed 'fuzzy' duplicates, or nearly identical articles, from our data. Fuzzy duplicates typically take the form of drafts, minor revisions, updates, summaries, or overviews

of original articles, as well as slightly altered advertisements republished multiple times. Our approach was to delete only those duplicates that were published within the same month as the original article, adhering to the principle that news media readers generally do not consume almost identical articles multiple times.

To identify these duplicates, we used Gensim (Rehurek & Sojka, 2011), a popular Python library. Each article was parsed into individual words, with punctuation, accents, and numbers stripped, and the text converted to lowercase. We then created a Bag of Words (BOW) representation, substituting words with numerical identifiers. Cosine similarity was calculated between all pairs of texts to detect duplicates. A threshold of 93% cosine similarity was set as the criterion for duplication based on visual examination.

In cases where both the original article and its duplicate were published on the same day, we decided to delete the shorter of the two. Conversely, when the original and its duplicate were published on different days, we preserved the article that was published first. Due to the limitations of cosine similarity in distinguishing lengthy articles, we selectively retained several long articles in the Handelsblatt corpus that were incorrectly classified as duplicates. Through this process, a total of 105,048 duplicates were removed from the dpa corpus, 991 from Handelsblatt, 4,986 from SZ, and 1,132 from Welt.

9. *Exclusion of Articles with a High Proportion of German Names*: We remove texts with a name density of at least 15% relative to the total word count, excluding numbers. This is important to ensure that sufficient content remains for topic analysis after the names are eliminated. First names are retrieved from Script[11] and beliebte-vornamen.de[12], while surnames are obtained from the Digital Dictionary of Surnames in Germany[13]. Based on this criterion, we excluded 212 articles from the dpa archive, 27 from Handelsblatt, 59 from SZ, and 24 from Welt, primarily comprising lists of politicians' or company managers' names.

10. *Irrelevant Text Removal*: We selectively removed portions of text from our articles that were either uninformative or posed challenges for topic and sentiment analysis. As a first step, we excluded website names and full URLs. However, we made exceptions for

[11]https://script.byu.edu/german-handwriting/tools/given-names
[12]https://www.beliebte-vornamen.de
[13]http://www.namenforschung.net/en/dfd/dictionary/list-of-all-published-entries/

specific website names like 'amazon.com', 'Booking.com', or 'Bild.de'. These are retained as they are important for maintaining context and sentence structure, as well as for topic identification, particularly since internet companies are primarily known by their website names. Additionally, we excluded physical addresses, e-mail addresses, telephone and fax numbers, contact information of news media, names of journalists, references to media sources, names of stocks or other assets at stock exchanges (e.g., '<DEMUS.FX1>'), case numbers from court decisions, data-driven lists, detailed voting results, references to supplementary information and page numbers, additional reading recommendations, promotional content, fact boxes, copyright notices, and any information included in the article text and not intended for publication, such as internal editorial notes, contact information for follow-up, and background details related to the article.

We removed irrelevant text portions from 1,937,646 articles in dpa, 87,008 articles in Handelsblatt, 33,504 articles in SZ, and 27,860 articles in Welt. After this, any texts shorter than 100 words were also excluded, which resulted in the removal of 60,573 articles from dpa, 140 from Handelsblatt, 275 from SZ, and 51 from Welt.

Steps Specific to Each Source:

Handelsblatt:

1. For Handelsblatt, we generally exclude articles shorter than 100 words, except when they are likely to be either the continuation of an article from a previous page or the beginning of an article that extends to the next page. These articles are later combined with their respective continuing parts, forming complete single entries. In this step, we kept 282 short articles that would otherwise have been removed.

2. Additionally, we removed 14 articles from the dataset due to umlaut encoding issues or excessive non-systematic errors.

Welt:

In sections with titles containing the term 'kompakt' or 'Kompakt' (e.g., 'Wirtschaft kompakt'), we encountered 19,406 instances where multiple articles were aggregated into a single entry. We separated these aggregated articles into 54,410 distinct entities and processed each individually to accurately capture their unique sentiments and topics. This step led to an increase in the number of articles by 35,004.

dpa:

For the dpa dataset, the first six steps addressed specific types of duplicates: news corrections, updates, summaries, overviews, repeated articles, and advance notifications. These dpa-specific duplicates were closely related to their originals and could be identified using article metadata or text. Each type had distinct characteristics, requiring tailored treatment, which we outline below. Together with the fuzzy duplicate removal discussed earlier, these steps allowed us to effectively focus on the unique content within the dpa archive.

1. The first type of dpa-specific duplicates we examined was *news corrections*. These articles typically involved only minor changes to the original texts, such as the alteration of a few facts. Initially, we planned to treat all corrected articles as duplicates and remove them. However, further analysis revealed that, starting from 2012, dpa often deleted the original articles associated with these corrections. To avoid discarding unique content, we revised our approach and removed only the corrected articles published before 2012. This adjustment resulted in the exclusion of 22,319 corrected articles from our corpus.

2. The second type of duplicate articles we addressed in our analysis of dpa corpus were *news updates*. Unlike news corrections, which typically involve minor changes, news updates add substantial news content to original articles and, therefore, were handled differently. If an updated news text and its original version were both published on the same day, we removed the original article to prioritize the most recent information. However, when the original and updated articles were published on different days, we retained both. This approach underscored the value of the original article's timeliness, capturing the moment when the news was first received and potentially impacted market participants, while the updated article often provided additional, crucial information. For instances of multiple updates to a single news item, we retained only the latest update, as it generally provided the most comprehensive and relevant information available. As a result of this step, 1,004 articles were removed from our dataset.

3. Another type of article in our dataset that frequently posed duplication issues was *summaries*. These often appeared alongside the original articles and were published on the same day. We identified two primary variations of summaries: those that included the original text with an additional segment, resembling updated articles, and those that con-

densed the original into a brief version highlighting its key points. Regardless of the type, the content of these summaries was typically very similar, if not identical, to the corresponding original articles. As a result, when an original article and its summary shared matching titles and publication dates, we treated the shorter version as a duplicate. This approach resulted in the removal of 49,773 such duplicates from our corpus.

4. Next, we turn our attention to *overviews*, an article type that also required special treatment due to duplication concerns. Overviews typically contain all available information about a specific event on a given day. Often, journalists would write an initial article in the morning and later publish an overview in the evening, expanding on the original by incorporating the latest developments and additional details. These evening overviews closely resemble updated news texts. In our dataset, we identified 4,675 instances where overviews shared the same title as original articles and were published on the same day. In these cases, to avoid redundancy, we removed the shorter article.

5. Another category we carefully examined was *repeated articles*, uniquely identified by their titles, which combined the word 'Repeat' with the exact title of the original article. In our dataset, we found 1,482 such articles, all published on the same day as their corresponding initial versions. These repeated articles usually turned out to be either identical to their originals or contained only minor corrections. We removed the shorter version from each pair of a repeated article and its initial publication.

6. The final type of duplicate articles we addressed was *advance notifications*, which preview events scheduled to occur in the future. In our dataset, we identified 606 such articles that shared titles with longer, more detailed updates published on the same day. We treated the shorter article in each pair as a duplicate and removed it.

7. Similar to the situation in the Welt archive, the dpa dataset contained 104,292 articles that were compilations of short pieces on various topics. These compilations fell into three categories: overviews of important daily news, brief economic updates, and stock rating analyses. The overviews covered a range of subjects, including political and economic developments as well as international news. The economic updates focused on key economic events of the day, while the stock rating analyses provided recommendations (buy, hold, or sell) for specific companies, accompanied by explanations of the reasoning behind each

recommendation.

We separated these compilations to accurately capture the distinct topics and sentiments of the individual pieces within them. This separation resulted in 899,484 individual articles. However, in line with our criterion of excluding articles shorter than 100 words, only 87,524 articles met this length requirement and were retained in the dataset. These retained articles replaced the original compilations, ultimately reducing the corpus size by 16,768 articles.

Category 3: Text Standardization

The third category of pre-processing steps focuses on improving the quality of the texts without affecting the size of the corpus. Specifically, these steps involve restoring correct umlauts in older articles, resolving issues introduced by OCR technology, separating erroneously merged words and numbers, addressing encoding problems, and restoring proper casing. Together, these adjustments ensure consistent and high-quality input for our sentiment and topic models.

Steps Applied Universally or to the Majority of Sources:

1. *Umlaut Normalization*: In articles from the 1990s and early 2000s (up to and including 2001), German umlauts (ö, ä, ü, ß, Ö, Ä, Ü) were often replaced with 'oe', 'ae', 'ue', 'ss', 'OE', 'AE', and 'UE'. For example, 'Nürnberg' would appear as 'Nuernberg'. To standardize word representation and restore correct umlauts, we utilized the Python library 'PyHunSpell' [14].

   Our methodology was specifically tailored for texts that lacked umlauts and were published before 2002. We spellchecked only words containing umlaut replacements. If 'PyHunSpell' indicated incorrect spelling, we generated a list of suggestions limited to those containing umlauts and selected the first one as the most likely correction.

   However, in cases where the spellchecker identified an error but provided no umlaut-containing suggestions, we manually replaced 'AE', 'OE', or 'UE' with 'Ä', 'Ö', or 'Ü', except in certain exceptions like 'OECD'. After making replacements, we spellchecked the word again. If the spelling remained incorrect and no suggestions were available, we proceeded to replace 'ae' and 'oe' with 'ä' and 'ö' for potentially multi-umlaut words, followed

[14] https://github.com/pyhunspell/pyhunspell

by another spellcheck. If the issue persisted without suggestions, we manually replaced ‘oe’, ‘ae’, ‘ue’ with the corresponding umlauts, avoiding changes to ‘ss’. For example, the word ‘UEberschusseinkuenfte’ (meaning ‘surplus income’) was corrected to ‘Überschusseinkünfte’. This approach, confirmed through visual inspection, yielded optimal results.

If a word did not contain ‘AE’, ‘OE’, ‘UE’ but was still misspelled without suggestions, we applied the same method of manual replacement and rechecking. The correct spelling, when achieved, was used as the final token version. Through this process, umlauts were restored in 90,627 articles from dpa, 70,108 articles from Handelsblatt, and 25,990 articles from SZ.

Additionally, we identified 206 articles within the Welt and SZ archives where umlauts were incorrectly encoded as HTML entities (e.g., ‘ä’, ‘ü’, ‘ö’, ‘Ä’, ‘Ü’, and ‘Ö’). These encodings were systematically replaced with their correct umlaut representations—‘ä’, ‘ü’, ‘ö’, ‘Ä’, ‘Ü’, and ‘Ö’. Specifically for SZ, two texts exhibited systematic misrepresentations of ‘ö’ as ‘|’, and ‘ü’ as ‘}’, seen in examples like ‘}ber’ and ‘|ffentlichen’. These were also corrected.

2. *Correction of OCR-Induced ‘O’ and ‘0’ Confusion*: In the archives of Handelsblatt and SZ, where OCR (Optical Character Recognition) technology was employed for digitization, we encountered a specific issue. The technology often struggles to differentiate between the digit ‘0’ and the letters ‘O’ or ‘o’. To address this, we systematically identified patterns such as ‘1OO’ or ‘2 OOO’ and replaced them with their correct numeric values, ‘100’ and ‘2 000’ respectively. This correction is important for accurately separating merged words and numbers in the next step. Moreover, it ensures that strings like ‘1OO’ are correctly identified as numeric values. We corrected this mistake in 177 articles from Handelsblatt and 123 texts from SZ.

3. *Separation of Merged Words and Numbers*: We corrected instances where numbers and words were erroneously merged. This separation into distinct tokens is particularly beneficial for cases where hyphen-separated words appear without hyphens (e.g., ‘20Jährige’, translated as ‘20-year-old’, instead of ‘20-Jährige’). In later pre-processing steps for topic modeling and sentiment analysis, hyphens are replaced with spaces. Therefore, by separating number-word pairs (e.g., ‘20Jährige’ to ‘20 Jährige’), we ensure that they are

treated similarly to their hyphen-separated counterparts, thus improving consistency in data preparation.

Additionally, this approach separates numbers from currency names (e.g., '100DM' into '100 DM', where 'DM' stands for 'Deutsche Mark') and from units of time, weight, or distance (e.g., '100km' into '100 km' and '16Uhr' into '16 Uhr', meaning '16 o'clock'). It also assists in rectifying simple errors (e.g., '30bis 40' corrected to '30 bis 40', meaning '30 to 40', or '10Fahrzeuge' corrected to '10 Fahrzeuge', meaning '10 vehicles') and fixing enumerations at sentence beginnings (e.g., '10Welche' corrected to '10 Welche', meaning '10 Which', and '8Wie' to '8 Wie', meaning '8 How').

Mindful of exceptions, we retain original forms for company names (e.g., '1822direkt', '3Sat', '4MBO'), model names of smartphones, airplanes, satellites (e.g., '4S', '328Jet'), and specific noun and adjective forms (e.g., '90er', meaning '90s', or '21st').

This pre-processing step affected 95,578 articles in dpa, 21,468 in Handelsblatt, 23,192 in SZ, and 3,797 in Welt.

Steps Specific to Each Source:

Handelsblatt:

We addressed unicode errors in 131 articles where specific characters like "Å," misrepresented the intended letters, such as "ł".

Welt:

We also corrected 178 articles affected by unique encoding issues in the Welt archive. Corrections included transforming 'Ha{ring}kann' to 'Håkan', 'u{cech}' to 'ŭ', and 'c{ogon}' to 'ç', among others. These and similar misencoded sequences were systematically replaced with their accurate representations.

dpa:

Finally, we encountered specific issues with a small subset of dpa articles published between 1991 and 2001, where the casing was incorrect and umlauts were missing. Specificaully, there were two types of problems with the casing. The first type, encompassing 77,135 articles, involved articles whose main text lacked capital letters. The second type, consisting of 50 articles, had 'Ae', 'Ue', and 'Oe' as replacements for capitalized umlauts (Ä, Ü, Ö).

This deviation from the standard practice of using 'AE', 'UE', 'OE' poses challenges for correct umlaut normalization.

To address these issues, we employed a truecasing model [15], developed by Reimers (2016) and inspired by the work of Lita et al. (2003). We trained this model using 1,000,000 dpa articles in which umlauts were replaced with non-umlaut equivalents. To evaluate the truecaser's effectiveness, we tested it on 100 randomly chosen sentences from 1,000 articles not involved in the training. The model achieved a 99.32% accuracy on these sentences, effectively resolving the casing issues in the dpa article subset.

Appendix B Effect of Pre-processing on the Dataset

This appendix illustrates the significant impacts of pre-processing steps on the dataset used in the study. For each of our sources—dpa, SZ, Handelsblatt, and Welt—the figures display the 30-day backward moving average of the daily number of articles published before and after pre-processing. In all figures, the blue line represents daily publications of the dataset before pre-processing, and the black line indicates daily publications of the pre-processed dataset. The X-axis corresponds to days.

Figure 8 clearly demonstrates that excluding 'dpa-AFX' articles from the dpa dataset prevents a misleading upward trend in the daily number of publications (green line)—attributable more to the expansion of a new product rather than a true increase in significant news coverage. The pre-processed dataset (black line) shows consistent publication levels over time.

In the case of SZ, as illustrated in Figure 9, pre-processing not only mitigates the sudden surge in article counts caused by the inclusion of regional news in 2006 (green line) but also addresses spikes and anomalies in the daily publication patterns, especially notable in the late 1990s (evident from the comparison of the blue and black lines).

As depicted in Figure 10, pre-processed data for Handelsblatt shows a more subdued downward trend (black line), suggesting improved consistency. In the case of Welt, as Figure 11 indicates, pre-processing significantly impacts the early part of the data, resulting in a smoother series of daily publications.

Overall, this analysis shows how pre-processing improves the quality of the corpus by con-

[15] https://github.com/nreimers/truecaser

centrating on homogeneous content consistently covered over time and minimizing data irregularities. This approach is critical for capturing topic changes driven by economic events rather than organizational shifts within the news media, as also emphasized by Bybee et al. (2024).

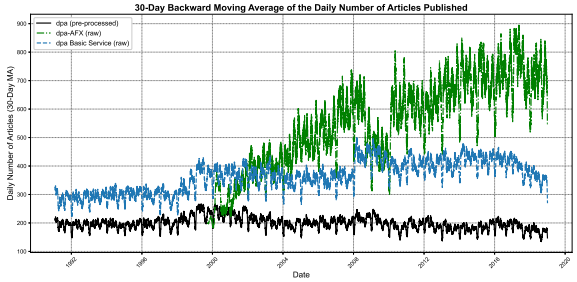


Figure 8: The blue line represents the dpa Basic Service before pre-processing, the green line indicates daily publications for dpa-AFX (articles removed during pre-processing), and the black line depicts the daily publications of the pre-processed dataset.

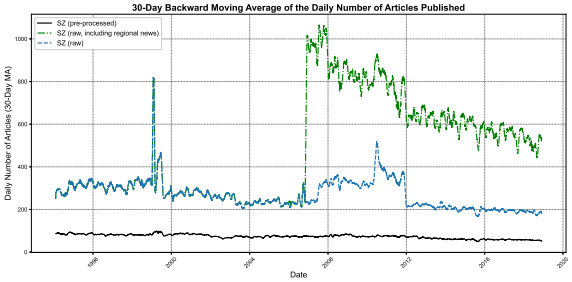


Figure 9: The blue line shows daily publications of the SZ dataset, excluding regional news, before pre-processing. The green line includes regional news (also from the original dataset), and the black line represents daily publications of the pre-processed dataset.

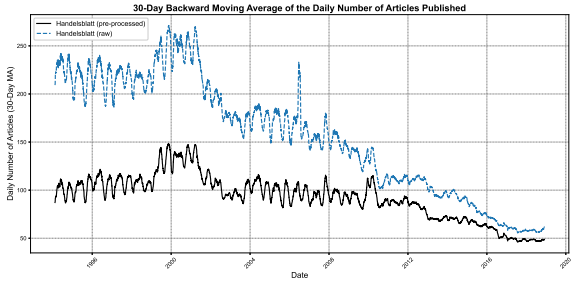


Figure 10: The blue line shows daily publications for Handelsblatt before pre-processing, and the black line shows the publications after pre-processing.

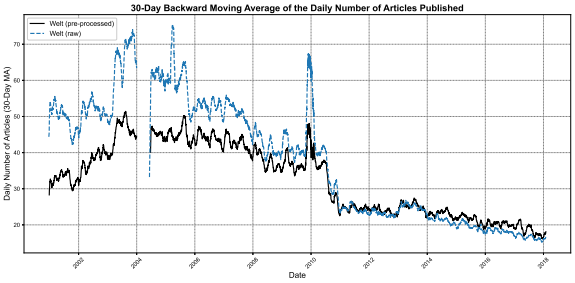


Figure 11: The blue line indicates daily publications of the original Welt dataset, while the black line represents publications after pre-processing.

Appendix C The MTI Dataset

This appendix provides additional details about the MTI dataset, specifically explaining how downloaded articles were matched with their sentiment annotations and pre-processed.

Appendix C.1 Matching with Sentiment Annotations

To ensure accurate matching of downloaded articles with sentiment annotations from the MTI dataset, we implemented several corrections. First, we adjusted some titles and publication dates within the MTI dataset because the titles slightly varied from those of the downloaded articles or contained orthographic errors, and there were a few discrepancies in publication dates compared to the downloaded articles.[16] These issues possibly resulted from manual data entry by annotators in the MTI dataset. Second, we corrected spelling and punctuation errors in several titles of the downloaded articles. Third, we normalized the titles of our downloaded articles and those in the MTI dataset by converting them to lowercase, removing certain punctuation, and standardizing spaces.

Appendix C.2 Dataset Pre-Processing

Some of the article texts required general pre-processing that was not specific to the sentiment model applied. In cases where an article was a compilation of several pieces—common in publications like BILD and BamS—we manually isolated the section annotated by MTI and removed the rest. We also removed non-essential content at the end of some articles, such as photo captions, editor's notes, source information, and unrelated background details. For articles from Capital, we corrected some lead-ins that had punctuation issues to ensure proper formatting. Moreover, we removed any duplicates from the dataset.

[16] Specifically, some of the online articles of Focus had a different publication date than the corresponding print articles used in the MTI dataset. In these cases, we identified the matches by the metadata and used the dates of the print version.

Appendix D Methodology

Appendix D.1 Pre-processing for word2vec model

Before estimating the word2vec model, we perform several standard pre-processing steps:

1. **Convert to Lowercase:** All text is transformed to lowercase to prevent the model from treating the same word differently due to case variations.

2. **Remove Punctuation:** Punctuation marks are removed, as they typically do not contribute meaningful information.

3. **Eliminate Non-Alphabetic Characters:** We strip away any non-alphabetic characters to focus the analysis purely on the words themselves.

4. **Normalize Whitespace:** Multiple spaces are reduced to a single space, ensuring a clean and consistent tokenization process.

5. **Tokenize Text:** The text is broken down into individual words, creating the tokens required as input for the word2vec model.

6. **Remove Single-Letter Tokens:** Single-letter tokens, which often lack significant meaning, are excluded from the dataset.

7. **Filter Rare Words:** Words appearing five times or fewer are removed to reduce noise and improve the quality of the vector representations.

Appendix D.2 Sampling techniques

To address the computational complexity of the word2vec model, we applied three sampling techniques originally proposed by Mikolov et al. (2013b), which are described below.

Subsampling discards words that appear frequently across various contexts but add little semantic value, such as common articles and prepositions. These words are less informative for understanding the meaning of more specific terms like 'business cycle conditions'. For each word w_i, we discard it with a probability given by:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}\,, \tag{12}$$

where t is a threshold parameter, set to 0.00001, and $f(w_i)$ is the frequency of word w_i in the corpus. The threshold $t = 0.00001$ is a standard choice in the literature (Mikolov et al., 2013b).

The second technique involves randomly shrinking the context window size C during training. By varying the window size within the range $\{1, 2, \ldots, C\}$, this method emphasizes words closer to the target word, as they generally have a stronger influence on its meaning.

The final approach we use is negative sampling, which improves training efficiency by limiting parameter updates to a small subset of words. Instead of adjusting embeddings for every word in the vocabulary for each target-context pair as required by equation (2), only the embeddings of the true context word and five "negative samples"—words not present in the context—are updated. Words are chosen for negative sampling based on the distribution that has demonstrated strong empirical performance (Mikolov et al., 2013b):

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{k=1}^{V} f(w_k)^{3/4}}\,. \tag{13}$$

Appendix D.3 Estimation details for word2vec model

The table below summarizes the key estimation details for our word2vec model:

Model Configuration	
Algorithm	Skip-gram
Embedding dimension	256
Initialization of embedding matrices $\mathbf{U}$ and $\mathbf{V}$	Uniform distribution [-1, 1]
Context window size C	10
Number of negative samples	5
Training Details	
Epochs	10
Batch size	128
Threshold for subsampling	0.00001
Optimization Settings	
Optimizer	Adam
Learning rate	0.0001

Table 14: Estimation details for the word2vec model.

Appendix D.4 t-SNE visualization of word embeddings

t-Distributed Stochastic Neighbor Embedding (t-SNE) was introduced by van der Maaten and Hinton (2008) as a method for visualizing high-dimensional data in a lower-dimensional space. In this study, we use t-SNE to visualize the word embeddings estimated by our word2vec model.

We begin with a set of word embeddings $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in a 256-dimensional space, where $N = 1,000$. The objective of t-SNE is to map these high-dimensional vectors into a two-dimensional space $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, such that the pairwise similarities in $\mathbf{Y}$ closely mirror those in the original high-dimensional space $\mathbf{X}$.

The similarity between two word embeddings $\mathbf{x}_i$ and $\mathbf{x}_j$ is modeled using a conditional probability distribution:

$$P_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)}, \tag{14}$$

where σ_i is the standard deviation of the Gaussian distribution, controlling the size of the neighborhood around $\mathbf{x}_i$. Higher probabilities are assigned to points that are closer in the 256-dimensional space. To compute the cost function, we need joint probabilities rather than conditional ones. These joint probabilities P_{ij} are defined as:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2N} \,. \tag{15}$$

In the lower-dimensional space, the similarity between points $\mathbf{y}_i$ and $\mathbf{y}_j$ is modeled using a Student's t-distribution:

$$Q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}} \,. \tag{16}$$

The optimal mapping is achieved by minimizing the Kullback-Leibler divergence between the joint probability distributions P and Q:

$$C = \sum_i \sum_{j \neq i} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \,. \tag{17}$$

By minimizing the cost function 17, t-SNE ensures that word embeddings that are close in the original 256-dimensional space remain close in the two-dimensional space. This enables us to effectively visualize the semantic relationships between words in two dimensions and evaluate the quality of the word embeddings.

Figure 12 presents a t-SNE visualization of 1,000 embeddings corresponding to the words most closely related to 'business cycle conditions', based on cosine similarity. Each point represents one of the 1,000 words, with 'business cycle conditions' highlighted in red. Ideally, words that are close to each other in the 2-dimensional space should also be closely semantically related, reflecting the quality of the embeddings. The closest words to 'business cycle conditions' include 'economic upswing', 'recession', 'business cycle', 'economic dynamics', and 'weak phase'—all of which are directly associated with the business cycle. Other nearby words correspond to important economic concepts such as economic sentiment, research institutes, export growth prospects, consumption, the labor market, and investments. Words that are further away pertain to financial markets, uncertainty, and economic policy. While these concepts are relevant to the business cycle, their greater distances are understandable and expected. Overall, the visualization indicates that our embeddings successfully capture key semantic relationships and are capable of identifying terms closely linked to the concept of interest.

Figure 12: t-SNE visualization of 1,000 words related to 'business cycle conditions'. The red point represents 'business cycle conditions'. Points are color-coded to show words grouped by concepts: price dynamics (e.g., 'price levels'), economic research institutes (e.g., 'ifo'), economic sentiment (e.g., 'business climate'), export growth prospects (e.g., 'export opportunities'), business cycle (e.g., 'recession'), consumption (e.g., 'consumer demand'), investments (e.g., 'investment activities'), labor market (e.g., 'unemployment'), commodity prices (e.g., 'oil prices'), economic turning points (e.g., 'economic turnaround'), inflation (e.g., 'inflation rates'), financial markets (e.g., 'credit crunch'), crises (e.g., 'economic crisis'), uncertainty/expectations (e.g., 'uncertainty'), economic policy (e.g., 'structural reforms'), and performance (e.g., 'boom'). All terms were translated from German using DeepL.

Appendix D.5 Identification of related terms with K-means clustering

We applied the K-means algorithm to independently cluster two sets of 1,000 word embeddings: those most cosine-similar to 'business cycle conditions' and those closest to 'economy'. The algorithm minimizes within-cluster variance through an initialization phase and two iterative steps: assignment and update.

1. **Initialization:** Randomly choose K initial cluster centers $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ from the set of 1,000 word embeddings.
2. **Assignment Step:** Assign each word embedding $\mathbf{u}_{w_i}$ to the cluster with the nearest center $\mathbf{c}_k$, for $k \in \{1, \ldots, K\}$. The cluster assignment C_k for embedding $\mathbf{u}_{w_i}$ is defined as:

$$C_k = \left\{ \mathbf{u}_{w_i} : \|\mathbf{u}_{w_i} - \mathbf{c}_k\|^2 \leq \|\mathbf{u}_{w_i} - \mathbf{c}_j\|^2 \text{ for all } j \neq k \right\}, \tag{18}$$

   where $\|\mathbf{u}_{w_i} - \mathbf{c}_k\|^2$ is the squared Euclidean distance between the embedding $\mathbf{u}_{w_i}$ and the cluster center $\mathbf{c}_k$.
3. **Update Step:** Recalculate the cluster centers $\mathbf{c}_k$ as the means of all points assigned to each cluster:

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{\mathbf{u}_{w_i} \in C_k} \mathbf{u}_{w_i}, \tag{19}$$

   where $|C_k|$ is the number of points in cluster C_k.
4. **Convergence:** Repeat the assignment and update steps until cluster assignments do not change.

Selecting K:

To determine the optimal number of clusters K, we calculate the Silhouette score, which evaluates how well the word embeddings $\mathbf{u}_{w_i}$ are clustered. For each embedding $\mathbf{u}_{w_i}$, the score $s(\mathbf{u}_{w_i})$ is given by:

$$s(\mathbf{u}_{w_i}) = \frac{b(\mathbf{u}_{w_i}) - a(\mathbf{u}_{w_i})}{\max\{a(\mathbf{u}_{w_i}), b(\mathbf{u}_{w_i})\}}, \tag{20}$$

where $a(\mathbf{u}_{w_i})$ is the average Euclidean distance between $\mathbf{u}_{w_i}$ and all other embeddings within the same cluster, and $b(\mathbf{u}_{w_i})$ is the average Euclidean distance from $\mathbf{u}_{w_i}$ to embeddings in the nearest different cluster.

The Silhouette score ranges from -1 to 1, with higher values indicating better clustering quality. To determine the optimal number of clusters, K, we tested values between 2 and 25 and selected the K that maximized the average Silhouette score, ensuring that the resulting clusters are both cohesive and well-separated. For 'business cycle conditions', the optimal K was 3, yielding 279 related terms, while for 'economy', the optimal K was 2, resulting in 653 related terms. These related terms are those found in the same cluster as the respective term of interest.

More details about K-means can be found in Hastie et al. (2001), and for Silhouette score, see Rousseeuw (1987).

Appendix D.6 Terms related to business cycle conditions

In this section, we present the 279 German terms related to the word 'business cycle conditions' and their English translations, followed by the 100 German terms related to 'economy' and their translations. Together, they represent the words associated with our aspect of interest.

Appendix D.6.1 Terms related to the word 'business cycle conditions'

Table 15: Terms related to 'Business Cycle Conditions'

German Term	English Translation	German Term	English Translation
konjunktur	business cycle conditions	standortbedingungen	site conditions
konjunkturaufschwung	economic upswing	angebotsseite	supply side
konjunkturaufschwungs	economic upswing	binnennachfrage	domestic demand
konjunkturdelle	economic downturn	inlandsnachfrage	domestic demand
konjunkturbelebung	economic recovery	konsumnachfrage	consumer demand
binnenkonjunktur	domestic economy	exportnachfrage	export demand
exportkonjunktur	export economy	nachfrageseite	demand side
konjunkturlage	economic situation	baunachfrage	construction demand
konjunkturverlauf	business cycle	wirtschaftswachstum	economic growth
konjunkturmotor	economic engine	wirtschaftswachstums	economic growth
konjunkturlokomotive	economic locomotive	wachstumsprognose	growth forecast
konjunkturforscher	economic researcher	wachstumspfad	growth path
konjunkturprognose	economic forecast	wachstumsdynamik	growth dynamics
konjunkturprognosen	economic forecasts	wachstumsimpulse	growth impulses
konjunkturexperte	economic expert	wachstumskräfte	growth forces
konjunkturexperten	economic experts	wachstumsbremse	brake on growth
konjunkturimpulse	economic stimulus	wachstumsschwäche	weak growth
konjunkturprogramm	economic stimulus program	nullwachstum	zero growth
konjunkturprogramme	economic stimulus programs	exportwachstum	export growth
konjunkturprogrammen	economic stimulus programs	produktivitätswachstum	productivity growth
konjunkturpaket	economic stimulus package	beschäftigungswachstum	employment growth
konjunkturpakete	economic stimulus packages	wachstumsbeitrag	growth contribution
konjunkturstütze	economic support	konsum	consumption
konjunkturtest	economic test	konsums	consumption

Continued on next page

Table 15 – *Continued from previous page*

German Term	English Translation	German Term	English Translation
konjunkturumfrage	business climate survey	privatkonsum	private consumption
konjunkturell	economically	investitionsschwäche	investment weakness
konjunkturelle	economic	investitionstätigkeit	investment activities
konjunkturellen	economic	ausrüstungsinvestitionen	investments in equipment
konjunktureller	economic	unternehmensinvestitionen	corporate investments
konjunkturbedingte	cyclical	bauinvestitionen	construction investments
wirtschaft	economy	anlageinvestitionen	capital investments
volkswirtschaft	economy	investitionsneigung	propensity to invest
gesamtwirtschaft	overall economy	investitionsbereitschaft	willingness to invest
weltwirtschaft	global economy	investitionsklima	investment climate
binnenwirtschaft	domestic economy	investitionsausgaben	investment expenditure
exportwirtschaft	export economy	investitionsquote	investment rate
wirtschaftsbereiche	economic sectors	dynamik	dynamics
wirtschaftsaufschwung	economic upswing	aufschwung	upswing
wirtschaftsaufschwungs	economic upswing	aufschwungs	upswing
wirtschaftsentwicklung	economic development	aufholprozess	catching-up process
wirtschaftslage	economic situation	auftriebskräfte	upswing
wirtschaftsleistung	economic output	abwärtsspirale	downward spiral
wirtschaftsdynamik	economic dynamics	abflachung	slowdown
wirtschaftsbelebung	economic recovery	schrumpfung	shrinkage
wirtschaftstätigkeit	economic activity	dämpfung	dampening
wirtschaftsforschung	economic research	verschlechterung	deterioration
wirtschaftsforscher	economic researcher	aufwertung	appreciation
wirtschaftsforschern	economic researchers	geldentwertung	devaluation
wirtschaftsweise	economic experts	anspringen	pick up
wirtschaftsweisen	economic experts	dämpfen	dampen
wirtschaftsinstitute	economic institutes	abwürgen	stall
wirtschaftsforschungsinstitut	economic research institute	abgewürgt	stalled
wirtschaftsforschungsinstitute	economic research institutes	absinken	sink
wirtschaftsforschungsinstituten	economic research institutes	verschlechtern	deteriorate
jahreswirtschaftsbericht	annual economic report	verpuffen	fizzle out
volkswirtschaftliche	economic	abgekoppelt	decoupled
volkswirtschaftlichen	economic	dämpfende	damping
gesamtwirtschaftlich	macroeconomic	dämpfenden	damping
gesamtwirtschaftliche	macroeconomic	lahmende	sluggish
gesamtwirtschaftlichen	macroeconomic	niedrigere	lower
außenwirtschaftliche	foreign economic	nachhaltige	sustainable
außenwirtschaftlichen	foreign economic	nachhaltigen	sustainable
binnenwirtschaftlichen	domestic economic	allmähliche	gradual
weltwirtschaftliche	global economic	allmählichen	gradual
weltwirtschaftlichen	global economic	moderate	moderate
ökonomen	economists	maßvollen	moderate
makroökonomische	macroeconomic	selbsttragenden	self-sustaining
makroökonomischen	macroeconomic	spürbare	noticeable
reale	real	durchgreifende	thorough
realen	real	durchgreifenden	thorough
reales	real	forschungsinstitute	research institutes
realer	real	ifo	ifo Institute
nominale	nominal	ifoinstitut	ifo Institute
nominalen	nominal	ifoinstituts	ifo Institute
währungsraum	currency area	ifochef	ifo head
gesamtdeutschland	Germany as a whole	ifopräsident	ifo president
eurogebiet	euro area	hanswerner	Sinn Hans-Werner
eurostaaten	eurozone countries	ifw	IfW

Continued on next page

Table 15 – *Continued from previous page*

German Term	English Translation	German Term	English Translation
mehrwertsteuererhöhung	increase in VAT	diw	DIW
steuereinnahmen	tax revenues	rwi	RWI
abgabenerhöhungen	tax increases	iwh	IWH
abgabenbelastung	tax burden	iw	IW
fiskalpolitik	fiscal policy	hüther	Michael Hüther
staatshaushalte	national bungets	hwwa	HWWA
staatsausgaben	government spending	straubhaar	Thomas Straubhaar
defizit	deficit	wochenbericht	weekly report
verschuldung	debt	frühjahrsgutachten	spring report
staatsverschuldung	national debt	herbstgutachten	fall report
staatsdefizit	government deficit	mittelfristige	medium-term
staatsdefizite	government deficits	mittelfristigen	medium-term
haushaltsdefizite	budget deficits	kreditklemme	credit crunch
defizitquote	deficit rate	exporte	exports
konsolidierungskurs	consolidation course	exporten	exports
finanzierungsbedingungen	financing conditions	exportboom	export boom
strukturprobleme	structural problems	exportchancen	export opportunities
strukturreformen	structural reforms	exportweltmeister	export world champion
inflationsbekämpfung	fighting inflation	außenbeitrag	net exports
strukturell	structural	außenhandel	foreign trade
strukturelle	structural	außenhandels	foreign trade
strukturellen	structural	bga	BGA
struktureller	structural	bgapräsident	BGA president
stimulierung	stimulation	börner	Anton Börner
stimulieren	stimulate	gewerbe	industry
ankurbelung	stimulation	gewerbes	industry
anzukurbeln	stimulate	verarbeitende	manufacturing
gegensteuern	counteract	verarbeitenden	manufacturing
expansiv	expansive	dienstleistungssektor	service sector
expansive	expansive	baubranche	construction industry
expansiven	expansive	wirtschaftsbau	commercial construction
auswirke	effect	bauwirtschaft	construction industry
wirkungen	effects	preisniveaus	price levels
effekte	effects	energiepreise	energy prices
arbeitsmarkt	labor market	preissteigerung	price increase
arbeitsmarktentwicklung	labor market development	preissteigerungsrate	price increase rate
arbeitsmarktlage	labor market situation	preissteigerungsraten	price increase rates
arbeitslosigkeit	unemployment	inflationsrate	inflation rate
arbeitslosenrate	unemployment rate	teuerungsrate	inflation rate
arbeitslosenquote	unemployment rate	preisliche	pricewise
arbeitslosenzahl	number of unemployed	rahmendaten	fundamentals
arbeitslosenzahlen	unemployment figures	bip	GDP
beschäftigungssituation	employment situation	bruttoinlandsprodukt	GDP
beschäftigungslage	employment situation	bruttoinlandprodukts	GDP
beschäftigungsentwicklung	employment development	bruttoinlandsprodukts	GDP
beschäftigungszuwachs	employment growth	bruttoinlandsproduktes	GDP
beschäftigungsaufbau	employment growth	sozialprodukt	national product
beschäftigungsabbau	reduction in employment	sozialprodukts	national product
reallöhne	real wages	verbrauch	consumption
lohnsteigerungen	wage increases	verbrauchs	consumption
lohnentwicklung	wage developments	kaufkraft	purchasing power
lohnpolitik	wage policy	realeinkommen	real income
lohnzurückhaltung	wage restraint	sparquote	savings rate
lohnabschlüsse	wage agreements	rate	rate

Continued on next page

Table 15 – *Continued from previous page*

German Term	English Translation	German Term	English Translation
lohnabschlüssen	wage agreements	niveaus	levels
tarifabschlüsse	collective agreements	prozentpunkt	percentage point
tarifabschlüssen	collective agreements	jahresschnitt	annual average
lohnstückkosten	unit labor costs	jahresdurchschnitt	annual average
arbeitskosten	labor costs	saisonbereinigte	seasonally adjusted
arbeitsproduktivität	labor productivity	saisonbereinigten	seasonally adjusted
produktivität	productivity	ungleichgewichte	imbalances
produktivitätszuwachs	increase in productivity		

Note: All terms have been translated from German to English using DeepL. The color coding is subjective, intended solely to improve readability and to distinguish groups of words that naturally cluster together.

Appendix D.6.2 Terms related to the word 'economy'

Table 16: Terms related to 'Economy'

German Term	English Translation	German Term	English Translation
wirtschaft	economy	konjunktur	business cycle conditions
wirtschafts	economic	binnenkonjunktur	domestic economy
volkswirtschaft	economy	konjunkturprogramm	economic stimulus program
weltwirtschaft	global economy	konjunkturprogramme	economic stimulus programs
ökonomie	economy	wachstumskräfte	growth forces
wirtschaftswachstum	economic growth	wachstumsschwäche	weak growth
wirtschaftswachstums	economic growth	dynamik	dynamics
wirtschaftsaufschwung	economic upswing	aufschwung	upswing
wirtschaftsentwicklung	economic development	aufschwungs	upswing
wirtschaftsleistung	economic output	entwicklung	development
wirtschaftskraft	economic power	aufholprozess	catching-up process
marktwirtschaft	market economy	stärker	stronger
außenwirtschaft	global trade	wichtiger	more important
privatwirtschaft	private sector	industrie	industry
wirtschaftsverbände	business associations	unternehmern	entrepreneurs
wirtschaftsforschung	economic research	investitionsklima	investment climate
wirtschaftsforscher	economic researcher	binnennachfrage	domestic demand
wirtschaftsforschungsinstitute	economic research institutes	wettbewerbsfähigkeit	competitiveness
wirtschaftsinstitute	economic institutes	bdi	BDI
wirtschaftsexperten	economic experts	bdipräsident	BDI President
ökonom	economist	thumann	Jürgen Thumann
ökonomen	economists	dihk	DIHK
wirtschaftspolitik	economic policy	diht	DIHT
wirtschaftliche	economic	handelskammertag	chamber of commerce
wirtschaftlichen	economic	handelskammertages	chamber of commerce
wirtschaftlicher	economic	wansleben	Martin Wansleben
gesamtwirtschaftliche	macroeconomic	börner	Anton Börner
gesamtwirtschaftlichen	macroeconomic	arbeitsmarkt	labor market
wirtschaftspolitische	economic policy	arbeitsmarkts	labor market
wirtschaftspolitischen	economic policy	arbeitsmarktes	labor market
ökonomische	economic	arbeitsmärkte	labor markets
ökonomischen	economic	arbeitsmärkten	labor markets
finanz	finance	arbeitsmarktpolitik	labor market policy
finanzpolitik	fiscal policy	arbeitsmarktreformen	labor market reforms
steuerpolitik	tax policy	beschäftigung	employment

Continued on next page

Table 16 – *Continued from previous page*

German Term	English Translation	German Term	English Translation
steuersenkungen	tax cuts	vollbeschäftigung	full employment
staatsfinanzen	public finances	arbeitslosigkeit	unemployment
staatsquote	public spending ratio	massenarbeitslosigkeit	mass unemployment
staatsausgaben	government spending	arbeitslosenzahlen	unemployment figures
staatsverschuldung	national debt	lohnpolitik	wage policy
haushaltskonsolidierung	budget consolidation	oecd	OECD
defizite	deficits	sachverständigenrat	expert council
reformen	reforms	forschungsinstitute	research institutes
strukturreformen	structural reforms	ifw	IfW
strukturwandel	structural change	diw	DIW
strukturelle	structural	iwh	IWH
strukturellen	structural	straubhaar	Thomas Straubhaar
deregulierung	deregulation	sorgen	ensure/concerns
stimulierung	stimulation	rahmenbedingungen	general conditions
ankurbelung	stimulation	schaffen	create/manage

Note: All terms have been translated from German to English using DeepL. The color coding is subjective, intended solely to improve readability and to distinguish groups of words that naturally cluster together.

Appendix D.7 Examples of filtered articles by sentiment class

The table below provides three examples of filtered MTI articles in their original German, with one example from each sentiment class: negative, no clear tone, and positive.

Table 17: Sentences Retained for Sentiment Analysis

Sentiment	Retained Sentences
Negative	**Wirtschaft**: Seit zehn Jahren hat Frankreich kaum Wachstum, dazu kommt ein sattes **Defizit** im **Außenhandel**, hohe **Staatsverschuldung** (97% der **Wirtschaftsleistung**). **Arbeitslosigkeit**: Mit 10 Prozent ist die **Arbeitslosenquote** fast doppelt so hoch wie in Deutschland, dramatisch ist die Jugendarbeitslosigkeit (aktuell 23,7%, mehr als in Rumänien). Innere Zerrissenheit: In Frankreich grassiert die Angst vor der **Arbeitslosigkeit**.
No clear tone	Darin heißt es: Das **Wirtschaftswachstum** könnte um bis zu drei Prozent höher ausfallen, würden Umwelt- und Menschenrechtsgruppen nicht gegen Kohleabbau und Atomkraft lobbyieren.
Positive	Berlin - Die deutsche **Wirtschaft** wächst! Das sagen die **Wirtschaftsweisen** in ihrer Prognose für 2014 voraus. Demnach dürfte das **Bruttoinlandsprodukt** 2014 um 1,9 Prozent und damit **stärker** als bisher angenommen steigen. Neben steigendem privaten **Konsum** beflügeln auch stetig wachsende Firmen-Investitionen in neue Anlagen und Maschinen die **Wirtschaft**.

Note: These examples are drawn from the MTI dataset and display only the sentences that were retained. A sentence is included if it contains at least one term related to business cycle conditions; these terms are highlighted in bold for clarity.

Appendix D.8 LSTM: mathematical details

To briefly explain the mathematical details of the LSTM model, let's consider an input sequence of length T, where each word in the sequence $x^1, \ldots, x^T$ is represented by an embedding vector of length I. The model begins processing the sequence at $t = 1$ and iteratively applies a set of update equations until $t = T$. At each time step t, the LSTM cell (illustrated in Figure 13) takes in three inputs: the previous cell state b_c^{t-1}, the previous hidden state b_h^{t-1}, and the current word embedding x^t. For simplicity, all the equations described here pertain to one unit within the LSTM cell, corresponding to one element in the hidden state and cell state vectors, as well as in all the gates. For example, with 32 units, the input gate is a vector of dimension 32, but our formula represents each individual element a_ι^t of this vector. For a more general discussion of LSTM cells, please refer to Nasekin and Chen (2020). In the case of multiple layers, the process remains the same; however, instead of using the input x^t, the hidden state b_h^t from the previous LSTM layer is used as input to the subsequent layer.

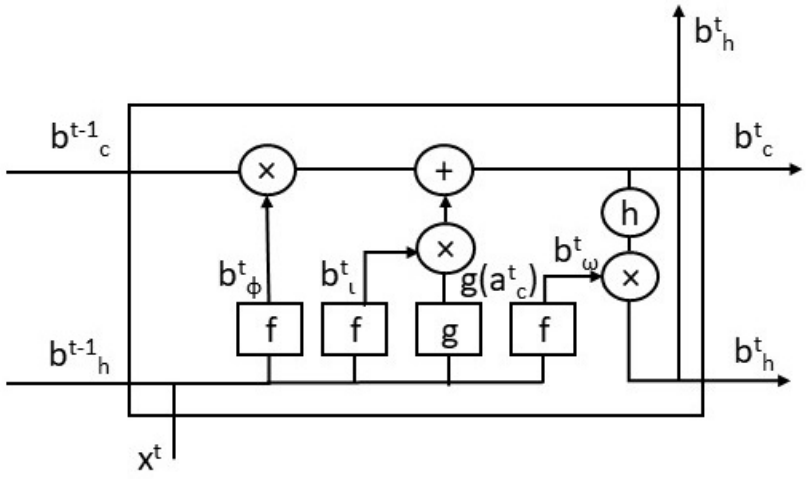


Figure 13: Structure of an LSTM cell.

The first gate in the LSTM cell, which determines what new information will be stored in the cell state, is known as the "input gate". This gate relies on the current input and the previous hidden state:

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1}, \tag{21}$$

where $w_{i\iota}$ represents the weight of the connection from the i-th unit in the input vector to the ι-th unit in the input gate, and $w_{h\iota}$ is the weight from the h-th unit in the hidden state to the ι-th unit in the input gate. Here, H denotes the number of units in the hidden layer.

The activation of the units in the input gate is given by:

$$b_\iota^t = f(a_\iota^t), \tag{22}$$

where f is the activation function of the gates, typically the sigmoid function[17]. This function outputs a value between 0 and 1, where 1 implies that the information is fully retained in the cell state, and 0 suggests that the information is completely discarded.

The second gate involved in updating the cell state is the "forget gate", which decides what information should be removed:

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1}, \quad b_\phi^t = f(a_\phi^t)\,. \tag{23}$$

The input and forget gates together modify the cell state. This process involves two main operations. First, a candidate for the cell state, representing new information, is computed as:

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}. \tag{24}$$

The actual cell state b_c^t is then updated by combining the previous cell state, scaled by the forget gate, with the candidate state, scaled by the input gate:

$$b_c^t = b_\phi^t b_c^{t-1} + b_\iota^t g(a_c^t), \tag{25}$$

where g is the cell candidate activation function, typically the hyperbolic tangent (tanh) function[18].

Finally, the hidden state needs to be updated. This is achieved using the "output gate", which is defined by:

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1}, \quad b_\omega^t = f(a_\omega^t). \tag{26}$$

[17]The sigmoid function is defined as $f(x) = \frac{1}{1+e^{-x}}$.

[18]The tanh function is defined as $g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, and it outputs values between -1 and 1.

The output gate is then multiplied by the updated cell state, which is passed through an activation function h, typically chosen to be the tanh:

$$b_h^t = b_\omega^t h(b_c^t). \tag{27}$$

Thus, the updated hidden state, representing the final output of the LSTM cell, is a filtered version of the updated cell state. This hidden state is then passed to the output layer, which applies a sigmoid activation function:

$$y = \sigma(a), \quad \text{where} \quad a = \sum_{h=1}^{H} w_h b_h^t. \tag{28}$$

Here, y represents the probability that the sentiment is positive or neutral. For the final time step T, this probability determines the sentiment classification: if y is 0.5 or greater, the sentiment of the article is classified as positive or no clear tone; otherwise, it is classified as negative.

All the weights in the LSTM are updated using backpropagation through time (BPTT), where gradients are accumulated over all T steps of the sequence before applying gradient descent. For detailed equations and further explanation, please refer to Graves (2012a) and Nasekin and Chen (2020).

Appendix D.9 Examples from the 'No clear tone' sentiment class

The table below provides three examples of articles classified under the 'No clear tone' category in their original German.

Table 18: Examples of Articles Annotated as 'No Clear Tone'

Type	Retained Sentences
Neutral sentiment	So hat sich die **Wirtschaft** seit dem entwickelt: **Arbeitslosigkeit**: vor der Abstimmung waren 1,64 Millionen arbeitslos, heute (Stand: Okt. 2016) sind es 16 000 weniger. Handel: die **Exporte** stiegen, das Handelsdefizit sank von 11,4 Milliarden (Stand Juni) auf 11 Milliarden Pfund (Stand Dezember). **Bruttoinlandsprodukt**: stieg von Juli bis September um 0,5 %.
Mixed sentiment	Kanzlerin Angela Merkel (59, CDU) empfing am späten Nachmittag die Chefs der weltweit mächtigsten **Wirtschaftsverbände**, u. a. Christine Lagarde (Währungsfonds IWF) und Angel Gurria (**OECD**). Die gute Nachricht: die **Weltwirtschaft** wird in diesem Jahr um 3,6 % wachsen, 2015 um 3,9 %. Die hohen **Staatsdefizite** vieler Länder könnten den **Aufschwung** gefährden, waren sich die **Wirtschaftsexperten** einig.
Limited information on the aspect	Als wichtigste europäische **Volkswirtschaft** müssen wir unserer globalen Rolle gerecht werden. Wenn die **Industrie** Kapazitäten für die Belieferung der Streitkräfte vorhalten soll, braucht es dazu ein klares Bekenntnis der Politik: für eine **nachhaltige** Finanzplanung.

Note: This table provides examples of articles classified under the 'No clear tone' category, illustrating different cases such as neutral sentiment, mixed sentiment, and articles that barely discuss business cycle conditions. These examples are drawn from the MTI dataset and display only the sentences that were retained. A sentence is included if it contains at least one term related to business cycle conditions; these terms are highlighted in bold for clarity.

Appendix D.10 Pre-processing for LSTM Model

Before estimating the LSTM model, we apply several standard pre-processing steps to the article texts:

1. **Convert text to lowercase:** All text is transformed to lowercase to ensure uniformity in case-sensitive words.

2. **Remove URLs:** Any URLs present in the text are removed to eliminate irrelevant content.

3. **Eliminate punctuation marks and non-alphabetic characters:** All punctuation marks and non-alphabetic characters are removed, leaving only word-based data.

4. **Normalize whitespaces:** Multiple consecutive spaces are reduced to a single space.

5. **Exclude single-letter tokens:** Tokens consisting of single letters are excluded from the text, as they typically represent words that contribute little to the overall meaning.

6. **Remove metadata from the articles:** Metadata includes information about the article that is not part of the main text.

7. **Filter words based on embeddings:** Words that do not have a corresponding embedding in the pre-trained word2vec model are excluded.

8. **Tokenize articles:** Each unique word is mapped to a unique integer, and article texts are converted into lists of integers. Our vocabulary consists of 29,240 unique words.

9. **Remove short articles:** Articles containing 20 or fewer words are excluded from the dataset.

10. **Standardize article length:** To handle variability in article length, each article is standardized to a fixed length of 200 words. Shorter articles are padded with zeros, while longer articles are truncated to include only the first 200 words. This ensures uniform input size for the LSTM model, allowing it to process articles of different lengths consistently.

Appendix D.11 Estimation Details for LSTM Model

The table below summarizes the key estimation details for our LSTM model:

Model Configuration	
Embedding layer	Converts input word indices into word embeddings
Embedding dimension	256 (pre-trained with word2vec)
Freeze embeddings	Yes (using pre-trained embeddings for faster training)
Hidden units per LSTM layer	32
Number of LSTM layers	2
Output layer	1 unit, sigmoid activation function
Dropout	50%, between LSTM layers and before the output layer
Initialization of hidden and cell states	Zero vectors
Maximum sequence length	200 words
Training Details	
Batch size	32
Number of epochs	40
Data shuffling	Yes, at the start of each epoch
Weight derivatives calculation	Truncated BPTT, 100-word chunks
Clipping threshold	5
Optimization Settings	
Optimizer	Adam
Learning rate	0.0001
Loss function	BCELoss (binary cross-entropy loss)

Table 19: Estimation details for the LSTM model.

Appendix D.12 Comparison with Alternative Sentiment Approaches

To provide a clearer interpretation of the LSTM model's performance, we also estimated a Linear Support Vector Machine (LSVM) model using the same 1,920 articles from the training set and evaluated it on the same 256 test articles. LSVM was selected as a benchmark because it is one of the most commonly used methods for text classification in economic and financial literature (see Kumar and Ravi, 2016), known for its strong performance. Our results show that the LSVM achieved an overall accuracy of 66.4%, which is very close to the LSTM's 66.8%. On the one hand, this similarity in accuracy is reassuring, indicating that our selected LSTM architecture avoids overfitting and generalizes well to unseen articles. On the other hand, it suggests that the potential advantages of neural networks, such as improved performance with larger datasets, might be limited by the relatively small size of our training set.

In addition to the LSVM benchmark, we compared the LSTM model's performance with

a lexicon-based approach, which we previously discussed as an alternative to machine learning methods. In our case, we applied the dictionary by Bannier et al. (2019) (henceforth BPW), a German adaptation of the Loughran and McDonald (2011) lexicon. This dictionary is specifically tailored for business communication and has been shown to produce sentiment indices that strongly correlate with economic and financial variables. To calculate sentiment for the test set articles, we computed the difference between the proportion of positive words and negative words identified in the dictionary. If the sentiment score was negative, the article was classified as having a negative sentiment towards business cycle conditions; otherwise, it was classified as positive or having no clear tone. The lexicon-based approach achieved an overall accuracy of 62.9%. While the LSTM model outperformed it, confirming the advantages of our methodology, the BPW dictionary still provided a robust benchmark, which explains its widespread use in the field.

Appendix D.13 Pre-processing for LDA Model

Prior to estimating the LDA model on our corpus, we applied several pre-processing steps to the article texts:

1. **Combine collocations into single tokens:** Collocations are meaningful sequences of two or three words, such as "business cycle". A token, in this context, represents a sequence of characters treated as a single unit by the model. To identify these collocations, we tagged each word in the articles using a Part-of-Speech (POS) tagger trained on the TIGER corpus (Brants et al., 2004). We then considered a sequence of words to be a collocation if it satisfied the POS patterns defined by Lang et al. (2018), such as Adjective-Noun, Noun-Noun, Noun-Preposition-Noun, Noun-Determiner-Noun, and Adjective-Adjective-Noun.

   To optimize computational performance, we limited our focus to the 2,000 most frequent two-word collocations and the 1,000 most common three-word collocations. Examples include *Angela_Merkel* (Germany's former chancellor), *IG_Metall* (Germany's largest metalworkers' union), and *Institut_für_Wirtschaftsforschung* (Institute for Economic Research). This transformation allows the model to capture the meaning of these phrases as a whole, rather than interpreting each word individually.

2. **Convert text to lowercase:** All text was converted to lowercase to ensure consistent

treatment of words, regardless of case.

3. **Remove apostrophes:** Apostrophes were removed to treat words as single tokens, preventing them from being split.

4. **Tokenization and token filtering:** The text was split into individual tokens. Non-alphabetic characters (such as numbers, punctuation, and currency symbols) and single-character words were removed to reduce noise, but tokens with underscores, representing collocations, were kept.

5. **Eliminate stopwords:** Frequently occurring words with little informational value, such as 'but', 'on', and 'he', were excluded to focus on more meaningful content. We used the Snowball stopword list for this step.[19]

6. **Exclude common names:** Frequent German first names and surnames were filtered out to avoid generating topics dominated by personal names, which would add little value to our analysis. The lists of common German first names and surnames used for this purpose are provided in Step 12 of the general dataset pre-processing.

7. **Stemming:** Tokens were stemmed using the Porter Stemmer for the German language.[20] This method reduces different grammatical forms of a word to a common stem, effectively standardizing the text. For example, *'kategorisch'* (categorical), *'kategorische'*, and *'kategorischen'* are all reduced to *'kategor'*.

8. **Remove stopwords after stemming:** After stemming, we performed another stopword removal to eliminate any stopwords that may have appeared during the process.

9. **Filter tokens by tf-idf:** Tokens with the lowest term frequency-inverse document frequency (tf-idf) scores were discarded, similar to the approach used by Hansen et al. (2018). These low-scoring tokens are either too rare or too common across the corpus, making them less useful for analysis. The tf-idf score for each token v is computed using the following formula:

$$\text{tf-idf}_v = \log(1 + N_v) \times \log\left(\frac{D}{D_v}\right) \tag{29}$$

[19] http://snowball.tartarus.org/algorithms/german/stop.txt
[20] http://snowball.tartarus.org/algorithms/german/stemmer.html

where N_v represents the number of times token v appears in the corpus, and D_v denotes the count of documents that contain the token.

Appendix D.14 Cross-validation Results for LDA

To determine the optimal number of topics, we applied 10-fold cross-validation with perplexity as the evaluation metric. Perplexity is used to assess how effectively a topic model can generalize to unseen data and and is calculated as follows:

$$\exp\left[-\frac{\sum_{d=1}^{D}\sum_{v=1}^{V} n_{d,v}\log\left(\sum_{k=1}^{K}\hat{\theta}_d^k\hat{\beta}_k^v\right)}{\sum_{d=1}^{D} N_d}\right], \tag{30}$$

where $n_{d,v}$ indicates how many times word v appears in document d.

The plot below shows that perplexity decreases as the number of topics increases, leading us to choose 200 topics as the optimal number.

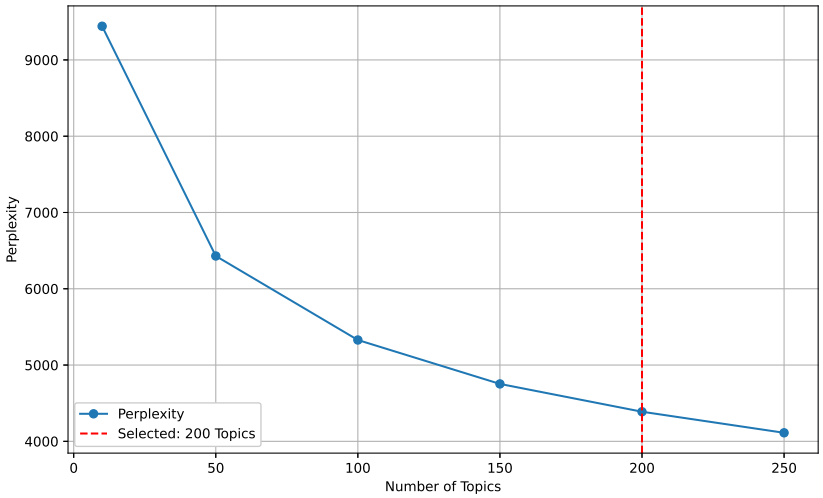


Figure 14: This plot shows the average perplexity values (Y axis) calculated on the test data for varying numbers of topics (X axis: 10, 50, 100, 150, 200, and 250). As the number of topics increases, perplexity decreases, indicating improved model performance. We selected 200 topics as the optimal point, where further gains in fit diminish, while computational demands become too high for larger topic numbers due to the dataset's size.

Appendix E Examples of the Estimated Topics

ID	Label	Most Probable Words
T0	Automotive Industry	autos (cars), fahrzeug (vehicle), herstell (manufacturer), pkw (passenger car), autoindustri (car industry), kfz (motor vehicle), vda (VDA - German Association of the Automotive Industry), diesel
T1	Stock Market Analysis and Investment Strategies	akti (stock), analyst, anleg (investor), bors (stock exchange), dax (DAX - German stock index), aktienmarkt (stock market), kurs (course), investor
T2	Corporate Governance and Financial Transparency	standard, transparenz (transparency), regeln (rules), bilanz (balance sheet), information, kontroll (control), pruf (check), wirtschaftspruf (auditor)
T3	Negotiations and Agreements	verhandl (negotiation), kompromiss (compromise), vereinbar (agreement), losung (solution), vereinbart (agreed), streit (dispute), scheit (fail), treff (meeting)
T4	Foreign Elections and Political Parties	partei (party), wahl (election), parlament (parliament), opposition, demokrat (democrat), konservativ (conservative), premi (premiere), sozialist (socialist), regierungschef (head of government), parlamentswahl (parliamentary election), kommunist (communist)
T5	Business Consulting and Management	mitarbeit (employee), berat (consultation), manag (manage), management, partn (partner), unternehmensberat (business consulting), erfahr (experience), geschaftsfuhr (business leader)
T6	Demonstrations and Protests	prot (protest), demonstration, aktion (action), polizei (police), demonstrant (demonstrator), strass (street), gewalt (violence), tausend (thousand), unruh (unrest)
T7	Culture, Arts and Literature	buch (book), kultur (culture), kunst (art), geschicht (history), bild (picture), les (read), jahrhundert (century), autor (author), art
T8	Stock Market and Financial Indices	punkt (point), index, wall_street, bors (stock exchange), akti (share), dow (Dow Jones), dollar, nasdaq, fiel (fell), schloss (closed), new_york
T9	Economic Indicators and Consumer Sentiment	punkt (point), ifo (ifo Institute), stimmung (mood), index, aktuell (current), indikator (indicator), zew (Centre for European Economic Research), verbessert (improved), konjunktur (business cycle conditions), optimist

Table 20: Labels for the First 10 Estimated Topics. The 'Most Probable Words' column includes original German stems alongside their English translations. Labels are selected subjectively, based on the most probable words and the articles with the highest share of each corresponding topic.

Appendix F Topics Selected for Forecasting

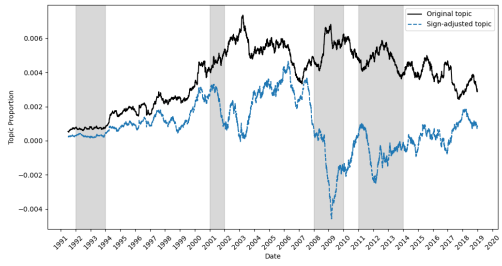
This section provides a description of the selected topics and their correlations with GDP growth. It also includes visualizations of the daily topic series and their sign-adjusted versions, as well as the results of robustness checks for the methodology used to construct sign-adjusted topics.

Appendix F.1 Overview of Selected Topics

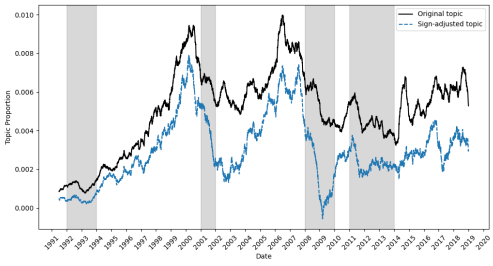
ID	Correlation	Label	Top 10 Count	Most probable words
T27	0.587/0.599	Economic Crises and Recessions	34	crisis, recession, economic crisis, deep, financial crisis, severe, dramatic, collapse, economic stimulus program, bad
T127	0.562/0.560	Major Banks and Investment Banking	34	Commerzbank, Deutsche Bank, institution, Dresdner Bank, investment bank, business, major bank, HypoVereinsbank, bank
T11	0.557/0.560	Mergers and Acquisitions	34	takeover, merger, corporation, business, competitor, partner, acquisition, strategy, subsidiary
T81	0.521/0.540	Corporate Restructuring and Job Cuts in Germany	34	employee, workplace, employed, Opel, plant, location, works council, dismissal, General Motors, reduction
T77	0.512/0.474	Private Investment	19	investor, fund, yield, investment, assets, real estate, saving, stock, long-term, capital
T74	0.495/0.497	Concerns about Economic Bubbles and Recessions	34	american, economist, America, danger, fear, boom, global, past, upswing, recession, soon, United States, world economy, bubble
T52	0.478/0.498	German Automobile Industry and Major Manufacturers	34	VW, Daimler, BMW, Chrysler, Ford, Porsche, Volkswagen, model, vehicle, group, cars
T131	0.468/0.462	German Investments in Emerging Markets	20	India, investment, investor, China, Indian, company, invest, engagement, invested, emerging country, Asia
T138	0.463/0.501	Financial and Economic Performance	32	billion, million, increase, last year, volume, share, rose, expects, revenue, increased
T100	0.456/0.478	Market Reactions to News	29	yesterday, pressure, so far, surprise, previously, recently, come under pressure, known, reacted, afterwards, announcement, announced, remained, prospect, signal, unexpectedly

Table 21: Selected topics and their correlations with the annualized quarterly GDP growth. The "Correlation" column lists the correlation with GDP growth for the first vintage and the average correlation across 34 vintages. The "Top 10 Count" indicates the number of vintages where the topic was among the 10 most correlated variables with GDP growth.

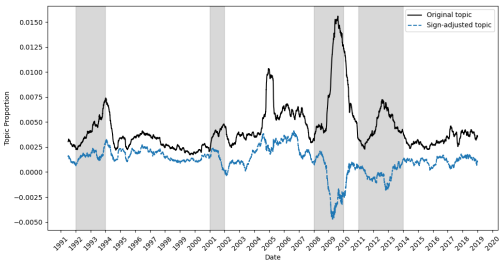
Appendix F.2 Daily Topics and Their Sign-Adjusted Versions



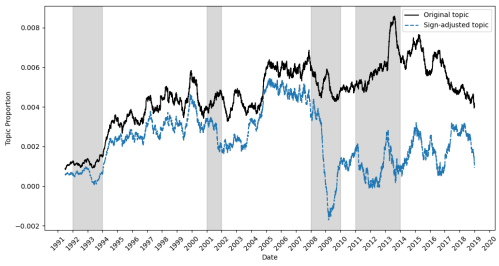
(a) Topic 127: Banking



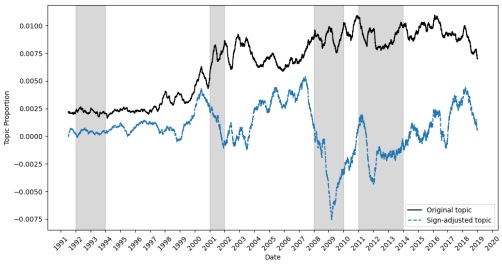
(b) Topic 11: M&As



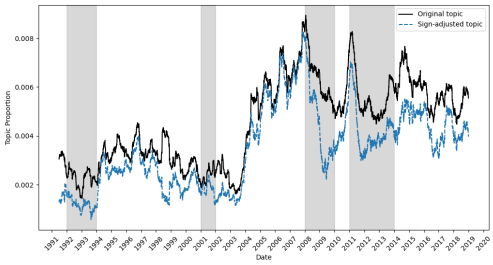
(c) Topic 81: Job Cuts



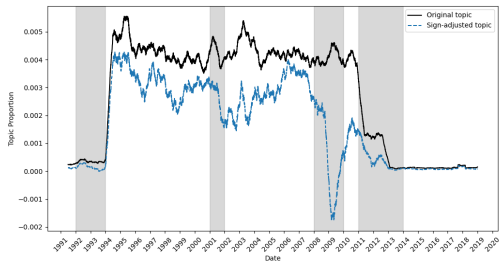
(d) Topic 77: Private Investment



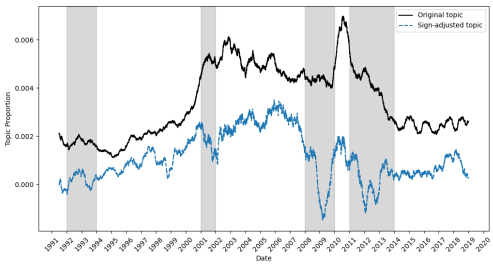
(e) Topic 74: Economic Bubbles



(f) Topic 131: Emerging Markets



(g) Topic 138: Economic Performance



(h) Topic 100: Market Reactions to News

Figure 15: 180-day backward rolling mean of daily topic series (black) and sign-adjusted topic series (blue) for the topics used in the out-of-sample forecasting experiment. The Y-axis shows the 180-day backward moving average of the daily topic proportion or sign-adjusted topic proportion, while the X-axis represents the specific day. Shaded areas indicate periods of severe recessions in Germany.

Appendix F.3 Robustness Analysis

In this subsection, we present the results of two robustness checks designed to assess the reliability of our methodology. The first robustness check evaluates the impact of adjusting the sign using 9 and 7 articles instead of the default 11. To analyze the effect of this modification, we plotted the standardized 180-day backward rolling mean of sign-adjusted topics for these different parameter values. Figure 16 presents the series for Topic 11 ("Mergers and Acquisitions"), though we conducted the same analysis for all selected topics.[21] The results indicate that our findings are robust to variations in the number of articles. We attribute this to the fact that our dataset includes four media sources, and the selected topics are consistently discussed in sufficient volume. On days when these topics are actively covered, each source likely contributes at least 2-3 relevant articles, ensuring adequate sentiment information. On days with low topic proportions, the sign becomes less critical, as the topic's overall influence is minimal. The only period where this might differ is 1991-1993, when only dpa was available, but for simplicity, we applied the 11-article rule across the entire period.

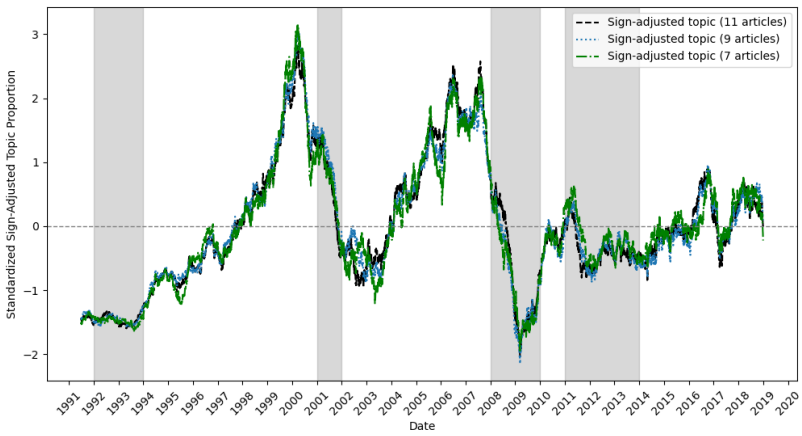


Figure 16: Standardized 180-day backward rolling mean of sign-adjusted topic proportions for Topic 11 ("Mergers and Acquisitions"). The plot compares results based on sentiment determined using 7, 9, and 11 articles. The Y-axis represents the standardized 180-day backward moving average of the sentiment-adjusted topic proportion, while the X-axis shows the specific day. Shaded areas indicate periods of severe recessions in Germany.

The second robustness check examined an alternative approach to adjusting topic proportions. Rather than relying on a majority vote for sign adjustment, we computed the average

[21] The analysis for all selected topics is available here: https://github.com/MashenkaOkuneva/newspaper_analysis/tree/main/topics/selected_topics_plots_number_of_articles.

sentiment across the 11 articles.[22] However, after standardizing the sign- and sentiment-adjusted topic proportions, we found no significant difference between the two approaches. For example, this can be seen with Topic 127 in Figure 17.[23] In conclusion, the results demonstrate that our methodology remains robust to the tested modifications.

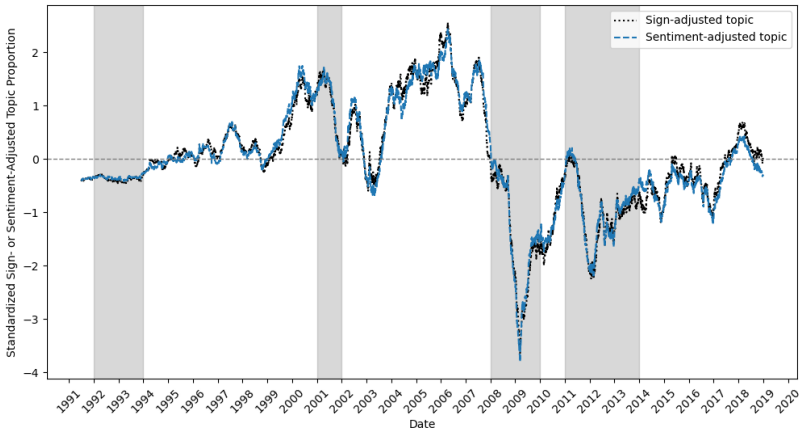


Figure 17: Standardized 180-day backward rolling mean of sign- and sentiment-adjusted topic proportions for Topic 127 ("Major Banks and Investment Banking"). The Y-axis represents the standardized 180-day backward moving average of the sign- or sentiment-adjusted topic proportion, while the X-axis shows the specific day. Shaded areas indicate periods of severe recessions in Germany.

[22] Here, we refer to topic series multiplied by the average topic-specific sentiment as "sentiment-adjusted topic proportions" to clearly distinguish them from sign-adjusted topic proportions.

[23] Results for other selected topics are available here: https://github.com/MashenkaOkuneva/newspaper_analysis/tree/main/topics/selected_topics_plots_sentiment_adjustment.

Appendix G Mixed-Frequency Structure of the DFM

The DFM's mixed-frequency structure is built upon the methodology of Mariano and Murasawa (2003). We assume that log GDP, Y_t^Q, is observed on the last business days of consecutive quarters, represented by $t_1(Q)$, $t_2(Q)$, and so on. Given that log GDP is a flow variable, the quarterly Y_t^Q is related to the unobserved daily log GDP, X_t, as:

$$Y_t^Q = \frac{1}{k_t} \sum_{s=t-k_t+1}^{t} X_s, \quad t = t_1(Q), t_2(Q), \dots , \tag{31}$$

where k_t denotes the number of business days in the quarter concluding on day t. Hence, for the quarterly growth rates, given by $y_t^Q = Y_t^Q - Y_{t-k_t}^Q$, the following holds:

$$y_t^Q = \sum_{s=t-k_t+1}^{t} \frac{t+1-s}{k_t} x_s + \sum_{s=t-k_t-k_{t-k_t}+2}^{t-k_t} \frac{s-t+k_t+k_{t-k_t}-1}{k_{t-k_t}} x_s, \quad t = t_1(Q), t_2(Q), \dots . \tag{32}$$

Equation (32) expresses the quarterly growth rates as a function of the daily growth rates x_s. By assuming that daily growth rates follow the same factor model as the daily variables y_t^D, we can formulate the measurement equation for $(y_t^{D\prime}\, y_t^{Q\prime})'$ using the aggregators of the daily factors, f_t^{QA} and f_t^{QP}:

$$\begin{pmatrix} y_t^D \\ y_t^Q \end{pmatrix} = \begin{pmatrix} \Lambda_D & 0 & 0 \\ 0 & \Lambda_Q & 0 \end{pmatrix} \begin{pmatrix} f_t \\ f_t^{QA} \\ f_t^{QP} \end{pmatrix} + \begin{pmatrix} \varepsilon_t^D \\ \varepsilon_t^Q \end{pmatrix}, \tag{33}$$

where y_t^Q is not observed for $t \neq t_1(Q), t_2(Q) \dots$. To explain how the aggregators bridge the observed quarterly GDP growth with the unobserved daily factors, we first give the formula for f_t^{QA}:

$$\begin{aligned} f_t^{QA} &= \sum_{s=t-k_t+1}^{t} \frac{t+1-s}{k_t} f_s + \sum_{s=t-k_t-k_{t-k_t}+2}^{t-k_t} \frac{s-t+k_t+k_{t-k_t}-1}{k_{t-k_t}} f_s \\ &= \sum_{s=t-k_t+1}^{t} W_s^C f_s + \sum_{s=t-k_t-k_{t-k_t}+2}^{t-k_t} W_s^P f_s, \end{aligned} \tag{34}$$

where W_s^C and W_s^P denote the weights assigned to the daily factors of the current and pre-

vious quarters, respectively. To implement this aggregation within a state-space framework, the inclusion of the second aggregator for the previous quarter, f_t^{QP}, is necessary. The construction of these aggregators is conditional on the time index:

1. When t corresponds to the first day of a quarter, i.e., $t = t_1(Q) + 1, t_2(Q) + 1, \ldots$:

$$f_t^{QA} = f_{t-1}^{QP} + W_t^C f_t \,,$$
$$f_t^{QP} = 0 \,;$$

2. For all other days:

$$f_t^{QA} = f_{t-1}^{QA} + W_t^C f_t \,,$$
$$f_t^{QP} = f_{t-1}^{QP} + W_t^P f_t \,.$$

For a deeper understanding of the state space representation, especially its adaptation for stock variables, we recommend referring to the original paper by Bańbura et al. (2011).

Appendix H Hard Data

Table 22: Economic Data

Name	**Frequency**	**Transformation**	**Code**	**Group**
Gross domestic product, chain index^{s,c}	quarterly	log, diff	BBKRT.Q.DE.Y.A.AG1.CA010.A.I	activity
Production in main construction industry ^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IP1.AA031.C.I	activity
Industrial production index^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IP1.ACM01.C.I	activity
New orders for main construction industry^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IO1.AA031.C.I	activity
New orders for industry^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IO1.ACM01.C.I	activity
Main construction industry turnover^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IT1.AA031.V.A	activity
Industry turnover^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.I.IT1.ACM01.V.I	activity
Consumer price index^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.P.PC1.PC100.R.I	prices
Consumer price index, excluding energy^{s}	monthly	log, diff	BBKRT.M.DE.S.P.PC1.PC110.R.I	prices
Producer price index^{s}	monthly	log, diff	BBKRT.M.DE.S.P.PP1.PP100.R.I	prices
Producer price index, excluding energy^{s}	monthly	log, diff	BBKRT.M.DE.S.P.PP1.PP200.R.I	prices
Hours worked: manufacturing^{s,c}	monthly	log, diff	BBKRT.M.DE.Y.L.BE2.AA022.H.I	labor market
Hours worked: construction^{s,c}	monthly	log, diff	BKRT.M.DE.Y.L.BE2.AA031.H.A	labor market
Federal notes yield (5-year)	daily	diff	LSEG Datastream	financial
Government bond yields (10-year)	daily	diff	LSEG Datastream	financial
Nominal effective exchange rate, narrow	daily	log, diff	BBEE1.D.I9.AAA.XZE012.A.AABAN.M00	financial
Nominal effective exchange rate, broad	daily	log, diff	BBEE1.D.I9.AAA.XZE022.A.AABAN.M00	financial
DAX performance-index	daily	log, diff	^GDAXI (Yahoo finance)	financial

Appendix I Dimensionality Reduction Techniques

The Unrestricted MIDAS model was estimated using several methods capable of handling many predictors: Ridge (Hoerl & Kennard, 1970), LASSO (Tibshirani, 1996), PCA (Stock & Watson, 2002), and Random Forests (Breiman, 2001). In this appendix, we provide a brief explanation of each approach.

Let H_t denote the N_H-dimensional vector of predictors, consisting of both text-based and hard data series, along with their lags. The selection of lags included in the model is discussed in the main text (see Equations (10) and (11)). Moreover, the parameters associated with the text and hard data series—denoted by β and θ, respectively—are combined into a single vector of parameters, μ.

Appendix I.1 Ridge and LASSO Regressions

Ridge and LASSO regressions are shrinkage methods designed to address overfitting in high-dimensional predictor space by minimizing a penalized residual sum of squares, formulated as follows:

$$\hat{\mu} = \arg\min_{\mu} \sum_{t=1}^{T} (y_{t+h} - \alpha_{h+1} - H_t\mu)^2 + \lambda \sum_{j=1}^{N_H} |\mu_j|^q, \tag{35}$$

where the parameter q determines the type of penalty applied. When $q = 1$, the method corresponds to LASSO with an L_1-norm penalty, while $q = 2$ corresponds to Ridge regression with an L_2-norm penalty. The parameter λ, common to both methods, controls the amount of shrinkage, with larger values leading to greater shrinkage. We selected the value of λ using 10-fold cross-validation.

While both methods effectively handle a large number of regressors and potential non-linearities, LASSO has the unique ability to perform variable selection by shrinking some coefficients (μ_j) exactly to zero, effectively excluding irrelevant predictors from the model. Ridge regression, on the other hand, shrinks coefficients towards zero without setting them to exactly zero, making it more suitable for scenarios where all variables are expected to contribute to the outcome, albeit to varying degrees.

Appendix I.2 Principal Component Analysis (PCA)

Another approach to reduce dimensionality is Principal Component Analysis (PCA). Let S_t denote the dataset that combines contemporaneous values of all text-based series X_t and all hard data series Z_t. Instead of using the original predictors H_t in the MIDAS model, we first extract r static factors from S_t and then use these factors and their K lags as regressors.

Following Stock and Watson (2006), the factors are estimated by solving the following optimization problem:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1} \sum_{t=1}^{T} (S_t - \Lambda F_t)^\top (S_t - \Lambda F_t), \tag{36}$$

subject to the constraints $\Lambda^\top \Lambda = I_r$, and $\Sigma_F = \mathbb{E}(F_t F_t^\top)$ being diagonal. Here, Λ is the matrix of loadings, and F_t is a vector of r factors, both treated as unknown parameters to be estimated.

The solution to this optimization problem is obtained by setting $\hat{\Lambda}$ to the first r eigenvectors of the sample variance matrix $\hat{\Sigma}_S = T^{-1} \sum_{t=1}^{T} S_t S_t^\top$. The estimated factors are then given by:

$$\hat{F}_t = \hat{\Lambda}^\top S_t, \tag{37}$$

which represent the first r principal components of S_t.

When S_t contains missing observations, the factors are estimated using the Expectation-Maximization (EM) algorithm, as described in Stock and Watson (2002). This algorithm iteratively imputes missing values in S_t and updates the factor estimates. Finally, the extracted factors, along with their K lags, are used in the original MIDAS regression in place of the predictors H_t, and the estimation is carried out via ordinary least squares (OLS).

Appendix I.3 Random Forests

Finally, we consider Random Forests, a non-linear technique for dimensionality reduction. This method relies on regression trees, which partition the predictor space H_t into distinct regions by identifying splits that minimize prediction error. Random Forests combine multiple trees trained on random subsets of the data to improve performance and reduce model variance. Below, we provide further details on both approaches.

Appendix I.3.1 Regression Trees

Following Hastie et al. (2001), regression trees divide the predictor space H_t into two regions by selecting a variable j and a split point s that minimize the sum of squared errors within those regions. Specifically, the algorithm solves the following optimization problem:

$$\min_{j,s} \left[\min_{c_1} \sum_{H_t \in R_1(j,s)} (y_{t+h} - c_1)^2 + \min_{c_2} \sum_{H_t \in R_2(j,s)} (y_{t+h} - c_2)^2 \right], \tag{38}$$

where the two regions are defined as:

$$R_1(j,s) = \{H_t \mid H_{tj} \leq s\}, \quad R_2(j,s) = \{H_t \mid H_{tj} > s\}. \tag{39}$$

For a given split, the optimal predictions $\hat{c}_1$ and $\hat{c}_2$ are the mean values of y_{t+h} within their respective regions:

$$\hat{c}_1 = \text{ave}(y_{t+h} \mid H_t \in R_1(j,s)), \quad \hat{c}_2 = \text{ave}(y_{t+h} \mid H_t \in R_2(j,s)). \tag{40}$$

Once the optimal split is identified, the data is divided, and the process continues recursively within each resulting region until a stopping criterion—here, a minimum of 5 observations per node—is reached. The final prediction is based on the average value of y_{t+h} in the terminal regions:

$$\hat{f}(H_t) = \sum_{m=1}^{M} \hat{c}_m \cdot \mathbb{I}(H_t \in R_m), \tag{41}$$

where M is the number of terminal nodes, and $\hat{c}_m$ is the predicted value of the dependent variable in region R_m.

Appendix I.3.2 Random Forests

Regression trees are simple and interpretable but can suffer from high variance. Random Forests address this limitation by using two key techniques: bootstrap aggregation and random feature selection.

First, the algorithm generates B bootstrap samples by sampling from the original dataset with replacement. For each sample, a deep regression tree is grown, and the final prediction is obtained by averaging the predictions of all trees. This aggregation process reduces the variance

of the model.

Second, at each split in a tree, the algorithm considers only a randomly selected subset of p predictors rather than evaluating all predictors. This reduces the correlation between individual trees, as even weak predictors may contribute to splits in some trees.

In this paper, we use 500 bootstrap samples and select 1/3 of the predictors at each split, which are standard choices in the literature.

Appendix J Additional Forecasting Experiment Results

This appendix provides the optimal weights assigned to the text-based model in forecast combinations for both the DFM and MIDAS. Additionally, it presents the out-of-sample forecasting results for the MIDAS model with its best-performing specifications.

Appendix J.1 Optimal Weights

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
DFM	0.32	0.66	0.61	0.51	0.87	0.35	0.66	1	1	0.64
MIDAS	0.36	0.75	0.42	0.29	0.14	0.70	0.24	0.47	0.69	0.59

Table 23: Optimal weights for the text-based model in forecast combinations. The table shows the weight assigned to the text-based model when linearly combining DFM and MIDAS forecasts from text-only and hard-data-only models. Weights are optimized to minimize the MSE of the combined forecast over the evaluation period, for each forecasting horizon (backcast, nowcast, 1-step-ahead, and 2-step-ahead) separately. Optimization is conducted subject to the constraint that weights are non-negative and sum to unity. For MIDAS, results rely on Ridge regression with $K = 3$. Forecasts are generated 30, 60, and 90 days into the quarter.

Appendix J.2 MIDAS with Best-Performing Specifications

Next, we present the results for the best-performing MIDAS specifications. First, we provide a table summarizing the best-performing models, selected from a set of 24 competing specifications that varied K (1 to 6) and applied different dimensionality reduction techniques. These models were chosen ex-post based on their overall RMSFE performance across all horizons.

Next, we compare the performance of the best-performing MIDAS models against the AR(1) and SPF benchmarks, as well as the relative performance of forecast combinations compared to the hard-data-only model. To construct the combined forecasts, we used the best-performing hard-data model and the best-performing text-only model. We also report the optimal weight assigned to the text-based model in these combinations. The results are qualitatively similar to those based on Ridge regression with $K = 3$ and discussed in the main text.

Forecast Timing	Data Type	Model	K
30 days	Text	Ridge	3
	Hard	LASSO	3
	Text and Hard	Ridge	3
60 days	Text	LASSO	3
	Hard	Ridge	3
	Text and Hard	Ridge	1
90 days	Text	PCA	3
	Hard	Ridge	2
	Text and Hard	Ridge	4

Table 24: Summary of the best-performing MIDAS models for forecasts produced 30, 60, or 90 days into the quarter. The selected models had the best performance (measured via RMSFE) when considering all forecasting horizons together. Model selection was conducted ex-post, using the benefit of hindsight on actual GDP growth during the evaluation period, from a set of 24 competing specifications that varied in K (1 to 6) and applied different dimensionality reduction techniques, including LASSO, Ridge, PCA, and RF.

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Only text										
MIDAS	**0.87**	**0.82***	**0.89**	**0.91**	1.11	**0.93**	**0.91**	1.00	1.04	1.00
Only hard data										
MIDAS	**0.75**	**0.88**	**0.78**	**0.60****	1.00	1.06	**0.93**	1.02	1.17	1.03
Text and hard data										
MIDAS	**0.74**	**0.85**	**0.92**	**0.70**	1.52	**0.92***	**0.88**	1.03	1.29	1.14
Forecast combination (optimal weights)										
MIDAS	**0.68**	**0.77***	**0.69***	**0.57****	0.99	**0.92**	**0.88**	0.97	1.04	1.00
Forecast combination (equal weights)										
MIDAS	**0.69***	**0.77**	**0.69***	**0.63****	1.02	**0.94**	**0.88**	0.97	1.07	1.00

Table 25: Relative RMSFE Scores: MIDAS Models vs AR(1). This table presents relative RMSFEs for MIDAS models estimated using text data only, hard data only, and models integrating both sources. The results are based on the best-performing specifications. Relative RMSFEs for forecast combinations of hard-only and text-only models using optimal and equal weights are also included. All values are expressed relative to the RMSFE of the AR(1) benchmark. Bold entries indicate RMSFEs at least 5% lower than that of the AR(1). Asterisks denote statistical significance based on one-sided DM test (* 10%, ** 5%, *** 1%).

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Only text										
MIDAS	1.39	**0.92**	1.07	1.09	1.11	**0.88**	**0.87***	0.96	1.00	0.96*
Only hard data										
MIDAS	1.20	0.98	**0.94**	**0.72***	0.99	1.01	**0.89**	0.98	1.12	0.99
Text and hard data										
MIDAS	1.18	**0.95**	1.11	**0.84**	1.51	**0.88****	**0.84****	0.99	1.24	1.09
Forecast combination (optimal weights)										
MIDAS	1.09	**0.86**	**0.82**	**0.69****	0.99	**0.87**	**0.84***	**0.93****	1.00	0.96*
Forecast combination (equal weights)										
MIDAS	1.11	**0.86**	**0.83**	**0.75***	1.01	**0.89***	**0.84***	**0.93****	1.03	0.96

Table 26: Relative RMSFE Scores: MIDAS Models vs SPF. This table presents relative RMSFEs for MIDAS models estimated using text data only, hard data only, and models integrating both sources. The results are based on the best-performing specifications. Relative RMSFEs for forecast combinations of hard-only and text-only models using optimal and equal weights are also included. All values are expressed relative to the RMSFE of the SPF benchmark (Reuters Poll). Bold entries indicate RMSFEs at least 5% lower than that of the SPF. Asterisks denote statistical significance based on one-sided DM test (* 10%, ** 5%, *** 1%).

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
Optimal Weights										
MIDAS	**0.91**	**0.88**	**0.88***	**0.95**	1.00	**0.86**	**0.94**	**0.95***	**0.89**	0.97
Equal Weights										
MIDAS	**0.92**	**0.88**	**0.88**	1.04	1.02	**0.88**	**0.94**	**0.95***	**0.91**	0.98

Table 27: Relative RMSFE Scores: Forecast Combinations vs Hard-Data Models. This table presents relative RMSFEs for forecast combinations of MIDAS hard-only and text-only models using both optimal and equal weights. The results are based on the best-performing specifications. All values are reported relative to the RMSFEs of the models estimated using hard data only. Bold entries indicate RMSFEs that are at least 5% lower than those of the hard-only models. Asterisks denote statistical significance based on one-sided DM tests (* 10%, ** 5%, *** 1%).

	Backcast	Nowcast			1 Step			2 Steps		
Model	30	30	60	90	30	60	90	30	60	90
MIDAS	0.37	0.60	0.40	0.21	0.13	0.78	0.55	0.56	0.91	1.00

Table 28: Optimal weights for the text-based model in forecast combinations. The table shows the weight assigned to the text-based model when linearly combining MIDAS forecasts from text-only and hard-data-only models. Weights are optimized to minimize the MSE of the combined forecast over the evaluation period, for each forecasting horizon (backcast, nowcast, 1-step-ahead, and 2-step-ahead) separately. Optimization is conducted subject to the constraint that weights are non-negative and sum to unity. The combinations use the best-performing hard-only and text-only models. Forecasts are generated 30, 60, and 90 days into the quarter.