

Macroeconomic Forecasting and Market Analysis with Newspaper Articles

Inaugural-Dissertation

zur Erlangung des akademischen Grades einer Doktorin
der Wirtschafts- und Sozialwissenschaften
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

M.Sc.

Mariia Okuneva

aus Kungur, Perm Gebiet

Kiel, 2025

Gedruckt mit Genehmigung der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

Dekan:
Prof. Dr. Carsten Schultz

Erstberichterstattender:
Prof. Dr. Kai Carstensen

Zweitberichterstattender:
Prof. Dr. Matei Demetrescu

Tag der Abgabe der Arbeit:
5. Dezember 2025

Tag der mündlichen Prüfung:
17. Februar 2026

Acknowledgements

Writing this dissertation has been a long, challenging, and deeply meaningful journey, and I am sincerely grateful to all those who have supported and encouraged me along the way.

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Kai Carstensen. He taught me econometrics and showed me how to conduct research with rigor and curiosity. His guidance, support, and thoughtful advice have shaped every stage of my academic development. I am especially grateful to him for trusting me with the most interesting project of my career to date on the use of text data for economic forecasting and for ensuring that I had the computational resources necessary to carry out my work.

I am also very grateful to my second supervisor, Prof. Dr. Matei Demetrescu. The statistical knowledge I gained from him enabled me to understand and meaningfully apply the machine learning models used in this dissertation. I also had the opportunity to explore statistical research under his supervision, and I greatly enjoyed teaching together, particularly the “Data Mining” course, which remains one of my favourites. His support, openness, and enthusiasm made a lasting impact on my development as a researcher.

My sincere thanks go to Julian Schröder, whose help was indispensable for securing the computational resources required for my computationally demanding projects. Without his support, parts of this research would simply not have been possible.

I would also like to thank my co-authors—Kai Carstensen, Philipp Hauber, Jasper Bär, Clemens Knoppe, and Mikaella Zitti—for the great collaboration, insightful discussions, and their substantial contributions to this dissertation. Working together, including our many late nights in the office, has been both inspiring and rewarding.

I am also very grateful to Media Tenor International for providing the professionally annotated dataset used in the first chapter of this dissertation. I furthermore wish to thank the Deutsche Bundesbank, whose financial support enabled the acquisition of the dpa data used in that chapter, as well as the Fritz Thyssen Foundation, whose funding made it possible to obtain the news article datasets employed in both the first and second chapters of this thesis.

I am deeply thankful to my colleagues at the Institute for Statistics and Econometrics at Kiel University. Their kindness, the warm atmosphere, and the pleasure of teaching alongside them made my time at the institute truly enjoyable. I am equally grateful to my students; teaching them has been one of the highlights of my time in Kiel, and I learned a great deal from their questions and discussions.

I also want to thank all my friends for their support throughout this journey and for making Kiel feel like a second home. Special thanks go to Alina Lenhardt, Alisha Singh,

Johannes Scheuerer, Le Nga Tran, Mikaella Zitti, Clemens Knoppe, Salomon Fiedler, Uliana Zaspá, Rouven Lindenau, and Richard Schnorrenberger.

Above all, I am profoundly grateful to my family—my parents and my brother, Misha—for their unwavering love, encouragement, and belief in me. Their support cannot be put into words. Without them, this dissertation would not have been possible, and I dedicate this work to them.

Contents

List of Abbreviations	VII
List of Tables	X
List of Figures	XIII
Introduction	1
1 Nowcasting German GDP with text data	4
1.1 Introduction	5
1.2 Text data	8
1.3 Sentiment analysis	11
1.3.1 Training set	12
1.3.2 Sentiment Extraction Methodology	15
1.3.3 Daily sentiment index	24
1.4 Sign-adjusted topics	26
1.4.1 Latent Dirichlet Allocation	27
1.4.2 Selected sign-adjusted topics	30
1.5 Forecast encompassing analysis	33
1.6 Out-of-sample forecasting experiment	35
1.6.1 Dynamic Factor Model	35
1.6.2 Unrestricted MIDAS	37
1.6.3 Empirical results	40
1.7 Conclusion	44
Appendix 1	47
2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty	94
2.1 Introduction	95
2.2 Text indices	99
2.3 Forecasting experiment	102
2.4 Empirical results	103
2.4.1 GDP	103

Contents

2.4.2	Investment	116
2.4.3	Consumption	123
2.5	Conclusion	130
3	Salmon stock returns around market news	157
3.1	Introduction	158
3.2	Data & pre-processing	160
3.2.1	Financial data	160
3.2.2	Text data	161
3.3	Methodologies	165
3.3.1	Topic modelling	165
3.3.2	Sentiment approach	173
3.4	Empirical results	174
3.4.1	The impact of industry-specific topics	174
3.4.2	Incorporating lexicon-based sentiment	178
3.4.3	Resolving limitations of the Loughran-McDonald dictionary	186
3.4.4	Out-of-sample forecasting exercise	192
3.5	Discussion	194

List of Abbreviations

AR(1) first-order autoregressive process

BAKKA Bakkafrost, salmon producer

BamS Bild am Sonntag, German newspaper

BCS Business Cycle Sentiment

BoW Bag-of-Words representation

BPTT backpropagation through time

COVID-19 the global health crisis caused by the coronavirus

DFM Dynamic Factor Model

DM Diebold-Mariano test

dpa Deutsche Presse-Agentur, German news agency

EM Expectation-Maximization algorithm

EMH Efficient Market Hypothesis

ESI Economic Sentiment Indicator

Factiva Dow Jones Factiva, global news and business information database

FAS Frankfurter Allgemeine Sonntagszeitung, German newspaper

GDP Gross Domestic Product

GMT Greenwich Mean Time

GSF Grieg Seafood, salmon producer

HAC Heteroskedasticity- and autocorrelation-robust

IMF International Monetary Fund

Contents

LDA Latent Dirichlet Allocation algorithm

LM Loughran–McDonald sentiment dictionary

LSTM Long Short-Term Memory neural network

LSVM Linear Support Vector Machine

LSG Lerøy Seafood Group, salmon producer

MIDAS Mixed Data sampling

ML Maximum Likelihood

MOWI Marine Harvest, salmon producer

MSE Mean Squared Error

MTI Media Tenor International

NRS Norwegian Royal Salmon, salmon producer

NTS NTS ASA, aquaculture and salmon farming company

OECD Organisation for Economic Co-operation and Development

OBSI Overall Business Sentiment Index

OLS Ordinary Least Squares

PCA Principal Component Analysis

POS Part-of-Speech

RF Random Forests

q/q quarter-on-quarter

R&D research and development

RMSFE Root Mean Squared Forecast Error

RNN Recurrent Neural Network

ROW rest of the world

SALM SalMar, salmon producer

SPF Survey of Professional Forecasters

SPI Share Price Index

SZ Süddeutsche Zeitung

tanh hyperbolic tangent activation function

tf-idf term frequency-inverse document frequency

t-SNE t-distributed Stochastic Neighbour Embedding

VAR Vector Autoregression

WamS Welt am Sonntag, German newspaper

List of Tables

1.1	Summary of the original dataset	9
1.2	Descriptive statistics of the final dataset	10
1.3	Articles matching MTI annotation	14
1.4	Sentences retained for sentiment analysis	18
1.5	Examples of articles annotated as ‘no clear tone’	21
1.6	Article length statistics by source	23
1.7	LSTM performance	24
1.8	Descriptive statistics after LSTM-specific pre-processing	25
1.9	Correlations between GDP growth and selected sign-adjusted topics	34
1.10	Encompassing regressions across forecast horizons	35
1.11	Relative RMSFE scores: DFM and MIDAS models vs AR(1)	41
1.12	Relative RMSFE scores: DFM and MIDAS models vs SPF	42
1.13	Relative RMSFE scores: forecast combinations vs hard-data models	44
1.14	Estimation details for the word2vec model	64
1.15	Terms related to ‘business cycle conditions’	68
1.16	Terms related to ‘economy’	70
1.17	Sentences retained for sentiment analysis	71
1.18	Examples of articles annotated as ‘no clear tone’	74
1.19	Estimation details for the LSTM model	75
1.20	Labels for the first 10 estimated topics	80
1.21	Selected topics and their correlations with the annualized quarterly GDP growth	81
1.22	Economic data	86
1.23	Optimal weights for the text-based model in forecast combinations	90
1.24	Summary of the best-performing MIDAS models	91
1.25	Relative RMSFE scores: MIDAS models vs AR(1)	91
1.26	Relative RMSFE scores: MIDAS models vs SPF	92
1.27	Relative RMSFE scores for MIDAS models: forecast combinations vs hard- data models	92
1.28	Optimal weights for the text-based model in forecast combinations	93

List of Tables

2.1	Correlations of selected plain topics with quarterly GDP growth (first release) and selected surveys	105
2.2	Correlations of selected economic lexicon-adjusted topics with quarterly GDP growth (first release) and selected surveys	106
2.3	Correlations of selected BCS-adjusted topics with quarterly GDP growth and selected surveys	107
2.4	Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with GDP growth	108
2.5	Relative root mean squared forecast errors for GDP growth nowcasts across subperiods	113
2.6	Correlations of plain topics with quarterly GDP growth (full sample and excluding the Financial Crisis, 2008-2009)	115
2.7	Correlations of BCS-adjusted topics with quarterly GDP growth (full sample and excluding the Financial Crisis, 2008–2009)	115
2.8	Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with investment growth	117
2.9	Correlations of selected economic lexicon-adjusted topics with quarterly investment growth (first release) and selected surveys	117
2.10	Correlations of selected BCS-adjusted topics with quarterly investment growth (first release) and selected surveys	119
2.11	Relative root mean squared forecast errors for investment growth nowcasts across subperiods	122
2.12	Correlations of plain topics with quarterly investment growth (full sample and excluding the Financial Crisis, 2008–2009)	123
2.13	Correlations of BCS-adjusted topics with quarterly investment growth (full sample and excluding the Financial Crisis, 2008–2009)	124
2.14	Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with consumption growth	125
2.15	Correlations of selected BCS-adjusted topics with quarterly consumption growth (first release) and selected surveys	126
2.16	Relative root mean squared forecast errors for Consumption growth nowcasts across subperiods	129
2.17	Correlations of plain topics with quarterly consumption growth (full sample and excluding the Financial Crisis, 2008–2009)	131
2.18	Correlations of BCS-adjusted topics with quarterly consumption growth (full sample and excluding the Financial Crisis, 2008–2009)	132
2.19	Labels for the 200 estimated topics based on the most probable words . . .	134
2.20	Hard data and surveys used in the forecasting experiment	145

List of Tables

2.21	Correlations of selected general lexicon-adjusted topics with quarterly GDP growth (first release) and selected surveys	146
2.22	Correlations of uncertainty-adjusted topics with quarterly GDP growth and selected surveys	147
2.23	Correlations of selected plain topics with quarterly investment growth (first release) and selected surveys	149
2.24	Correlations of selected BCS-adjusted topics with quarterly investment growth (first release) and selected surveys	150
2.25	Correlations of selected general lexicon-adjusted topics with quarterly investment growth (first release) and selected surveys	151
2.26	Correlations of selected uncertainty-adjusted topics with quarterly investment growth (first release) and selected surveys	151
2.27	Correlations of selected plain topics with quarterly consumption growth (first release) and selected surveys	153
2.28	Correlations of selected economic lexicon-adjusted topics with quarterly consumption growth (first release) and selected surveys	153
2.29	Correlations of selected general lexicon-adjusted topics with quarterly consumption growth (first release) and selected surveys	154
2.30	Correlations of selected uncertainty-adjusted topics with quarterly consumption growth (first release) and selected surveys	154
2.31	Correlations of selected BCS-adjusted topics with quarterly consumption growth (first release) and selected surveys	155
3.1	Notations in LDA	167
3.2	Forecasting exercise results	194
3.3	Estimated topics with most probable words	197
3.4	Topic components and top-correlated topics	201
3.5	Components based on LM sentiment-weighted topics and top-correlated topics	202
3.6	Components based on topics adjusted with extended sentiment and top-correlated topics	203

List of Figures

1.1	Daily publications across datasets	11
1.2	Business cycle and labor market sentiment	13
1.3	Architecture of the LSTM model used for sentiment analysis	20
1.4	Daily publications across datasets after LSTM pre-processing	25
1.5	Daily Overall Business Sentiment Index across media outlets	26
1.6	Daily topic and sign-adjusted topic for Topic 27 (“Economic Crises and Recessions”)	32
1.7	Daily topic and sign-adjusted topic for Topic 52 (“German Automobile Industry and Major Manufacturers”)	32
1.8	Daily publications: dpa	60
1.9	Daily publications: SZ	60
1.10	Daily publications: Handelsblatt	60
1.11	Daily publications: Welt	60
1.12	t-SNE visualization of 1,000 words related to ‘business cycle conditions’	66
1.13	Structure of an LSTM cell	72
1.14	Perplexity across different numbers of topics	79
1.15	Selected topics and their sign-adjusted counterparts	82
1.16	Robustness of sign-adjusted Topic 11 to the number of articles used	83
1.17	Sign- and sentiment-adjusted topic dynamics for Topic 127	84
2.1	Robustness of sentiment adjustment to the number of articles for T29 (“Banking”) using the economic lexicon	101
2.2	Crisis topic in plain and sentiment-/uncertainty-adjusted forms (T50, and T27 in the BCS-adjusted case)	109
2.3	Economic Growth topic in plain and sentiment-/uncertainty-adjusted forms (T120)	110
2.4	Mean squared forecast errors (MSFEs) for different nowcasting models of GDP growth (2008–2018)	111
2.5	Mean squared forecast errors (MSFEs) for different nowcasting models of GDP growth, 2008–2010 and 2011–2018	112

2.6	Nowcasts and actual GDP growth (first release) for different models, 2008–2010 and 2011–2018	114
2.7	Mean squared forecast errors (MSFEs) for different nowcasting models of investment growth (2008–2018)	120
2.8	Mean squared forecast errors (MSFEs) for different nowcasting models of investment growth, 2008–2010 and 2011–2018	121
2.9	Mean squared forecast errors (MSFEs) for different nowcasting models of consumption growth (2008–2018)	128
2.10	Mean squared forecast errors (MSFEs) for different nowcasting models of consumption growth, 2008–2010 and 2011–2018	130
2.11	Nowcasts and actual GDP growth (first release) for different models, 2008–2010 and 2011–2018	148
2.12	Nowcasts and actual investment growth (first release) for different models, 2008–2010 and 2011–2018	152
2.13	Nowcasts and actual consumption growth (first release) for different models, 2008–2010 and 2011–2018	156
3.1	Share Price Index (SPI) for daily price data	161
3.2	Investors’ Reaction or SPI logarithmic returns	162
3.3	Number of articles per month, 5-month moving average	163
3.4	LDA algorithm schematic	166
3.5	Perplexity of the 100-topic LDA model estimated on the training data . . .	170
3.6	Illustrative example of topic modelling: Topic 36	171
3.7	Perplexity for different numbers of topics	172
3.8	Cumulated impulse responses for selected components	179
3.9	Cumulated impulse response of stock returns to innovation in component 1	181
3.10	Cumulated impulse responses of stock returns to innovations in components 3 and 4	182
3.11	Impulse responses for LM sentiment and topic 36 (Covid) multiplied with LM sentiment	183
3.12	Impulse responses for topic interactions with LM sentiment	185
3.13	Impulse responses for extended sentiment measures	187
3.14	Impulse responses for topic–sentiment interactions using the extended dictionary	189
3.15	Impulse response for component 1 based on topics multiplied with extended sentiment	191
3.16	Impulse responses for components 4 and 7 based on topics multiplied with extended sentiment	192

List of Figures

3.17 Robustness check: IRFs for components based on topics multiplied with
extended sentiment 204

Introduction

The overarching theme of this dissertation is whether a particular type of alternative data, namely text data, can improve forecasts of key macroeconomic aggregates or help explain the financial market dynamics. Large-scale text data have become available only relatively recently and have attracted substantial attention in the empirical macroeconomic and financial literature. My central research question concerns which dimensions of news data are informative for macroeconomic forecasting and financial analysis, and how these dimensions should be adapted to the variable of interest.

The existing literature has predominantly focused on three types of text-based indicators. The first strand constructs sentiment measures that quantify the positive or negative tone of news articles and shows that such measures can predict macroeconomic activity. A second strand uses topic indicators, capturing variation in the intensity with which different themes are covered in the media. A third, more recent strand combines these approaches by assigning sentiment to individual topics, and therefore produces topic-specific sentiment measures that reflect both the thematic focus and the tone of news coverage.

My work adopts this hybrid perspective but departs from standard off-the-shelf methods by asking whether commonly used text dimensions are appropriate for the specific forecasting question at hand. Rather than relying directly on existing lexicons or unsupervised topics, I adapt sentiment analysis and topic modelling so that the resulting indicators capture the economically relevant information for the variable being forecast, whether GDP, consumption, investment, or stock returns.

Chapter 1: Nowcasting German GDP with text data

The first chapter, “*Nowcasting German GDP with text data*,” is joint work with Philipp Hauber, Kai Carstensen, and Jasper Bär. In this chapter, we analyse a large and unique corpus of 12.4 million German news articles spanning 1991–2018. We develop a Business Cycle Sentiment measure using aspect-based sentiment analysis. Unlike general-purpose or economics-specific sentiment lexicons, which are not originally designed for macroeconomic forecasting, we train a neural network on a professionally annotated dataset in which articles were coded in real time with a focus on the business cycle.

We then combine this sentiment with aspect-based topics estimated using the Latent Dirichlet Allocation (LDA) algorithm on a subset of articles that contain information on the business cycle. This ensures that both sentiment and topics are concentrated on cyclical dynamics. The resulting topics adjusted with the Business Cycle Sentiment are particularly informative for themes that have a relatively constant representation over time. For example, the raw topic share for the automobile industry is fairly stable and

therefore uninformative for nowcasting, but the sentiment-adjusted topic captures the substantial contraction in this sector during the Financial Crisis.

This chapter shows that these sentiment-adjusted topics correlate strongly with GDP growth. To assess whether they contain information useful for real-time forecasting, we use forecast-encompassing tests and a real-time out-of-sample forecasting experiment. The encompassing tests reveal that sentiment-adjusted topics contain statistically significant information beyond professional forecasts. In the forecasting exercise, we find that combining Dynamic Factor Model (DFM) forecasts derived separately from hard data and text information consistently outperforms a benchmark DFM based solely on hard data, with the greatest gains seen in the nowcasts. The main contribution of this chapter is the development of new text indicators focused on the business cycle that prove valuable in a simple forecasting experiment.

Chapter 2: Economic Forecasting with Topics, Sentiment, and Uncertainty

The second chapter, “*Economic Forecasting with Topics, Sentiment, and Uncertainty*,” expands upon the joint work of Chapter 1 and systematically evaluates which text-based measures carry the most predictive power for three macroeconomic aggregates published with a substantial lag: GDP, investment, and consumption. I compare a benchmark DFM based on hard data and surveys with DFMs augmented by a single additional text factor. This factor is constructed either from: (i) plain topics estimated on the full corpus; (ii) the same topics adjusted using two standard sentiment lexicons; (iii) topics adjusted with the share of uncertainty terms; or (iv) aspect-based topics adjusted with the Business Cycle Sentiment (BCS) from Chapter 1.

The key question is whether it matters which text dimension is prioritised for macroeconomic forecasting and whether this depends on economic conditions. The results show that plain topics perform best during the Financial Crisis (2008–2010) because they effectively act as crisis dummies and capture the onset of the recovery more quickly. In contrast, BCS-adjusted topics perform better in the more stable post-crisis period (2011–2018) for GDP and investment, because they are able to capture cyclical direction more effectively.

This pattern may also help explain why the literature often finds improved forecasting performance during turbulent economic times: the correct sentiment adjustment matters for tracking macroeconomic fluctuations outside crisis periods. The chapter also documents that correlations between plain topics and GDP, consumption, and investment weaken substantially once the crisis years are excluded, while BCS-adjusted topics remain correlated during both expansions and recessions for GDP and investment, and track commonly used survey indicators quite closely. For consumption, correlations are weaker, which I attribute to the need to capture sentiment toward household-relevant dimensions, for example labour market conditions.

Chapter 3: Salmon Stock Returns Around Market News

The final chapter, “*Salmon Stock Returns Around Market News*,” is joint work with Clemens Knoppe and Mikaella Zitti. This chapter examines how media coverage influences asset prices in the salmon market and demonstrates the importance of domain-specific sentiment. Using more than 6,000 industry-specific news articles, we show that sentiment based on a standard economic lexicon produces counterintuitive effects on stock returns, indicating that domain-agnostic sentiment does not adequately capture the nuances of industry reporting.

To address this limitation, we expand the lexicon with industry-specific terminology, capturing information unique to the salmon sector. This domain-specific sentiment measure, particularly when combined with news topics and adjusted for market competition, produces economically meaningful relationships with both the level and the volatility of stock returns and improves out-of-sample predictions of stock price movements. This chapter highlights that the effectiveness of sentiment analysis critically depends on tailoring the lexicon to the specific market under study.

Summary

Together, the three chapters demonstrate the substantial forecasting value of text-based information across distinct domains. They show that newspaper narratives, whether in general economic reporting or industry-specific news, contain timely signals that can improve predictions of both macroeconomic aggregates and financial market outcomes. The findings highlight the importance of adapting general text-learning methodologies to the specific economic or financial question, as such adjustments substantially influence forecasting results.

More broadly, my research contributes to several strands of the macroeconomic and financial literature, including work on sentiment and expectations, the effects of uncertainty on macroeconomic outcomes, and the interaction between information frictions and market behaviour. Across these areas, the dissertation moves the literature from the question “*Can text improve forecasts?*” toward the more interesting question: “*Which textual dimensions matter, in which regimes, and for which macroeconomic or financial variables?*”

I believe future research would benefit from combining unsupervised and supervised text learning methods with economic theory, for example by developing text-based indicators that capture sentiment toward economically relevant dimensions such as labour market conditions, monetary policy, fiscal policy, or inflation.

Chapter 1

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.

Keywords: Textual analysis, Topic Modeling, Sentiment Analysis, Macroeconomic News, Machine Learning, Forecasting.

JEL classification: C530, C550, E370.

*This study is coauthored by Philipp Hauber, Kai Carstensen, and Jasper Bär.
It is published as the CESifo Working Paper No 11587.*

1.1 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 handling missing observations at the start and end of the sample. As a benchmark, we use the Mixed Data Sampling (MIDAS) model (Forni 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 cy-

cle 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 1.2 provides an overview of the dataset and discusses the pre-processing steps applied. In Section 1.3, we outline the sentiment analysis methodology, including details on the training set. Section 1.4 focuses on the extraction and sentiment adjustment of news topics, as well as the selection of topics for forecasting. Section 1.5 presents the results of the encompassing test. Section 1.6 describes the out-of-sample forecasting experiment. Finally, Section 1.7 concludes the paper.

1.2 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 1.1, 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 Genios¹ and span from Monday to Saturday between January 1994 and November 2018. The *Handelsblatt* cor-

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

Table 1.1: 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%

Notes: 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.

pus, 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 1.1 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 1.A and documented in the accompanying code repository². 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% 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

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

Table 1.2: 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%

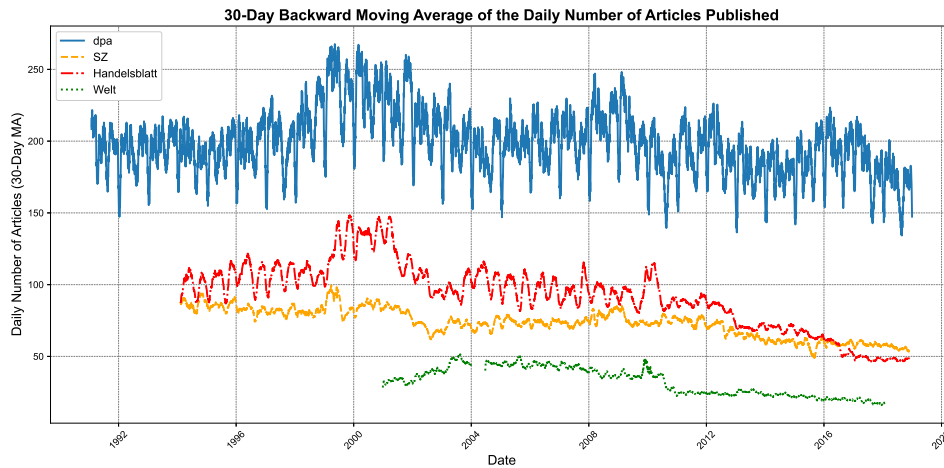
Notes: The “Articles per Day” and “Articles per Month” columns report the average number of articles per day and per month, respectively, in our final data set. The “Articles total” column represents the total number of articles and the “Share” column shows the share of each source.

(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, 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.

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 1.2. 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

Figure 1.1: Daily publications across datasets



Notes: 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.

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 1.B.

To summarize our findings, Figure 1.1 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.

1.3 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 fa-

vored 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 1.3.1, our methodology for sentiment extraction in Subsection 1.3.2, and the resulting daily sentiment index in Subsection 1.3.3.

1.3.1 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.

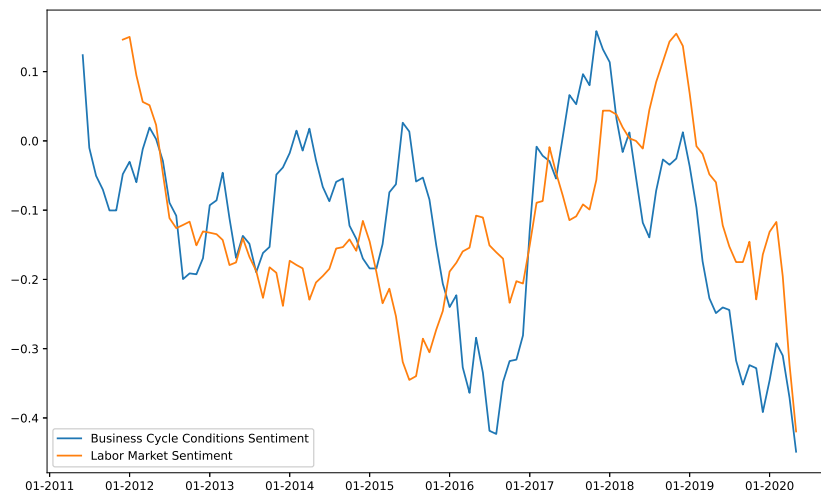
The dataset we received from MTI contains details on each article’s publication date, source (newspaper or journal), title, sentiment aspect (e.g., business cycle conditions), and the number of annotators who evaluated an article as negative, having no clear tone, or positive towards this particular 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 1.2. 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 1.2: Business cycle and labor market sentiment



Notes: This figure shows business cycle sentiment (blue line) and labor market sentiment (orange line), both constructed using Media Tenor data. The vertical axis represents the 6-month rolling mean of the difference between the proportions of positive and 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 1.3 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 archives³. For details, please refer to our repository.⁴

Table 1.3: 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 1.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 error, especially in cases where the sentiment was unclear.

The final training set, following the general pre-processing steps outlined in Appendix 1.C.2, covers the period from January 2011 until May 2020, with an average of 29 annotations

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

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

per month. From Table 1.3, 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.

1.3.2 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.

1.3.2.1 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:

$$L = \frac{1}{P} \sum_{p=1}^P \sum_{-C \leq j \leq C, j \neq 0} \log p(w_{p+j}|w_p), \quad (1.1)$$

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:

$$p(w_{p+j}|w_p) = \frac{\exp(v_{w_{p+j}}^\top u_{w_p})}{\sum_{i=1}^V \exp(v_i^\top u_{w_p})}, \quad (1.2)$$

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 (1.2), 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 (1.1). This optimization is achieved by iteratively updating the entries in matrices \mathbf{U} and \mathbf{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 \mathbf{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 1.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 \mathbf{U} and \mathbf{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 1.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 1.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 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 1.D.4.

After confirming that words most similar to the estimated embeddings of ‘business cycle

⁵Cosine 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 $j = 1, 2, \dots, V$). 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.

⁶Gerhard Schröder was German chancellor from 1998 until 2005.

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 1.D.5.

For ‘business cycle conditions’, the clustering approach proved effective, identifying 279 related terms, which are listed in Appendix 1.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 1.D.6.2.

Although economy and business cycle conditions are distinct concepts, there are two reasons 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 1.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 1.4: 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.

Notes: 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 1.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 1.4 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.

1.3.2.2 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 1.3. 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 1.D.8.

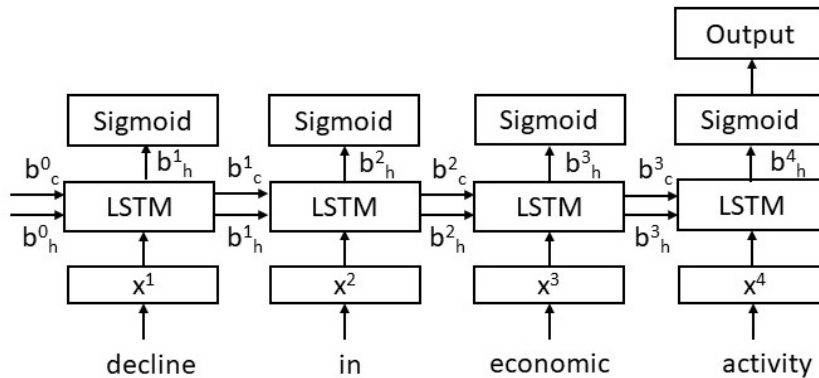


Figure 1.3: Architecture of the LSTM model used for sentiment analysis

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 1.5 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 with-

out 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 1.5: 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 Gurría (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.

Notes: 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 1.D.9.

Before estimating the LSTM model on the filtered MTI articles, we performed model-specific pre-processing to prepare the data (see Appendix 1.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 1.6, 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 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

Table 1.6: 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

Notes: 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.

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 1.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 1.D.10 and the architecture detailed in Appendix 1.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 1.7, 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.

We compared the LSTM model’s performance with two alternative sentiment approaches:

Table 1.7: LSTM performance

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

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 1.D.12.

1.3.3 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 1.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 1.8 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 1.2) to 47%, while Handelsblatt’s share has risen from 17% to 28%.

This shift is further illustrated in Figure 1.4, 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 1.4 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.

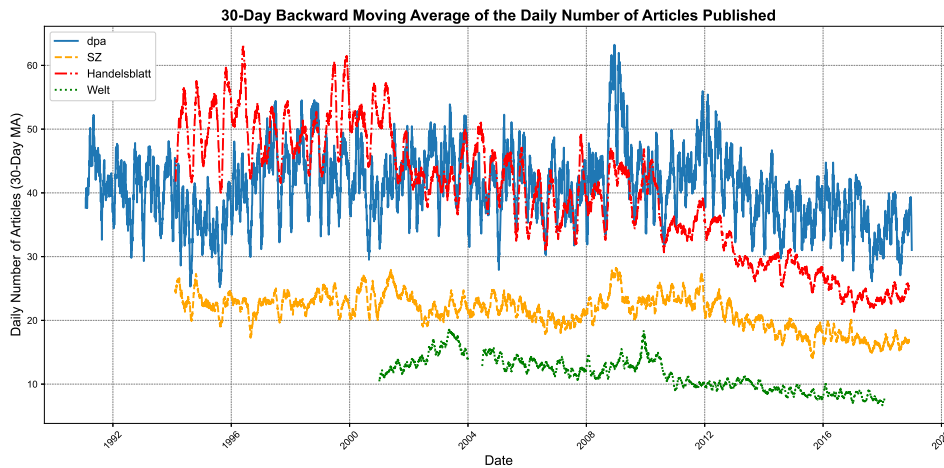
While the next phase of our analysis will focus on sentiment predictions for individual articles, we also constructed a daily Overall Business Sentiment Index (OBSI) for each

Table 1.8: 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%

Notes: 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.

Figure 1.4: Daily publications across datasets after LSTM pre-processing

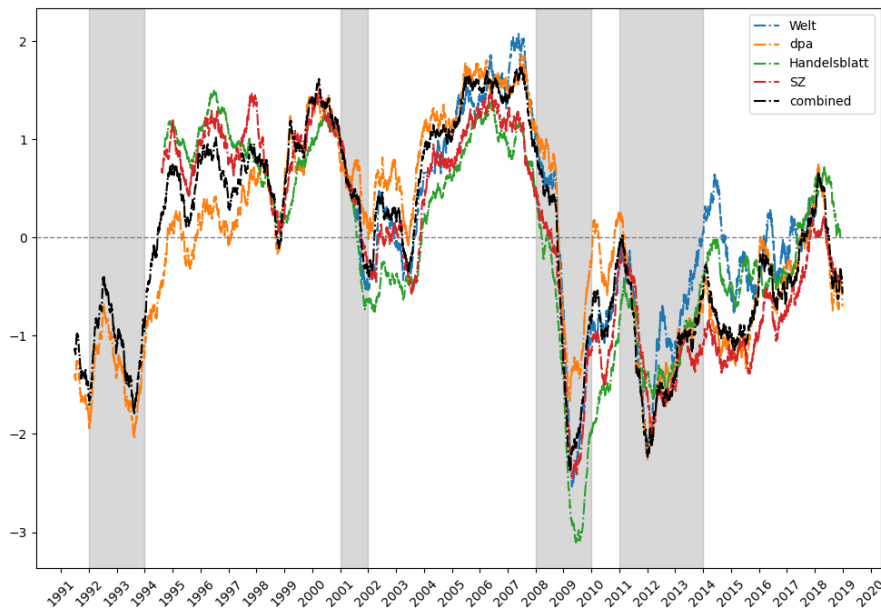


Notes: This graph displays the 30-day backward moving average of the daily number of articles published by dpa (blue), SZ (orange), Handelsblatt (red), and Welt (green) in the final dataset, after LSTM-specific pre-processing steps were applied. The Y-axis shows this moving average, and the X-axis corresponds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

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 1.5 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 post-reunification recession (1992-1993), the dot-com bubble (2001), the Great Recession (2008-2009), and the European sovereign debt crisis (2011-2013). This

Figure 1.5: Daily Overall Business Sentiment Index across media outlets



Notes: This figure shows the 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.

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 1.5 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.

1.4 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.

1.4.1 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 characterized 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 $k = 1, \dots, K$, 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 $n = 1, \dots, N_d$, 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 $z_{n,d} \in \{1, \dots, K\}$. 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 1.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:

$$Pr(z_{n,d} = k | z_{-(n,d)}, w, \alpha, \eta) \propto \frac{m_{v,-(n,d)}^k + \eta}{\sum_v m_{v,-(n,d)}^k + V\eta} \times (m_{k,-n}^d + \alpha), \quad (1.3)$$

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, $m_{v,-(n,d)}^k$ 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, $m_{k,-n}^d$ 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 (1.3) 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 1.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 1.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:

$$\hat{\theta}_d^k = \frac{m_k^d + \alpha}{\sum_{k=1}^K (m_k^d + \alpha)}, \quad (1.4)$$

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:

$$\hat{\beta}_k^v = \frac{m_v^k + \eta}{\sum_{v=1}^V (m_v^k + \eta)}, \quad (1.5)$$

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 1.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:

$$Pr\left(\tilde{z}_{n,\tilde{d}} = k \mid \tilde{z}_{-(n,\tilde{d})}, \tilde{w}, \alpha, \eta\right) \propto \hat{\beta}_k^v \left(m_{k,-n}^{\tilde{d}} + \alpha\right), \quad (1.6)$$

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

⁷See 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.

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 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.

1.4.2 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 1.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 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,

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

refer to Appendix 1.F.1.

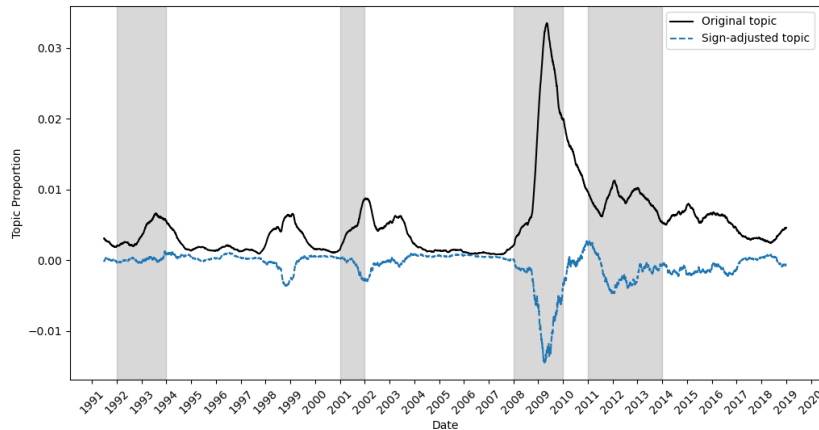
More importantly, one of the main strengths of our approach is that each topic comes with a clear narrative, allowing us to combine statistical evidence with economic reasoning when selecting the most relevant topics. For instance, T179, “Financial Crises and Market Regulation”, had a strong correlation 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 1.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 1.F.1). Figure 1.6 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 1.7), 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 industries.

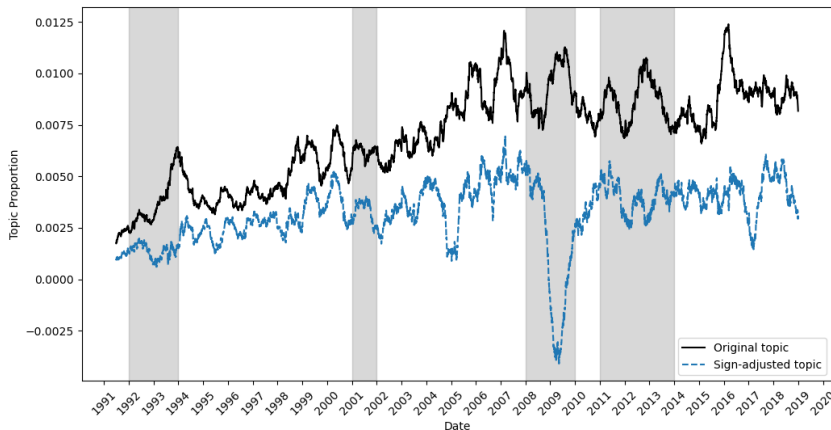
Other selected topics cover banking, mergers and acquisitions, job cuts, investments, financial and economic performance, as well as market reactions to news. The most probable words associated with each topic, along with their labels, are presented in Appendix 1.F.1, while Appendix 1.F.2 provides visualizations of the original and sign-adjusted series. Overall, the behavior of these topics and their sign-adjusted counterparts

Figure 1.6: Daily topic and sign-adjusted topic for Topic 27 (“Economic Crises and Recessions”)



Notes: The figure shows the 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.

Figure 1.7: Daily topic and sign-adjusted topic for Topic 52 (“German Automobile Industry and Major Manufacturers”)



Notes: The figure shows the 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.

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 1.F.3 for details) and produces time series that are strongly correlated with the variable we intend to forecast.

1.5 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 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:

$$\epsilon_{t+h|t} = \alpha + \beta f_{t+h|t} + \gamma s_t + u_{t+h}, \quad (1.7)$$

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 (1.7), we use Newey-West heteroscedasticity and autocorrelation robust (HAC) standard errors, with the lag selected automatically as in Newey and West (1994).

Table 1.9: Correlations between GDP growth and selected sign-adjusted topics

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

Notes: This table reports the contemporaneous correlations between the first release of quarterly GDP growth and ten selected sign-adjusted topics over the period 2006–2018.

The first interesting result of our analysis is that, as seen in Table 1.9, 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.

Moreover, the results of the encompassing test, presented in Table 1.10, 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.

Table 1.10: Encompassing regressions across forecast horizons

		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***

Notes: This table reports the encompassing regression results 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 * $p < 0.1$.

1.6 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.

1.6.1 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 $y_t^D = [y_{1,t}, y_{2,t}, \dots, y_{n_D,t}]'$ 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:

$$y_t^D = \Lambda_D f_t + \varepsilon_t^D, \quad \varepsilon_t^D \sim \text{i.i.d.} N(0, \text{diag}(\sigma_1^2, \dots, \sigma_{n_D}^2)), \quad (1.8)$$

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 :

$$f_t = A_1 f_{t-1} + \dots + A_p f_{t-p} + u_t, \quad u_t \sim \text{i.i.d.} N(0, Q), \quad (1.9)$$

where A_1, \dots, 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 1.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 1.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 1.9. 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.

1.6.2 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 (Froni 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 $h = -1$:

$$y_{t+h} = \alpha_{h+1} + \sum_{i=0}^{K-1} \beta_{i,h+1} \cdot x_{t+h-i/m} + \sum_{j=1}^K \theta_{j-1,h+1} \cdot z_{t+h-j/m} + \epsilon_{t+h}, \quad (1.10)$$

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:

$$y_{t+h}^v = \alpha_{h+1} + \sum_{i=0}^{K-1} \beta_{i,h+1} \cdot x_{t-(1-v)-i/m} + \sum_{j=2}^{K+1} \theta_{j-2,h+1} \cdot z_{t-(1-v)-j/m} + \epsilon_{t+h}^v, \quad (1.11)$$

where y_{t+h}^v 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 1.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 performance 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.

1.6.3 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 1.11 and 1.12 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 1.J, the results remain qualitatively similar when using the best-performing specifications. MIDAS with monthly data serves as a benchmark to assess whether 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 quar-

Table 1.11: Relative RMSFE scores: DFM and MIDAS models vs AR(1)

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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%).

ters) 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 1.11, 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 compared to the SPF (Table 1.12), 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

Table 1.12: Relative RMSFE scores: DFM and MIDAS models vs SPF

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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%).

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 aggregates 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 1.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.

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 1.13). We focus on combinations rather than

Table 1.13: Relative RMSFE scores: forecast combinations vs hard-data models

Model	Backcast	Nowcast			1 Step			2 Steps		
	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*

Notes: 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%).

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.

1.7 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 ap-

plied 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.

Appendix 1

1.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 available⁹.

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.

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

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

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 economic 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’ library¹⁰ 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

¹⁰<https://pypi.org/project/langdetect/>

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%.

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 Script¹¹ and *beliebte-vornamen.de*¹², while surnames are obtained from the Digital Dictionary of Surnames in Germany¹³. 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.

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/>

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 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 condensed 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 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’,

¹⁴<https://github.com/pyhunspell/pyhunspell>

‘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 “1”.

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 ‘ů’, 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. Specifically, 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 model¹⁵, 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.

1.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 1.8 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.

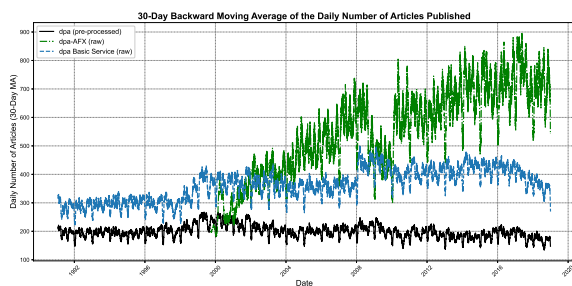
¹⁵<https://github.com/nreimers/truecaser>

In the case of SZ, as illustrated in Figure 1.9, 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 1.10, pre-processed data for Handelsblatt shows a more subdued downward trend (black line), suggesting improved consistency. In the case of Welt, as Figure 1.11 indicates, pre-processing significantly impacts the early part of the data, resulting in a smoother series of daily publications.

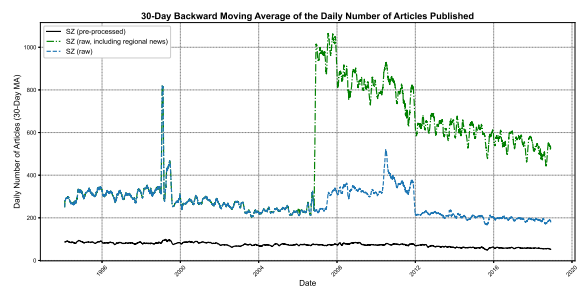
Overall, this analysis shows how pre-processing improves the quality of the corpus by concentrating on homogeneous content consistently covered over time and minimizing data irregularities. This approach is critical for capturing topic changes driven by economic events rather than organizational shifts within the news media, as also emphasized by Bybee et al. (2024).

Figure 1.8: Daily publications: dpa



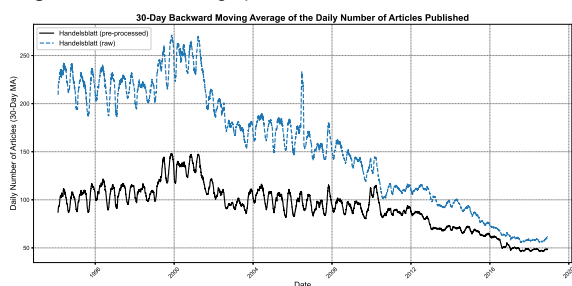
Notes: The blue line represents the dpa Basic Service before pre-processing, the green line indicates daily publications for dpa-AFX (articles removed during pre-processing), and the black line depicts the daily publications of the pre-processed dataset.

Figure 1.9: Daily publications: SZ



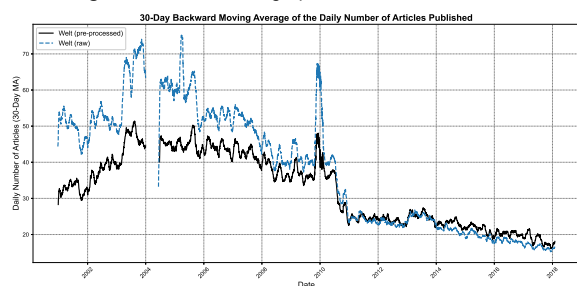
Notes: The blue line shows daily publications of the SZ dataset, excluding regional news, before pre-processing. The green line includes regional news (also from the original dataset), and the black line represents daily publications of the pre-processed dataset.

Figure 1.10: Daily publications: Handelsblatt



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

Figure 1.11: Daily publications: Welt



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

1.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.

1.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.

1.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.

¹⁶Specifically, 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.

1.D Methodology

1.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.

1.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:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}, \quad (1.12)$$

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 $\{1, 2, \dots, C\}$, 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 (1.2), 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):

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{k=1}^V f(w_k)^{3/4}}. \quad (1.13)$$

1.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 1.14: Estimation details for the word2vec model

1.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 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ 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 $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$, 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:

$$P_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, \quad (1.14)$$

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:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2N}. \quad (1.15)$$

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

$$Q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}. \quad (1.16)$$

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

$$C = \sum_i \sum_{j \neq i} P_{ij} \log \frac{P_{ij}}{Q_{ij}}. \quad (1.17)$$

By minimizing the cost function 1.17, 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 1.12 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 1.12: t-SNE visualization of 1,000 words related to 'business cycle conditions'



Notes: 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.

1.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 $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$ 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 $k \in \{1, \dots, K\}$. The cluster assignment C_k for embedding \mathbf{u}_{w_i} is defined as:

$$C_k = \{\mathbf{u}_{w_i} : \|\mathbf{u}_{w_i} - \mathbf{c}_k\|^2 \leq \|\mathbf{u}_{w_i} - \mathbf{c}_j\|^2 \text{ for all } j \neq k\}, \quad (1.18)$$

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:

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{\mathbf{u}_{w_i} \in C_k} \mathbf{u}_{w_i}, \quad (1.19)$$

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:

$$s(\mathbf{u}_{w_i}) = \frac{b(\mathbf{u}_{w_i}) - a(\mathbf{u}_{w_i})}{\max\{a(\mathbf{u}_{w_i}), b(\mathbf{u}_{w_i})\}}, \quad (1.20)$$

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).

1.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.

1.D.6.1 Terms related to the word ‘business cycle conditions’

Table 1.15: 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
konjunkturopernte	economic expert	wachstumskräfte	growth forces
konjunkturopernten	economic experts	wachstumsschwäche	weak growth
konjunkturimpulse	economic stimulus	nullwachstum	zero growth
konjunkturprogramm	economic stimulus program	exportwachstum	export growth
konjunkturprogramme	economic stimulus programs	produktivitätswachstum	productivity growth
konjunkturprogrammen	economic stimulus programs	beschäftigungswachstum	employment growth
konjunkturpaket	economic stimulus package	wachstumsbeitrag	growth contribution
konjunkturpakete	economic stimulus packages	konsum	consumption
konjunkturstütze	economic support	konsums	consumption
konjunkturtest	economic test	privatkonsum	private consumption
konjunkturumfrage	business climate survey	investitionschwäche	investment weakness
konjunkturrell	economically	investitionsstätigkeit	investment activities
konjunkturrelle	economic	ausrüstungsinvestitionen	investments in equipment
konjunkturrellen	economic	unternehmensinvestitionen	corporate investments
konjunkturreller	economic	bauinvestitionen	construction investments
konjunkturbedingte	cyclical	anlageinvestitionen	capital investments
wirtschaft	economy	investitionsneigung	propensity to invest
volkswirtschaft	economy	investitionsbereitschaft	willingness to invest
gesamtwirtschaft	overall economy	investitionsklima	investment climate
weltwirtschaft	global economy	investitionsausgaben	investment expenditure
binnenwirtschaft	domestic economy	investitionsquote	investment rate
exportwirtschaft	export economy	dynamik	dynamics
wirtschaftsbereiche	economic sectors	aufschwung	upswing
wirtschaftsaufschwung	economic upswing	aufschwungs	upswing
wirtschaftsaufschwungs	economic upswing	aufholprozess	catching-up process
wirtschaftsentwicklung	economic development	auftriebskräfte	upswing
wirtschaftslage	economic situation	abwärtsspirale	downward spiral
wirtschaftsleistung	economic output	abflachung	slowdown
wirtschaftsdynamik	economic dynamics	schrumpfung	shrinkage
wirtschaftsbelebung	economic recovery	dämpfung	dampening
wirtschaftstätigkeit	economic activity	verschlechterung	deterioration
wirtschaftsforschung	economic research	aufwertung	appreciation
wirtschaftsforscher	economic researcher	geldentwertung	devaluation
wirtschaftsforschern	economic researchers	anspringen	pick up
wirtschaftsweise	economic experts	dämpfen	dampen
wirtschaftsweisen	economic experts	abwürgen	stall
wirtschaftsinstitute	economic institutes	abgewürgt	stalled
wirtschaftsforschungsinstitut	economic research institute	absinken	sink
wirtschaftsforschungsinstitute	economic research institutes	verschlechtern	deteriorate
wirtschaftsforschungsinstituten	economic research institutes	verpu en	fizzle out
jahreswirtschaftsbericht	annual economic report		

Continued on next page

Chapter 1 Nowcasting German GDP with text data

Table 1.15 – Continued from previous page

German Term	English Translation	German Term	English Translation
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
ökonomien	economists	maßvollen	moderate
makroökonomische	macroeconomic	selbsttragenden	self-sustaining
makroökonomischen	macroeconomic	spurbare	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	ifoинstituts	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	ITW
mehrwertsteuererhöhung	increase in VAT	diw	DTW
steuereinnahmen	tax revenues	rwi	RWI
abgabenerhöhungen	tax increases	iwh	IWH
abgabenbelastung	tax burdens	iw	IW
fiskalpolitik	fiscal policy	hüther	Michael Hüther
staatshaushalte	national budgets	hwwa	HWWA
staatsausgaben	government spending	straubhaar	Thomas Straubhaar
defizit	deficit	wochenbericht	weekly report
verschuldung	debt	frühjahrgutachten	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
strukturereformen	structural reforms	exportweltmeister	export world champion
inflationbekämpfung	fighting inflation	außenbeitrag	net exports
strukturell	structural	außenhandel	foreign trade
strukturelle	structural	außenhandels	foreign trade
strukturellen	structural	bga	BGA
struktureller	structural	bgapäsident	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	effects	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	bruttoinlandsprodukts	GDP
beschäftigungsentwicklung	employment development	bruttoinlandsproduktes	GDP
beschäftigungszuwachs	employment growth	sozialprodukt	national product
beschäftigungsaufbau	employment growth	sozialprodukts	national product
beschäftigungsabbau	reduction in employment	verbrauch	consumption
reallöhne	real wages	verbrauchs	consumption
lohnsteigerungen	wage increases	kaufkraft	purchasing power
lohnentwicklung	wage developments	realeinkommen	real income
lohnpolitik	wage policy	sparquote	savings rate
lohnzurückhaltung	wage restraint	rate	rate
lohnabschlüsse	wage agreements	niveaus	levels
lohnabschlüssen	wage agreements	prozentpunkt	percentage point
tarifabschlüsse	collective agreements	jahresschnitt	annual average
tarifabschlüssen	collective agreements	jahresdurchschnitt	annual average
lohnstückkosten	unit labor costs	saisonbereinigte	seasonally adjusted
arbeitskosten	labor costs		

Continued on next page

Chapter 1 Nowcasting German GDP with text data

Table 1.15 – *Continued from previous page*

German Term	English Translation	German Term	English Translation
arbeitsproduktivität	labor productivity	saisonbereinigten	seasonally adjusted
produktivität	productivity	ungleichgewichte	imbalances
produktivitätszuwachs	increase in productivity		

Notes: 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.

1.D.6.2 Terms related to the word ‘economy’

Table 1.16: 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
ökonomien	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
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
strukturereformen	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	scha en	create/manage

Notes: 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.

1.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 1.17: 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.

Notes: 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.

1.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, \dots, 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 1.13) 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.

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:

$$a_\iota^t = \sum_{i=1}^I w_{i\iota} x_i^t + \sum_{h=1}^H w_{h\iota} b_h^{t-1}, \quad (1.21)$$

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

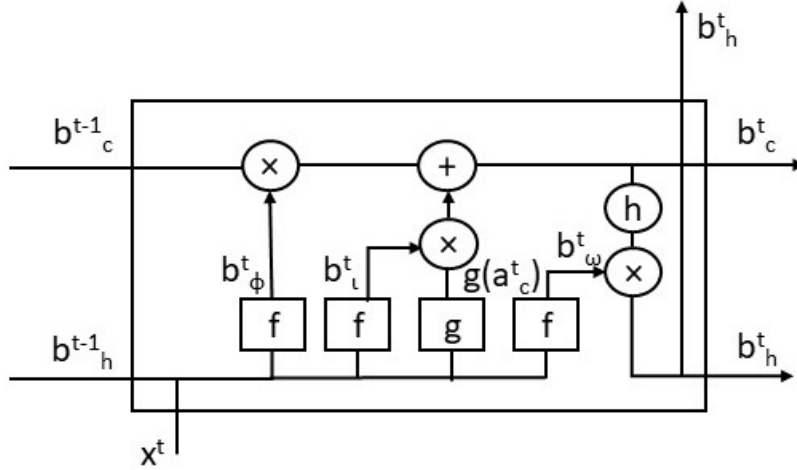


Figure 1.13: Structure of an LSTM cell

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:

$$b_\iota^t = f(a_\iota^t), \quad (1.22)$$

where f is the activation function of the gates, typically the sigmoid function¹⁷. 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:

$$a_\phi^t = \sum_{i=1}^I w_{i\phi} x_i^t + \sum_{h=1}^H w_{h\phi} b_h^{t-1}, \quad b_\phi^t = f(a_\phi^t). \quad (1.23)$$

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:

$$a_c^t = \sum_{i=1}^I w_{ic} x_i^t + \sum_{h=1}^H w_{hc} b_h^{t-1}. \quad (1.24)$$

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:

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

$$b_c^t = b_\phi^t b_c^{t-1} + b_i^t g(a_c^t), \quad (1.25)$$

where g is the cell candidate activation function, typically the hyperbolic tangent (tanh) function¹⁸.

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

$$a_\omega^t = \sum_{i=1}^I w_{i\omega} x_i^t + \sum_{h=1}^H w_{h\omega} b_h^{t-1}, \quad b_\omega^t = f(a_\omega^t). \quad (1.26)$$

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:

$$b_h^t = b_\omega^t h(b_c^t). \quad (1.27)$$

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:

$$y = \sigma(a), \quad \text{where} \quad a = \sum_{h=1}^H w_h b_h^t. \quad (1.28)$$

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).

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

1.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 1.18: 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 Gurría (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.

Notes: 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.

1.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.

1.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 1.19: Estimation details for the LSTM model

1.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.

1.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.

¹⁹<http://snowball.tartarus.org/algorithms/german/stop.txt>

²⁰<http://snowball.tartarus.org/algorithms/german/stemmer.html>

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:

$$\text{tf-idf}_v = \log(1 + N_v) \times \log\left(\frac{D}{D_v}\right) \quad (1.29)$$

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.

1.D.14 Cross-validation results for LDA

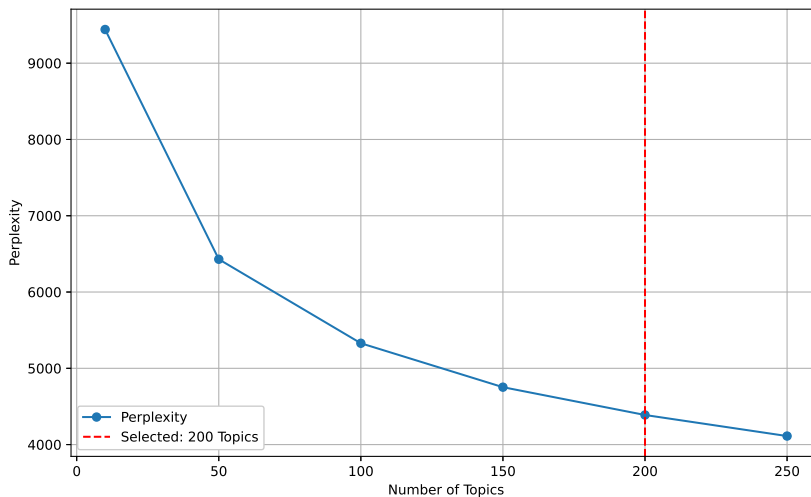
To determine the optimal number of topics, we applied 10-fold cross-validation with perplexity as the evaluation metric. Perplexity is used to assess how effectively a topic model can generalize to unseen data and is calculated as follows:

$$\exp \left[- \frac{\sum_{d=1}^D \sum_{v=1}^V n_{d,v} \log \left(\sum_{k=1}^K \hat{\theta}_d^k \hat{\beta}_k^v \right)}{\sum_{d=1}^D N_d} \right], \quad (1.30)$$

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 1.14: Perplexity across different numbers of topics



Notes: 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.

1.E Examples of the estimated topics

Table 1.20: Labels for the first 10 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), scheid (fail), tre (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), geschäftsfuhr (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

Notes: 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.

1.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.

1.F.1 Overview of selected topics

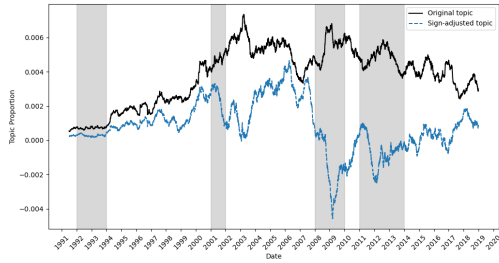
Table 1.21: Selected topics and their correlations with the annualized quarterly GDP growth

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

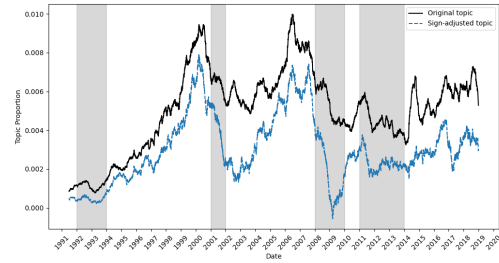
Notes: The "Correlation" column lists the correlation with GDP growth for the first vintage and the average correlation across 34 vintages. The "Top 10 Count" indicates the number of vintages where the topic was among the 10 most correlated variables with GDP growth.

1.F.2 Daily topics and their sign-adjusted versions

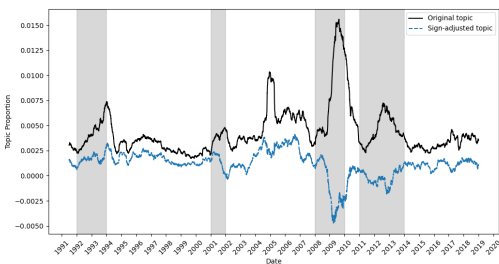
Figure 1.15: Selected topics and their sign-adjusted counterparts



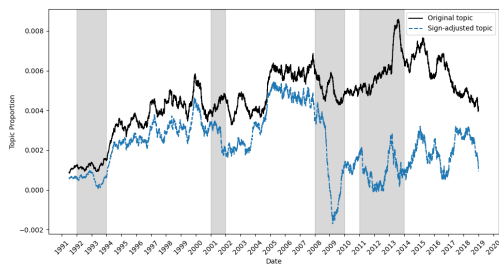
(a) Topic 127: Banking



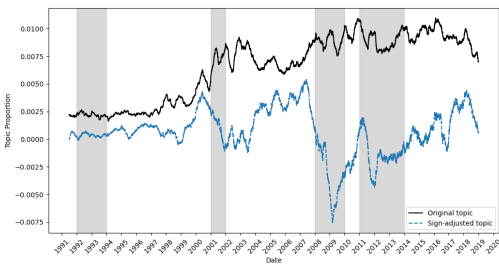
(b) Topic 11: M&As



(c) Topic 81: Job Cuts



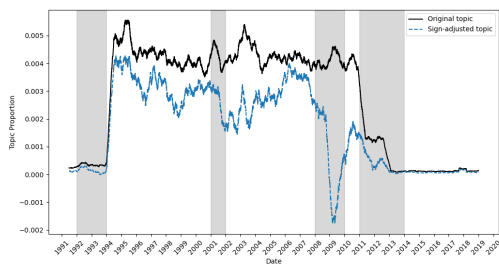
(d) Topic 77: Private Investment



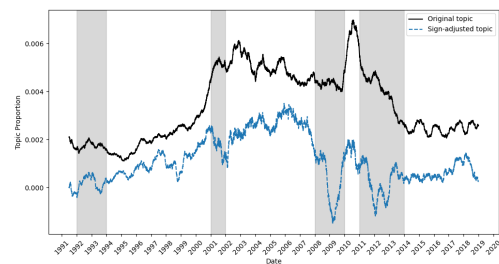
(e) Topic 74: Economic Bubbles



(f) Topic 131: Emerging Markets



(g) Topic 138: Economic Performance



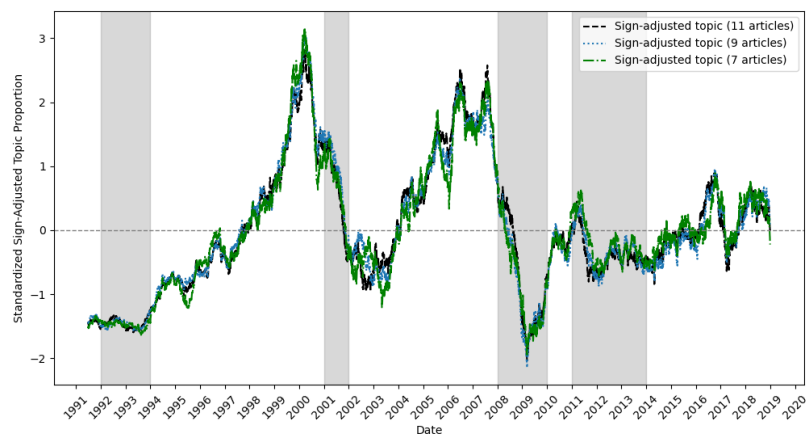
(h) Topic 100: Market Reactions to News

Notes: Each panel shows the 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.

1.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 1.16 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 1.16: Robustness of sign-adjusted Topic 11 to the number of articles used



Notes: The figure displays the 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.

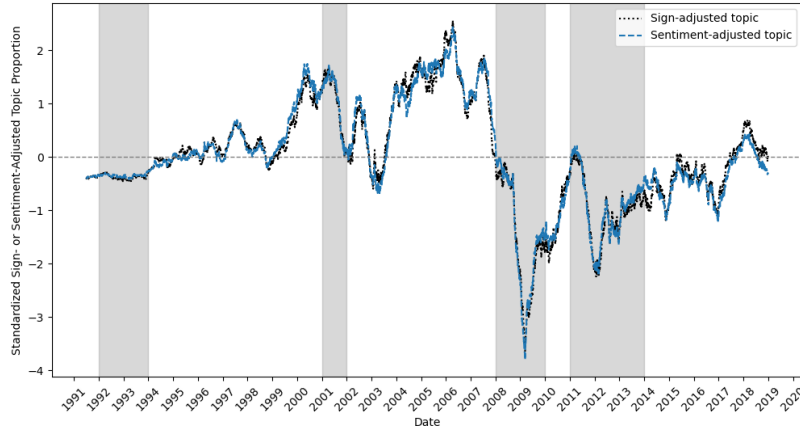
The second robustness check examined an alternative approach to adjusting topic proportions. Rather than relying on a majority vote for sign adjustment, we computed the average sentiment across the 11 articles.²² However, after standardizing the sign- and

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

²²Here, we refer to topic series multiplied by the average topic-specific sentiment as “sentiment-adjusted topic proportions”

sentiment-adjusted topic proportions, we found no significant difference between the two approaches. For example, this can be seen with Topic 127 in Figure 1.17.²³ In conclusion, the results demonstrate that our methodology remains robust to the tested modifications.

Figure 1.17: Sign- and sentiment-adjusted topic dynamics for Topic 127



Notes: The figure shows the 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.

1.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:

$$Y_t^Q = \frac{1}{k_t} \sum_{s=t-k_t+1}^t X_s, \quad t = t_1(Q), t_2(Q), \dots, \quad (1.31)$$

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:

$$y_t^Q = \sum_{s=t-k_t+1}^t \frac{t+1-s}{k_t} x_s + \sum_{s=t-k_t-k_t+2}^{t-k_t} \frac{s-t+k_t+k_t-k_t-1}{k_t-k_t} x_s, \quad t = t_1(Q), t_2(Q), \dots \quad (1.32)$$

to clearly distinguish them from sign-adjusted topic proportions.

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

Equation (1.32) 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 $(y_t^{D'} y_t^{Q'})'$ using the aggregators of the daily factors, f_t^{QA} and f_t^{QP} :

$$\begin{pmatrix} y_t^D \\ y_t^Q \end{pmatrix} = \begin{pmatrix} \Lambda_D & 0 & 0 \\ 0 & \Lambda_Q & 0 \end{pmatrix} \begin{pmatrix} f_t \\ f_t^{QA} \\ f_t^{QP} \end{pmatrix} + \begin{pmatrix} \varepsilon_t^D \\ \varepsilon_t^Q \end{pmatrix}, \quad (1.33)$$

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} :

$$\begin{aligned} f_t^{QA} &= \sum_{s=t-k_t+1}^t \frac{t+1-s}{k_t} f_s + \sum_{s=t-k_t-k_{t-k_t}+2}^{t-k_t} \frac{s-t+k_t+k_{t-k_t}-1}{k_{t-k_t}} f_s \\ &= \sum_{s=t-k_t+1}^t W_s^C f_s + \sum_{s=t-k_t-k_{t-k_t}+2}^{t-k_t} W_s^P f_s, \end{aligned} \quad (1.34)$$

where W_s^C and W_s^P denote the weights assigned to the daily factors of the current and previous 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., $t = t_1(Q) + 1, t_2(Q) + 1, \dots$:

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

2. For all other days:

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

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).

1.H Hard data

Table 1.22: Economic data

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

1.1 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 (1.10) and (1.11)). 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 .

1.1.1 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:

$$\hat{\mu} = \arg \min_{\mu} \sum_{t=1}^T (y_{t+h} - \alpha_{h+1} - H_t \mu)^2 + \lambda \sum_{j=1}^{N_H} |\mu_j|^q, \quad (1.35)$$

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 nonlinearities, 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.

1.1.2 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:

$$\min_{F_1, \dots, F_T, \Lambda} T^{-1} \sum_{t=1}^T (S_t - \Lambda F_t)^\top (S_t - \Lambda F_t), \quad (1.36)$$

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:

$$\hat{F}_t = \hat{\Lambda}^\top S_t, \quad (1.37)$$

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).

1.1.3 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.

1.1.3.1 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:

$$\min_{j,s} \left[\min_{c_1} \sum_{H_t \in R_1(j,s)} (y_{t+h} - c_1)^2 + \min_{c_2} \sum_{H_t \in R_2(j,s)} (y_{t+h} - c_2)^2 \right], \quad (1.38)$$

where the two regions are defined as:

$$R_1(j, s) = \{H_t \mid H_{tj} \leq s\}, \quad R_2(j, s) = \{H_t \mid H_{tj} > s\}. \quad (1.39)$$

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:

$$\hat{c}_1 = \text{ave}(y_{t+h} \mid H_t \in R_1(j, s)), \quad \hat{c}_2 = \text{ave}(y_{t+h} \mid H_t \in R_2(j, s)). \quad (1.40)$$

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:

$$\hat{f}(H_t) = \sum_{m=1}^M \hat{c}_m \cdot \mathbb{I}(H_t \in R_m), \quad (1.41)$$

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

1.1.3.2 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.

1.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.

1.J.1 Optimal weights

Table 1.23: Optimal weights for the text-based model in forecast combinations

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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.

1.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.

Table 1.24: Summary of the best-performing MIDAS models

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

Notes: The table reports the best-performing MIDAS models for forecasts produced 30, 60, and 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.

Table 1.25: Relative RMSFE scores: MIDAS models vs AR(1)

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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%).

Table 1.26: Relative RMSFE scores: MIDAS models vs SPF

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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%).

Table 1.27: Relative RMSFE scores for MIDAS models: forecast combinations vs hard-data models

Model	Backcast	Nowcast			1 Step			2 Steps		
	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

Notes: 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%).

Table 1.28: Optimal weights for the text-based model in forecast combinations

	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

Notes: 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 (back-cast, 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.

Chapter 2

Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Abstract

This paper examines whether text-based information extracted from a large corpus of German news articles can improve nowcasts of German GDP, investment, and consumption, and whether forecasting performance depends on the particular type of text-based series employed. I construct plain topics using Latent Dirichlet Allocation (LDA) and adjust them using several sentiment and uncertainty measures: an economic sentiment lexicon, a general sentiment lexicon, and the share of uncertainty-related terms. In addition, I compare their performance with aspect-based topics derived from articles explicitly containing information on business cycle conditions and adjusted with Business Cycle Sentiment (BCS). Forecasting performance is evaluated in a real-time setting using a monthly Dynamic Factor Model, with a specification relying solely on hard data and surveys serving as the benchmark. The results show that plain topics provide the most reliable improvements during the Financial Crisis, offering earlier gains within the quarter for all three variables. In contrast, during the more stable period, BCS-adjusted topics clearly outperform the alternative text specifications when nowcasting GDP and investment, reflecting their stronger and more stable correlations with these variables across economic regimes. For consumption, however, none of the existing sentiment adjustments yields systematic gains in calmer times, suggesting that sentiment measures tailored to household-relevant factors, such as labour market expectations or policy-related sentiment, may be required. Taken together, the results indicate that text-based indicators can improve macroeconomic nowcasting, but the most effective adjustment depends both on the variable being forecast and on prevailing economic conditions.

Keywords: Text Mining; Topic Modeling; Sentiment Analysis; Uncertainty; Forecasting

JEL classification: C530, C550, E370.

This study is a single-authored paper. It is published as the SSRN Working Paper No 5804384.

2.1 Introduction

The increasing availability of unstructured text data from digitalized news media archives has generated substantial interest in its potential for macroeconomic analysis and forecasting. From such data, researchers typically extract three main dimensions: sentiment, uncertainty, and topics. This article examines whether topics alone or topics adjusted with sentiment or uncertainty provide more accurate forecasts of key macroeconomic variables for Germany.

Among these dimensions, sentiment analysis has received the most attention. The simplest and most widely applied method is the lexicon-based approach, which relies on pre-defined word lists with associated sentiment scores or labels to quantify the tone of a text. This method is easy to implement and interpret, transferable across countries, and applicable to diverse text sources. Economists have employed both general-purpose dictionaries, such as Harvard IV (e.g., Tetlock, 2007), Afinn (Nielsen, 2011), and word lists derived from product reviews (Hu and Liu, 2004), as well as domain-specific lexicons designed for economics and finance, including the financial stability dictionary (Correa et al., 2017), lists derived from 10-K reports (Loughran and McDonald, 2011), and a recent economics-specific dictionary (Barbaglia et al., 2025). In addition, some studies combine multiple dictionaries (e.g., Shapiro et al., 2022). Empirical results are generally encouraging: the dictionary by Correa et al., 2017 was particularly successful in predicting macroeconomic variables during the Financial Crisis (Kalamara et al., 2022), while general-purpose dictionaries performed better in forecasting GDP growth during the COVID-19 pandemic, a crisis of primarily non-economic origin (Ashwin et al., 2024). Overall, sentiment-based measures appear most effective for forecasting macroeconomic aggregates in turbulent times. A plausible explanation is that sentiment indices, by averaging sentiment across all covered topics, provide a clear signal in crisis periods when most topics share a negative tone. In contrast, such averaging may be less informative during calmer periods, when sentiment varies substantially across topics.

The second dimension is uncertainty. A growing literature shows that uncertainty shocks can have sizable effects on the real economy. For example, Baker et al., 2016 document that higher economic policy uncertainty is associated with declines in investment and output, while Alexopoulos and Cohen, 2015 provide evidence that increases in general economic uncertainty reduce economic activity. Despite its popularity in the literature, however, economic uncertainty has often proven less effective for forecasting macroeconomic aggregates than sentiment (see e.g. Kalamara et al., 2022), possibly due to the relative sparsity of this measure.

The third important dimension is topics. One of the first studies to examine the macroeconomic effects of news using topic modeling is Larsen and Thorsrud (2019), who apply

Latent Dirichlet Allocation (LDA) to extract topics from a large corpus of news articles. They show that certain topics, particularly those related to financial markets, are highly predictive of fluctuations in GDP, consumption, and investment. Similarly, Bybee et al. (2024) estimate topics from economic news using LDA and employ LASSO regression to identify those that comove most closely with macroeconomic and financial aggregates. A topic labeled “Recession” emerges as the strongest explanatory variable for a wide range of indicators, including industrial production. Both studies, however, indicate the importance of sentiment adjustment. Larsen and Thorsrud (2019) emphasize that accounting for the sign of each topic improves forecast accuracy, while Bybee et al. (2024) show that when pairs of topics are included alongside individual topics in their LASSO analysis, no single topic is selected in isolation; instead, most of the selected interactions involve the “Recession” topic. Since “Recession” clearly carries a negative tone, this suggests that combining topic information with sentiment may improve predictive performance.

More recently, Ellingsen et al. (2022) conduct an out-of-sample forecasting exercise, comparing the predictive power of LDA-based topics derived from a large news corpus with that of the FRED-MD dataset for forecasting GDP, consumption, and investment. Their results show that topics improve forecasts, but their contribution is particularly strong during the Financial Crisis. This pattern again points to the relevance of sentiment adjustment: topics by themselves are informative when they are consistently associated with a particular sentiment (e.g., the topic “Recession”), but in calmer periods, when many topics are reported at a steady rate while their sentiment shifts, combining them with sentiment is important for accurate forecasting.

While each of the three text dimensions discussed above (sentiment, uncertainty, and topics) has proven useful for economic analysis and forecasting, their limitations motivate the approach taken in this article. Specifically, I combine topics with sentiment or uncertainty. This has two main advantages. First, it improves interpretability: rather than relying on a single sentiment measure averaged across many topics, I examine which type of sentiment (or uncertainty) drives the results (e.g., sentiment regarding the automotive industry). Second, it ensures that topics reported at a steady rate but with time-varying sentiment are represented more accurately in the analysis.

A few studies have explored this combined approach. Thorsrud (2016) adjust LDA-estimated topics using a general sentiment lexicon and show that the resulting sentiment-adjusted topics improve GDP forecasts relative to a survey of professional forecasters, particularly during the Financial Crisis. Similarly, van Dijk and de Winter (2023) modify LDA topics with an economics- and finance-specific dictionary and also report improved forecasting performance, again most notably in turbulent economic periods. Ardia et al. (2019) apply several sentiment lexicons to adjust topics retrieved via LexisNexis and demonstrate improved out-of-sample accuracy for industrial production forecasts when

using sentiment-adjusted topics.

One of the closest studies to this research is Aprigliano et al. (2023), who adjust Factiva-classified topics using an economics-specific sentiment lexicon and an economic policy uncertainty index. They find that text-based series reduce forecast uncertainty for GDP, consumption, and investment, especially during recessions, and that sentiment-adjusted topics are selected more frequently than uncertainty-adjusted topics in forecasting models. My approach differs from theirs in two key respects. First, I estimate topics using LDA rather than relying on provider-defined categories, which allows for a more data-driven classification of news content. Second, I adjust topics not only with sentiment and uncertainty lexicons, but also with the Business Cycle Sentiment extracted using a supervised machine learning approach.

This article contributes to the literature on text-based economic forecasting by systematically evaluating whether plain topics or topics adjusted with sentiment or uncertainty provide more accurate nowcasts of German GDP, consumption, and investment. The novelty lies in adjusting LDA-estimated topics using several conceptually distinct sentiment and uncertainty measures. In line with earlier studies, I employ off-the-shelf, theoretically grounded German lexicons: SentiWS, a general-purpose sentiment dictionary (Remus et al., 2010); a German translation of the economic lexicon by Loughran and McDonald, 2011 (Banner et al., 2019); and the share of uncertainty-related terms. The central question, however, is whether the Business Cycle Sentiment, which is theoretically more relevant for macroeconomic forecasting but also more difficult to extract (see Okuneva et al., 2024), can improve nowcasts when combined with aspect-based topics derived by applying LDA solely to articles that explicitly contain information about the business cycle. I evaluate whether these text-based measures provide additional predictive power beyond traditional hard data and survey indicators when nowcasting variables that are characterized by long publication lags.

To address this question, I estimate a monthly Dynamic Factor Model (Bańbura et al., 2010) with factors extracted from hard data and surveys, and a separate factor extracted from text-based measures. This model is widely used in central banks for macroeconomic forecasting and is particularly suited to evaluating whether text data improve forecasting performance, and if so, which types of text-based information are most useful. The out-of-sample period covers 2008Q1–2018Q4 and includes both the Financial Crisis and the European Debt Crisis.

Business Cycle Sentiment is an example of aspect-based sentiment, pioneered in the economic literature by Barbaglia et al. (2023). There are two main reasons to expect gains from this approach. First, the chosen aspect is broad enough to capture meaningful variation in the data, yet directly relevant to macroeconomic aggregates. Second, the sentiment measure is extracted using a supervised machine learning model trained on a large,

domain-specific dataset annotated by professional coders. While supervised approaches typically yield more accurate sentiment measures, they remain rarely applied in economic forecasting due to the high costs of producing labeled data.

Unlike lexicon-based adjustments, which are applied to topics derived from the full corpus of 3.3 million articles, the aspect-based sentiment is combined with topics estimated from a restricted set of articles specifically related to the business cycle. This ensures that the resulting topics are closely aligned with the sentiment dimension of interest and remain economically meaningful.

The empirical results reveal a clear dependence of forecasting performance on the broader economic regime. During the Financial Crisis, models augmented with plain topics deliver the most systematic improvements for all three variables and do so earlier within the quarter than sentiment-adjusted specifications. This pattern reflects the fact that topic prevalence responds more quickly than sentiment in the immediate post-crisis recovery. At the same time, it highlights a limitation of plain topics: many of them behave essentially as crisis indicators. Once the crisis ends, their correlations with macroeconomic variables weaken considerably.

In calmer periods, this pattern reverses. BCS-adjusted topics consistently outperform all other text specifications for GDP and investment, yielding sizable and statistically significant gains relative to the benchmark. Their advantage stems from the stability of their correlations across regimes, which allows them to take cyclical direction into account. For consumption, by contrast, none of the existing sentiment adjustments yields systematic improvements in stable periods, suggesting that sentiment measures targeting household-relevant dimensions, such as labour market or inflation expectations, would be more appropriate. Across all variables and economic environments, uncertainty-adjusted topics perform the weakest, providing little additional information beyond the benchmark.

The type of adjustment also affects both the magnitude and the interpretation of the resulting relationships. For instance, the “Crisis” topic adjusted with Business Cycle Sentiment is more strongly correlated with GDP growth than when adjusted with lexicons, and it exhibits the sharpest decline during the Financial Crisis followed by an overshooting in the recovery, closely mirroring GDP dynamics. More broadly, BCS adjustment produces topic series that are predominantly economic in nature, capturing crises, growth dynamics, financial conditions, and developments in key industries and the broader economy, which substantially improves interpretability. Financial topics become even more prominent for investment, consistent with the central role of financial conditions in investment decisions. This contrasts with general sentiment adjustments, which tend to yield political or general topics that are less directly connected to macroeconomic activity.

Finally, BCS-adjusted topics correlate strongly with widely used survey indicators such as the Ifo Business Climate Index and the European Commission’s Economic Sentiment

Indicator, further validating their relevance. Across all specifications (plain, economic lexicon-adjusted, and BCS-adjusted topics) the highest correlations consistently arise with the Ifo Business Expectations Index, suggesting that text-based indicators reflect expectations about future economic conditions. For consumption, while overall results are weaker and many of the selected topics relate to politics or policy issues, several of the most correlated topics are promising and relate to labour market conditions and reporting from economic research institutes. Nonetheless, a more targeted sentiment adjustment tailored to these topics would likely be required to improve forecasting performance in future work.

The remainder of the paper is organized as follows. Section 2.2 explains the construction of the text indices. Section 2.3 outlines the out-of-sample forecasting experiment. Section 2.4 presents the empirical results. Section 2.5 concludes.

2.2 Text indices

As outlined in the introduction, the text-based indices used in the forecasting experiment are either plain topics extracted from news articles using LDA (Blei et al., 2003) or combinations of these topics with sentiment or uncertainty. This section explains their construction. I first describe how topics are estimated, then how sentiment and uncertainty are quantified at the article level, and finally how these measures are combined with the topic series to obtain sentiment- and uncertainty-adjusted indices.

The text corpus and its pre-processing follow Okuneva et al. (2024). In short, the dataset contains 3,336,299 pre-processed articles from the largest German news agency, dpa, and three leading newspapers: Süddeutsche Zeitung, Handelsblatt, and Welt. Topics used either on their own or adjusted with sentiment lexicons and the share of uncertainty terms are estimated using the LDA algorithm on articles published before 2008 (2,070,035 articles), thereby avoiding look-ahead bias. Topics later adjusted with the Business Cycle Sentiment are estimated separately, using 887,300 articles published before 2010, restricted to those containing at least some information on the aspect of interest. Although this set includes two years from the out-of-sample forecasting period, the results show that these topics, when adjusted for the Business Cycle Sentiment, do not outperform the plain topics estimated without out-of-sample data during the Financial Crisis. For this reason, I do not view the overlap as problematic and retain the model specification from Okuneva et al. (2024), where these indices were originally developed.

Both sets of topics are estimated using the collapsed Gibbs sampling algorithm of Griffiths and Steyvers (2004). The number of topics is set to 200, as Okuneva et al. (2024) demonstrate that this choice is optimal based on perplexity and on the interpretability of the resulting topics. Topics are labeled using the final sample, and those discussed in the results are listed in Appendix 2.A. A full list of topic labels for the 200 topics estimated on

the 887,300 articles is available online.¹ After obtaining document-level topic distributions for the training set, I re-estimate them for each article in the test set. To construct daily time series, all articles from a given day are pooled into a single document and its topic distribution is re-sampled.

I next compute article-level sentiment using two lexicons: SentiWS (Remus et al., 2010) and the German adaptation of the Loughran and McDonald (2011) dictionary by Bannier et al. (2019), hereafter referred to as BPW. SentiWS is a general-purpose sentiment dictionary, whereas BPW is domain-specific, designed for economics and finance. Both have been successfully used in GDP forecasting: SentiWS by Shrub et al. (2022) and BPW by Kalamara et al. (2022). Since both dictionaries are available in the original German, no translation is required.

SentiWS offers broader coverage, with 17,806 negative and 16,400 positive word forms, compared with 10,147 negative and 2,223 positive word forms in BPW. The advantage of BPW lies in its domain focus: terms such as ‘tax’, ‘costs’, ‘expense’, and ‘liability’ are often labeled negative in general dictionaries but treated as neutral in financial texts. Another difference concerns scoring: SentiWS assigns each word a value between -1 and 1 , while BPW uses a binary scheme, classifying words as either negative (-1) or positive ($+1$). When a word appears on both the positive and negative lists of SentiWS, I take the average of the two values. For each article, sentiment is then calculated as the sum of all word scores divided by the total number of words in the article.

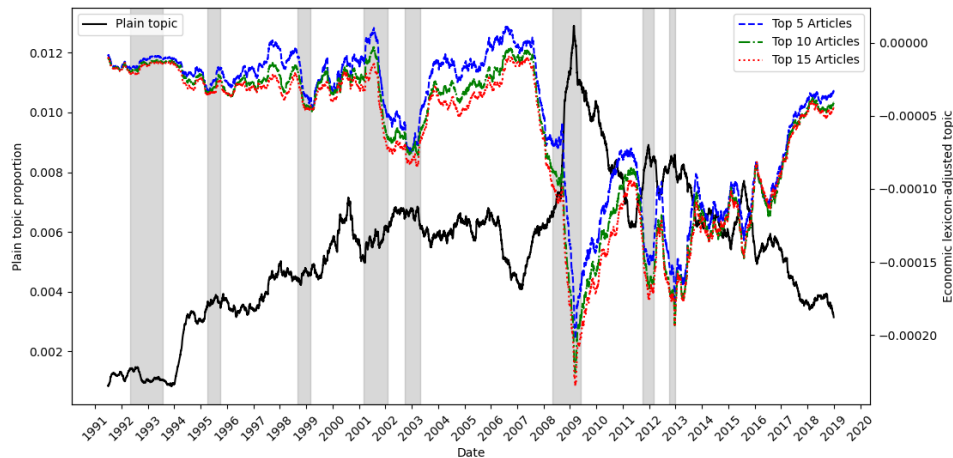
In addition to sentiment, I construct an article-level uncertainty measure. For each article, uncertainty is calculated as the ratio of uncertainty-related words to the total word count. The vocabulary consists of 46 terms that appear in the corpus at least 20 times, including both general words such as ‘Unsicherheit’ (uncertainty) and compound forms like ‘Unsicherheitsfaktor’ (uncertainty factor).

Finally, to adjust topics with sentiment and uncertainty lexicons, I proceed in three steps. First, for each topic and day, I identify the ten articles with the highest proportion of that topic. Second, I compute the average sentiment or uncertainty score across these articles, which represents the topic-specific sentiment or uncertainty on that day. Third, I adjust the daily topic value by multiplying it with the corresponding sentiment or uncertainty measure.

To assess robustness, I also experiment with using the top five and the top fifteen articles instead of ten. Figure 2.1 illustrates this using the “Banking” topic (T29). The black line shows the plain topic, while the blue, green, and red lines correspond to sentiment-adjusted series using the economic BPW lexicon with 5, 10, and 15 articles, respectively. As expected, using fewer articles (five) yields a more volatile series. By contrast, results for 10 and 15 articles are very similar, with both series exhibiting a stronger and eco-

¹https://github.com/MashenkaOkuneva/newspaper_analysis/blob/main/topics/Topic_labels.pdf

Figure 2.1: Robustness of sentiment adjustment to the number of articles for T29 (“Banking”) using the economic lexicon



Notes: The figure shows the plain “Banking” topic (T29) and three variants of its economic lexicon-adjusted counterpart. The black line displays the 180-day backward moving average of the daily topic proportion. The blue, green, and red lines show the corresponding economic lexicon-adjusted topic series (also 180-day backward moving averages), where sentiment is calculated using the 5, 10, and 15 articles with the highest share of T29 on each day, respectively. The left axis refers to the raw topic proportion, while the right axis refers to the economic lexicon-adjusted topic proportion. The X-axis reports the calendar day. Shaded areas denote recession periods in Germany as identified by Carstensen et al. (2020).

nomically meaningful decline during the Financial Crisis than the topic adjusted with five articles. Based on this evidence, I select 10 articles for adjustment: this provides a good compromise between reducing noise and ensuring that only the most relevant articles are captured. The choice is also supported by the large size of the corpus, with an average of 326 articles per day.

The construction of the Business Cycle Sentiment and its use in adjusting topics is described in detail in Okuneva et al. (2024). Here, I briefly summarize the main idea. An LSTM (Long Short-Term Memory) model is trained on a Media Tenor dataset of 3,286 articles annotated by professional coders, who read each article in full and classified its sentiment toward business cycle conditions as positive, negative, or lacking a clear tone. For training, I retain only sentences that include some information on the business cycle, distinguishing between negative tone and positive or no clear tone.

The trained LSTM model is then applied to the main corpus, restricted to 887,300 articles that contain at least some information on the aspect of interest. Each article is classified either as negative (−1) or as positive/no clear tone (+1). For each topic, I identify the eleven articles with the highest share of that topic and assign it the prevailing sentiment sign (+1 or −1). Finally, the daily topic proportion is multiplied by this prevailing Business Cycle Sentiment sign to obtain the adjusted series.

In the next section, I describe how the sentiment- and uncertainty-adjusted daily topic series are incorporated into the out-of-sample forecasting experiment.

2.3 Forecasting experiment

The forecasting exercise focuses on three key macroeconomic variables with long publication lags: GDP, consumption, and investment. Forecasts are generated using the monthly Dynamic Factor Model (DFM) of Bańbura et al. (2010). This model has become a standard tool for macroeconomic forecasting, widely applied in central banks, including the New York Fed Staff Nowcast (see Bok et al., 2018). Its main advantages are the ability to handle large datasets through dimensionality reduction, to accommodate both monthly and quarterly series, and to address the non-synchronicity of data releases. Since the model is well established in the literature, I refer to Bańbura et al. (2010) for methodological details and provide only the implementation specifics here.

My real-time dataset consists of 18 monthly economic indicators, 10 financial variables, 8 survey series, 10 text-based indices (which vary depending on the text approach and the forecast target), and one quarterly variable (quarter-on-quarter annualized GDP, consumption, or investment growth). The full list of hard data and survey series is provided in Appendix 2.B. I do not include more disaggregated macroeconomic indicators, as they generally do not provide substantial gains in forecasting accuracy (see, e.g., Bok et al., 2018). Data vintages for economic and financial variables come from the Deutsche Bundesbank Real-Time Database and are supplemented with surveys (the ifo Business Cycle Index and its components, the GfK Consumer Climate Indicator and its components, and the European Commission’s Economic Sentiment Indicator, ESI).

Before estimating the DFM, the dataset was pre-processed to ensure stationarity. Transformations for hard data and surveys are reported in Appendix 2.B. For the daily text-based indices, I applied a 30-day backward-looking moving average filter to reduce noise, detrended the resulting series with a biweight filter (bandwidth of 1,200 days) to remove long-run trends that are uninformative for short-term forecasting (Stock and Watson, 2016), and aggregated them to monthly frequency by averaging. All transformations were carried out in real time. Finally, all series (hard data, surveys, and text indices) were standardized prior to estimation.

The main objective of this forecasting experiment is to assess whether text-based information improves forecasts of key macroeconomic aggregates relative to a benchmark model containing only hard data and surveys. If text proves useful, the next question is which representation is most effective: plain topics, topics adjusted with sentiment lexicons or the share of uncertainty terms, or topics adjusted with the more advanced Business Cycle Sentiment measure based on supervised machine learning. To this end, I estimate two factors from hard data and surveys for GDP and investment, and one factor for consumption (based on empirical performance), alongside a separate factor that loads only on the text series and the forecasted variable. For the text block, I select the 10 plain

topics (or sentiment/uncertainty-adjusted topics) most highly correlated with the target variable, as this specification delivered the best forecasting performance in Okuneva et al. (2024). This block structure allows me to isolate the contribution of text-based information and evaluate whether it improves forecasts of macroeconomic aggregates.

The design of the forecasting experiment follows standard practice in the real-time forecasting literature. The estimation sample spans 1991Q2 to 2007Q4, while the out-of-sample period runs from 2008Q1 to 2018Q4. This period covers both the Financial Crisis and the European Debt Crisis. I focus on nowcasts of quarterly variables, which are updated seven times per quarter. The first nowcast for each quarter is produced on its first day, using all information available up to the end of the previous quarter (e.g., the 2008Q1 nowcast uses data up to December 31, 2007). Subsequent updates are generated on the 16th day of the first month, the 1st and 16th days of the second month, and the 1st and 16th days of the third month, with the final nowcast produced on the 1st day of the subsequent quarter. Each update incorporates all data releases available up to that day. Estimation is carried out using the `DynamicFactorMQ` class from the `Statsmodels` package in Python.² The lag order of the VAR process for the factors is set to three. The idiosyncratic component follows an AR(1) process. The model is estimated in a state-space framework using the Kalman filter, with maximum likelihood inference facilitated by the EM algorithm.

2.4 Empirical results

This section reports the results of the real-time out-of-sample forecasting exercise for GDP, investment, and consumption. The analysis pursues two main objectives. First, it examines whether the selected text-based indices are strongly correlated with the forecast variables and with popular survey indicators. Second, it evaluates whether incorporating a separate factor constructed from text-based information improves forecasting performance relative to a benchmark DFM that relies only on hard data and surveys.

2.4.1 GDP

2.4.1.1 Descriptive analysis

The first forecast variable is GDP growth³. For the topics estimated from the baseline corpus of 3.3 million articles (“plain topics”), these same topics adjusted with the economy-specific BPW lexicon (“economic lexicon-adjusted topics”), the general SentiWS lexicon

²https://www.statsmodels.org/devel/generated/statsmodels.tsa.statespace.dynamic_factor_mq.DynamicFactorMQ.html

³The GDP series is taken from the Bundesbank real-time database under the code BBKRT. Q. DE. Y. A. AG1. CA010. A. I, which corresponds to the chain-linked volume index of GDP, calendar- and seasonally adjusted.

(“general lexicon-adjusted topics”), and the share of uncertainty terms (“uncertainty-adjusted topics”), I select the ten topics most strongly correlated (in absolute value) with GDP growth to construct the separate text factor. For the aspect-based topics derived from the business cycle corpus of 887,300 articles and adjusted with business cycle sentiment (“BCS-adjusted topics”), I select the ten topics that exhibit stable correlations with GDP growth across 34 real-time vintages during the period 2010-2018 (see Okuneva et al., 2024 for details).

Tables 2.1, 2.2, and 2.3 report the correlations between GDP growth, the selected topics (plain and adjusted with economic lexicon or BCS), and two leading survey indicators commonly used to track GDP dynamics — the Ifo Business Climate Index (including its components, the Ifo Business Situation and Ifo Business Expectations) and the European Commission’s Economic Sentiment Indicator (ESI). To conserve space, the corresponding results for the general lexicon- and uncertainty-adjusted topics are provided in Appendix 2.C. All tables report correlation coefficients and significance levels based on *t*-statistics from univariate OLS regressions, with Newey–West standard errors used to correct for autocorrelation.

This analysis is important for three reasons. First, it clarifies which topics (and their sentiment- or uncertainty-adjusted counterparts) are selected for the forecasting experiment and whether this selection aligns with economic intuition (for example, whether adjustment with BCS shifts attention toward topics that are more informative about GDP dynamics). Second, it allows me to assess how the correlations of a given topic change across different adjustments and which specification yields the strongest association with GDP growth. Third, it examines whether text-based indicators contain information similar to that in established survey indicators. Strong correlations with surveys would suggest that text series capture economically meaningful signals, whereas correlations that are excessively high could indicate redundancy.

Table 2.1 shows that the correlations of the ten plain topics most closely associated with GDP growth range from 0.231 to 0.498 in absolute value. The signs of the coefficients are economically plausible: most selected topics display negative correlations with GDP growth, indicating that their increased prominence in the news tends to coincide with weaker economic activity. Two topics, T150 (“Corporate Growth”) and T59 (“Commodity Markets”), exhibit positive correlations, consistent with the notion that these themes attract greater attention during periods of expansion.

Several topics were excluded from consideration due to their limited economic relevance (T121, “Literature and Arts”), excessive specificity to a particular company or political party (T56, “Performance and Continental AG”; T78, “CSU and Leadership Dynamics”), or focus on historical episodes unlikely to inform current forecasts (T31, “Governmental Reorganization”; T172, “German Reunification and Economic Transition”). The remaining

Table 2.1: Correlations of selected plain topics with quarterly GDP growth (first release) and selected surveys

ID	Label	GDP	ifo Climate	ifo Situation	ifo Expectations	ESI
T50	Crisis	-0.498***	-0.524***	-0.349***	-0.712***	-0.493***
T150	Corporate Growth	0.438***	0.479***	0.347***	0.593***	0.508***
T29	Banking	-0.424**	-0.404***	-0.325***	-0.439***	-0.342***
T21	Policy Measures	-0.352*	-0.180	-0.059	-0.364**	-0.148
T38	Problem Solving	-0.351**	-0.183	-0.052	-0.387**	-0.183
T108	US Politics	-0.340**	-0.254**	-0.129	-0.429***	-0.220*
T59	Commodity Markets	0.314***	0.283**	0.171	0.423***	0.262*
T120	Economic Growth	-0.250*	-0.323***	-0.277***	-0.316**	-0.319***
T91	Media Coverage of Plans and Rumors	-0.240***	-0.366***	-0.272**	-0.436***	-0.388***
T134	Steel Industry Restructuring and Downsizing	-0.231	-0.350**	-0.308**	-0.330**	-0.374***

Notes: The selected topics show the strongest correlations with GDP growth among all estimated plain topics. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Situation” denotes the ifo Business Situation for industry & trade (balances); “ifo Expectations” denotes the ifo Business Expectations for industry & trade (balances); “ESI” denotes the Economic Sentiment Indicator of European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with GDP growth data. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order 4).

topics were retained because they allow for a clearer economic interpretation and appear more relevant for forecasting purposes. Notably, several of these topics are also strongly correlated with survey indicators: for instance, T50 (“Crisis”) exhibits a correlation of -0.712 with the ifo Business Expectations Index.

Turning to the economic lexicon-adjusted topics (Table 2.2), I find that nearly all correlations with GDP growth and survey indicators are statistically significant, with the only exception being the correlation between T154 (“Electronics Industry”) and the ifo Business Situation Indicator. The magnitudes are also higher than for the plain topics, ranging from 0.375 to 0.522 with GDP growth. Only one topic, T56 (“Performance and Continental AG”), was excluded despite its high correlations, as it was narrowly focused on a single company. Notably, all correlation signs are positive, consistent with the interpretation that more favorable sentiment about these topics is associated with stronger economic activity. Correlations with survey indicators are particularly high: T50 (“Crisis”) correlates with the ifo Business Expectations Index at 0.709, and T120 (“Economic Growth”) at 0.789.

Finally, the aspect-based topics adjusted with BCS, designed to emphasize economically relevant content, show even stronger correlations with GDP growth, ranging from 0.381 to 0.530 (Table 2.3). Nearly all coefficients are statistically significant, with the exception of T100 and its correlation with the ifo Business Situation Index. Correlations with survey

Table 2.2: Correlations of selected economic lexicon-adjusted topics with quarterly GDP growth (first release) and selected surveys

ID	Label	GDP	ifo Climate	ifo Situation	ifo Expectations	ESI
T50	Crisis	0.522***	0.494***	0.310***	0.709***	0.463***
T183	Financial and Economic Performance	0.486**	0.521***	0.402***	0.599***	0.522***
T120	Economic Growth	0.479***	0.544***	0.337***	0.789***	0.502***
T29	Banking	0.446**	0.492***	0.366***	0.595***	0.455***
T150	Corporate Growth	0.442***	0.594***	0.464***	0.668***	0.606***
T154	Electronics Industry	0.415**	0.289**	0.121	0.531***	0.318***
T167	Corporate Financial Performance	0.413***	0.586***	0.500***	0.579***	0.579***
T21	Policy Measures	0.412**	0.378***	0.218**	0.582***	0.389***
T112	Trade Fairs	0.394**	0.586***	0.474***	0.632***	0.607***
T7	Mergers and Acquisition	0.375**	0.406***	0.282***	0.525***	0.433***

Notes: The selected topics show the strongest correlations with GDP growth among all estimated plain topics adjusted with economic lexicon. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Situation” denotes the ifo Business Situation for industry & trade (balances); “ifo Expectations” denotes the ifo Business Expectations for industry & trade (balances); “ESI” denotes the Economic Sentiment Indicator of European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with GDP growth data. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order 4).

indicators are also high, though slightly lower than in the economic lexicon-adjusted case: the strongest is between T138 (“Financial and Economic Performance”) and the ifo Business Expectations Index at 0.684, compared with 0.789 for the economic lexicon-adjusted topics. This pattern suggests that BCS-adjusted topics remain closely aligned with survey-based assessments of economic conditions while at the same time capturing somewhat distinct information. Notably, across all specifications (plain, economic lexicon-adjusted, and BCS-adjusted topics) the highest correlations consistently arise with the ifo Business Expectations Index, suggesting that text-based indicators reflect expectations about future economic conditions.

Results for the plain topics adjusted with the general lexicon and for these same topics adjusted with the share of uncertainty terms are reported in Appendix 2.C. Correlations of GDP with the general lexicon-adjusted topics are relatively strong, although the weakest coefficient is 0.301 — lower than for economic lexicon- or BCS-adjusted topics. By contrast, correlations between GDP and the uncertainty-adjusted topics are much weaker in absolute value, which may reflect the fact that many articles contain no uncertainty-related terms, leaving the resulting measure relatively sparse. Overall, topics adjusted with the economic lexicon and with BCS sentiment exhibit the strongest contemporaneous correlations with GDP growth and therefore appear to be the most promising candidates for nowcasting this variable.

Table 2.3: Correlations of selected BCS-adjusted topics with quarterly GDP growth and selected surveys

ID	Label	GDP	ifo Climate	ifo Situation	ifo Expectations	ESI
T27	Economic Crises and Recessions	0.530***	0.433***	0.276***	0.616***	0.393***
T81	Corporate Restructuring and Job Cuts in Germany	0.507***	0.368***	0.253***	0.486***	0.344***
T127	Major Banks and Investment Banking	0.466**	0.463***	0.315***	0.615***	0.442***
T138	Financial and Economic Performance	0.461**	0.503***	0.336***	0.684***	0.515***
T11	Mergers and Acquisitions	0.449***	0.473***	0.350***	0.571***	0.525***
T52	German Automobile Industry and Major Manufacturers	0.440***	0.409***	0.290***	0.527***	0.383***
T100	Market Reactions to News	0.434***	0.320***	0.138	0.581***	0.305***
T74	Concerns about Economic Bubbles and Recessions	0.427**	0.471***	0.334***	0.601***	0.442***
T131	German Investments in Emerging Markets	0.390**	0.381***	0.294***	0.437***	0.370***
T77	Private Investment	0.381**	0.416***	0.307***	0.506***	0.419***

Notes: The selected topics show the strongest correlations with GDP growth among all estimated aspect-based topics adjusted with BCS. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Situation” denotes the ifo Business Situation for industry & trade (balances); “ifo Expectations” denotes the ifo Business Expectations for industry & trade (balances); “ESI” denotes the Economic Sentiment Indicator of European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with GDP growth data. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order 4).

To summarize the results and assess how sentiment and uncertainty adjustments influence the selection of topics most closely correlated with GDP growth, Table 2.4 provides an overview. A consistent finding across all specifications is that crisis-related topics consistently exhibit some of the strongest correlations with GDP growth, regardless of the adjustment applied. For the plain topics, as well as for those adjusted with sentiment lexicons or the share of uncertainty terms, this corresponds to T50 (“Crisis”). In the BCS-adjusted specification, the corresponding topic is T27 (“Economic Crises and Recessions”), which represents a more economically focused version of the crisis theme. Another recurring topic across the different adjustments is T120 (“Economic Growth”), which also aligns with economic intuition, as its prominence typically increases during episodes of economic stress.

At the same time, the choice of adjustment clearly influences the broader set of topics selected. Economic lexicon-adjusted topics, which incorporate sentiment tailored to business communication, place greater emphasis on financial themes and business activity. By contrast, adjustment with the general lexicon brings forward broader themes such as T38 (“Problem Solving”) and topics related to politics and policy. Adjustment based on the

Table 2.4: Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with GDP growth

Thematic category	Topics (plain)	Economic lexicon-adjusted	BCS-adjusted	General lexicon-adjusted	Uncertainty-adjusted
Crisis/Growth	T50 Crisis	T50 Crisis	T27 Economic Crises and Recessions	T50 Crisis	T50 Crisis
	T120 Economic Growth	T120 Economic Growth	T74 Concerns about Economic Bubbles and Recessions	T120 Economic Growth	T120 Economic Growth
	T150 Corporate Growth	T150 Corporate Growth			
Financial Topics	T29 Banking	T29 Banking	T100 Market Reactions to News	T29 Banking	—
	T59 Commodity Markets	T167 Corporate Financial Performance T183 Financial and Economic Performance	T127 Major Banks and Investment Banking T138 Financial and Economic Performance		
Industries	T134 Steel Industry Restructuring and Downsizing	T154 Electronics Industry	T52 German Automobile Industry and Major Manufacturers	T134 Steel Industry Restructuring and Downsizing	T134 Steel Industry Restructuring and Downsizing T154 Electronics Industry T155 Automotive Industry
Politics/Policies	T21 Policy Measures T108 US Politics	T21 Policy Measures	—	T21 Policy Measures T108 US Politics T124 Elections and Office Succession	T21 Policy Measures
More general topics	T38 Problem Solving T91 Media Coverage of Plans and Rumors	—	—	T38 Problem Solving T91 Media Coverage of Plans and Rumors T98 Public Appearances and Reactions	T139 Interviews and Opinions
Business Activity	—	T7 Mergers and Acquisitions T112 Trade Fairs	T11 Mergers and Acquisitions	—	T7 Mergers and Acquisitions T14 Corporate Structure and M&A
Labor Market	—	—	T81 Corporate Restructuring and Job Cuts in Germany	—	T190 Savings and Retirement Planning
Investment	—	—	T77 Private Investment T131 German Investments in Emerging Markets	—	—

Notes: The table summarizes the plain topics and their sentiment- or uncertainty-adjusted counterparts (economic lexicon-adjusted, general lexicon-adjusted, and uncertainty-adjusted topics), as well as the aspect-based topics adjusted with BCS sentiment, that exhibit the strongest correlations with GDP growth. Topics are grouped into broader thematic categories based on subjective interpretation to illustrate how different adjustments influence the selection of topics most closely associated with GDP growth.

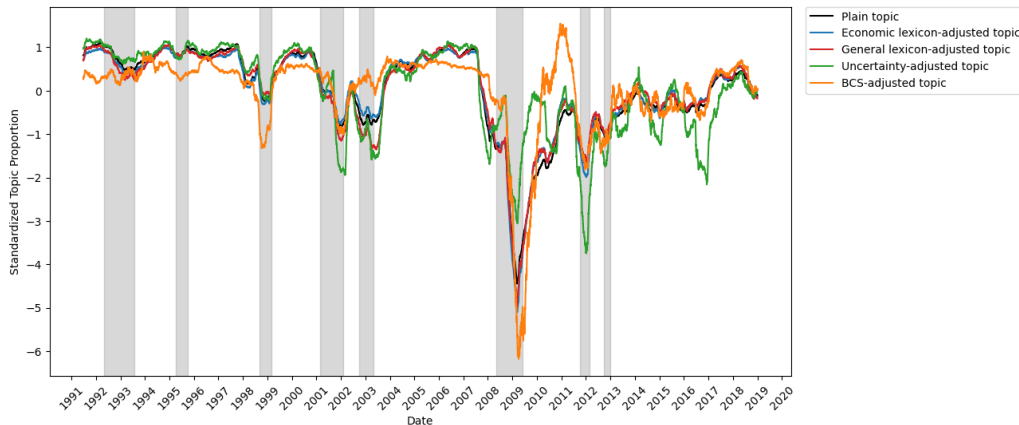
proportion of uncertainty terms leads to the selection of T21 (“Policy Measures”), consistent with the findings of Baker et al. (2016), along with other topics in which uncertainty is frequently discussed, such as mergers and acquisitions or industry-specific developments. Most importantly, adjustment with BCS sentiment yields a set of topics that appears most economically relevant: in addition to crisis-related themes, the selection includes financial topics, the automotive industry, business activity, the labor market, and investment—all central dimensions of the economy.

Taken together, these results indicate that the choice of sentiment or uncertainty adjustment substantially affects which news topics receive greater weight, with important implications for their economic interpretation and relevance for forecasting GDP growth.

Moreover, even when the same topic is selected in both plain and sentiment- or uncertainty-adjusted form, its correlation with GDP growth depends strongly on the type of adjustment applied. For instance, the crisis-related topic (T50, “Crisis”, in the plain, economic lexicon-adjusted, general lexicon-adjusted, and uncertainty-adjusted versions, and T27, “Economic Crises and Recession” in the BCS-adjusted specification) appears in all cases. Yet its correlation with GDP growth ranges from only 0.244 in absolute value for the uncertainty-adjusted version to as high as 0.530 when the aspect-based crisis topic is adjusted with BCS sentiment.

This difference is clearly visible in Figure 2.2. To make the series comparable, I reverse

Figure 2.2: Crisis topic in plain and sentiment-/uncertainty-adjusted forms (T50, and T27 in the BCS-adjusted case)

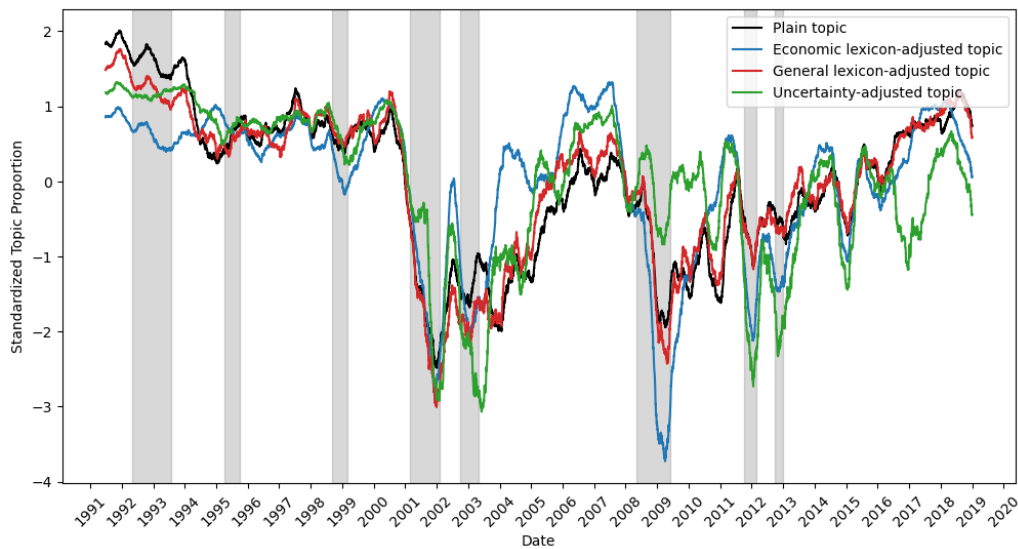


Notes: The figure shows the plain “Crisis” topic (T50) and its sentiment- or uncertainty-adjusted counterparts, with T27 (“Economic Crises and Recessions”) serving as the corresponding aspect-based topic in the BCS-adjusted specification. The Y-axis shows the standardized 180-day backward moving average of the respective daily text series, and the X-axis shows the corresponding calendar day. The plain topic (T50) is shown in black, the economic lexicon-adjusted topic in blue, the general lexicon-adjusted topic in red, the uncertainty-adjusted topic in green, and the BCS-adjusted topic (T27) in orange. For comparability, the signs of the plain topic and the uncertainty-adjusted topic have been reversed, and all series are standardized. Shaded areas denote recession periods in Germany as identified by Carstensen et al. (2020).

the sign of both the plain topic and the uncertainty-adjusted series. While all variants of the crisis topic display a pronounced decline during the Financial Crisis, the BCS-adjusted version shows the sharpest drop, capturing the depth of the downturn more accurately. In addition, this series exhibits a pronounced overshooting in sentiment after the crisis, reflecting overly optimistic recovery assessments that characterized public discourse at the time and consistent with overshooting dynamics highlighted in the uncertainty literature (see, e.g., Bloom, 2009). Finally, the uncertainty-adjusted topic reacts more strongly to the European Debt Crisis than to the Financial Crisis, consistent with Baker et al. (2016), which highlights how strongly the choice of adjustment shapes the economic signal extracted from news.

Another illustrative case is topic T120 (“Economic Growth”). Here, I compare the plain topic series with the versions adjusted using the economic and general lexicons, as well as the share of uncertainty terms. As before, the signs of the plain and uncertainty-adjusted series are reversed for comparability. The most striking differences again emerge during the Financial Crisis: although all series decline over this period, the drop is most pronounced in the economic lexicon-adjusted series, reflecting the BPW lexicon’s focus on business communication. Consequently, this adjustment yields the largest contemporaneous correlation with GDP growth (0.479). This example highlights that both dimensions of the news signal matter—what topics are discussed and the sentiment or uncertainty expressed in connection with them.

Figure 2.3: Economic Growth topic in plain and sentiment-/uncertainty-adjusted forms (T120)



Notes: The figure shows the plain “Economic Growth” topic (T120) and its sentiment- or uncertainty-adjusted counterparts. The Y-axis shows the standardized 180-day backward moving average of the respective daily text series, and the X-axis the corresponding calendar day. The plain topic (T120) is shown in black, the economic lexicon-adjusted topic in blue, the general lexicon-adjusted topic in red, and the uncertainty-adjusted topic in green. For comparability, the signs of the plain topic and the uncertainty-adjusted topic have been reversed, and all series are standardized. Shaded areas denote recession periods in Germany as identified by Carstensen et al. (2020).

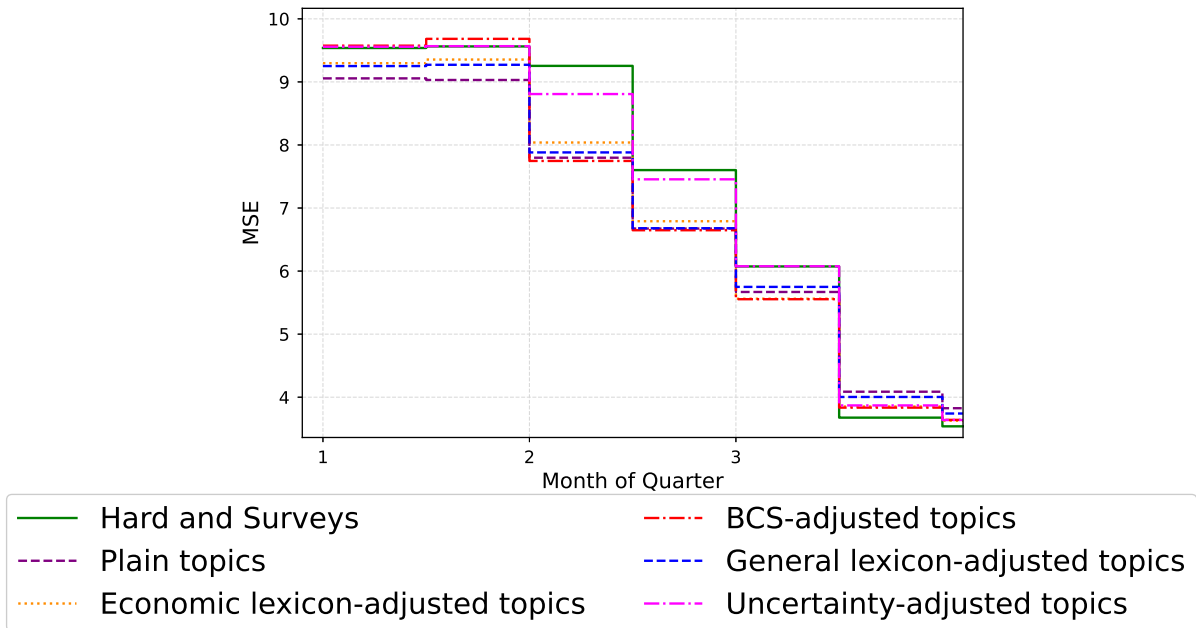
2.4.1.2 Results of the forecasting experiment

Finally, I turn to the results of the out-of-sample forecasting experiment. The purpose of this exercise is to assess whether text-based information improves nowcasts of GDP growth compared with a benchmark model that relies only on traditional financial and macroeconomic series, as well as survey indicators.

Figures 2.4 and 2.5 report mean squared forecast errors (MSFEs) for a benchmark Dynamic Factor Model (DFM) with two factors extracted from hard and survey data (green, solid line) and five competing models that additionally include one text-based factor. The text factor is constructed from either plain topics (purple, dashed line), economic lexicon-adjusted topics (orange, dotted line), BCS-adjusted topics (red, dash-dotted line), general lexicon-adjusted topics (blue, dashed line), or topics adjusted by the share of uncertainty terms (pink, dash-dotted line). Figure 2.4 shows results for the full evaluation period (2008–2018), while Figure 2.5 distinguishes between the Financial Crisis (2008–2010) and the subsequent period of calmer economic conditions that nevertheless includes the European Debt Crisis (2011–2018).

Separating these two periods is important, since forecast errors are markedly higher during the Financial Crisis and would otherwise dominate the full-sample results. Moreover, different text adjustments may be informative in times of economic distress than in more stable conditions. The GDP growth series is taken from the Bundesbank real-time database (first release), and all models are estimated using an expanding-window

Figure 2.4: Mean squared forecast errors (MSFEs) for different nowcasting models of GDP growth (2008–2018)



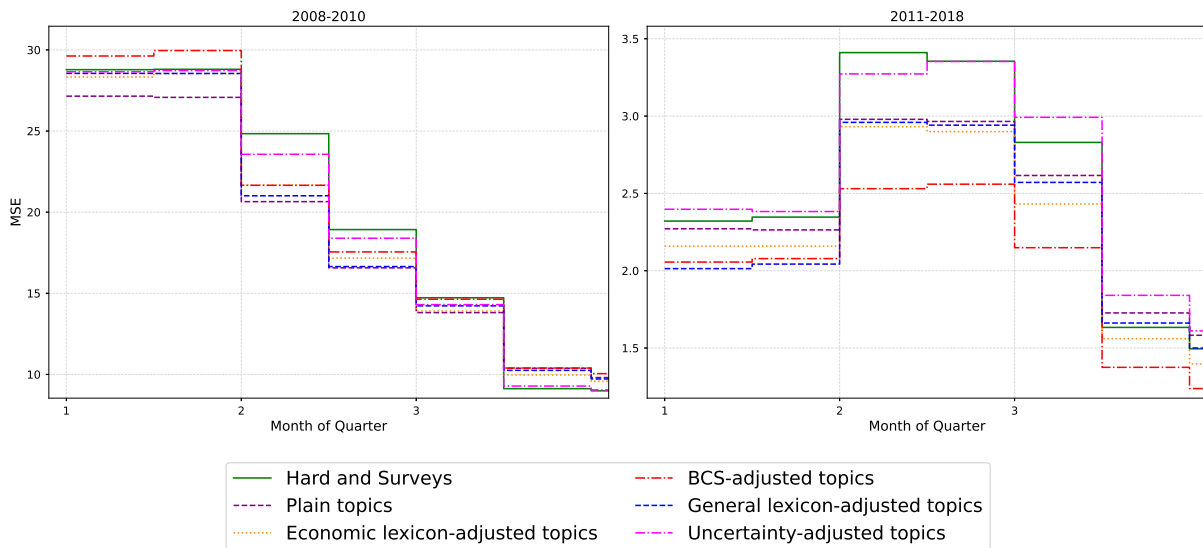
Notes: The figure reports mean squared forecast errors (MSFEs) for different Dynamic Factor Models (DFMs) evaluated over the period 2008–2018. The benchmark model includes two factors extracted from hard and survey data only (shown in green, solid line). Competing models augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The horizontal axis indicates the nowcast day, covering days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. Results are shown for the full evaluation sample (2008–2018).

approach.

Figure 2.5 shows that during the Financial Crisis (2008–2010, left panel), the specification that most clearly outperforms the benchmark is the one using plain topics, with the largest gains in the first two months of the quarter. Models that include sentiment-adjusted topics (economic lexicon, BCS, or general lexicon) deliver improvements of a similar magnitude, although the BCS-adjusted version performs slightly less well at the very beginning of the quarter. By contrast, the uncertainty-adjusted specification performs somewhat worse than the other text-based models in the second month of the quarter, precisely when the gains of the other specifications are the largest. By the third month of the quarter, and on the first day of the subsequent quarter, the gains from incorporating text information diminish and the performance of the text-based DFMs converges toward, or in some cases falls below, that of the benchmark model.

The strongest improvements occur in the first two months of the quarter, when the model incorporates only text information from the final month of the previous quarter or the first month of the current one. This pattern is consistent with earlier evidence

Figure 2.5: Mean squared forecast errors (MSFEs) for different nowcasting models of GDP growth, 2008–2010 and 2011–2018



Notes: The figure reports mean squared forecast errors (MSFEs) of different Dynamic Factor Models (DFMs) on the vertical axis against the day of the nowcast on the horizontal axis. The benchmark DFM includes two factors extracted from hard and survey data only (green, solid line). The competing DFMs augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The nowcast days include days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. The left panel reports results for the Financial Crisis period (2008–2010), and the right panel for the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

(see, e.g., Ashwin et al., 2024) showing that text data is most informative early in the quarter, when few hard data releases are available. A novel finding here is that the type of adjustment matters: plain topics yield the largest gains.

Table 2.5 (Panel B) reports the corresponding relative RMSFE statistics. Bold entries indicate at least a 5% improvement over the benchmark, and asterisks denote statistical significance based on a one-sided Diebold–Mariano (DM; Diebold and Mariano, 1995) test with Newey–West standard errors. The table confirms the graphical evidence: relative RMSFEs are lowest for the model using plain topics up to and including the beginning of the third month of the quarter, with statistically significant gains of about 9% at the start of the second month. By contrast, the BCS-adjusted specification outperforms the benchmark only in the second month of the quarter, and these improvements are not statistically significant.

Turning to the 2011–2018 period (right panel of Figure 2.5), the results look markedly different. In this calmer period, the clear leader throughout the quarter is the model that incorporates BCS-adjusted topics developed in Okuneva et al. (2024). While the specifica-

Table 2.5: Relative root mean squared forecast errors for GDP growth nowcasts across subperiods

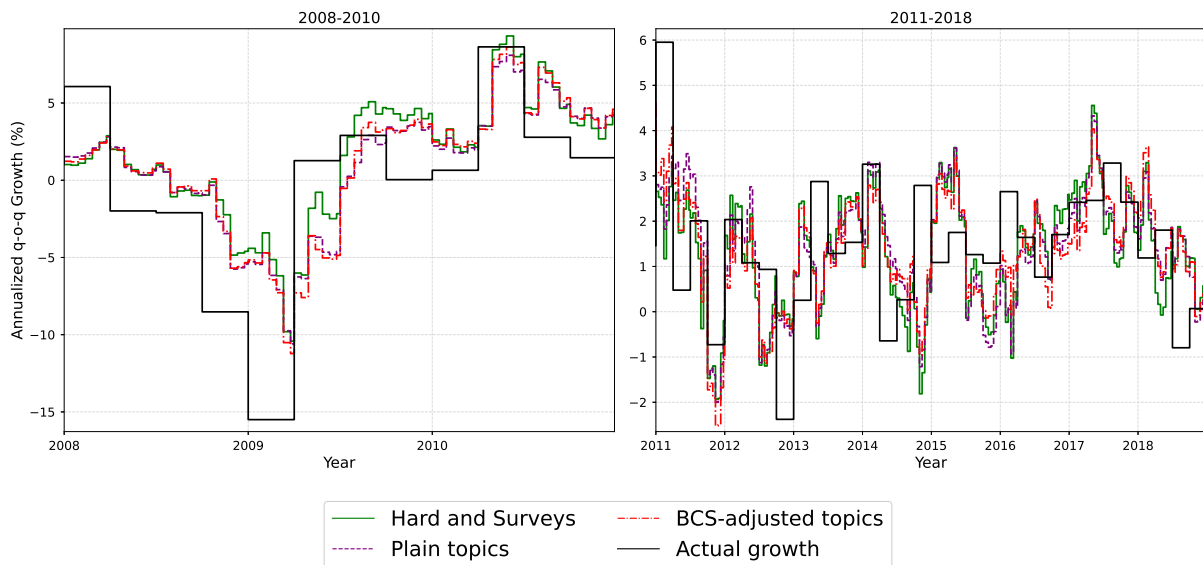
Text Factor	M1-01	M1-16	M2-01	M2-16	M3-01	M3-16	M4-01
Panel A: 2008–2018 (Full Sample)							
Plain topics	0.97	0.97	0.92*	0.94*	0.97	1.05	1.04
Economic lexicon-adj. topics	0.99	0.99	0.93**	0.95*	0.96	1.02	1.01
BCS-adj. topics	1.00	1.01	0.91**	0.94	0.96	1.02	1.01
General lexicon-adj. topics	0.98	0.98	0.92*	0.94*	0.97	1.04	1.03
Uncertainty-adj. topics	1.00	1.00	0.98*	0.99	1.00	1.03	1.01
Panel B: 2008–2010 (Financial Crisis)							
Plain topics	0.97	0.97	0.91*	0.94	0.97	1.07	1.04
Economic lexicon-adj. topics	0.99	1.00	0.93*	0.95	0.97	1.05	1.03
BCS-adj. topics	1.01	1.02	0.93	0.96	1.00	1.07	1.06
General lexicon-adj. topics	1.00	1.00	0.92*	0.94	0.98	1.06	1.04
Uncertainty-adj. topics	1.00	1.00	0.97*	0.99	0.99	1.01	1.00
Panel C: 2011–2018 (Post-crisis, incl. European Debt Crisis)							
Plain topics	0.99	0.98	0.93**	0.94*	0.96*	1.03	1.03
Economic lexicon-adj. topics	0.96	0.96*	0.93***	0.93**	0.93***	0.98	0.97
BCS-adj. topics	0.94**	0.94*	0.86***	0.87**	0.87**	0.92*	0.91*
General lexicon-adj. topics	0.93**	0.93**	0.93**	0.94**	0.95**	1.01	1.00
Uncertainty-adj. topics	1.02	1.01	0.98	1.00	1.03	1.06	1.04

Notes: The table reports root mean squared forecast errors (RMSFEs) of a Dynamic Factor Model (DFM) that augments a benchmark specification with one text factor. The benchmark DFM contains two factors extracted from hard and survey data. In the competing models, the additional factor is constructed from plain topics, economic lexicon-adjusted topics, general lexicon-adjusted topics, BCS-adjusted topics, or topics adjusted with the share of uncertainty terms. All entries are expressed as RMSFEs relative to the benchmark model. Values in bold denote improvements of at least 5% compared with the benchmark. Asterisks indicate statistical significance of the improvement according to a one-sided Diebold–Mariano (DM) test (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). Columns indicate the month and day within the quarter when the nowcast is made. Specifically, M1-01 is the first day of the first month, M1-16 is day 16 of the first month and so on. M4-01 is the first day of the fourth month, i.e., the first day of the subsequent quarter.

tions using plain topics and other sentiment-adjusted variants also yield improvements—particularly the general-lexicon adjustment, which performs well up to the middle of the third month—it is the BCS-adjusted topics that deliver the largest gains, reaching up to 14% relative to the benchmark. As shown in Table 2.5 (Panel C), these improvements are statistically significant across the entire quarter. This indicates that the way sentiment adjustment is implemented matters: applying sentiment specifically tailored to business cycle conditions to topics extracted from economic articles is particularly effective in more stable periods. In this context, the more targeted adjustment clearly pays off.

Overall, the results for the full evaluation period (Figure 2.4 and Panel A of Table 2.5) indicate that plain topics and the various sentiment-adjusted specifications yield the largest gains relative to the benchmark, particularly when forecasting in the second month of the quarter. However, as discussed above, these full-sample improvements are largely

Figure 2.6: Nowcasts and actual GDP growth (first release) for different models, 2008–2010 and 2011–2018



Notes: The figure compares nowcasts from several Dynamic Factor Models (DFMs) with the first-release estimate of quarterly GDP growth. The benchmark specification includes two factors extracted from hard and survey data only (green, solid line). Competing models augment this benchmark with one additional text factor, constructed from either plain topics (purple, dashed line) or BCS-adjusted topics (red, dash-dotted line), which were the best-performing specifications in the MSFE evaluation. Nowcasts are shown against the realized quarterly GDP growth series (black, solid line). The left panel displays the results for the Financial Crisis period (2008–2010), while the right panel covers the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

driven by the elevated forecast errors during the Financial Crisis.

Why do plain topics perform better than BCS-adjusted topics when forecasting in the first month of the quarter during the Financial Crisis? Figure 2.6 helps answer this question by plotting actual GDP growth (first release, black) alongside nowcasts from the benchmark model (green) and the two best-performing text-based specifications: the model with plain topics (purple) and the one with BCS-adjusted topics (red). Results are shown separately for 2008–2010 and 2011–2018, and nowcasts from all models are reported in Appendix 2.D. Focusing on the left panel for the Financial Crisis period, the nowcasts based on plain and BCS-adjusted topics track actual GDP growth closely and remain broadly similar. However, even though the BCS-adjusted topics capture the depth of the downturn slightly better at the end of 2009Q1, it is the first month of 2009Q2 that proves decisive: in this period, the plain topic specification captures the recovery more rapidly than the BCS-adjusted version. Figure 2.2 helps explain this difference. While coverage of the crisis topic begins to rebound in the second quarter of 2009, the BCS-adjusted series recovers more slowly because sentiment remains negative for longer.

Why do BCS-adjusted topics perform better during the calmer period from 2011 to

Table 2.6: Correlations of plain topics with quarterly GDP growth (full sample and excluding the Financial Crisis, 2008-2009)

ID	Label	Full sample	Without Financial Crisis
T50	Crisis	-0.498***	-0.241***
T150	Corporate Growth	0.438***	0.270***
T29	Banking	-0.424**	-0.116
T21	Policy Measures	-0.352*	-0.151
T38	Problem Solving	-0.351**	-0.278**
T108	US Politics	-0.340**	-0.254**
T59	Commodity Markets	0.314***	0.284***
T120	Economic Growth	-0.250*	-0.123*
T91	Media Coverage of Plans and Rumors	-0.240***	-0.227**
T134	Steel Industry Restructuring and Downsizing	-0.231	-0.089

Notes: The table reports correlations between plain topics and quarterly GDP growth (first release). Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t -statistics from OLS regression with Newey-West standard errors (maximum lag order = 4).

Table 2.7: Correlations of BCS-adjusted topics with quarterly GDP growth (full sample and excluding the Financial Crisis, 2008-2009)

ID	Label	Full sample	Without Financial Crisis
T27	Economic Crises and Recessions	0.530***	0.235***
T81	Corporate Restructuring and Job Cuts in Germany	0.507***	0.263***
T127	Major Banks and Investment Banking	0.466**	0.227***
T138	Financial and Economic Performance	0.461**	0.245***
T11	Mergers and Acquisitions	0.449***	0.282***
T52	German Automobile Industry and Major Manufacturers	0.440***	0.172**
T100	Market Reactions to News	0.434***	0.259***
T74	Concerns about Economic Bubbles and Recessions	0.427**	0.225***
T131	German Investments in Emerging Markets	0.390**	0.269***
T77	Private Investment	0.381**	0.141

Notes: The table reports correlations between BCS-adjusted topics and quarterly GDP growth (first release). Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t -statistics from OLS regression with Newey-West standard errors (maximum lag order = 4).

2018? First, the right panel of Figure 2.6 shows that the BCS-adjusted specification captures the strong increase in GDP growth in early 2011 more accurately than the alternatives. Second, the correlation analysis helps clarify this finding. As reported in Tables 2.6 and 2.7, BCS-adjusted topics not only display higher and more significant correlations with GDP growth in the full sample, as discussed earlier, but they also remain more strongly correlated with GDP growth when the Financial Crisis period is excluded. In contrast, several plain topics lose statistical significance when the crisis is excluded: for example, the correlations of the “Banking” and “Policy Measures” topics become insignificant, the “Economic Growth” topic drops to marginal significance with a small coefficient, and the correlation for the “Steel Industry Restructuring and Downsizing” topic approaches zero.

Taken together, these results indicate that BCS-adjusted topics yield more stable and economically meaningful correlations with GDP growth in calmer times. This stability, in turn, translates into more reliable forecasting performance when economic conditions are less turbulent.

2.4.2 Investment

2.4.2.1 Descriptive analysis

The next variable considered in the forecasting experiment is investment growth⁴. As a first step, I examine which topics are most closely associated with this variable and whether they resemble the topics identified for GDP. Table 2.8 provides a summary.

The overlap is substantial. As in the case of GDP, crisis- and growth-related topics appear among the most relevant, which is consistent with economic intuition. More broadly, the topics most strongly correlated with investment growth cover finance, politics, industry-specific developments, business activity, the labor market, and several more general themes. In addition, one technology-related topic also appears among the most relevant.

One important difference, however, is that financial topics play an even more important role for investment than for GDP. Among the plain topics and the economic lexicon-adjusted topics, three of the most strongly correlated topics fall into this thematic category, while four such topics are selected in the BCS-adjusted specification and two in both the general lexicon-adjusted and the uncertainty-adjusted specifications. Newly selected topics include T23 (“Insolvency and Financial Rescue”), T63 (“Commodity Markets and Precious Metals”), and T168 (“Derivatives and Financial Instruments”). This is consistent with economic intuition, as financial conditions are a central driver of investment dynamics.

⁴The investment series is taken from the Bundesbank real-time database under the code BBKRT.O.DE.Y.A.CE1.BAA00.A.I., which corresponds to the chain-linked volume index of private sector investment, calendar- and seasonally adjusted.

Table 2.8: Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with investment growth

Thematic category	Topics (plain)	Economic lexicon-adjusted	BCS-adjusted	General lexicon-adjusted	Uncertainty-adjusted
Crisis/Growth	T50 Crisis	T50 Crisis	T27 Economic Crises and Recessions T74 Concerns about Economic Bubbles and Recessions	T50 Crisis	T50 Crisis
	T120 Economic Growth	T120 Economic Growth		T120 Economic Growth	
	T150 Corporate Growth	T150 Corporate Growth			
Financial topics	T23 Insolvency and Financial Rescue T29 Banking	T29 Banking	T63 Commodity Markets and Precious Metals T127 Major Banks and Investment Banking T138 Financial and Economic Performance T168 Derivatives and Financial Instruments	T23 Insolvency and Financial Rescue T29 Banking	T23 Insolvency and Financial Rescue T59 Commodity Markets
	T59 Commodity Markets	T167 Corporate Financial Performance T183 Financial and Economic Performance			
Politics/Policies	T21 Policy Measures T108 US Politics	T21 Policy Measures	—	T21 Policy Measures T108 US Politics	T21 Policy Measures
Technology	T110 Technological Innovation	—	—	—	—
More general topics	—	—	—	T38 Problem Solving T91 Media Coverage of Plans and Rumors T182 Announcements and Reactions	T18 Agreements and Cooperation T123 Media and Newspapers
Industries	T134 Steel Industry Restructuring and Downsizing	T154 Electronics Industry	T52 German Automobile Industry and Major Manufacturers	T134 Steel Industry Restructuring and Downsizing	T134 Steel Industry Restructuring and Downsizing T154 Electronics Industry
Business Activity	—	T9 Small and Medium-Sized Enterprises T112 Trade Fairs	T48 Business Success and Economic Resilience T118 Corporate Governance and Executive Management	—	T7 Mergers and Acquisitions T9 Small and Medium-Sized Enterprises
Labor Market	—	—	T81 Corporate Restructuring and Job Cuts in Germany	—	—

Notes: The table summarizes the plain topics and their sentiment- or uncertainty-adjusted counterparts (economic lexicon-adjusted, general lexicon-adjusted, and uncertainty-adjusted topics), as well as the aspect-based topics adjusted with BCS sentiment, that exhibit the strongest correlations with investment growth. Topics are grouped into broader thematic categories based on subjective interpretation to illustrate how different adjustments influence which topics are most closely associated with investment dynamics.

Table 2.9: Correlations of selected economic lexicon-adjusted topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Current	ifo Expectations	ESI
T112	Trade Fairs	0.585***	0.586***	0.474***	0.632***	0.607***
T150	Corporate Growth	0.564***	0.594***	0.464***	0.668***	0.606***
T167	Corporate Financial Performance	0.554***	0.586***	0.500***	0.579***	0.579***
T154	Electronics Industry	0.532***	0.289**	0.121	0.531***	0.318***
T183	Financial and Economic Performance	0.517***	0.521***	0.402***	0.599***	0.522***
T50	Crisis	0.513***	0.494***	0.310***	0.709***	0.463***
T120	Economic Growth	0.511***	0.544***	0.337***	0.789***	0.502***
T21	Policy Measures	0.485***	0.378***	0.218**	0.582***	0.389***
T29	Banking	0.469***	0.492***	0.366***	0.595***	0.455***
T9	Small and Medium-Sized Enterprises	0.456**	0.492***	0.402***	0.522***	0.467***

Notes: The selected topics show the strongest correlations with quarterly investment growth among all plain topics adjusted with the economic lexicon. "ifo Climate" denotes the ifo Business Climate for industry & trade (balances); "ifo Current" denotes the ifo Current Business Situation; "ifo Expectations" denotes the ifo Business Expectations; "ESI" denotes the European Commission's Economic Sentiment Indicator. Monthly survey indicators are aggregated to the quarterly frequency for comparability with investment growth data. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

The two adjustments that produce the highest correlations with investment growth are, as expected, the economic lexicon-adjusted topics and the BCS-adjusted topics. The

economic lexicon adjustment yields topics that are strongly related to investment growth, with coefficients ranging from 0.456 to 0.585 (see Table 2.9). These values are even higher than those observed for GDP, and they are both sizable and statistically significant. Moreover, the corresponding topics are also highly correlated with the Ifo Business Climate Index, its components, and the European Commission’s Economic Sentiment Indicator, making them promising candidates for the forecasting exercise. The most strongly correlated topics in this case are T112 (“Trade Fairs”), T150 (“Corporate Growth”), and T167 (“Corporate Financial Performance”), in contrast to the crisis-related topic that was most closely associated with GDP. These topics place greater emphasis on business and financial activities, which is consistent with the nature of investment dynamics.

Two economic lexicon-adjusted topics that were highly correlated with investment growth, T56 (“Performance and Continental AG”) and T179 (“Consumer Goods”), were excluded from further consideration. The first is too narrowly focused on a specific company, while the second appears more relevant for forecasting consumption rather than investment. Overall, the ten economic lexicon-adjusted topics retained for analysis appear economically meaningful and plausibly linked to fluctuations in investment.

Topics adjusted with BCS sentiment perform very strongly, with correlations ranging from 0.461 to 0.608 (see Table 2.10). All coefficients are positive, large in magnitude, and statistically significant, as expected. One topic, T15 (“General Commentary”), was excluded from further consideration because of its generic nature. Interestingly, the topics identified as most relevant for GDP are also highly correlated with investment (see Table 2.24 in Appendix 2.E). At the same time, selection based on correlations with investment highlights additional themes, particularly in finance (T63, T168) and business activity (T48, T118), which are expected to be especially important for this variable across different economic regimes.

Correlations with plain topics and those adjusted using the general lexicon are relatively strong, albeit somewhat lower than for the two specifications discussed above, ranging from 0.308 to 0.537 in the former case and from 0.366 to 0.523 in the latter. All coefficients are statistically significant and economically meaningful. By contrast, the lowest correlations are consistently observed for the uncertainty-adjusted topics, which range from 0.223 to 0.464 in absolute value. Further details on these correlation coefficients are provided in Appendix 2.E.

Moreover, as in the case of GDP, even when the same underlying theme is selected across different specifications, its correlation with investment growth varies depending on the adjustment applied. For example, topic T59 (“Commodity Markets”) is among the most strongly correlated topics in both the plain and the uncertainty-adjusted specifications, and the corresponding aspect-based topic T63 (“Commodity Markets and Precious Metals”) is likewise selected under the BCS adjustment. Yet the associated correlation

Table 2.10: Correlations of selected BCS-adjusted topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Current	ifo Expectations	ESI
T27	Economic Crises and Recessions	0.608***	0.433***	0.276***	0.616***	0.393***
T138	Financial and Economic Performance	0.576***	0.503***	0.336***	0.684***	0.515***
T81	Corporate Restructuring and Job Cuts in Germany	0.539***	0.368***	0.253***	0.486***	0.344***
T48	Business Success and Economic Resilience	0.479***	0.535***	0.400***	0.640***	0.494***
T52	German Automobile Industry and Major Manufacturers	0.477***	0.409***	0.290***	0.527***	0.383***
T168	Derivatives and Financial Instruments	0.476***	0.482***	0.365***	0.570***	0.434***
T127	Major Banks and Investment Banking	0.475**	0.463***	0.315***	0.615***	0.442***
T63	Commodity Markets and Precious Metals	0.471***	0.445***	0.318***	0.564***	0.437***
T74	Concerns about Economic Bubbles and Recessions	0.461***	0.471***	0.334***	0.601***	0.442***
T118	Corporate Governance and Executive Management	0.461***	0.418***	0.317***	0.491***	0.374***

Notes: The selected topics show the strongest correlations with quarterly investment growth among all aspect-based topics adjusted with BCS sentiment. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Current” denotes the ifo Current Business Situation; “ifo Expectations” denotes the ifo Business Expectations; “ESI” denotes the European Commission’s Economic Sentiment Indicator. Monthly survey indicators are aggregated to the quarterly level for comparability with investment growth data. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

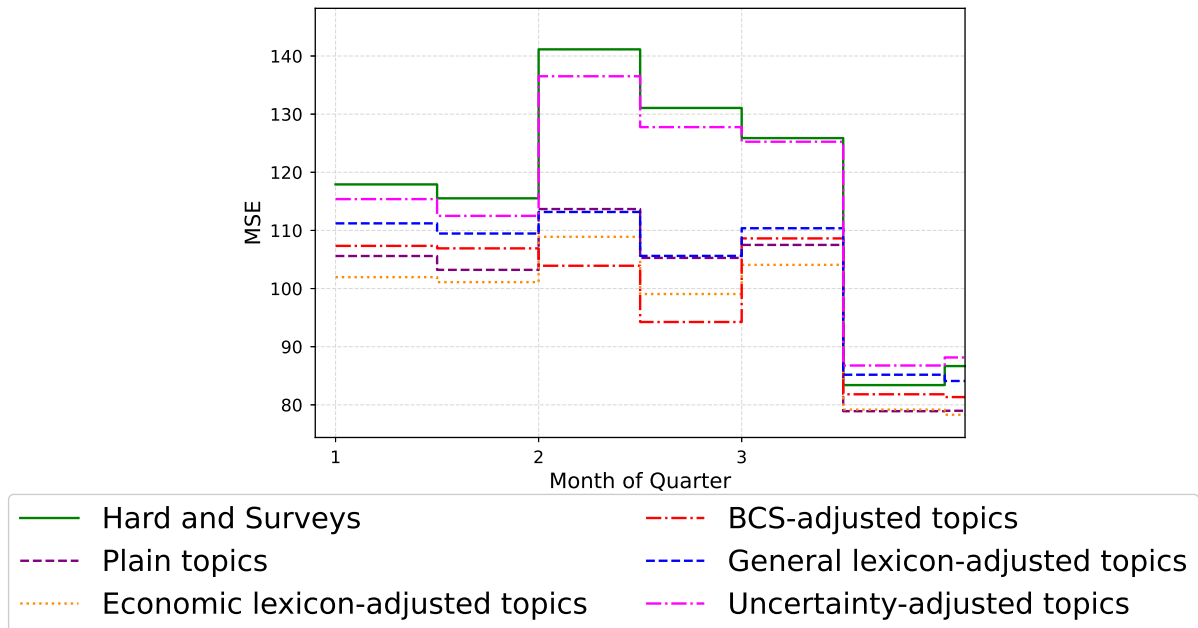
coefficients differ substantially: from 0.306 in the uncertainty-adjusted case, to 0.349 for the plain topic, and up to 0.471 when combined with BCS sentiment. This again highlights that the choice of adjustment significantly influences the economic signal extracted from news.

Overall, while all topic specifications (plain, sentiment-adjusted, and uncertainty-adjusted) appear suitable candidates for the forecasting exercise, the most economically interpretable sets are the plain topics and those adjusted with the economic lexicon or BCS sentiment. In contrast, adjustments based on the general lexicon or the share of uncertainty terms tend to emphasize broader themes.

2.4.2.2 Results of the forecasting experiment

The main question, of course, is whether plain and sentiment- or uncertainty-adjusted topics improve out-of-sample nowcasts of investment. Figure 2.7 reports mean squared forecast errors (MSFEs) for the benchmark model with two factors extracted from hard

Figure 2.7: Mean squared forecast errors (MSFEs) for different nowcasting models of investment growth (2008–2018)



Notes: The figure reports mean squared forecast errors (MSFEs) for different Dynamic Factor Models (DFMs) evaluated over the period 2008–2018. The benchmark model includes two factors extracted from hard and survey data only (shown in green, solid line). Competing models augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The horizontal axis indicates the nowcast day, covering days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. Results are shown for the full evaluation sample (2008–2018).

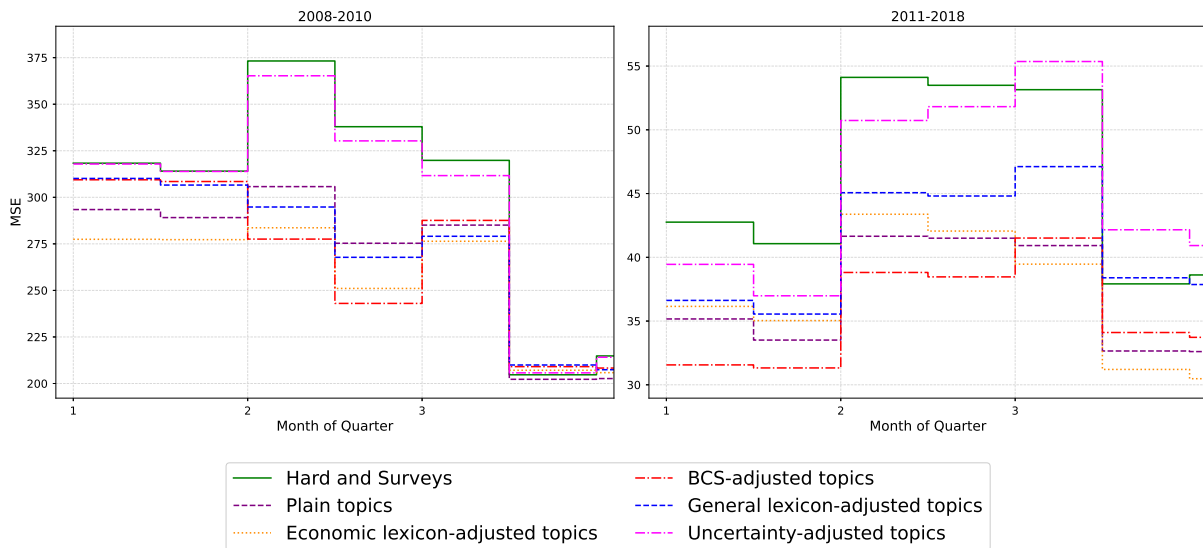
data and surveys (green), alongside specifications that include an additional text factor constructed from plain topics (purple) or from topics adjusted with the economic lexicon (orange), BCS sentiment (red), the general lexicon (blue), or the share of uncertainty terms (pink). The figure displays results for the full evaluation sample (2008–2018).

Three specifications consistently outperform the benchmark throughout the quarter: plain topics, economic lexicon-adjusted topics, and BCS-adjusted topics. These are also the topic sets that are most highly correlated with investment growth and the most economically intuitive. By contrast, the weakest performance is observed for the uncertainty-adjusted specification.

These results are examined in greater detail in Panel A of Table 2.11. For the plain topics specification, the reductions in MSFE relative to the benchmark are statistically significant up to the middle of the third month of the quarter. For the sentiment-adjusted specifications, the improvements are more concentrated in the second month of the quarter and at the beginning of the third.

As before, nowcasting during the Financial Crisis is substantially more difficult, with

Figure 2.8: Mean squared forecast errors (MSFEs) for different nowcasting models of investment growth, 2008–2010 and 2011–2018



Notes: The figure reports mean squared forecast errors (MSFEs) of different Dynamic Factor Models (DFMs) on the vertical axis against the day of the nowcast on the horizontal axis. The benchmark DFM includes two factors extracted from hard and survey data only (green, solid line). The competing DFMs augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The nowcast days include days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. The left panel reports results for the Financial Crisis period (2008–2010), and the right panel for the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

much larger error magnitudes. To avoid mixing crisis-specific dynamics with performance under normal conditions, I therefore separately analyze MSFEs for the Financial Crisis period (2008–2010, left panel of Figure 2.8) and the subsequent period of more stable economic conditions (2011–2018, right panel).

Compared with GDP, forecasting errors for investment are considerably larger in both subperiods. During the Financial Crisis (left panel), the only model that performs noticeably worse than the others is the uncertainty-adjusted specification. All remaining models—plain topics and sentiment-adjusted topics—deliver improvements over the benchmark up to the middle of the third month.

Panel B of Table 2.11 shows that the plain topics specification exhibits statistically significant gains beginning in the second half of the first month and lasting until the middle of the third month. For the sentiment-adjusted specifications, the gains are somewhat larger and achieve statistical significance in the second month of the quarter, while for the general-lexicon adjustment, significance also extends into the beginning of the third month. Overall, plain topics alone provide statistically significant gains, though

Table 2.11: Relative root mean squared forecast errors for investment growth nowcasts across subperiods

Text Factor	M1-01	M1-16	M2-01	M2-16	M3-01	M3-16	M4-01
Panel A: 2008–2018 (Full Sample)							
Plain topics	0.95**	0.95**	0.90***	0.90***	0.92***	0.97	0.95
Economic lexicon-adj. topics	0.93	0.94	0.88**	0.87**	0.91*	0.97	0.95
BCS-adj. topics	0.95	0.96	0.86**	0.85**	0.93*	0.99	0.97
General lexicon-adj. topics	0.97	0.97	0.90*	0.90**	0.94*	1.01	0.99
Uncertainty-adj. topics	0.99	0.99*	0.98*	0.99	1.00	1.02	1.01
Panel B: 2008–2010 (Financial Crisis)							
Plain topics	0.96	0.96*	0.91**	0.90***	0.94**	0.99	0.97
Economic lexicon-adj. topics	0.93	0.94	0.87**	0.86**	0.93	1.01	0.98
BCS-adj. topics	0.99	0.99	0.86*	0.85**	0.95	1.01	0.98
General lexicon-adj. topics	0.99	0.99	0.89*	0.89*	0.93*	1.01	0.98
Uncertainty-adj. topics	1.00	1.00	0.99	0.99	0.99*	1.00	1.00
Panel C: 2011–2018 (Post-crisis, incl. European Debt Crisis)							
Plain topics	0.91**	0.90**	0.88***	0.88**	0.88**	0.93	0.92
Economic lexicon-adj. topics	0.92	0.92	0.90*	0.89*	0.86*	0.91	0.89
BCS-adj. topics	0.86***	0.87***	0.85***	0.85***	0.88*	0.95	0.93
General lexicon-adj. topics	0.93	0.93	0.91*	0.92	0.94	1.01	0.99
Uncertainty-adj. topics	0.96	0.95*	0.97	0.98	1.02	1.05	1.03

Notes: The table reports root mean squared forecast errors (RMSFEs) of a Dynamic Factor Model (DFM) that augments a benchmark specification with one text factor. The benchmark DFM contains two factors extracted from hard and survey data. In the competing models, the additional factor is constructed from plain topics, economic lexicon-adjusted topics, BCS-adjusted topics, general lexicon-adjusted topics, or topics adjusted with the share of uncertainty terms. All entries are expressed as RMSFEs relative to the benchmark model. Values in bold denote improvements of at least 5% compared with the benchmark. Asterisks indicate statistical significance of the improvement according to a one-sided Diebold–Mariano (DM) test (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). Columns indicate the month and day within the quarter when the nowcast is made. Specifically, M1-01 is the first day of the first month, M1-16 is day 16 of the first month and so on. M4-01 is the first day of the fourth month, i.e., the first day of the subsequent quarter.

sentiment-adjusted versions perform slightly better in the second month of the quarter.

During the calmer economic period (2011–2018, right panel of Figure 2.8), the results closely mirror those obtained for GDP. The clear leader throughout the quarter is the BCS-adjusted specification, which yields improvements relative to the benchmark for all nowcast days, with statistically significant gains up to the middle of the third month. The plain topics specification remains competitive but is consistently outperformed by the BCS-adjusted model. Notably, none of the other sentiment-adjusted specifications performs as well as the BCS-adjusted version. This pattern again highlights that the value of adjustment varies across economic regimes: applying an appropriate sentiment adjustment becomes particularly informative when economic conditions are less turbulent.

Why is this the case? Correlations between plain and BCS-adjusted topics and invest-

Table 2.12: Correlations of plain topics with quarterly investment growth (full sample and excluding the Financial Crisis, 2008–2009)

ID	Label	Full sample	Without Financial Crisis
T150	Corporate Growth	0.537***	0.455***
T50	Crisis	-0.505***	-0.326***
T134	Steel Industry Restructuring and Downsizing	-0.370**	-0.241*
T29	Banking	-0.356**	-0.057
T110	Technological Innovation	0.354***	0.369***
T59	Commodity Markets	0.349***	0.326***
T21	Policy Measures	-0.345**	-0.174
T23	Insolvency and Financial Rescue	-0.313**	-0.122
T108	US Politics	-0.309**	-0.275**
T120	Economic Growth	-0.308**	-0.219*

Notes: The table reports correlations between plain topics and quarterly investment growth (first release). Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t -statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

ment growth provide additional insight. As shown in Tables 2.12, plain topics are highly and significantly correlated with investment growth in the full sample. However, once the Financial Crisis period is excluded, several of these correlations decline sharply in magnitude: five of the ten topics either lose statistical significance or remain only marginally significant.

By contrast, as seen in Table 2.13, the BCS-adjusted topics remain strongly and significantly correlated with investment growth both in the full sample and when the crisis years are removed. As with GDP, this pattern suggests that many plain topics behave primarily as crisis indicators, whereas BCS-adjusted topics capture meaningful cyclical variation across both turbulent and more stable economic regimes.

Nowcasts from all models and their comparison to actual investment growth are presented in Appendix 2.F for both periods (2008–2010 and 2011–2018).

2.4.3 Consumption

2.4.3.1 Descriptive analysis

Finally, I turn to the nowcasting results for consumption. As with GDP and investment, the first step is to examine which plain and sentiment- or uncertainty-adjusted topics are most closely associated with consumption growth.⁵ The key question is whether

⁵The consumption series is taken from the Bundesbank real-time database under the code BBKRT. Q. DE. Y. A. CA1. BA100. V. A, which corresponds to nominal private consumption, calendar- and seasonally adjusted.

Table 2.13: Correlations of BCS-adjusted topics with quarterly investment growth (full sample and excluding the Financial Crisis, 2008–2009)

ID	Label	Full sample	Without Financial Crisis
T27	Economic Crises and Recessions	0.608***	0.329***
T138	Financial and Economic Performance	0.576***	0.445***
T81	Corporate Restructuring and Job Cuts in Germany	0.539***	0.295***
T48	Business Success and Economic Resilience	0.479***	0.391***
T52	German Automobile Industry and Major Manufacturers	0.477***	0.252***
T168	Derivatives and Financial Instruments	0.476***	0.357***
T127	Major Banks and Investment Banking	0.475**	0.255***
T63	Commodity Markets and Precious Metals	0.471***	0.322***
T74	Concerns about Economic Bubbles and Recessions	0.461***	0.284***
T118	Corporate Governance and Executive Management	0.461***	0.292***

Notes: The table reports correlations between BCS-adjusted topics and quarterly investment growth (first release). BCS-adjusted topics combine topic proportions with business cycle sentiment. Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t -statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

the selected topics are economically interpretable and whether their correlations with consumption growth are sufficiently strong to justify their inclusion in the forecasting exercise.

As before, all selected topics are summarized in Table 2.14. A first striking difference compared with GDP and investment is that a much larger share of the topics most strongly related to consumption are political or policy-oriented. Examples include T45 (“Diplomatic Visits”), T78 (“CSU and Leadership Dynamics”), and T100 (“Protests and Demonstrations”). This pattern is particularly pronounced for the BCS-adjusted and uncertainty-adjusted specifications, which select four and six topics from the “Politics/Policies” category, respectively. Moreover, under BCS adjustment two additional topics relate directly to economic policy—T102 (“Taxation and Fiscal Policy”) and T191 (“Monetary Policy and Central Banking”)—which I group separately in Table 2.14. This stands in contrast to GDP and investment, where the most relevant topics predominantly capture crises, growth dynamics, or economic and financial developments rather than political or policy issues.

A second notable difference concerns the magnitude of the correlations. For both plain

Table 2.14: Summary of plain, sentiment-adjusted, and uncertainty-adjusted topics most strongly correlated with consumption growth

Thematic category	Topics (plain)	Economic lexicon-adjusted	BCS-adjusted	General lexicon-adjusted	Uncertainty-adjusted
Politics/Policies	T45 Diplomatic Visits	T45 Diplomatic Visits	T45 Political Dynamics of the FDP and Coalition Government	T13 CDU/CSU-FDP Coalition Politics	T1 Elections
	T78 CSU and Leadership Dynamics T82 Economic and Monetary Union	T78 CSU and Leadership Dynamics	T58 American Politics and Presidents T73 International Relations and Diplomacy T180 International Summits and Conferences	T45 Diplomatic Visits T78 CSU and Leadership Dynamics	T45 Diplomatic Visits T78 CSU and Leadership Dynamics T100 Protests and Demonstrations T131 Health Policy T194 Public Sector
Business Activity	T9 Small and Medium-Sized Enterprises	—	—	T9 Small and Medium-Sized Enterprises T14 Corporate Structure and M&A	—
Industries	T84 Insurance Industry	T134 Steel Industry Restructuring and Downsizing	—	T134 Steel Industry Restructuring and Downsizing	T46 German Maritime Sector
	T134 Steel Industry Restructuring and Downsizing T155 Automotive Industry	T155 Automotive Industry	—	—	T134 Steel Industry Restructuring and Downsizing
Economic Forecasting	T83 Research Institutes	T83 Research Institutes	—	T83 Research Institutes	—
More general topics	T139 Interviews and Opinions	T139 Interviews and Opinions	T54 Media and Journalism T59 Law Enforcement and Crime Prevention	—	T38 Problem Solving T117 Future Vision
Labor Market	T142 Employment Contracts and Severance Rights	T142 Employment Contracts and Severance Rights T178 Labor Market and Unemployment	T81 Corporate Restructuring and Job Cuts in Germany	T142 Employment Contracts and Severance Rights	—
Financial topics	—	T167 Corporate Financial Performance T183 Financial and Economic Performance	T147 Business Financing and Credit Solutions	T75 Accounting Standards, Reporting, and Risk Management T147 Stock Trading and Financial Markets	—
Monetary Policy	—	—	T191 Monetary Policy and Central Banking	—	—
Fiscal Policy	—	—	T102 Taxation and Fiscal Policy	—	—

Notes: The table summarizes the plain topics and their sentiment- or uncertainty-adjusted counterparts (economic lexicon-adjusted, general lexicon-adjusted, and uncertainty-adjusted topics), as well as the aspect-based topics adjusted with BCS sentiment, that exhibit the strongest correlations with consumption growth. Topics are grouped into broader thematic categories based on subjective interpretation to illustrate how different adjustments influence which topics are most closely associated with consumption dynamics.

and adjusted topics, the absolute values are markedly lower—and less frequently statistically significant—than in the cases of GDP and investment. Plain topics exhibit the strongest correlations across all approaches, with absolute values ranging from 0.218 to 0.410 (see Appendix 2.G). Among the sentiment-adjusted specifications, the general lexicon yields the highest correlations, ranging from 0.205 to 0.381. By contrast, the BCS-adjusted topics display comparatively weak associations with consumption growth, with correlations ranging from 0.172 to 0.298 (see Table 2.15).

Interestingly, when examining correlations between the BCS-adjusted topics that are most strongly associated with GDP and are economically well-interpretable, their relationship with consumption is weak. However, these same topics are relatively strongly correlated with the Business Cycle Expectations component of the GfK survey (see Table 2.31 in Appendix 2.G), which provides a useful validation of the adjustment approach: it was explicitly designed to emphasize sentiment toward business cycle conditions. Nevertheless, correlations with actual consumption growth remain low.

Among the BCS-adjusted topics selected specifically for consumption, only six out of ten coefficients are statistically significant (Table 2.15), and the correlations with GfK survey series are generally weak, with the strongest values observed for the question on business cycle expectations. Three topics (T38, T105, and T125) were excluded, as other topics were judged more economically relevant. Moreover, in contrast to the

Table 2.15: Correlations of selected BCS-adjusted topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T73	International Relations and Diplomacy	0.298	0.256***	0.128	0.066	0.150
T102	Taxation and Fiscal Policy	-0.268***	0.019	-0.040	0.060	-0.092
T58	American Politics and Presidents	0.232**	0.269***	0.114	0.014	0.111
T191	Monetary Policy and Central Banking	-0.227	0.345***	0.150	0.083	0.212*
T180	International Summits and Conferences	0.220*	0.216*	0.081	-0.040	0.107
T147	Business Financing and Credit Solutions	0.205	0.362***	0.092	-0.022	0.044
T81	Corporate Restructuring and Job Cuts in Germany	0.194*	0.302***	0.093	0.036	0.223**
T59	Law Enforcement and Crime Prevention	0.184*	0.307***	0.114	0.027	0.127
T54	Media and Journalism	0.175**	0.224**	0.058	0.020	0.179
T45	Political Dynamics of the FDP and Coalition Government	-0.172	-0.137	-0.123	-0.084	-0.160

Notes: The selected topics show the strongest correlations with consumption growth among all estimated aspect-based topics adjusted with BCS sentiment. "GfK BCE" denotes the GfK Business Cycle Expectations indicator; "GfK IE" denotes GfK Income Expectations; "GfK WtB" denotes GfK Willingness-to-Buy; "GfK CCI" denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly level for consistency with consumption growth data. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

GDP and investment cases, where sentiment-adjusted topics displayed uniformly positive correlations, three of the ten coefficients here are negative. For instance, more favorable sentiment regarding taxation and fiscal policy (T102) is associated with lower consumption growth. One possible explanation is that policy-related topics tend to be reported in a negative context; thus, a decline in sentiment may correspond to increased coverage of such issues, which in turn coincides with stronger consumption. More broadly, this pattern may suggest that for consumption, a sentiment measure oriented toward political or policy themes would be more appropriate.

Overall, the evidence for consumption is less compelling. Both plain and adjusted topics yield weaker correlations, many of the most relevant topics are concentrated on politics rather than economic activity, and some of the correlation signs raise interpretability concerns. At the same time, several findings indicate that text-based measures have potential for nowcasting consumption, provided that the adjustment is tailored to this variable.

As shown in Table 2.14, topics from categories that proved important for GDP and investment, such as business activity, industries, and financial topics, still appear among

the most relevant for consumption. In addition, new and economically meaningful patterns emerge. A larger number of topics related to the labor market are selected for consumption than for the other variables. For example, T142 (“Employment Contracts and Severance Rights”) appears under the plain, economic lexicon, and general lexicon specifications, while T178 (“Labor Market and Unemployment”) is significantly correlated with consumption under the economic lexicon adjustment. This is consistent with the idea that household consumption decisions are closely linked to labor market expectations. Moreover, since none of the existing sentiment measures target labor market conditions, I believe that a tailored labor market sentiment adjustment could strengthen these correlations.

Another noteworthy finding is that T83 (“Research Institutes”) is selected under the plain, economic-lexicon, and general-lexicon specifications. This topic groups together reporting from German research institutes on the state of the economy. It is plausible that households respond to such information. Under the general-lexicon adjustment, its correlation with consumption growth is the highest in magnitude (0.301), and under the economic lexicon adjustment it is also strongly correlated with the GfK Business Cycle Expectations component (see Appendix 2.G). Since these reports typically cover broad business cycle conditions as well as developments in specific industries and the labor market, concentrating on consumer-relevant dimensions may help strengthen the signal for forecasting consumption. Taken together, the initial results are promising but also indicate that off-the-shelf sentiment adjustments are not sufficient for consumption.

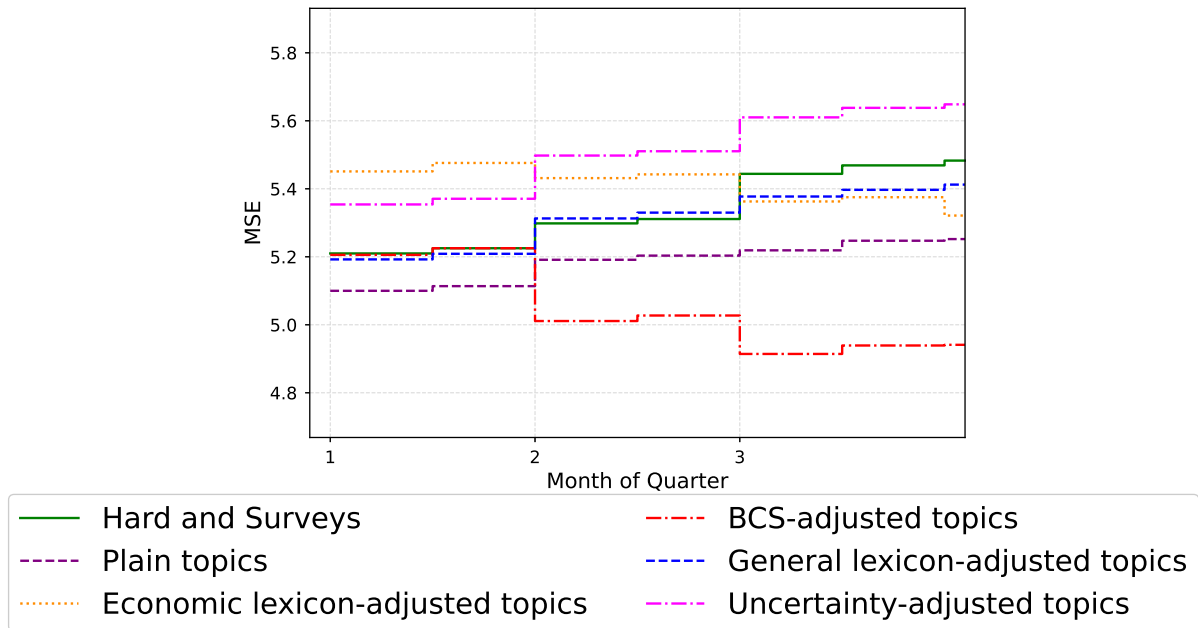
2.4.3.2 Results of the forecasting experiment

To summarize the forecasting results for consumption, I report mean squared forecast errors (MSFEs) for the benchmark model (green), which is estimated with one factor extracted from hard data and surveys (adding a second factor did not improve performance). These are compared with five alternative specifications that augment the benchmark with an additional text-based factor constructed from plain topics (purple) or topics adjusted with the economic lexicon (orange), BCS sentiment (red), the general lexicon (blue), or the share of uncertainty terms (pink). Results are shown for the full evaluation period (2008–2018) in Figure 2.9.

Across the full sample, two specifications yield improvements relative to the benchmark: plain topics and BCS-adjusted topics. The plain topics specification provides gains over the entire quarter, whereas the BCS-adjusted specification begins to outperform the benchmark only from the second month onward. As shown in Panel A of Table 2.16, however, these gains become statistically significant for both models only from the beginning of the third month of the quarter.

As before, nowcasting performance differs markedly across economic regimes. To ac-

Figure 2.9: Mean squared forecast errors (MSFEs) for different nowcasting models of consumption growth (2008–2018)



Notes: The figure reports mean squared forecast errors (MSFEs) for different Dynamic Factor Models (DFMs) evaluated over the period 2008–2018. The benchmark model includes one factor extracted from hard and survey data only (shown in green, solid line). Competing models augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The horizontal axis indicates the nowcast day, covering days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. Results are shown for the full evaluation sample (2008–2018).

count for this, I separately analyze the Financial Crisis period (2008–2010) and the subsequent period of more stable conditions (2011–2018). The left panel of Figure 2.10 shows that the full sample pattern is largely driven by the crisis years. Consistent with the findings for GDP and investment, plain topics and BCS-adjusted topics deliver the strongest improvements. While BCS-adjusted topics yield larger gains from the second month onward, Panel B of Table 2.16 shows that statistically significant reductions in root mean squared forecast errors (RMSFEs) over the entire quarter are obtained only for the plain topics specification. The BCS-adjusted model delivers statistically significant improvements only from the beginning of the third month. Overall, in line with the existing literature, topics appear able to improve consumption forecasts during the crisis period, although the magnitude of these gains is modest.

When turning to the calmer economic period (2011–2018, right panel of Figure 2.10), a different pattern emerges. In this case, the only specification yielding noticeable improvements is the one based on the economic-lexicon adjustment. This finding is consistent with the topic selection results, where the economic lexicon-adjusted topics appeared most

Table 2.16: Relative root mean squared forecast errors for Consumption growth nowcasts across subperiods

Text Factor	M1-01	M1-16	M2-01	M2-16	M3-01	M3-16	M4-01
Panel A: 2008–2018 (Full Sample)							
Plain topics	0.99	0.99	0.99	0.99	0.98**	0.98**	0.98**
Economic lexicon-adj. topics	1.02	1.02	1.01	1.01	0.99	0.99	0.99
BCS-adj. topics	1.00	1.00	0.97	0.97	0.95*	0.95*	0.95*
General lexicon-adj. topics	1.00	1.00	1.00	1.00	0.99	0.99	0.99
Uncertainty-adj. topics	1.01	1.01	1.02	1.02	1.02	1.02	1.01
Panel B: 2008–2010 (Financial Crisis)							
Plain topics	0.98*	0.98*	0.98*	0.98*	0.96**	0.96**	0.96**
Economic lexicon-adj. topics	1.08	1.08	1.06	1.06	0.99	0.99	0.98
BCS-adj. topics	0.99	0.99	0.93	0.93	0.90*	0.90*	0.90*
General lexicon-adj. topics	1.00	1.00	1.00	1.00	0.98	0.98	0.98
Uncertainty-adj. topics	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel C: 2011–2018 (Post-crisis, incl. European Debt Crisis)							
Plain topics	0.99	0.99	1.00	1.00	0.99	0.99	0.99
Economic lexicon-adj. topics	0.97	0.97	0.96	0.96	0.99	0.99	0.99
BCS-adj. topics	1.01	1.01	1.02	1.01	1.00	1.00	1.00
General lexicon-adj. topics	0.99	0.99	1.00	1.00	1.01	1.01	1.01
Uncertainty-adj. topics	1.03	1.03	1.03	1.03	1.03	1.03	1.03

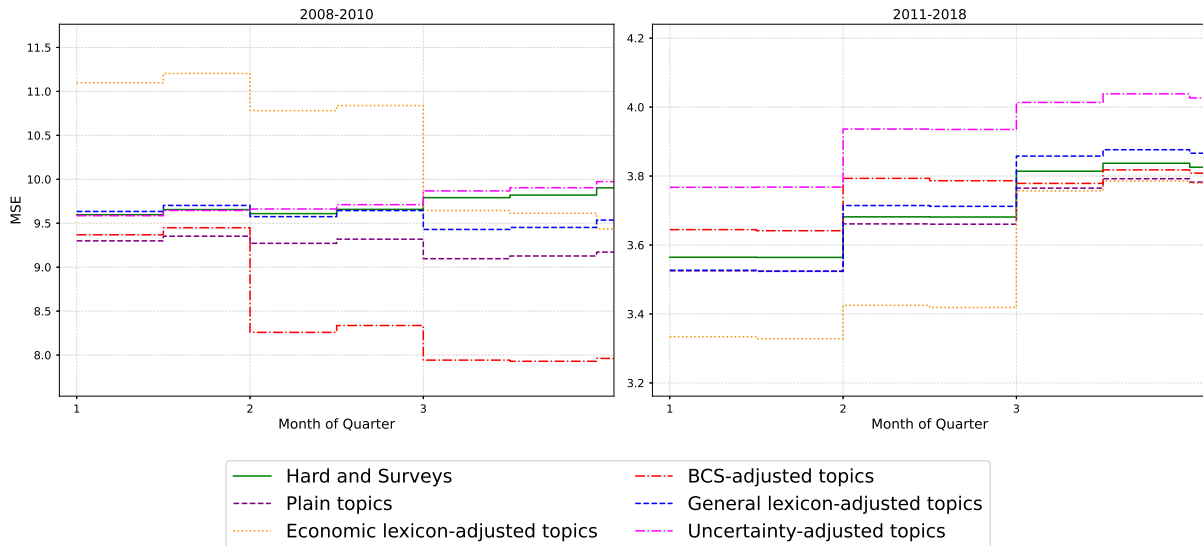
Notes: The table reports root mean squared forecast errors (RMSFEs) of a Dynamic Factor Model (DFM) that augments a benchmark specification with one text factor. The benchmark DFM contains one factor extracted from hard and survey data. In the competing models, the additional factor is constructed from plain topics, economic lexicon-adjusted topics, BCS-adjusted topics, general lexicon-adjusted topics, or topics adjusted with the share of uncertainty terms. All entries are expressed as RMSFEs relative to the benchmark model. Values in bold denote improvements of at least 5% compared with the benchmark. Asterisks indicate statistical significance of the improvement according to a one-sided Diebold–Mariano (DM) test (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). Columns indicate the month and day within the quarter when the nowcast is made. Specifically, M1-01 is the first day of the first month, M1-16 is day 16 of the first month and so on. M4-01 is the first day of the fourth month, i.e., the first day of the subsequent quarter.

economically meaningful. However, none of the differences relative to the benchmark is statistically significant (Panel C of Table 2.16).

Why are the results for consumption considerably weaker than those for GDP and investment? Tables 2.17 and 2.18 provide additional insight. In contrast to GDP and investment, plain topics exhibit stronger correlations with consumption growth than BCS-adjusted topics, both in the full sample and when the crisis years are excluded. Yet, once the Financial Crisis period is removed, the same pattern observed for the other variables reappears: for plain topics, half of the ten topics either lose statistical significance or remain only marginally significant.

This suggests that many plain topics again behave more like crisis indicators. However, unlike in the case of GDP and investment, the BCS adjustment does not restore strong

Figure 2.10: Mean squared forecast errors (MSFEs) for different nowcasting models of consumption growth, 2008–2010 and 2011–2018



Notes: The figure reports mean squared forecast errors (MSFEs) of different Dynamic Factor Models (DFMs) on the vertical axis against the day of the nowcast on the horizontal axis. The benchmark DFM includes one factor extracted from hard and survey data only (green, solid line). Competing DFMs augment the benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); or (v) BCS-adjusted topics (red, dash-dotted line). The nowcast days include days 1 and 16 of months 1, 2, and 3 of each quarter, as well as day 1 of the subsequent quarter. The left panel reports results for the Financial Crisis period (2008–2010), and the right panel reports results for the period of calmer economic conditions (2011–2018), which nevertheless includes the European Debt Crisis.

and stable correlations outside the crisis period. Taken together with the topic selection results, this indicates that the business cycle sentiment adjustment used in this paper is not well suited for consumption. A sentiment measure more closely tied to household considerations—such as political sentiment, labor market sentiment, or sentiment about inflation—may be more appropriate and could offer greater value in future work.

2.5 Conclusion

This paper has examined whether text-based information extracted from a large corpus of German news articles can improve nowcasts of GDP, investment, and consumption. More specifically, it addressed whether forecasting performance depends on the particular type of text-based series employed. To this end, I considered both plain topics obtained from an unsupervised LDA model and topics adjusted with sentiment or uncertainty measures. In particular, I compared adjustments based on an economic sentiment lexicon, a general sentiment lexicon, the share of uncertainty terms, and a novel Business Cycle Sentiment

Table 2.17: Correlations of plain topics with quarterly consumption growth (full sample and excluding the Financial Crisis, 2008–2009)

ID	Label	Full sample	Without Financial Crisis
T45	Diplomatic Visits	0.410***	0.394***
T78	CSU and Leadership Dynamics	-0.353***	-0.304***
T139	Interviews and Opinions	-0.282**	-0.210**
T83	Research Institutes	-0.278**	-0.264**
T9	Small and Medium-Sized Enterprises	-0.253**	-0.208*
T134	Steel Industry Restructuring and Downsizing	-0.238**	-0.167
T84	Insurance Industry	-0.229**	-0.229**
T82	Economic and Monetary Union	0.223*	0.177
T155	Automotive Industry	-0.221**	-0.121
T142	Employment Contracts and Severance Rights	-0.218	-0.112

Notes: The table reports correlations between plain topics and quarterly consumption growth (first release). Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t -statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

(BCS) measure developed in Okuneva et al. (2024), which was extracted using a machine-learning approach and a large training set. Forecasting performance was evaluated in a real-time setting using a monthly dynamic factor model estimated with an expanding-window approach, with a specification based solely on hard data and surveys serving as the benchmark.

Two main findings emerge. First, the correlation analysis confirms that text-based indicators are informative, but the nature and strength of these relationships vary across macroeconomic variables. For GDP and investment, correlations are strong, economically interpretable, and dominated by topics capturing crises, growth dynamics, financial conditions, and developments in key industries and the general economy. These relationships are highly robust: sentiment adjustments—especially those based on the economic lexicon and the BCS—tend to preserve or even strengthen them.

By contrast, for consumption the correlations are considerably weaker and more frequently centred on political or policy themes. Although some economically meaningful patterns emerge—particularly topics related to industries, business activity, and the labour market—the overall evidence suggests that the existing sentiment adjustments are not well aligned with consumption dynamics. This points to the need for sentiment measures more closely tailored to household-relevant dimensions, such as labour market expectations, economic policies, or inflation-related sentiment.

Second, forecasting performance exhibits a clear and robust regime dependence across

Table 2.18: Correlations of BCS-adjusted topics with quarterly consumption growth (full sample and excluding the Financial Crisis, 2008–2009)

ID	Label	Full sample	Without Financial Crisis
T73	International Relations and Diplomacy	0.298	0.299
T102	Taxation and Fiscal Policy	-0.268***	-0.261***
T58	American Politics and Presidents	0.232**	0.072
T191	Monetary Policy and Central Banking	-0.227	-0.306**
T180	International Summits and Conferences	0.220*	0.233*
T147	Business Financing and Credit Solutions	0.205	-0.150
T81	Corporate Restructuring and Job Cuts in Germany	0.194*	-0.088
T59	Law Enforcement and Crime Prevention	0.184*	0.052
T54	Media and Journalism	0.175**	0.119
T45	Political Dynamics of the FDP and Coalition Government	-0.172	-0.190

Notes: The table reports correlations between BCS-adjusted topics and quarterly consumption growth (first release). BCS-adjusted topics combine topic proportions with business cycle sentiment. Full sample refers to the 1991–2018 period. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t -statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

all variables. During the Financial Crisis, models augmented with plain topics improve forecasts most consistently. While sentiment-adjusted series also yield gains relative to the benchmark, models using plain topics tend to perform better earlier within the quarter. As the GDP example illustrates, this is because in the first recovery quarter after the Financial Crisis topic prevalence adjusts more quickly than sentiment.

In calmer periods, by contrast, BCS-adjusted topics emerge as the clear leader for both GDP and investment. This pattern is fully consistent with the correlation results: BCS-adjusted topics incorporate cyclical direction and exhibit stable, economically meaningful correlations across regimes, whereas many plain topics function largely as crisis indicators whose informativeness diminishes once extreme conditions subside. For consumption, however, no specification yields consistent improvements in the calmer period, again suggesting that this variable likely requires a sentiment adjustment tailored to household-relevant dimensions.

Overall, the findings demonstrate the potential of text data to complement traditional hard and survey indicators in real-time forecasting. At the same time, the results emphasize that the value of combining topics with sentiment depends on the economic envi-

ronment, and that choosing an appropriate sentiment adjustment is crucial, as it shapes both the interpretability and the predictive power of text-based measures.

Future research could build on this approach by developing alternative sentiment adjustments tailored to specific variables (for instance, consumption) and by testing text-based series in daily or non-linear forecasting models, which may deliver further improvements, particularly during turbulent economic times, beyond those obtained with the monthly dynamic factor model considered here. Applications to other key macroeconomic variables, such as inflation and unemployment, also represent promising directions.

Appendix 2

2.A Estimated topics

Table 2.19: Labels for the 200 estimated topics based on the most probable words

ID	Label	Most probable words
T0	Uncertainty and Outlook	darub (about it), o (unclear), hinaus (beyond), klar (clear), stellt (clarifies/questions), fest (fixed), steht (certain), stell (clarify/question), bleib (remain), zunach (initially), bleibt (remains), zukunft (future), unklar (unclear), gestellt (completed), bislang (so far), wann (when), genau (exactly), vollig (completely), liess (let), konkret (concrete), moglicherweise (possibly), aussicht (prospect)
T1	Elections	wahl (election), partei (party), stimm (vote), kandidat (candidate), sieg (victory), ergebnis (result), wahlkampf (campaign), sitz (seat), gewahlt (elected), parlamentswahl (parliamentary election), prozent_der_stimmen (percentage of votes), mehrheit (majority), kommunalwahl (local election), knapp (narrowly), umfrag (poll), demokrat (democrat), mandat (mandate)
T4	Dutch Business	niederland (Netherlands), amsterdam (Amsterdam), den_haag (The Hague), grossbritanni (Great Britain), holland (Holland), frankreich (France), belgi (Belgium), de (of), ing (ING), itali (Italy), hfl (Dutch guilder), sech (six), ausserd (besides), meint (says), gemacht (made), unilev (Unilever), bekannt (known), abn_amro (ABN AMRO), gehort (belongs)
T5	Pricing and Cost Trends	preis (price), kost (cost), pro (per), teuer (expensive), billig (cheap), erhoh (increase), niedrig (low), je (per), gebuhr (fee), gunstig (aordable), cent, verbrauch (consumption), durchschnitt (average), jahrlich (annually), steig (rise), zusatz (additional), kostet (costs), senk (reduce), pfennig (penny), erhoh (raised), preiserhoh (price hike), bezahl (pay), lieg (lie), teu (costly), gering (small), erheb (significantly), hoch (high)
T7	Mergers and Acquisitions	ubernahm (takeover), verkauf (sale), fusion (merger), konz (group), anteil (share), zusammenschluss (merger), partn (partner), beteilig (stake), ubernehm (takeover), konzern (corporation), milliarden_euro (billion euros), konkurrent (competitor), finanzinvestor (financial investor), investor, tocht (subsidiary)
T8	Organizations and Awards	mitglied (member), verein (association), stiftung (foundation), organisation (organization), vertret (representation), arbeit (work), gegründet (founded), list, nam (name), erhalt (receive), vorsitz (chairmanship), preis (prize), initiativ (initiative), gehort (belongs), grundung (foundation), grupp (group), person, engagement (commitment), aktiv (active), sitz (seat), ausgezeichnet (awarded)
T9	Small and Medium-Sized Enterprises	firm (company), mittelstand (SMEs), industri (industry), branch (industry), betrieb (firm), arbeitsplatz (workplace), dienstleist (service), ausland (abroad), investition (investment), besond (especially), bereich (area), meist (mostly), klein (small), beschafitigt (employed), gerad (especially), mittelstandl (medium-sized), standort (location), zunehm (increasing), deutsche_unternehmen (German companies)
T13	CDU/CSU-FDP Coalition Politics	cdu (CDU), fdp (FDP), union (Union), csu (CSU), vorsitz (leadership), koalition (coalition), schaubl (Schäuble), liberal, partei (party), fraktion (parliamentary group), mollemann (Möller), fraktionschef (faction leader), bundestag (Bundestag), generalsekretar (secretary general), bundestagsfraktion (Bundestag faction), westerwell (Westerwelle), wolfgang_schauble (Wolfgang Schäuble), merz (Merz), angela_merkel (Angela Merkel)
T14	Corporate Structure and M&A	gmbh (LLC), ag (public limited company), grupp (group), mitarbeit (employee), geschäftsfuhr (management), gesellschaft (company), beschafitigt (employed), co (Co.), tocht (subsidiary), umsatz (sales), gegründet (founded), firma (firm), ubernomm (acquired), werk (plant), holding
T16	Political Dynamics in Algeria	islam (Islam), demokrati (democracy), muslim (Muslim), demokrat (democrat), algeri (Algeria), militar (military), anhang (followers), religios (religious), algi (Algiers), islamist (Islamist), grupp (group), general, regim (regime), fuhr (led), radikal (radical), opposition, gewalt (violence), alger (Algerian)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T18	Agreements and Cooperation	vertrag (contract), gemeinsam (together), zusammenarbeit (cooperation), abkomm (agreement), vereinbar (agreement), kooperation (cooperation), unterzeichnet (signed), partn (partner), vereinbart (agreed), unterzeichn (to sign), seit (since), eng (close), partnerschaft (partnership), geschloss (concluded), erklar (declare), entsprech (corresponding), gegenseit (mutual), abschluss (conclusion), bezieh (relationship)
T19	Family and Bereavement	alt (old), famili (family), jahrig (year old), sohn (son), tod (death), vat (father), leb (life), brud (brother), jung (young), haus (house), mann (man), gestorb (deceased), damal (then), letzt (last), mutt (mother), kam (came), freund (friend), frau (woman), erb (inheritance)
T21	Policy Measures	massnahm (measure), programm (program), forder (funding), ziel (goal), notwend (necessary), mittel (funds), gefordert (demanded), unterstütz (support), sollt (should), ford (promote), zusatz (additional), beitrage (contribution), hilf (help), verbesser (improvement), rahm (framework), verstärkt (intensified), ausserd (besides), erford (required), initiativ (initiative), finanziell (financial), konkret (concrete), verbess (improve), insbesond (especially), scha (create), betont (emphasized)
T23	Insolvency and Financial Rescue	schuld (debt), pleit (default), insolvenz (insolvency), kredit (credit), finanziell (financial), glaubig (creditor), verlust (loss), bank, angeschlag (troubled), sanier (restructure), konkur (bankruptcy), millionen_euro (millions of euros), forder (claim), gestellt (filed), investor, rettung (rescue), geld (funds), erhalt (preservation), insolvenzverwalt (insolvency management)
T28	Foreign Exchange and Currency Markets	dollar, euro, yen, wahrung (currency), kostet (costs), gegenub (against), schwach (weak), kur (rate), fest (firm), handl (trade), notiert (quoted), mark, vortag (previous day), devis (foreign exchange), devisaenmarkt (foreign exchange market), leicht (slight), frankfurt (Frankfurt), referenzkur (reference rate)
T29	Banking	bank, frankfurt (Frankfurt), deutsche_bank (Deutsche Bank), commerzbank (Commerzbank), institut (institution), kund (customer), dresdner_bank (Dresdner Bank), deutschen_bank (Deutsche Bank), grossbank (major bank), geschaft (business), hypo (HypoVereinsbank), vereinsbank (HypoVereinsbank), investmentbank (investment bank), kreditinstitut (credit institution), banking, filial (branch), tocht (subsidiary), dresdn (Dresdner Bank)
T31	Governmental Reorganization	bonn (Bonn), bundesregier (federal government), kinkel (Klaus Kinkel, a former German Minister of Foreign Affairs), bundesrepubl (Federal Republic of Germany), minist (minister), bundeskanzler_helmut_kohl (Federal Chancellor Helmut Kohl), fdp (FDP, a political party in Germany), betont (emphasized), rexrodt (Günther Rexrodt, a German politician), bundestag (federal parliament), staatssekretar (State Secretary), regierungssprech (government spokesman), vorsitz (chairman), meint (says), gensch (Dietrich Genscher, a German politician), nannt (names), wort (word), eigener_bericht (own report), verwi (refer), wies (point out), kanzleramt (Chancellery)
T36	Global Development and Poverty	weltweit (worldwide), arm (poor), reich (rich), welt (world), entwickl (development), organisation, entwicklungsland (developing country), global, hilf (aid), armut (poverty), oecd (OECD), million, mensch (people), afrika (Africa), entwicklungshilf (development aid), aid, industrieland (industrialized country), kampf (struggle)
T37	Russian Politics	rusland (Russia), russisch (Russian), moskau (Moscow), jelzin (Boris Yeltsin), putin (Vladimir Putin), tschetscheni (Chechnya), kreml (Kremlin), russ (Russians), tschetschen (Chechen), georgi (Georgia), meldet (reports), tass (TASS), дума (Duma), grosny (Grozny), tschernomyrdin (Viktor Chernomyrdin), weissrusland (Belarus), präsidnt_boris_jelzin (President Boris Yeltsin), kaukasus (Caucasus), kommunist (communist), yukos (Yukos)
T38	Problem Solving	problem (problematic), schwierig (difficult), losung (solution), probl (problem), los (solve), schnell (quick), such (search), find, situation, gelost (solved), aufgab (task), angesicht (in light of), deshalb (therefore), notwend (necessary), allein (alone), gefund (found), notig (necessary), bring, dringend (urgent), rasch (quick)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T39	Scandals, Mistakes, and Accountability	fehl (mistake), verantwort (responsibility), o ent (public), falsch (wrong), gemacht (made), vertrau (trust), konsequenz (consequence), schwer (severe), schad (damage), zieh (draw), verhalt (behavior), gezogen (drawn), skandal (scandal), gerat (come under pressure), raumt (admits), erklar (explain), geword (become), o enbar (apparently), kritik (criticism), eindruck (impression), anseh (reputation)
T45	Diplomatic Visits	besuch (visit), tre (meeting), reis (travel), bezieh (relationship), aussenminist (foreign minister), besucht (visited), empfang (reception), ministerprasident (prime minister), abend (evening), einlad (invitation), zusamm (together), traf (met), anschliess (then), reist (travels), samstag (Saturday), o ziell (o cial), eingetro (arrived), delegation, verhaltnis (relation), gast (guest), tri t (meets), unterstutz (support), stand (stance), vertret (representation)
T46	German Maritime Sector	hamburg (Hamburg), schi (ship), schleswig (Schleswig), holstein (Holstein), kiel (Kiel), boot (boat), haf (harbor), werft (shipyard), kust (coast), lubeck (Lübeck), nord (north), bord (board), hansestadt (Hanseatic city), marin (navy), see (sea), meer (sea)
T48	UN Security Council	un (UN), uno (UNO), sicherheitsrat (Security Council), resolution, vereinten_nationen (United Nations), generalsekretar (Secretary-General), diplomat, annan (Kofi Annan), botschaft (embassy), welticherheitsrat (UN Security Council), mitglied (member), sanktion (sanction)
T50	Crisis	kris (crisis), tief (deep), schlecht (bad), folg (consequence), verlust (loss), schwer (severe), verlor (lost), dramt (dramatic), angesicht (in light of), massiv (massive), problem, situation, schwach (weak), trotz (despite), druck (pressure), gerat (come under pressure), auswirk (e ect), drastisch (drastic), schwierig (di cult), unsich (uncertain), hart (hard)
T51	French Politics	franzos (French), paris (Paris), frankreich (France), chirac (Jacques Chirac), sozialist (Socialist), premierminist (Prime Minister), sarkozy (Nicolas Sarkozy), le_monde (Le Monde), national (National Front), jospin (Lionel Jospin), mitterrand (François Mitterrand), le_figaro (Le Figaro), jacques_chirac (Jacques Chirac), alstom (Alstom), link (left), franc (Franc), liberation (Libération), nationalversammlung (National Assembly)
T53	Personal Background and Career	jahrig (year old), damal (then), mann (man), karri (career), alt (old), gilt (is considered), gebor (born), arbeitet (works), kam (came), gehort (belongs), spitz (top), anfang (beginning), begann (began), erfahr (experience), gern (likes), studiert (studied), nie (never), job, stet (always), gelernt (learned), erfolgreich (successful)
T56	Performance and Continental AG	ziel (goal), erreicht (achieved), erfolg (success), erreich (achieve), ergebnis (result), zeigt (shows), zufrieden (satisfied), trotz (despite), erzielt (achieved), erfolgreich (successful), betont (emphasized), schwierig (di cult), hervor (excellent), zuversicht (confidence), erstmal (for the first time), fest (firm), positiv (positive), reif (tyres), gelungen (successful), erwart (expected), voll (full), hob (raised), nannt (named), dennoch (nevertheless), continental (Continental AG), gesetzt (set), conti (Continental AG)
T59	Commodity Markets	gold, preis (price), nachfrag (demand), je (per), rohsto (commodity), tonn (ton), weltweit (worldwide), london (London), metall (metal), aluminium (aluminum), erreicht (reached), hoch (high), steigend (rising), technisch (technical), bestand (stock), kupf (copper), angebot (o er), erwart (expect)
T61	Libyan Politics	hiess (was stated), teilt (reports), bekannt (known), angekündigt (announced), gegeb (given), zunach (initially), zuvor (before), bestatigt (confirmed), sprecherin (spokeswoman), woll (wants), mitteil (report), kundigt (announces), liby (Libya), nachd (after), erneut (again), schritt (step), einzel (detail), mitgeteilt (reported), begrundet (justified)
T62	Corporate Restructuring	bereich (area), einheit (unit), aktivitat (activity), struktur (structure), ziel (goal), zentral (central), gemeinsam (together), eigenstand (independent), bisher (so far), dusseldorf (Düsseldorf), bleib (remain), viag (VIAG AG), dach (umbrella), veba (VEBA AG), zukunft (future), neb (in addition), regional, einzeln (individual), gehor (belongs), konzentri (concentrate), gefuhrt (led), degussa (Degussa AG), trennung (separation)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T68	International Financial Aid	iwf (IMF), prag (Prague), tschechisch (Czech), weltbank (World Bank), kredit (loan), tschechi (Czech Republic), finanzminist (Minister of Finance), milliarden_dollar (billion dollars), internationalen_waehrungsfonds (International Monetary Fund), washington (Washington), bulgari (Bulgaria), bulgar (Bulgarian), hilf (aid), finanzkris (financial crisis), havel (Václav Havel), kris (crisis), direktor (managing director), sudetendeutsch (Sudeten German), westlich (western), zentralbank (central bank), waehrungsfond (monetary fund)
T72	Corporate Financial Performance	mill (million), dm (DM, the German mark), ag (corporation), mrd (billion), umsatz (revenue), ergebnis (result), erhoht (increased), geschäftsjahr (financial year), vorjahr (previous year), verlust (loss), divid (dividend), stieg (rose), grupp (group), konz (group), jahresuberschuss (annual surplus), je (per), ertrag (revenue), verbessert (improved), gewinn (profit), vorstand (board), bereich (division), erzielt (achieved), erreicht (reached)
T74	Opinion Commentary	wer (who), darf (may), mag (like), burg (citizen), endlich (finally), recht (right), wirklich (really), lasst (lets), anzeig (advertiser), wohl (probably), bleibt (remains), ja (yes), gar (at all), lang (long), gewiss (certainly), gest (yesterday), eben (just), eigent (actual), statt (instead), niemand (no one), gerade (just), vielleicht (perhaps), tun (do), wund (wonder), vernunft (reason), braucht (needs), kaum (hardly)
T75	Accounting Standards, Reporting, and Risk Management	einzeln (individual), transparenz (transparency), unterschied (difference), standard, risik (risk), entsprech (corresponding), instrument, bewert (evaluation), insbesond (especially), bestimmt (certain), wesent (essential), grundsatz (principle), information, verschied (different), unabhing (independent), jeweil (respective), regel (rule), rahm (framework)
T78	CSU and Leadership Dynamics	csu (CSU, a German political party), munch (Munich), stoib (Edmund Stoiber, a German politician), bayer (Bavarian), bay (Bavaria), munchn (Munich), bayern (Bavaria's), waigel (Theo Waigel, a German politician), edmund_stoiber (Edmund Stoiber, a German politician), glos (Michael Glos, a German politician), vorsitz (chairman), seehof (Horst Seehofer, a German politician), freistaat (Free State), ministerpräsident (prime minister), beckstein (Günther Beckstein, a German politician)
T79	Afghanistan Conflict	afghanistan (Afghanistan), taliban (Taliban), kampf (fight), afghan (Afghan), kabul (Kabul), isaf (International Security Assistance Force), pakistan (Pakistan), provinz (province), anti, islam (Islam), gefuhrt (led), unterstutz (supported), amerikan (American), sud (south), schutztrupp (protective force), radikal (radical), hilf (aid), tadschikistan (Tajikistan), fuhr (led), terror
T80	India-Pakistan Relations	indi (India), pakistan (Pakistan), indisch (Indian), kaschmir (Kashmir), neu_delhi (New Delhi), ind (Indian), islamabad (Islamabad), musharraf (Musharraf), region, grenz (border), kampf (conflict), kam (came), delhi (Delhi), gefuhrt (led), konflikt (conflict), bundesstaat (federal state), vajpaye (Vajpayee), bislang (so far), unabhing (independent), krieg (war), unterstutz (support), nepal (Nepal), bangladesch (Bangladesh)
T82	Economic and Monetary Union	euro (euro), waehrungsunion (monetary union), waehrung (currency), finanzminist (finance minister), defizit (deficit), maastricht (Maastricht), stabilitatspakt (Stability Pact), kriteri (criteria), einfuhr (introduction), stabilitat (stability), national, vertrag (treaty), erfull (fulfilled), munz (coin), frankreich (France), start, eu (EU), januar (January), einhalt (compliance)
T83	Research Institutes	institut (institute), ifo, diw, erwart (expectation), wachstum (growth), investition (investment), gering (low), beschafft (employment), niedrig (low), real, okonom (economist), arbeitsmarkt (labor market), durfte (may), einschatz (assessment), wettbewerb (competitive), prognos (forecast), steig (increase), deutschen_wirtschaft (German economy)
T84	Insurance Industry	allianz (Allianz), versicher (insurance), schweiz (Switzerland), versich (insured), zurich (Zurich), sfr (CHF), kund (customer), gesellschaft (company), lebensversicher (life insurance), grupp (group), munch (Munich), münchener_rück (Munich Re), schad (damage), versichert (insured), bern (Bern), leb (life), frank (Franc), schweizer (Swiss), aach (Aachen), gerling (Gerling Group), prami (premium), ruck (Hannover Re), geschäft (business)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T87	Environment and Nuclear Energy	transport, umweltschutz (environmental protection), umwelt (environment), atomkraftwerk (nuclear power plant), klimaschutz (climate protection), anlag (facility), kernkraftwerk (nuclear power station), greenpeac (Greenpeace), reaktor (reactor), umweltminist (environment minister), okolog (ecologist), ausstieg (phase-out), gorleb (Gorleben), castor (Castor transport), kyoto (Kyoto Protocol)
T89	Independence and Post-Soviet Transition	sowjet (Soviet), republ (republic), sowjetunion (Soviet Union), kanad (Canadian), litau (Lithuanian), kanada (Canada), unabhng (independent), gus (CIS), gorbatschow (Gorbachev), udssr (USSR), ehemal (former), moskau (Moscow), kommunist (communist), gemeinschaft (community), estland (Estonia), lettland (Latvia), vietnam (Vietnam), westlich (Western), gebiet (territory), hilf (aid), west (West)
T91	Media Coverage of Plans and Rumors	bericht (report), berichtet (reports), zufolge (according to), bestatigt (confirmed), hiess (was said), zeitung (newspaper), information, wonach (according to which), spekulat (speculation), angeb (allegedly), kreis (circle), woll (want to), laut (according to), dementiert (denied), o enbar (apparently), gerucht (rumor), sprecherin (spokeswoman), o ziell (o cial), bestat (confirmed)
T93	Conferences and Summits	tre (meeting), berat (consultation), konferenz (conference), gipfel (summit), vertr (representation), teilnehm (participation), gemeinsam (together), thema (theme), minist (minister), regierungschef (head of government), sitzung (session), beginn (beginning), them (topic), gipfeltre (summit meeting)
T94	Middle East Diplomacy	gypt (Egypt), arab (Arab), israel (Israel), syri (Syria), libanon (Lebanon), kairo (Cairo), libanes (Lebanese), hisbollah (Hezbollah), saudi (Saudi), arabi (Arabia), syrisch (Syrian), nahost (Middle East), beirut (Beirut), jordani (Jordanian), nahen_osten (Middle East), ausseminist (foreign minister), region, damaskus (Damascus), fried (peace), mubarak (Husni Mubarak), jordan (Jordan), saudisch (Saudi), assad (Hafid al-Assad), sudlibanon (South Lebanon), friedensprozess (peace process)
T98	Public Appearances and Reactions	liess (let), gegn (against), kaum (hardly), wohl (probably), o ent (publicly), kampf (fight), alt (old), versucht (tried), ton (tone), scheint (seems), eher (rather), schliesslich (after all), machtig (powerful), mann (man), kritik (criticism), hart (harsh), o (openly), sogar (even), auftritt (appearance)
T99	Israeli-Palestinian Conflict	israel (Israel), palastinens (Palestinian), jerusal (Jerusalem), arafat (Yasser Arafat), hamas (Hamas), gaza (Gaza), nahost (Middle East), israelis (Israelis), westjordanland (West Bank), scharon (Ariel Scharon), gazastreif (Gaza Strip), plo (Palestine Liberation Organization), tel_aviv (Tel Aviv), friedensprozess (peace process), judisch (Jewish), netanjahu (Benjamin Netanyahu), siedlung (settlement), gebiet (territory), barak (Ehud Barak), ramallah (Ramallah), abbas (Mahmud Abbas), arab (Arab), fried (peace), rabin (Yitzhak Rabin), streif (Strip), nahen_osten (Middle East)
T100	Protests and Demonstrations	prot (protest), polizei (police), demonstrant (demonstrator), demonstration, mensch (people), aktion (action), tausend (thousand), demonstriert (demonstrated), samstag (Saturday), strass (street), kundgeb (rally), stadt (city), protestiert (protested), hundert (hundred), anhang (supporters), teilnehm (participants), polizist (police o cer)
T102	Trade and WTO	handel (trade), wto (World Trade Organization), verhandl (negotiation), export, eu (EU), zoll (tari), genf (Geneva), amerikan (American), import, gatt (GATT), o nung (opening), einfuhr (import), abkomm (agreement), liberalisier (liberalization), welthandelsorganisation (WTO), abbau (reduction), ausfuhr (export), freien (free), produkt (product), subvention (subsidy)
T104	Comparison, Contrast, and Analysis	andererseits (on the other hand), unterschied (di erence), bedeut (meaning), einerseits (on the one hand), wesent (essential), gegenub (compared to), besond (especially), gering (slight), entwickl (development), erheb (significant), insbesond (in particular), relativ (relative), tatsach (fact), betracht (consider), einzeln (individual), sowohl (both), daraus (from that), verhaltnis (relation), stellt (presents), folg (consequence), eher (rather), zunehmend (increasing), verander (change)
T106	Postal Services and Higher Education	post, student, hochschul (higher education), universitat (university), brief (letter), bonn (Bonn), bafog (BAföG - student financial aid), postbank (Postbank), studium (studies), wettbewerb (competition), professor, paket (package), studiengebuh (tuition fee), deutsche_post (Deutsche Post)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T108	US Politics	bush (George W. Bush), washington (Washington), amerikan (American), clinton (Bill Clinton), republikan (republican), demokrat (democrat), kongress (Congress), senat (Senate), präsident_bill_clinton (President Bill Clinton), senator (Senator), amerika (America), bundesstaat (federal state), weißen_haus (White House), vereinigten_staaten (United States)
T109	Financial Performance	millionen_dm (millions of DM), milliarden_dm (billions of DM), umsatz (sales), ag (public company), vorjahr (previous year), konz (group), ergebnis (result), stieg (rose), million, geschäftsjahr (fiscal year), munch (Munich), gewinn (profit), erhöht (increased), erzielt (achieved), plus, eigener_bericht (own report), vorstandsvorsitz (board chair)
T110	Technological Innovation	entwickl (develop), produkt (product), technisch (technical), technik (technology), kund (customer), entwickelt (developed), technologi (technology), bereich (area), dienstleist (service), how, modern, know, beispiel (example), zukunft (future), schnell (fast), innovation, system, qualitat (quality)
T111	Education and Youth	schul (school), lehr (teach), jugend (youth), bildung (education), kind (child), elt (parent), bess (better), lern (learn), unterricht (class), jung (young), ausbild (training), sprach (language), beruf (profession), sozial (social), pisa (PISA), alt (old), kultusminist (education minister), studi (study)
T112	Trade Fairs	mess (fair), branch (industry), ausstell (exhibitor), verband (association), frankfurt (Frankfurt), veranstalt (event), dusseldorf (Düsseldorf), besuch (visitor), ausland (abroad), herstell (manufacturer), industri (industry), koln (Cologne), maschinenbau (mechanical engineering), maschin (machine), geschäftsfuhr (managing director), trend, cebit (CeBIT, a former IT trade fair)
T115	Flight Schedules and Incidents	uhr (o'clock), stund (hour), abend (evening), morg (morning), samstag (Saturday), zunach (initially), uberblick (overview), nacht (night), maschin (engine), minut (minute), nachmittag (afternoon), flugzeug (aircraft), zuvor (before), sech (six), begann (began), berichtet (reported), zweit (second), letzt (last), gebracht (brought), spat (late), beginn (start), mittag (midday), vormittag (morning), kam (arrived), nah (near), wochen (weeks), gekommen (arrived)
T116	Japan's Economy	japan (Japan), tokio (Tokyo), yen, fuhrnd (leading), ldp (LDP), nikkei (Nikkei), mitsubishi (Mitsubishi), angesicht (in view of), marz (March), milliarden_yen (billions of yen), bill (billion), nippon (Nippon), april (April), hashimoto (Ryūtarō Hashimoto), billionen_yen (trillions of yen), asi (Asia), asiat (Asian), koizumi (Junichiro Koizumi)
T117	Future Vision	welt (world), zukunft (future), chanc (opportunity), gemeinsam (together), bess (better), erfolg (success), vision, strategi (strategy), seh (see), alt (old), europas (Europe's), erfolgreich (successful), scha (achieve), gerad (just), blick (view), herausforder (challenge)
T118	Transport Policy	fahr (driver), lkw (truck), verkehr (tra c), strass (road), kilomet (kilometer), maut (toll), autobahn (highway), transrapid (Transrapid, a German-developed high-speed train system), autofahr (motorist), auto (car), fahrzeug (vehicle), streck (route), lastwag (lorry), verkehrsmminist (transport minister), autos (cars), fahrt (journey), kfz (motor vehicle), einfuhr (import), pkw (passenger car), steu (tax), wissmann (Matthias Wissmann, a former Germany's Federal Minister of Transport), bundesverkehrsminist (Federal Minister of Transport)
T119	Kosovo Conflict	kosovo (Kosovo), belgrad (Belgrade), alban (Albanian), serbisch (Serbian), jugoslavi (Yugoslavia), serbi (Serbia), milosevic (Slobodan Milošević), jugoslaw (Yugoslavian), mazedoni (Macedonia), albani (Albania), serb (Serb), balkan (Balkan), osz (Organization for Security and Co-operation in Europe), pristina (Pristina), unabhng (independent), uck (Kosovo Liberation Army), montenegro (Montenegro), kfor (KFOR: Kosovo Force), sloweni (Slovenia)
T120	Economic Growth	wachstum (growth), konjunktur (business cycle conditions), prognos (forecast), aufschwung (upswing), schwach (weak), erwart (expectation), wirtschaftswachstum (economic growth), export, bip (GDP), erhol (recovery), bruttoinlandsprodukt (gross domestic product), anstieg (increase), rezession (recession)
T121	Literature and Arts	buch (book), autor (author), kultur (culture), kunst (art), geschicht (history), alt (old), musik (music), bild (picture), kunstl (artis), art, bekannt (well-known), star, verlag (publisher), theat (theater), film

Continued on next page

Table 2.19 continued from previous page

ID	Label	Most probable words
T122	Information Technology	microsoft (Microsoft), softwar (software), it (IT), sap (SAP), comput (computer), ibm (IBM), appl (Apple), tech (high-tech), system, windows (Windows), programm (program), pc (PC), high (high-tech), oracl (Oracle), internet (Internet), herstell (manufacturer), produkt (product), kund (customer), anbiet (provider), technologi (technology)
T123	Media and Newspapers	zeitung (newspaper), medi (media), hamburg (Hamburg), spiegel (Der Spiegel), interview, verlag (publisher), magazin (magazine), berichtet (reports), info, bild (Bild), blatt (paper), focus (Focus), vorab (in advance), bericht (report), samstag (Saturday), sperrfrist (embargo), tageszeit (daily newspaper), nachrichtenmagazin (news magazine)
T124	Elections and Office Succession	amt (office), nachfolg (succession), kandidat (candidate), jahrig (years-old), spitz (top), gilt (is considered), gewahlt (elected), post (position), bisher (so far), amtszeit (term), rucktritt (resignation), kandidatur (candidacy), stellvertret (deputy), bewerb (contest), tritt (enters), person, ehemal (former), offiziell (officially)
T128	Holocaust Remembrance	opfer (victim), jüdisch (Jewish), geschicht (history), histor (history), jud (Jew), jahrestag (anniversary), entschad (compensation), vergang (past), holocaust (Holocaust), damals (then), rede (speech), ns (National Socialism), ehemal (former), erinnert (reminded), nazi (Nazi), symbol
T131	Health Policy	arzt (doctor), kasse (insurance fund), krankenkasse (health insurance fund), patient, medizin (medicine), krankenhaus (hospital), versichert (insured person), gesundheitsreform (healthcare reform), klinik (clinic), gesundheitswesen (healthcare system), krank (sick), gesund (healthy), medikament (medication), apothek (pharmacy), behandlung (treatment)
T132	Korean Nuclear Tensions	nordkorea (North Korea), seoul (Seoul), sudkorea (South Korea), sudkorean (South Korean), nordkorean (North Korean), pjongjang (Pyongyang), korean (Korean), rakete (rocket), atomwaffe (nuclear weapon), korea (Korea), nordkoreas (North Korea's), nuklear (nuclear), test, sudkoreas (South Korea's), waffe (weapon), washington (Washington), nord (north), sud (south), atomar (atomic), kommunist (communist), abrüst (disarmament), amerikan (American)
T134	Steel Industry Restructuring and Downsizing	mitarbeit (employee), arbeitsplatz (workplace), beschäftigt (employed), stell (position), werk (plant), standort (location), thyssenkrupp (ThyssenKrupp - a German multinational conglomerate), abbau (downsizing), krupp (ThyssenKrupp), betriebsrat (works council), stahl (steel), konzern (group), belegschaft (workforce), betriebsrat (affected), entlass (dismissal), stellenabbau (job cuts), schliessung (closure), produktion (production), angekündigt (announced), abgebaut (reduced), konzern (corporation), gestrich (cancelled), personalabbau (personnel reduction)
T136	Expectations and Uncertainty	bleibt (remains), chancenlos (chance), hoffnung (hope), dürfte (is likely), bleibe (remain), letzt (last), lang (long), steht (stands), erfolg (success), kaum (hardly), kommt (comes), klar (clear), endlich (finally), wohl (probably), signal, hoffnung (hope), hand, bring, hart (hard), zumindest (at least), versprechen (promise), schwer (difficult), tisch (table)
T138	Right-Wing Extremism	polizei (police), npd (National Democratic Party of Germany), tat (act), recht (right), jugend (youth), jahrig (year-old), verbot (ban), rechtsextrem (right-wing extremism), rechtsextremist (right-wing extremist), gewalt (violence), mann (man), rechtsradikal (far-right), gewalttat (violent act), überfall (attack), schwer (severe), verfassungsschutz (domestic intelligence)
T139	Interviews and Opinions	sz (Süddeutsche Zeitung), ja (yes), sag (say), seh (see), tun (do), glaub (believe), natur (of course), nein (no), warum (why), richtig (right), klar (clear), herr (Mr.), leute (people), wiss (know), interview, brauch (need), gesagt (said), überhaupt (at all), denk (think), einfach (simply), genau (exactly), kommt (comes), wirklich (really), beispiel (example), deshalb (therefore), besser (better), vielleicht (maybe)
T140	UK Politics and Northern Ireland	britisch (British), london (London), grossbritannien (Great Britain), blair (Tony Blair, a British politician), brit (Brit), labour, pfund (pound), nordirland (Northern Ireland), irisch (Irish), england (England), konservativ (conservative), protestant (Protestant), irland (Ireland), bbc (BBC), tony_blair (Tony Blair), belfast (Belfast), the financial_times (Financial Times), katholik (Catholic), sinn_fein (Sinn Féin, an Irish political party)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T142	Employment Contracts and Severance Rights	erhalt (receive), anspruch (entitlement), regel (rule), gehalt (salary), kündigung (termination), arbeitnehm (employee), gilt (applies), arbeit (work), leistung (performance), bezug (reference), kassel (Kassel), gelt (valid), arbeitgeb (employer), entsprech (corresponding), mitarbeit (employee), gezahlt (paid), voll (full), angestellt (employed), zahlung (payment), abfind (severance), mindest (minimum), befristet (fixed-term)
T143	Financial Advice and Risks	wer (who), geld (money), meist (mostly), oft (often), beispiel (example), je (per), bess (better), einfach (simple), kommt (comes), gerad (just), kaum (hardly), häufig (frequently), schnell (quickly), lasst (lets), gilt (applies), deshalb (therefore), verdi (earn), schlecht (bad), genau (exactly), desto (the ... the), gleich (same), risiko (risk), gar (at all), sogar (even), selt (rarely), sollt (should), richtig (correct), gewinn (profit)
T145	Italian Politics and Industry	itali (Italian), italien (Italy), rom (Rome), mailand (Milan), berlusconi (Silvio Berlusconi), fiat (Fiat), la_repubblica (La Repubblica, italian newspaper), mitt (middle), lir (lira), della (Rizzoli Corriere della Sera), corri (Rizzoli Corriere della Sera), sera (Rizzoli Corriere della Sera), prodi (Romano Prodi), turin (Turin), link (left), silvio_berlusconi (Silvio Berlusconi), olivetti (Olivetti), banca (bank)
T147	Stock Trading and Financial Markets	bors (stock exchange), frankfurt (Frankfurt), handel (trading), akti (stock), anleg (investor), gehandelt (traded), option, deutsche_börse (Deutsche Börse), nasdaq (NASDAQ), geschäft (business), euronext (Euronext), wert (value), deutschen_börse (German stock exchange), derivat (derivative), investor, ise (London Stock Exchange), nys (NYSE), terminbors (futures exchange), futur (futures), london (London), wertpapi (securities), bank (bank), brok (broker), eurex (Eurex)
T148	Scheduling and Delays	januar (January), dezemb (December), oktob (October), novemb (November), septemb (September), termin (appointment), beginn (start), ursprung (origin), verschob (postponed), verschieb (postponement), voraussicht (expected), zunach (initially), februar (February), frist (deadline), angekündigt (announced), letzt (last), vorgeseh (planned), endgult (final), spatest (latest), anfang (beginning), verlängert (extended)
T149	Trading and Market Movements	dm (DM, the German mark), frankfurt (Frankfurt), main (Main), pfennig, leicht (slightly), zeigt (shows), plus, fixing, kass (cash), fest (stable), je (per), geschäft (business), handel (trade), stieg (rose), behauptet (held up), fiel (fell), tendiert (tends), va (preferred shares), schwach (weak), notiert (quoted), zog (rose), verlor (lost), dusseldorf (Düsseldorf), gehandelt (traded)
T150	Corporate Growth	geschäft (business), weltweit (worldwide), wachstum (growth), marktanteil (market share), umsatz (revenue), wachs (growth), wettbewerb (competition), branch (industry), ausbau (expansion), anbiet (provider), steig (increase), konz (group), gewinn (profit), konkurrent (competitor), marktfuhr (market leadership), konkurrenz (competition), asi (Asia), mark (market), bereich (sector), expansion
T151	China	china (China), chines (Chinese), peking (Beijing), hongkong (Hong Kong), chinas (China's), taiwan (Taiwan), shanghai (Shanghai), tibet (Tibet), ausland (abroad), asi (Asia), provinz (province), shanghai (Shanghai), volksrepubl (People's Republic), fuhrung (leadership), menschenrecht (human rights), yuan, berichtet (reports), chen (Chen Shui-bian), jiang_zemin (Jiang Zemin)
T152	Oil Prices and Energy Markets	ol (oil), opec (OPEC), olpreis (oil price), dollar, preis (price), lit (litre), shell (Shell), barrel, je (per), bp (BP), benzin (gasoline), tankstell (service station), rohol (crude oil), ra_neri (refinery), erdol (crude oil), pipelin (pipeline), gas, weltweit (worldwide), benzinpreis (gasoline price), produktion (production), forder (demand), cent, energi (energy), diesel, nachfrag (demand)
T153	People, Life Stories, and Emotions	mann (man), mal (once), bild (picture), mensch (person), hand, paar (couple), aug (eye), kopf (head), erzahlt (tells), leut (people), gesicht (face), rot (red), stund (hour), seh (see), haus (house), blau (blue), minut (minute), frau (woman), lieb (dear), leb (live), nacht (night), journalist, ja (yes), blick (glance), morg (morning), herr (Mr.), kolleg (colleague)
T154	Electronics Industry	herstell (manufacturer), gerat (device), infineon (Infineon), intel (Intel), sony (Sony), weltweit (worldwide), comput (computer), nokia (Nokia), philips (Philips), munch (Munich), pc (PC), handy (mobile phone), dell (Dell), konz (group), digital, bereich (sector), geschäft (business), motorola (Motorola), technologi (technology), elektron (electronic), chip

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T155	Automotive Industry	daiml (Daimler), bmw (BMW), chrysl (Chrysler), ford (Ford), fahrzeug (vehicle), autos (cars), opel (Opel), gm (GM), herstell (manufacturer), daimlerchrysl (DaimlerChrysler), motor (engine), modell (model), auto (car), autoherstell (automaker), stuttgart (Stuttgart), werk (plant), pkw (passenger car), general_motors (General Motors), absatz (sales)
T158	Kidnappings and Hostage Situations	botschaft (embassy), entfuhr (kidnap), geiseln (hostages), entfuhr (kidnapped), freilass (release), journalist, diplomat, mann (man), verschleppt (abducted), grupp (group), freigelass (released), gewalt (violence), verlass (left), berichtet (reports), aussenministerium (Foreign Ministry)
T161	Polish Politics and Coal Mining	pol (Poland), polnisch (Polish), warschau (Warsaw), ess (Essen), rag (RAG AG), kohle (coal), bergbau (mining), steinkohl (coal), solidaritat (solidarity), bergleut (miners), kaczyński (Jarosław Kaczyński), zech (colliery), zloty, walesa (Lech Wałęsa), vertrieb (sale), grenz (border), betont (emphasized), kumpel (miner), vertret (representative), recht (law), forder (demand), subvention (subsidy), ehemal (former)
T164	Data and Privacy	dat (data), syst (system), information, kart (card), kund (customer), elektron (electronic), automat (automated), telefon (telephone), comput (computer), mail, system, person, technisch (technical), nutz (use), test, zentral (central), datenschutz (data protection)
T165	German State Politics	cdu (Christian Democratic Union), pds (Party of Democratic Socialism), brandenburg (Brandenburg), anhalt (Anhalt), sachs (Saxony), potsdam (Potsdam), landtag (state parliament), partei (party), magdeburg (Magdeburg), ministerpräsident (Minister-president), stolp (Manfred Stolpe), landtagswahl (state election), schill (Ronald Schill), vorpomm (Vorpommern), gysi (Gregor Gysi), schwerin (Schwerin), rot (red), bundnis (alliance)
T167	Corporate Financial Performance	millionen_euro (million euros), umsatz (revenue), gewinn (profit), milliarden_euro (billion euros), ergebnis (result), konz (group), verlust (loss), stieg (rose), vorjahr (previous year), ersten_halfjahr (first half-year), ersten_quartal (first quarter), gesamtjahr (full year), steu (tax), legt (legte zu, increased), prognos (forecast), neun (nine), plus, quartal (quarter), geschäftsjahr (financial year)
T168	Housing	bau (construction), wohnung (apartment), miet (rent), haus (house), wohnungsbau (residential construction), million, o ent (public), bauwirtschaft (construction industry), sozial (social), forder (support), baustell (construction site), vermiet (rent out), bauarbeit (construction work), bauunternehmen (construction company), verband (association), topf (fund), baugewerb (building trade)
T169	Economic and Political Trends, Austria	osterreich (Austria), plus, gegenub (versus), minus, januar (January), statist (statistic), oktob (October), septemb (September), fpo (Freedom Party), ruckgang (decline), stieg (rose), novemb (November), februar (February), wiesbad (Wiesbaden), dezemb (December), ging (went), haid (Jörg Haider), ovp (People's Party), leicht (slightly), gestieg (increased), juli (July), vormonat (previous month), nahm (took), sank (fell)
T170	Economic and Social Policy	sozial (social), gesellschaft (society), okonom (economic), marktwirtschaft (market economy), gerecht (fair), globalisier (globalization), wirtschaftspolit (economic policy), burg (citizen), mensch (people), verantwort (responsibility), zukunft (future), arbeit (work), staatlich (state), aufgab (task), wettbewerb (competition), notwend (necessary), syst (system), sozialstaat (welfare state)
T172	German Reunification and Economic Transition	ost (east), west, sachs (Saxony), ostdeutsch (East German), ostdeutschland (East Germany), leipzig (Leipzig), dresd (Dresden), thuring (Thuringia), neuen_ländern (new states), neuen_bundesländern (new federal states), erfurt (Erfurt), sachsisch (Saxon), anhalt (Anhalt), westdeutsch (West German), hall (Halle), neuen_länder (new states), westdeutschland (West Germany), einheit (unity)
T174	GDR	ddr (GDR), stasi (Stasi), ehemal (former), mitarbeit (employee), akt (file), damal (at that time), sed (Socialist Unity Party), honeck (Erich Honecker), behord (authority), ex, grenz (border), bundesrepubl (Federal Republic), tatig (active), person, kontakt (contact), wend (Peaceful Revolution), gauck (Joachim Gauck)

Continued on next page

Chapter 2 Text-Based Economic Forecasting with Topics, Sentiment, and Uncertainty

Table 2.19 continued from previous page

ID	Label	Most probable words
T178	Labor Market and Unemployment	arbeitslos (unemployed), arbeitsmarkt (labor market), arbeit (work), million, job, beschafft (employ), arbeitslosengeld (unemployment benefit), ba (Federal Employment Agency), arbeitsplatz (job position), hartz (Hartz reforms), ii (Hartz II), sozialhilfe (social assistance), stell (position), alt (old), arbeitsamt (employment office), arbeitsmarktpolit (labor market policy), bundesanstalt_für_arbeit (Federal Employment Office), mensch (person), langzeitarbeitslos (long-term unemployed), beschäftigt (employed), arbeitslosenzahl (number of unemployed)
T179	Consumer Goods	nam (name), firma (company), alt (old), design, ide (idea), farb (color), mal (times), meist (mostly), welt (world), stück (piece), kommt (comes), restaurant, einfach (simple), kund (customer)
T182	Announcements and Reactions	gest (yesterday), überrasch (surprise), setzt (sets), zeigt (shows), reagiert (reacts), offenbar (apparently), erneut (again), kam (came), zuvor (before), druck (pressure), nachd (afterwards), zunach (initially), ankund (announce), liess (let), blieb (remained), angekündigt (announced)
T183	Financial and Economic Performance	mrd (billion), mill (million), eur (euro), knapp (just under), gest (rose), gegenub (compared to), anteil (share), volum (volume), dürfte (is likely), gesamt (total), vergangenen_jahr (last year), wert (value), allein (alone), liegt (lies), drittel (third), halft (half), steig (increase), rechnet (expects), erreicht (reached), entspricht (equivalent to), erhöht (increased), lieg (lie), einnahm (revenue)
T184	Church	kirch (church), papst (pope), kathol (Catholic), vatican (Vatican), kardinal (cardinal), christlich (Christian), mensch (human), christ (Christ), bischof (bishop), evangel (evangelical), rom (Rome), heilig (holy), gott (God), gläubig (believer), ekd (Evangelical Church in Germany), glaub (believe), priest, erzbischof (archbishop)
T189	Role and Influence	roll (role), spiel (play), spielt (plays), interest (interest), einfluss (influence), steht (stands), position, besond (especially), gegenub (compared to), dürfte (is likely), nehm (take), gespielt (played), jung (young), eng (close), deshalb (therefore), seh (see), gewicht (weight), wichtige_rolle (important role), verhältnis (relationship)
T190	Savings and Retirement Planning	privat (private), spar (saving), geld (money), erhalt (receive), betrag (amount), altersvorsorg (retirement provision), alt (old), zins (interest), vertrag (contract), wer (who), betrieb (company), zusatz (additional), einkomm (income), jährlich (yearly), liegt (lies), je (per), staatlich (state), mindest (minimum), steu (tax), zahlt (pays), gering (low), auszah (payout)
T191	European Union	eu (EU), brussel (Brussels), europäischen_union (European Union), kommission (Commission), beitritt (accession), gipfel (summit), erweiter (expansion), regierungschef (head of government), union (Union), aussenminist (foreign minister), europäische_union (European Union), europas (Europe's), verfass (constitution), mitgliedstaat (member state), gemeinschaft (Community)
T192	Iraq War	irak (Iraq), bagdad (Baghdad), krieg (war), amerikan (American), saddam (Saddam), saddam_husein (Saddam Hussein), kuwait (Kuwait), militar (military), washington (Washington), golfkrieg (Gulf War), angri (attack), trupp (troop), soldat (soldier), wa (weapon), bush (Bush), massenvernichtungswa (weapon of mass destruction), alliiert (ally), regim (regime)
T193	Bosnian War	bosni (Bosnia), serbisch (Serbian), serb (Serb), kroatisch (Croatian), bosnisch (Bosnian), uno (UNO, United Nations Organization), kroati (Croatia), sarajevo (Sarajevo), moslem (Muslim), zagreb (Zagreb), herzegowina (Herzegovina), un (UN - United Nations), kroat (Croat), jugoslavi (Yugoslavia), bosnischen_serben (Bosnian Serbs), stadt (city), belgrad (Belgrade), kriegsverbrech (war crime), trupp (troop), tribunal
T194	Public Sector	bund (federal government), kommun (municipal), beamt (civil servant), öffent (public), gemeind (municipality), stadt (city), kommunal (local), aufgab (task), zustand (condition), verwalt (administration), burg (citizen), bundesland (federal state), dien (service), finanziell (financial), sollte (should), übertrag (transfer), kost (cost), kompetenz (competence), stell (position), stadtetag (city budget)

Continued on next page

Table 2.19 continued from previous page

ID	Label	Most probable words
T197	Aerospace Industry	airbus (Airbus), boeing (Boeing), ead (EADS), flugzeug (aircraft), luft (air), dasa (DASA), maschin (machine), auftrag (order), projekt, industri (industry), flugzeugbau (aircraft manufacturing), mtu (MTU), raumfahrt (space travel), amerikan (American), bestell (order), satellit (satellite), jet, entwickl (development), bau (construction), munch (Munich), daiml (Daimler), partn (partner), zivil (civil), system

Notes: 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.

2.B Hard data and surveys

Table 2.20: Hard data and surveys used in the forecasting experiment

Name	Transformation	Group
Production in main construction industry	3	Activity
Industrial production index	3	Activity
New orders for main construction industry	3	Activity
New orders for industry	3	Activity
Main construction industry turnover	3	Activity
Industry turnover	3	Activity
Retail turnover excluding cars	3	Activity
Consumer price index	3	Prices
Consumer price index, excluding energy	3	Prices
Producer price index	3	Prices
Producer price index, excluding energy	3	Prices
Export price index	3	Prices
Import price index	3	Prices
Hours worked: manufacturing	3	Labor market
Hours worked: construction	3	Labor market
Employment	3	Labor market
Gross wages and salaries: manufacturing and mining	3	Labor market
Gross wages and salaries: construction	3	Labor market
CDAX	3	Financial
Government bond yields (1-year)	2	Financial
Government bond yields (5-years)	2	Financial
Government bond yields (10-years)	2	Financial
Nominal effective exchange rate (narrow)	3	Financial
Nominal effective exchange rate (broad)	3	Financial
Yields on debt securities issued by residents	2	Financial
Yields on bank debt securities	2	Financial
Yields on corporate debt securities	2	Financial
Yields on public debt securities	2	Financial
ifo: industry and trade, climate	2	Surveys
ifo: industry and trade, current situation	2	Surveys
ifo: industry and trade, expectations	2	Surveys
GfK: business cycle expectations	2	Surveys
GfK: income expectations	2	Surveys
GfK: willingness-to-buy	2	Surveys
GfK: consumer climate indicator	2	Surveys
Economics Sentiment Indicator	2	Surveys

Transformation 3 = difference of logarithms; Transformation 2 = first differences; Transformation 1 = levels.

2.C Additional correlation results: GDP

Table 2.21: Correlations of selected general lexicon-adjusted topics with quarterly GDP growth (first release) and selected surveys

ID	Label	GDP	ifo Climate	ifo Situation	ifo Expectations	ESI
T29	Banking	0.529***	0.473***	0.352***	0.570***	0.440***
T50	Crisis	0.511***	0.523***	0.340***	0.729***	0.500***
T21	Policy Measures	0.439**	0.378***	0.215**	0.590***	0.363***
T108	US Politics	0.393**	0.354***	0.203**	0.552***	0.332***
T120	Economic Growth	0.367**	0.424***	0.325***	0.493***	0.432***
T38	Problem Solving	0.361**	0.390**	0.235*	0.579***	0.397**
T134	Steel Industry Restructuring and Downsizing	0.323*	0.442***	0.360***	0.471***	0.472***
T124	Elections and Office Succession	0.308**	0.169	0.111	0.230*	0.200
T98	Public Appearances and Reactions	0.308**	0.373***	0.233**	0.536***	0.388***
T91	Media Coverage of Plans and Rumors	0.301***	0.477***	0.348***	0.584***	0.493***

Notes: The selected topics show the strongest correlations with GDP growth among all estimated plain topics adjusted with general lexicon. "ifo Climate" denotes the ifo Business Climate for industry & trade (balances); "ifo Situation" denotes the ifo Business Situation for industry & trade (balances); "ifo Expectations" denotes the ifo Business Expectations for industry & trade (balances); "ESI" denotes the Economic Sentiment Indicator of European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with GDP growth data. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order 4).

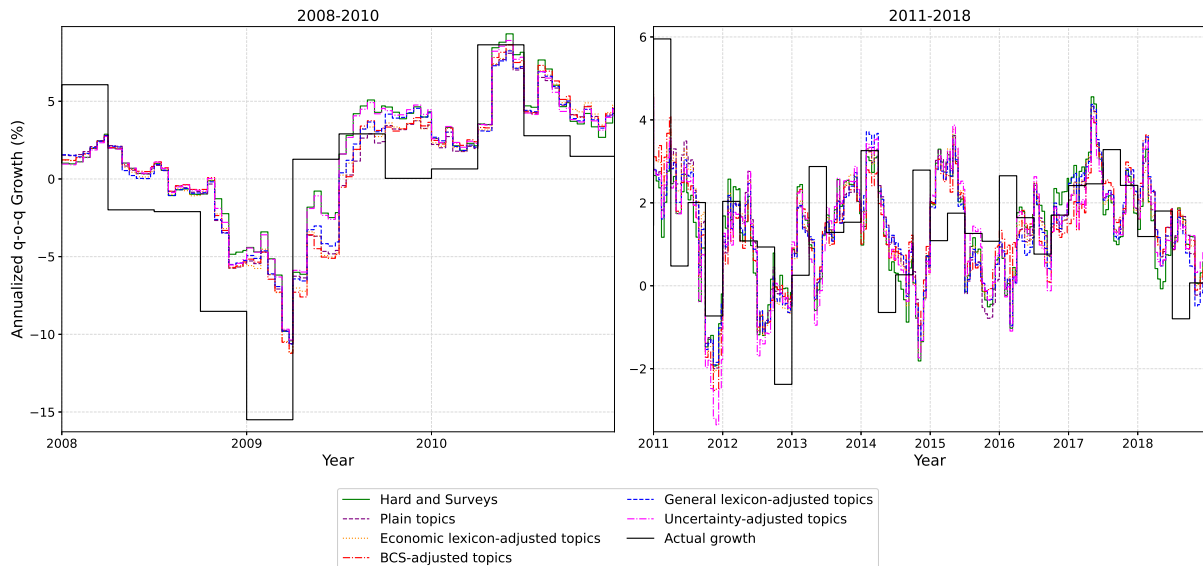
Table 2.22: Correlations of uncertainty-adjusted topics with quarterly GDP growth and selected surveys

ID	Label	GDP	ifo Climate	ifo Situation	ifo Expectations	ESI
T134	Steel Industry Restructuring and Downsizing	-0.440**	-0.369***	-0.228**	-0.535***	-0.391***
T21	Policy Measures	-0.262	-0.263*	-0.128	-0.451***	-0.246*
T14	Corporate Structure and M&A	-0.258**	-0.269***	-0.153*	-0.420***	-0.194*
T190	Savings and Retirement Planning	-0.256**	-0.249**	-0.171	-0.332***	-0.237**
T50	Crisis	-0.244*	-0.255**	-0.118	-0.450***	-0.259**
T154	Electronics Industry	-0.243**	-0.208***	-0.070	-0.419***	-0.197***
T7	Mergers and Acquisition	-0.237*	-0.100	-0.018	-0.232*	-0.084
T155	Automotive Industry	-0.235*	-0.310***	-0.196*	-0.441***	-0.300***
T120	Economic Growth	-0.233***	-0.260***	-0.147	-0.404***	-0.279***
T139	Interviews and Opinions	-0.221	-0.211	-0.164	-0.241	-0.195

Notes: The selected topics show the strongest correlations with GDP growth among all estimated plain topics adjusted with the share of uncertainty terms. "ifo Climate" denotes the ifo Business Climate for industry & trade (balances); "ifo Situation" denotes the ifo Business Situation for industry & trade (balances); "ifo Expectations" denotes the ifo Business Expectations for industry & trade (balances); "ESI" denotes the Economic Sentiment Indicator of European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with GDP growth data. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order 4).

2.D GDP forecasts: all models

Figure 2.11: Nowcasts and actual GDP growth (first release) for different models, 2008–2010 and 2011–2018



Notes: The figure compares nowcasts from several Dynamic Factor Models (DFMs) with the first-release estimate of quarterly GDP growth. The benchmark specification includes two factors extracted from hard and survey data only (green, solid line). Competing models augment this benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); and (v) BCS-adjusted topics (red, dash-dotted line). Nowcasts are shown against the realized quarterly GDP growth series (black, solid line). The left panel displays the results for the Financial Crisis period (2008–2010), while the right panel covers the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

2.E Additional correlation results: investment

Table 2.23: Correlations of selected plain topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Current	ifo Expectations	ESI
T150	Corporate Growth	0.537***	0.479***	0.347***	0.593***	0.508***
T50	Crisis	-0.505***	-0.524***	-0.349***	-0.712***	-0.493***
T134	Steel Industry Restructuring and Downsizing	-0.370**	-0.350**	-0.308**	-0.330**	-0.374***
T29	Banking	-0.356**	-0.404***	-0.325***	-0.439***	-0.342***
T110	Technological Innovation	0.354***	0.338**	0.231*	0.447***	0.427***
T59	Commodity Markets	0.349***	0.283**	0.171	0.423***	0.262*
T21	Policy Measures	-0.345**	-0.180	-0.059	-0.364**	-0.148
T23	Insolvency and Financial Rescue	-0.313**	-0.431***	-0.391***	-0.377***	-0.434***
T108	US Politics	-0.309**	-0.254**	-0.129	-0.429***	-0.220*
T120	Economic Growth	-0.308**	-0.323***	-0.277***	-0.316**	-0.319***

Notes: The table reports the strongest correlations between plain topics and quarterly investment growth (first release). “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Current” denotes the ifo Current Business Situation; “ifo Expectations” denotes the ifo Business Expectations; “ESI” denotes the European Commission’s Economic Sentiment Indicator. Monthly survey indicators are aggregated to the quarterly frequency for comparability with investment growth data. Topics T62, T99, and T104 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

Table 2.24: Correlations of selected BCS-adjusted topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Current	ifo Expectations	ESI
T27	Economic Crises and Recessions	0.608***	0.433***	0.276***	0.616***	0.393***
T138	Financial and Economic Performance	0.576***	0.503***	0.336***	0.684***	0.515***
T81	Corporate Restructuring and Job Cuts in Germany	0.539***	0.368***	0.253***	0.486***	0.344***
T52	German Automobile Industry and Major Manufacturers	0.477***	0.409***	0.290***	0.527***	0.383***
T127	Major Banks and Investment Banking	0.475**	0.463***	0.315***	0.615***	0.442***
T74	Concerns about Economic Bubbles and Recessions	0.461***	0.471***	0.334***	0.601***	0.442***
T77	Private Investment	0.445***	0.416***	0.307***	0.506***	0.419***
T100	Market Reactions to News	0.439***	0.320***	0.138	0.581***	0.305***
T11	Mergers and Acquisitions	0.413***	0.473***	0.350***	0.571***	0.525***
T131	German Investments in Emerging Markets	0.391**	0.381***	0.294***	0.437***	0.370***

Notes: The selected topics are the BCS-adjusted topics used in the GDP analysis and are those that exhibit the strongest correlations with quarterly GDP growth. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Current” denotes the ifo Current Business Situation; “ifo Expectations” denotes the ifo Business Expectations; “ESI” denotes the European Commission’s Economic Sentiment Indicator. Monthly survey indicators are aggregated to the quarterly frequency for comparability with investment growth data. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

Table 2.25: Correlations of selected general lexicon-adjusted topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Situation	ifo Expectations	ESI
T50	Crisis	0.523***	0.523***	0.340***	0.729***	0.500***
T21	Policy Measures	0.497***	0.378***	0.215**	0.590***	0.363***
T29	Banking	0.478***	0.473***	0.352***	0.570***	0.440***
T134	Steel Industry Restructuring and Downsizing	0.462***	0.442***	0.360***	0.471***	0.472***
T120	Economic Growth	0.439***	0.424***	0.325***	0.493***	0.432***
T38	Problem Solving	0.422**	0.390**	0.235*	0.579***	0.397**
T108	US Politics	0.415***	0.354***	0.203**	0.552***	0.332***
T182	Announcements and Reactions	0.382**	0.413***	0.326***	0.461***	0.438***
T23	Insolvency and Financial Rescue	0.376**	0.473***	0.402***	0.467***	0.484***
T91	Media Coverage of Plans and Rumors	0.366***	0.477***	0.348***	0.584***	0.493***

Notes: The selected topics show the strongest correlations with investment growth among all estimated plain topics adjusted with general lexicon. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Situation” denotes the ifo Business Situation for industry & trade (balances); “ifo Expectations” denotes the ifo Business Expectations for industry & trade (balances); “ESI” denotes the Economic Sentiment Indicator of the European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with investment growth data. Topics T56 and T99 were excluded. Significance levels: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Significance levels are based on t-statistics from OLS regression with Newey-West SEs (maximum lag order = 4).

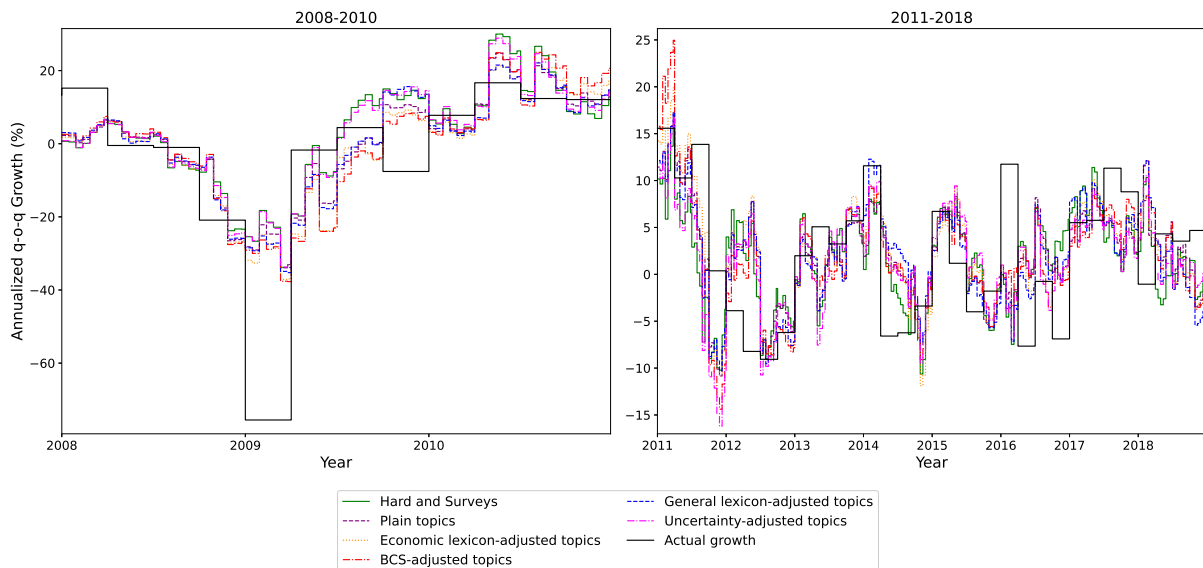
Table 2.26: Correlations of selected uncertainty-adjusted topics with quarterly investment growth (first release) and selected surveys

ID	Label	Investment	ifo Climate	ifo Situation	ifo Expectations	ESI
T134	Steel Industry Restructuring and Downsizing	-0.464**	-0.369***	-0.228**	-0.535***	-0.391***
T59	Commodity Markets	0.306***	0.169	0.124	0.211*	0.141
T123	Media and Newspapers	-0.292*	-0.223**	-0.167	-0.267***	-0.258**
T154	Electronics Industry	-0.289***	-0.208***	-0.070	-0.419***	-0.197***
T9	Small and Medium-Sized Enterprises	-0.266***	-0.259**	-0.147	-0.407***	-0.235**
T23	Insolvency and Financial Rescue	-0.255**	-0.391***	-0.330***	-0.394***	-0.376***
T18	Agreements and Cooperation	-0.253***	-0.210*	-0.122	-0.321**	-0.201
T21	Policy Measures	-0.247	-0.263*	-0.128	-0.451***	-0.246*
T50	Crisis	-0.233**	-0.255**	-0.118	-0.450***	-0.259**
T7	Mergers and Acquisition	-0.223*	-0.100	-0.018	-0.232*	-0.084

Notes: The selected topics show the strongest correlations with investment growth among all estimated plain topics adjusted with the share of uncertainty terms. “ifo Climate” denotes the ifo Business Climate for industry & trade (balances); “ifo Situation” denotes the ifo Business Situation for industry & trade (balances); “ifo Expectations” denotes the ifo Business Expectations for industry & trade (balances); “ESI” denotes the Economic Sentiment Indicator of the European Commission. Monthly survey indicators are aggregated to the quarterly level for consistency with the investment growth data. Topics T56 and T153 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey-West SEs (maximum lag order = 4).

2.F Investment forecasts: all models

Figure 2.12: Nowcasts and actual investment growth (first release) for different models, 2008–2010 and 2011–2018



Notes: The figure compares nowcasts from several Dynamic Factor Models (DFMs) with the first-release estimate of quarterly investment growth. The benchmark specification includes two factors extracted from hard and survey data only (green, solid line). Competing models augment this benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); and (v) BCS-adjusted topics (red, dash-dotted line). Nowcasts are shown against the realized quarterly investment growth (black, solid line). The left panel displays the results for the Financial Crisis period (2008–2010), while the right panel covers the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

2.G Additional correlation results: consumption

Table 2.27: Correlations of selected plain topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T45	Diplomatic Visits	0.410***	0.029	0.028	0.029	0.077
T78	CSU and Leadership Dynamics	-0.353***	-0.080	-0.082	-0.118	-0.136
T139	Interviews and Opinions	-0.282**	-0.054	0.032	0.047	-0.079
T83	Research Institutes	-0.278**	-0.012	-0.041	-0.020	-0.168
T9	Small and Medium-Sized Enterprises	-0.253**	0.133	0.026	0.021	-0.055
T134	Steel Industry Restructuring and Downsizing	-0.238**	-0.450***	-0.260***	0.025	-0.043
T84	Insurance Industry	-0.229**	-0.083	-0.092	-0.054	-0.337**
T82	Economic and Monetary Union	0.223*	-0.259***	-0.099	-0.076	-0.090
T155	Automotive Industry	-0.221**	-0.184*	-0.047	0.069	-0.023
T142	Employment Contracts and Severance Rights	-0.218	-0.098	-0.081	0.041	-0.038

Notes: The selected topics show the strongest correlations with consumption growth among all estimated plain topics. "GfK BCE" denotes the GfK Business Cycle Expectations indicator; "GfK IE" denotes GfK Income Expectations; "GfK WtB" denotes GfK Willingness-to-Buy; "GfK CCI" denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly level for consistency with consumption growth data. Topics T94, T169, and T174 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

Table 2.28: Correlations of selected economic lexicon-adjusted topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T78	CSU and Leadership Dynamics	0.343***	0.191*	0.081	0.096	0.064
T155	Automotive Industry	0.285***	0.438***	0.149*	0.034	0.060
T134	Steel Industry Restructuring and Downsizing	0.276**	0.542***	0.280***	0.059	0.100
T139	Interviews and Opinions	0.264**	0.402***	0.103	0.060	0.228*
T183	Financial and Economic Performance	0.237**	0.459***	0.263***	0.205*	0.389**
T142	Employment Contracts and Severance Rights	0.223	0.231**	0.142	-0.003	0.098
T178	Labor Market and Unemployment	0.220*	0.304***	0.178**	0.090	0.159**
T83	Research Institutes	0.218	0.560***	0.237**	0.133	0.241
T45	Diplomatic Visits	-0.206*	-0.052	-0.071	0.027	-0.073
T167	Corporate Financial Performance	0.194*	0.541***	0.221***	0.135	0.260**

Notes: The selected topics show the strongest correlations with consumption growth among all estimated plain topics adjusted with the economic lexicon. "GfK BCE" denotes the GfK Business Cycle Expectations indicator; "GfK IE" denotes GfK Income Expectations; "GfK WtB" denotes GfK Willingness-to-Buy; "GfK CCI" denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly level for consistency with consumption growth data. Topics T89 and T169 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

Table 2.29: Correlations of selected general lexicon-adjusted topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T78	CSU and Leadership Dynamics	0.381***	0.169	0.122	0.150	0.177
T45	Diplomatic Visits	-0.328***	0.212**	0.093	0.067	-0.022
T83	Research Institutes	0.301**	0.238**	0.166*	0.096	0.223
T9	Small and Medium-Sized Enterprises	0.276*	0.219*	0.128	0.130	0.223
T13	CDU/CSU–FDP Coalition Politics	0.241***	0.060	0.059	0.056	0.201
T14	Corporate Structure and M&A	0.240	-0.014	0.061	0.051	0.163
T142	Employment Contracts and Severance Rights	0.227	0.276***	0.159*	0.006	0.115
T134	Steel Industry Restructuring and Downsizing	0.216*	0.555***	0.306***	0.033	0.090
T147	Stock Trading and Financial Markets	0.207*	-0.250**	-0.140	-0.110	-0.153
T75	Accounting Standards, Reporting, and Risk Management	0.205*	-0.022	-0.019	0.031	0.210

Notes: The selected topics show the strongest correlations with consumption growth among all estimated plain topics adjusted with the general lexicon. “GfK BCE” denotes the GfK Business Cycle Expectations indicator; “GfK IE” denotes GfK Income Expectations; “GfK WtB” denotes GfK Willingness-to-Buy; “GfK CCI” denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly level for consistency with consumption growth data. Topics T161, T169, and T174 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

Table 2.30: Correlations of selected uncertainty-adjusted topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T78	CSU and Leadership Dynamics	-0.264**	-0.302***	-0.160	-0.084	-0.144
T134	Steel Industry Restructuring and Downsizing	-0.256***	-0.476***	-0.239**	-0.055	-0.180*
T100	Protests and Demonstrations	0.218*	-0.059	-0.020	0.022	0.024
T117	Future Vision	0.216*	-0.223**	-0.060	-0.040	-0.059
T45	Diplomatic Visits	0.199*	-0.079	-0.013	0.010	0.053
T38	Problem Solving	0.196	-0.249***	-0.084	-0.105	-0.128
T194	Public Sector	-0.190	-0.102	-0.084	-0.038	-0.092
T46	German Maritime Sector	-0.183	-0.274***	-0.116	-0.072	-0.051
T1	Elections	0.174	0.108	0.052	0.048	-0.009
T131	Health Policy	-0.174*	-0.120	-0.075	0.008	0.004

Notes: The selected topics show the strongest correlations with consumption growth among all estimated plain topics adjusted with the share of uncertainty terms. “GfK BCE” denotes the GfK Business Cycle Expectations indicator; “GfK IE” denotes GfK Income Expectations; “GfK WtB” denotes GfK Willingness-to-Buy; “GfK CCI” denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly level for consistency with consumption growth data. Topics T80, T89, T111, T172, and T192 were excluded. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West SEs (maximum lag order = 4).

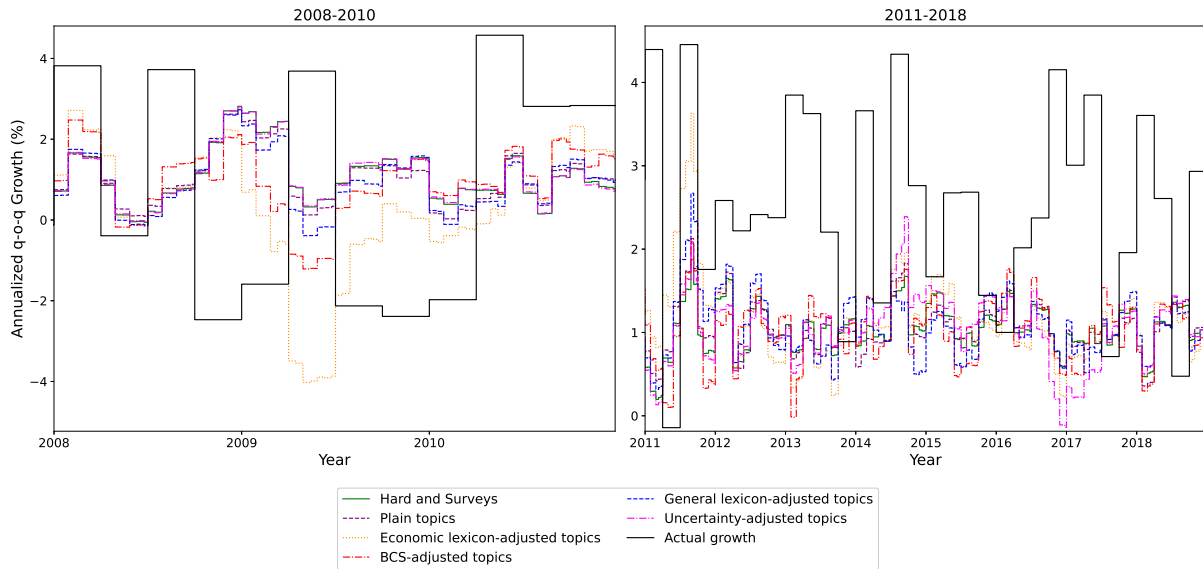
Table 2.31: Correlations of selected BCS-adjusted topics with quarterly consumption growth (first release) and selected surveys

ID	Label	Consumption	GfK BCE	GfK IE	GfK WtB	GfK CCI
T81	Corporate Restructuring and Job Cuts in Germany	0.194*	0.302***	0.093	0.036	0.223**
T127	Major Banks and Investment Banking	0.151	0.481***	0.188*	0.068	0.264*
T27	Economic Crises and Recessions	0.135	0.378***	0.124	0.009	0.096
T100	Market Reactions to News	0.129	0.346***	0.084	0.045	0.037
T11	Mergers and Acquisitions	0.119	0.496***	0.269***	0.244**	0.324**
T52	German Automobile Industry and Major Manufacturers	0.104	0.414***	0.158**	0.069	0.159**
T74	Concerns about Economic Bubbles and Recessions	0.090	0.472***	0.179*	0.092	0.218*
T77	Private Investment	0.053	0.376***	0.123	0.065	0.185
T138	Financial and Economic Performance	0.025	0.497***	0.148	0.078	0.170**
T131	German Investments in Emerging Markets	0.005	0.379***	0.102	0.038	0.211

Notes: The selected topics are the BCS-adjusted topics used in the GDP analysis and are those that exhibit the strongest correlations with quarterly GDP growth. "GfK BCE" denotes the GfK Business Cycle Expectations indicator; "GfK IE" denotes GfK Income Expectations; "GfK WtB" denotes GfK Willingness-to-Buy; "GfK CCI" denotes the GfK Consumer Climate Indicator. Monthly survey indicators are aggregated to the quarterly frequency for consistency with consumption growth data. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels are based on t-statistics from OLS regressions with Newey–West standard errors (maximum lag order = 4).

2.H Consumption forecasts: all models

Figure 2.13: Nowcasts and actual consumption growth (first release) for different models, 2008–2010 and 2011–2018



Notes: The figure compares nowcasts from several Dynamic Factor Models (DFMs) with the first-release estimate of quarterly consumption growth. The benchmark specification includes one factor extracted from hard and survey data only (green, solid line). Competing models augment this benchmark with one additional text factor, constructed from: (i) plain topics (purple, dashed line); (ii) economic lexicon-adjusted topics (orange, dotted line); (iii) general lexicon-adjusted topics (blue, dashed line); (iv) uncertainty-adjusted topics (pink, dash-dotted line); and (v) BCS-adjusted topics (red, dash-dotted line). Nowcasts are shown against the realized quarterly consumption growth (black, solid line). The left panel displays the results for the Financial Crisis period (2008–2010), while the right panel covers the subsequent period of calmer economic conditions, which nevertheless includes the European Debt Crisis (2011–2018).

Chapter 3

Salmon stock returns around market news

Abstract

We examine the relationship between media news and trading behaviour in the salmon market. For this, we create a share price index (SPI) based on five salmon aquaculture companies trading on the Oslo Stock Exchange (OSE). We use the Latent Dirichlet Allocation (LDA) algorithm to obtain news topics and a lexicon-based sentiment analysis. We find that topics relating to COVID-19 and sustainability have a significant negative impact on the salmon market, while topics on land-based aquaculture have a significant positive impact. The sentiment series based on the Loughran-McDonald lexicon is found to have a negative and insignificant effect on stock returns. Hence, we expand the lexicon with industry-specific words. A negative shock to sentiment within the news related to competitors foreshadows a significant increase in returns due to the markets' competitive nature. Through our out-of-sample forecasting experiment, we find that the incorporation of news data can improve the predictive performance.

Keywords: Salmon Market; Text Mining; Topics Modelling; LDA; Sentiment Analysis

JEL classification: C38; C55; C58; G14; G17; Q02; Q22.

This study is coauthored by Clemens Knoppe and Mikaella Zitti.

It is published in Marine Resource Economics, 40(2).

3.1 Introduction

The aquaculture industry has experienced rapid growth as a food production sector (FAO, 2018; Garlock et al., 2020), with global production surging from approximately 14.9 million tonnes in 1995 to 82.1 million tonnes in 2018 (FAO, 2020). Originating in the late 1960s, the Norwegian aquaculture industry has developed into one of the country's most vital export industries alongside oil and gas within a span of 50 years. Salmon has emerged as one of the most successful species in the industry (Asche, 2008), leading Norway to become the world's largest salmon producer, accounting for over half of the total production (Hersoug, 2021). The industry's expansion has facilitated its consolidation, simultaneously generating increased demand, elevated salmon prices, and augmented volatility (Asche et al., 2019; Bloznelis, 2016; Oglend, 2013; Zitti, 2024).

Guttormsen (1999) showed that salmon is a volatile commodity, with its volatility more than doubling and now surpassing that of most comparable commodities (Dahl & Oglend, 2014). Previous research has revealed that volatility exhibits clustering (Dahl & Yahya, 2019), and that it is partially explained by trends in other food commodities (Oglend, 2013). Building upon this literature, our study investigates the types of news that influence the salmon market. Given that investor behaviour is driven by news surrounding the companies they are invested in, we analyse the impact of salmon production-related news articles on stock prices, which have been found to be more reactive than commodity prices (Dahl et al., 2021).

The theoretical background for understanding the influence of news articles on stock returns can be anchored in the Efficient Market Hypothesis (EMH), which posits that all available information is always reflected in stock prices (Fama, 1965, 1970). In this context, news signifies a mechanism by which new information is disseminated and thus alters expectations, leading to changes in prices. EMH relies on the assumptions of rational expectations and the full incorporation of information by profit-maximizing investors. However, another strand of literature highlights the limits of arbitrage, challenging the notion of full market efficiency (De Long et al., 1990; Shleifer & Vishny, 1997). These inefficiencies create the potential for the effects of investor sentiment on stock returns, as demonstrated empirically by Brown and Cliff (2005), Kling and Gao (2008), and Verma et al. (2008). The role of news in financial market sentiment dynamics has further been studied by Gusev et al. (2015).

In light of this theoretical background, it is not surprising that the application of text mining methods in finance and macroeconomics has yielded promising results (Borup et al., 2023; Ellingsen et al., 2022). For instance, Hansen et al. (2018) employed the Latent Dirichlet Allocation (LDA) algorithm to assess the impact of transparency on monetary policy deliberations, while Larsen (2021) used LDA to classify news types and determine

the impact of various uncertainties on the economy. In the financial literature, topics identified through LDA have been utilized to predict stock movements (Nguyen et al., 2015). Our study builds on this approach by incorporating LDA to analyze news topics, providing an overview of primary themes in our dataset. Additionally, we use a lexicon-based approach for sentiment analysis, a method widely applied in stock market studies (Hanna et al., 2020; Karalevicius et al., 2018; Khedr, Yaseen, et al., 2017; Li et al., 2014, 2020). Specifically, we employ the lexicon developed by Loughran and McDonald (2011, henceforth: LM), which is tailored for financial applications. By integrating text mining techniques such as LDA and lexicon-based sentiment analysis, our goal is to elucidate how news topics and sentiment influence investor behavior in the salmon market.

While LDA remains a popular choice for topic modelling, recent advancements such as Non-Negative Matrix Factorization (NMF), Contextualized Topic Models (CTM), and Dynamic Topic Models (DTM) have been developed to address its limitations. However, for our specific application—analysing salmon market news—LDA is still the most theoretically suitable option. Unlike NMF (Lee & Seung, 1999), known for its efficiency, LDA offers a more intuitive analysis by quantifying the attention devoted to specific topics in precise percentage terms, facilitating straightforward comparisons and trend analyses over time. Additionally, while CTMs (Bianchi et al., 2021) improve topic coherence by integrating contextual embeddings with Bag-of-Words (BoW), the effectiveness of CTMs in domain-specific applications relies on robust, domain-relevant embeddings, which are challenging to develop with our relatively small dataset. Moreover, although DTM (Blei & Lafferty, 2006) tracks the evolution of topics over time—an important feature given our multi-year dataset—it struggles to identify entirely new topics (Di Caro et al., 2017), a critical aspect when unexpected issues like the COVID-19 pandemic arise. In conclusion, despite the known limitations of LDA, we believe our choice is theoretically sound and well-suited to meet the specific demands of our study.

Complementing text mining techniques, we utilise a Vector Autoregressive (VAR) model to investigate the relationship between investors' trading behaviour and salmon market-related financial news similar to (Kräussl & Mirgorodskaya, 2017). However, we first apply principal component analysis (PCA) to the topics estimated by the LDA algorithm to identify the most dominant themes shaping market news. These themes are orthogonal, i.e. uncorrelated, by construction, which simplifies the VAR analysis considerably. The inclusion of sentiment analysis allows us to consider the directional effects as well. One of our key contributions in this regard is the improvement of the typical sentiment calculation method by considering the competitive market structure. Moreover, we augment the sentiment dictionary with industry-specific terminology, which enables us to capture information that may not be effectively detected using the financially-oriented LM dictionary alone.

To capture investors' trading behaviour, we examine stock price data from the top five salmon market producer companies in terms of market capitalisation, listed on the Oslo Stock Exchange (OSE): Marine Harvest (MOWI), SalMar (SALM), Grieg Seafood (GSF), Lerøy Seafood Group (LSG), and Bakkafrost (BAKKA). Although futures contracts trading is available in the salmon market, its low liquidity and sporadic trades limit its usefulness (Andersen & de Lange, 2021; Bloznelis, 2018; Dahl et al., 2021; Ewald et al., 2022). Building upon Dahl et al. (2021)'s finding that stock prices assimilate salmon price information more swiftly than the salmon futures exchange market, we focus on stock prices. For our textual analysis, we employ news articles retrieved from Intrafish (<https://www.intrafish.com/>), spanning from January 2010 to July 2022, and filtered using the keywords "salmon", "finance", and "prices". This approach enables us to achieve a comprehensive understanding of the factors that drive investors' behaviour within the salmon market, extending beyond the interplay between supply and demand.

The remainder of this paper is organised as follows. Section 3.2 elaborates on the data sets incorporated and the pre-processing methods employed. Section 3.3 outlines the methodologies and techniques utilised to estimate the relationship between salmon market investors' behaviour and related financial news. Section 3.4 presents the results, and Section 3.5 discusses the primary conclusions.

3.2 Data & pre-processing

3.2.1 Financial data

We obtained daily stock prices from Refinitiv for the period spanning January 2016 to July 2022 for five salmon-producing companies with the highest market capitalizations trading on the Oslo Stock Exchange (OSL). These companies are Marine Harvest (MOWI), SalMar (SALM), Grieg Seafood (GSF), Lerøy Seafood Group (LSG), and Bakkafrost (BAKKA). We selected Norwegian companies because Norway is the largest producer of salmon globally. Focusing on this particular group allows us to obtain a relatively homogeneous sample, as these companies are subject to similar seasonal, regulatory, political, and economic events. We believe that this focus enhances the robustness of our empirical analysis.

We choose the salmon market stock prices over the salmon spot or futures contracts because Dahl et al. (2021) found that stock prices reflect salmon price information earlier than the Fish Pool Index, the primary price index of farmed salmon. We construct a Share Price Index (SPI) corresponding to an equally weighted portfolio of the five stocks. We normalise the price index by dividing each price P_t , where $t = 1, 2, 3, \dots, T$, with the first price of the series P_0 , to demonstrate the growth rate of the shared prices. We

constructed the SPI instead of using individual share prices because an index should be less sensitive to company-specific information and better reflect on general news related to the salmon market. The index is depicted in Figure 3.1.

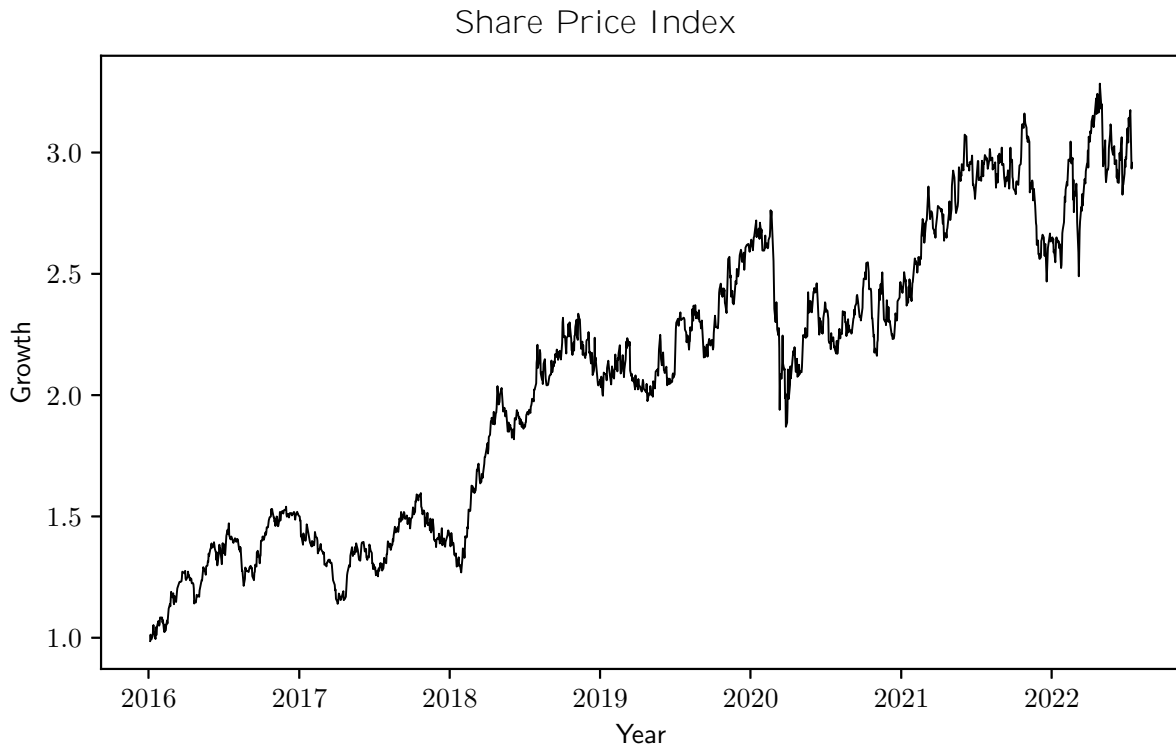


Figure 3.1: Share Price Index (SPI) for daily price data

To measure investors' behaviour we determine a reaction measure. The reaction is defined as the return from day to day, denoted as $Y_t = P_t/P_{t-1}$, where P_t is the SPI price at time t and P_{t-1} is the SPI price at time $t - 1$. To account for proportional changes in the returns, we calculate the logarithmic returns as follows:

$$R_t = \ln(P_t/P_{t-1}), \quad (3.1)$$

where R_t denotes the investors' reaction or logarithmic returns of SPI. The development of R_t is illustrated in Figure 3.2 where the impact of COVID-19 pandemic is evident.

3.2.2 Text data

We collected news articles that were published on IntraFish¹ between November 2012 and July 2022. Intrafish is one of the leading news sources in seafood, fisheries and aquaculture markets. It is based in Norway, but reports in English and thereby makes the news about companies such as those in our index accessible to international investors. Hence we

¹<https://www.intrafish.com/>

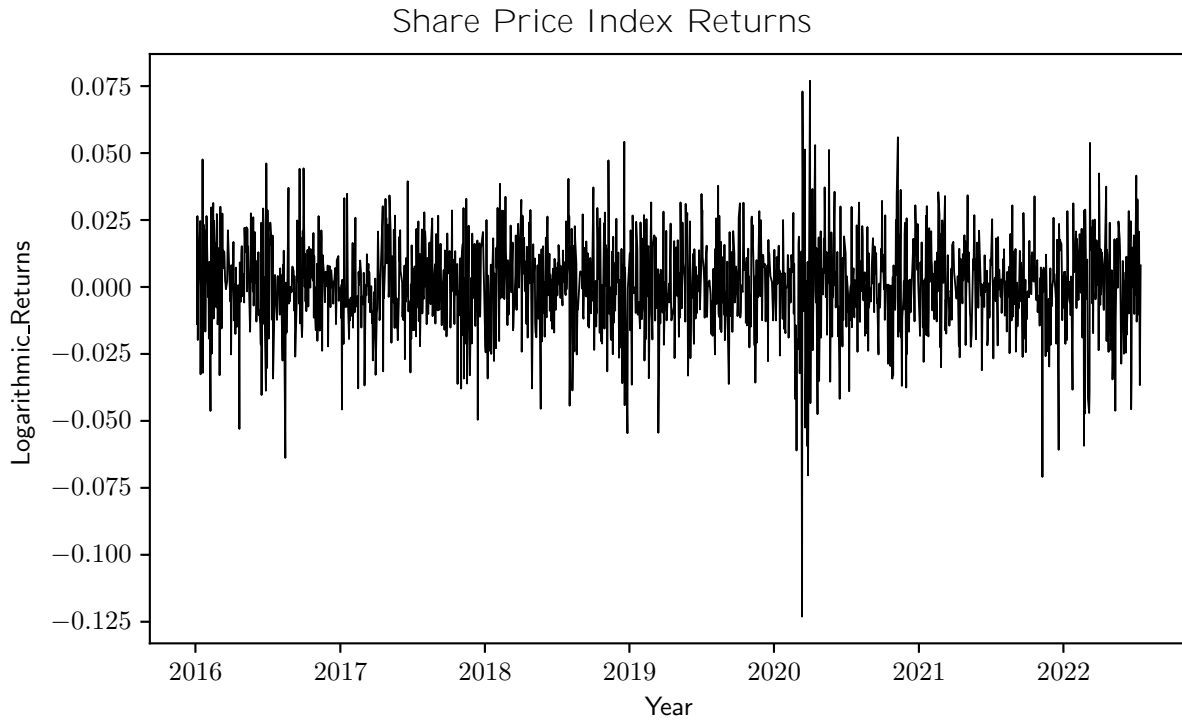


Figure 3.2: Investors' Reaction or SPI logarithmic returns

expect most, if not all relevant events to be reported here, while we also expect investors of these companies to be able to react to the respective news articles.

We performed a keyword-based search, filtering articles containing at least one of the following terms either in the text or in the metadata: “salmon”, “prices”, or “finance”. Our scope encompassed all articles relevant to the salmon market, as well as financial articles relating to salmon and other fish or seafood commodities, which are prominent topics on the website. Prices partially determine the revenue of salmon producers and are therefore of high relevance. As there is a lot of news about other types of seafood industries (wild catch, different species, etc.), we deemed it necessary to filter the articles before further analyses. Due to a low number of articles published per month between 2012 and 2016, we limited the scope to articles published from 1 January 2016 to 8 July 2022 (see Figure 3.3).

In our next step, we filtered the dataset to focus only on articles that provide meaningful insights. To do this, we removed articles that contained many short entries, which were difficult to analyse using our chosen text mining methods. Specifically, articles with titles containing strings such as “IntraFish Price Tracker”, “Top Headlines”, “Top Stories”, “LIVE Updates”, “Reports”, and “Conference Updates” were discarded.

Further, we removed articles that held no relevance to our research question, such as those carrying promotional content indicated by phrases like “IntraFish Podcast”, “IntraFish App”, etc., in their titles. Articles with little text accompanying videos and

photos, denoted by “VIDEO:” or “PHOTOS:” in the title, along with some other short articles that seemed irrelevant, were also excluded from the analysis.

Moreover, we disaggregated blog posts into individual entries based on their posting times, recognizing that each small article in a blog conveys unique sentiment and discusses a distinct topic. By employing these filtering and transforming strategies, we enhanced the quality of our dataset.

Article Frequency

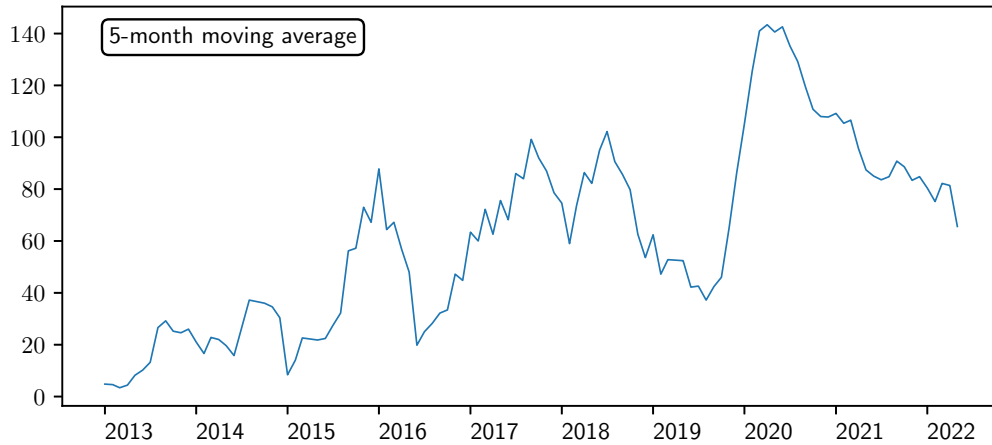


Figure 3.3: Number of articles per month, 5-month moving average

In the next step, we pre-processed the remaining articles. This involved resolving encoding issues and removing certain strings and text fragments that could potentially introduce noise to the data rather than contributing meaningful narratives. This included the removal of HTML tags, web addresses, and links to videos, photos, tweets, and ads. Text segments supporting graphs, photos, or tables within the articles were also discarded. We further omitted information that was quantitative in nature, such as the price per salmon weight class, export volume, and export earnings, as this information did not aid in either sentiment or topic identification. Frequently recurring text segments in articles that didn’t offer any new insights, such as the methodology for calculating the Nasdaq Salmon Index, were also eliminated. In the interest of brevity, we won’t list all the strings and text parts that were removed, but the guiding principle was to focus on text that is pertinent, provides new information to the market, and is compatible with our chosen text mining algorithms. Following this thorough cleaning process, articles with 20 or less words were deleted, as short text can hinder topic modelling effectiveness.

Before finalizing the text data pre-processing, we ensured consistency by converting all the articles to a uniform time zone - the Greenwich Mean Time (GMT). The final version of our news data, which comprises 6082 articles, includes the article’s title, the time it was posted, and the text content.

For alignment with the daily log returns data, we assigned dates to articles based on their potential impact on stock returns. Since the Oslo Stock Exchange (OSE) is open from 9:00 am to 4:20 pm Central European Summer Time (GMT+02:00), we perceived articles published after 2:20 pm GMT as potentially impacting the next trading day’s stock market. Additionally, articles published post 2:20 pm GMT on Fridays and over the weekends were attributed to influence the forthcoming trading day’s stock return. Likewise, articles released on public holidays - in accordance with the Oslo Stock Exchange’s operational hours on such days - were understood to bear an impact on the stock market on the subsequent trading day.

Following the completion of standard pre-processing and organisation of the textual data, we move on to more specialised pre-processing steps, tailored for topic modelling. Firstly, we transform collocations into single terms. Collocations are identified using part-of-speech patterns as proposed by Justeson and Katz (1995). These are combinations of two or three words that together hold a specific meaning, such as “salmon farming”, “United States”, or “earnings before interest”. In line with Hansen et al. (2018), we retain 194 two-word collocations with a frequency of at least 100 and 85 three-word collocations with a frequency of at least 50. This step effectively increases the vocabulary, as the input will now contain not only unigrams like “seafood”, “industry”, or “sales”, but also bigrams such as “seafood company”, “vice president”, and “salmon producers”, as well as trigrams like “Oslo Stock Exchange” and “Leroy Seafood Group”. The main goal of this approach is to ensure that important terminologies are accurately represented in the topic modelling process. For example, “earnings before interest” is a specific accounting term, which needs to be maintained as a single unit for accurate analysis.

Next, we transform all upper-case letters to lower-case and split contractions into their constituent words (e.g., ‘aren’t’ becomes ‘are not’). Subsequently, we carry out tokenisation, where each token represents a sequence of characters that are being treated as a group. We then remove non-alphabetic characters such as numbers, punctuation, currency symbols etc., as well as stop words. These stop words² are commonly used words that carry little standalone meaning (examples include ‘but’, ‘I’, ‘at’). Next, we remove IntraFish journalist names from each article as their high frequency within our corpus effectively categorises them as stop words.

A crucial step in our pre-processing is stemming. We utilise the Porter Stemmer, a popular algorithm for the English language, to reduce words of varying grammatical forms but with a common root to their base form or stem (for example, ‘consist’ and ‘consisted’ would be stemmed to ‘consist’). Additionally, this stemming process applies not only to single words but also to extracted collocations. For instance, the collocations “salmon producer” and “salmon producers” are both stemmed to “salmon produc”. By doing so,

²<http://snowball.tartarus.org/algorithms/english/stop.txt>

the same stem represents multiple forms of a word or phrase, significantly aiding in the standardization of the text data.

To manage dimensionality, we employ the Term Frequency-Inverse Document Frequency (TF-IDF) method as in Blei and Lafferty (2009) and Hansen et al. (2018). The TF-IDF score for each token v is calculated as:

$$\log(1 + N_v) \times \log\left(\frac{D}{D_v}\right) \quad (3.2)$$

where N_v is the count of the token v in the corpus, D_v is the number of articles that contain the term v , and D is the total number of articles in the corpus. Tokens appearing either very rarely or very frequently in all the articles have a lower TF-IDF score, so we discard the tokens with the lowest scores. At the end of this process, our corpus includes 11,666 unique stems, such as “outbreak”, “industri”, “seafood_industri”, and “norway_royal_salmon”, which are then used for the subsequent topic modelling.

3.3 Methodologies

3.3.1 Topic modelling

3.3.1.1 Latent Dirichlet Allocation (LDA)

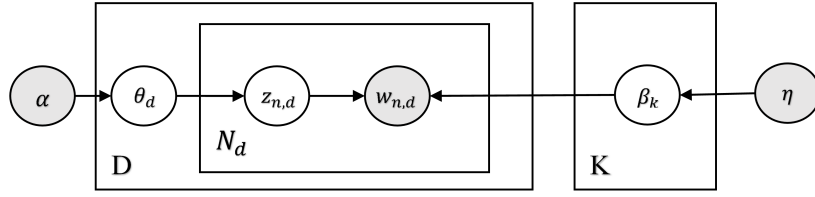
The Latent Dirichlet Allocation (LDA), an unsupervised generative probabilistic model, was initially developed by Blei et al. (2003). The central premise of LDA is that documents, in our case, news articles, are represented as random mixtures of latent topics. Each of these topics is defined as a particular distribution of words.

In a corpus of D articles, we can identify V unique words. An article d ($d \in 1, \dots, D$) is characterised by a collection of topics K . Each topic k ($k \in 1, \dots, K$) is a distribution $\beta_k \in \Delta^{V-1}$ over V unique words in the vocabulary. Note that our choice of notation is based on Hansen et al. (2018). The probability distributions of the K topics allow the same term to occur in different topics, potentially with varying weights. Thus, one can think of a topic as a list of words, each weighted to reflect their relevance to the topic. LDA is a mixed-membership model where each article is associated with multiple topics. Therefore, each article d is described by its own distribution $\theta_d \in \Delta^{K-1}$ over the K topics.

The Latent Dirichlet Allocation (LDA) model’s inner workings can be best understood by visualizing the generative process, which details how documents are created according to the model’s perspective. Given a document d , let N_d denote the number of words it contains. We also consider two single-value prior hyperparameters, α and η . The data generating process according to LDA is as follows:

1. For each topic $k = 1, \dots, K$, draw a distribution over words $\beta_k \sim \text{Dirichlet}(\eta)$,

Figure 3.4: LDA algorithm schematic



Notes: The schematic illustrates the structure of the LDA model. Shaded circles are observed ($w_{n,d}$) or hyperparameters of the Dirichlet priors (α, η), while white cells are latent, unobserved random variables. The underlying assumption is that a document d is a mixture over topics (θ_d), leading to topic assignments ($z_{n,d}$) of the word n in document d . Meanwhile, topics (β_k) are probability vectors over words that are observed in the corpus. The same words can have non-zero weights in several topics, but a word within a specific document is associated with one topic only.

independently.

2. For each document $d = 1, \dots, D$, draw a distribution over topics $\theta_d \sim \text{Dirichlet}(\alpha)$, independently.
3. For each document $d = 1, \dots, D$, and for each word $n = 1, \dots, N_d$ within the document:
 - a) Draw a topic assignment $z_{n,d} \sim \text{Multinomial}(\theta_d)$, where $z_{n,d} \in \{1, \dots, K\}$.
 - b) Draw a word $w_{n,d} \sim \text{Multinomial}(\beta_{z_{n,d}})$, where $w_{n,d} \in \{1, \dots, V\}$.

This conceptual generative process can be conveniently visualised as a directed acyclic graph (see Figure 3.4). Notations used in Figure 3.4 are shown in Table 3.1.

Within the LDA model, both θ_d and β_k are treated as random variables. The necessity for a Bayesian formulation of the model is primarily motivated by the extensive number of parameters to be estimated, namely $K \times V$ for β_k and $D \times K$ for θ_d . Dirichlet priors are an optimal choice due to their conjugacy with the multinomial distribution. The hyperparameters η and α govern the sparsity of β_k and θ_d , respectively. When α decreases, fewer topics exhibit a high probability, while the remaining topics maintain a small but positive probability.

Given the presumption that all hidden variables (topic-specific vocabulary distributions β_k , document-specific topic proportions θ_d , and per-word topic assignments $z_{n,d}$) and observed variables (words $w_{n,d}$, hyperparameters α and η) are known, the joint distribution of all these variables can be denoted as:

$$\Pr(B, T, z, w | \alpha, \eta) = \prod_{k=1}^K \Pr[\beta_k | \eta] \prod_{d=1}^D \Pr[\theta_d | \alpha] \prod_{n=1}^{N_d} \Pr[w_{n,d} | \beta_{z_{n,d}}] \Pr[z_{n,d} | \theta_d]. \quad (3.3)$$

Here, we have defined $B = (\beta_1, \dots, \beta_K)$, $T = (\theta_1, \dots, \theta_D)$, $z_d = (z_{1,d}, \dots, z_{N_d,d})$, $z = (z_1, \dots, z_D)$, and $w = (w_1, \dots, w_D)$.

However, the actual challenge lies in inferring the latent variables β_k , θ_d , and $z_{n,d}$ from the observed documents. Hence, the posterior of interest is $\Pr(B, T, z|w, \alpha, \eta)$.

LDA Parameters

Notation	Definition
α, η	Hyperparameters of the Dirichlet prior distributions
β_k	The distribution over words
K	The number of topics
θ_d	The article-specific topic distribution
$z_{n,d}$	The assignment of a word w_n ($n \in 1, \dots, N_d$) in an article d to a given topic
$w_{n,d}$	The observed word w_n in an article d
N_d	The number of words in an article d
D	The number of articles in the data set

Table 3.1: Notations in LDA

3.3.1.2 Estimation

To estimate the parameters of the LDA model, we apply collapsed Gibbs sampling, an approach introduced by Griffiths and Steyvers (2004). The essence of this method lies in the conjugacy of the Dirichlet prior to the multinomial distribution, enabling us to analytically marginalise the parameters B and T out of the joint distribution $\Pr(B, T, z, w|\alpha, \eta)$. Consequently, we can express the probability of the observed and latent variables as $\Pr(z, w|\alpha, \eta)$.

In practical terms, our goal is to determine the posterior $\Pr(z|w, \alpha, \eta)$, as the topic assignments z are not observed. Using conditional-marginal factorisation of the joint probability, this posterior can be expressed as:

$$\Pr(z|w, \alpha, \eta) = \Pr(z_{n,d} = k|z_{(-n,d)}, w, \alpha, \eta)\Pr(z_{(-n,d)}|w, \alpha, \eta). \quad (3.4)$$

Consequently, the computational task is simplified to sampling the topic assignments $z_{n,d}$ for each word in the corpus, given all the words w and other topic assignments $z_{(-n,d)}$. The advantage of this procedure is that it eliminates the need to sample topic proportions θ_d and topic-specific vocabulary distributions β_k .

A complete derivation of the conditional distribution $\Pr(z_{n,d} = k|z_{(-n,d)}, w, \alpha, \eta)$ can be found in the technical appendix of Hansen et al. (2018). The resulting distribution is expressed as:

$$\Pr(z_{n,d} = k|z_{(-n,d)}, w, \alpha, \eta) \propto \frac{m_{v,-(n,d)}^k + \eta}{\sum_v m_{v,-(n,d)}^k + V\eta} \times (m_{k,-n}^d + \alpha), \quad (3.5)$$

where $m_{k,-n}^d$ represents the count of words in document d assigned to topic k , excluding the current assignment $z_{n,d}$, and $m_{v,-(n,d)}^k$ is the number of occurrences of word $w_{n,d}$ assigned to topic k throughout the corpus, excluding the current assignment $z_{n,d}$.

Intuitively, the probability of assigning the current word $w_{n,d}$ to topic k increases if many other words in document d are also assigned to topic k and if the word $w_{n,d}$ has a high probability under topic k .

With the conditional distribution at hand, we can now proceed to detail the collapsed Gibbs sampling algorithm. Firstly, we initialise the topic assignment variables z to the values in $\{1, \dots, K\}$ by randomly drawing $z_{n,d}$ from a uniform distribution. For each document $d = 1, \dots, D$ and each word $n = 1, \dots, N_d$, we sequentially draw a new topic assignment $z_{n,d}$ through multinomial sampling using Eq. 3.5, based on all the updated topic assignments $z_{(-n,d)}$. We then repeat this procedure for iterations 2 to 4000 as part of a burn-in phase and again for iterations 4001 to 8000, keeping samples with a thinning interval of 50 to ensure that the autocorrelation between samples is low. To be precise, we retain 80 samples corresponding to iterations $\{4050, 4100, \dots, 8000\}$.

We need to choose three parameters to estimate the model: hyperparameters α and η , and the number of topics K . The values for hyperparameters are derived from Griffiths and Steyvers (2004) and are $\alpha = 50/K$, and $\eta = 200/V$. The number of topics is set to 100.

The collapsed Gibbs sampling procedure yields a set of samples with estimated topic assignments z . Yet, we do not gain direct insights about the primary parameters of interest β_k and θ_d . For each stored sample, we can estimate topic-specific vocabulary distributions and document-specific topic proportions using predictive distributions over new topics and new words. The probability that a new $(N_d + 1)$ -th word in a document d is assigned to topic k is given by

$$\hat{\theta}_d^k = Pr(z_{(N_d+1),d} = k | z_d) = \frac{m_k^d + \alpha}{\sum_{k=1}^K (m_k^d + \alpha)}, \quad (3.6)$$

where m_k^d is a count of words in document d assigned to topic k .

In a similar fashion, the predictive distribution over new words is expressed as

$$\hat{\beta}_k^v = Pr(w_{(N_d+1),d} = v | w, z) = \frac{m_v^k + \eta}{\sum_{v=1}^V (m_v^k + \eta)}, \quad (3.7)$$

where m_v^k is the count of times word $w_{n,d}$ is assigned to topic k in the entire corpus. We estimate β_k and θ_d for each iteration in $\{4050, 4100, \dots, 8000\}$.

We use a measure called perplexity to determine if the chain has converged. The formula for perplexity is given by

$$\exp \left[- \frac{\sum_{d=1}^D \sum_{v=1}^V n_{d,v} \log \left(\sum_{k=1}^K \hat{\theta}_d^k \hat{\beta}_k^v \right)}{\sum_{d=1}^D N_d} \right], \quad (3.8)$$

where $n_{d,v}$ is a count of word v in document d and $\hat{\theta}_d^k$ and $\hat{\beta}_k^v$ are introduced above.

This is a measure of how well the LDA model fits the data. Perplexity, often used in topic modelling, is monotonically decreasing in the log-likelihood of the unobserved documents. Therefore, a model that predicts the data well has a low perplexity. The first 4000 replications of the chain are characterised by rapidly decreasing perplexity values (see Figure 3.5 and note that we do not save perplexity for the first 500 iterations) and hence are discarded.

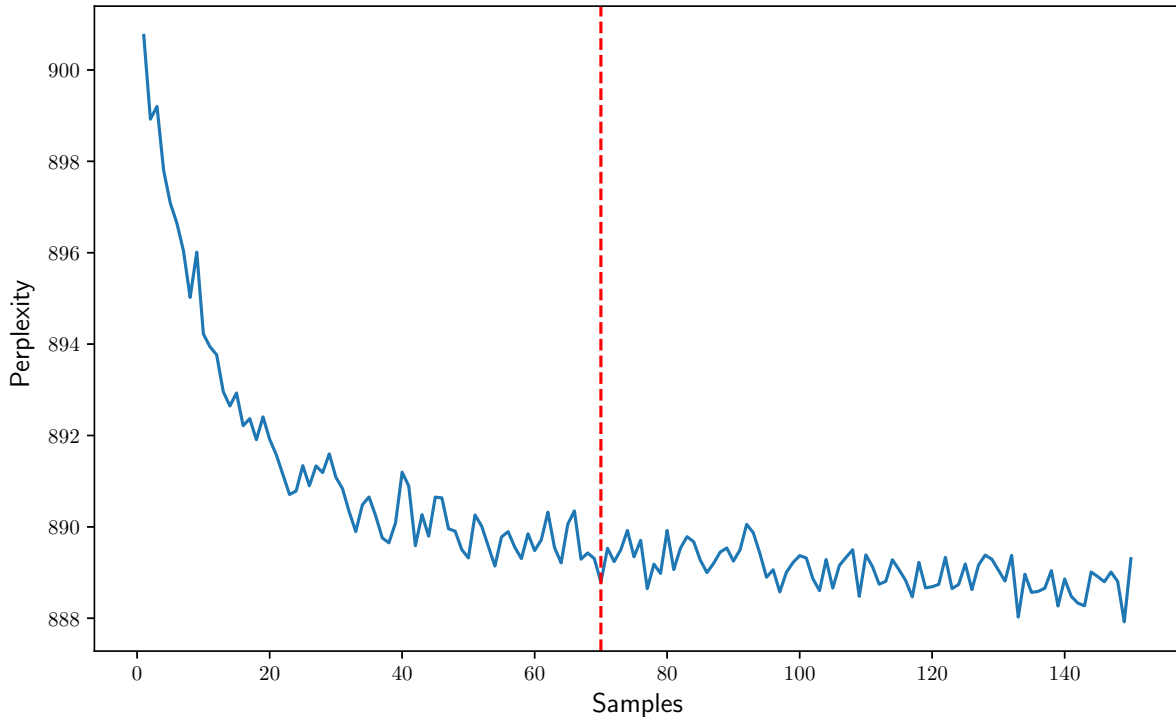
The predictive document-specific topic distribution $\hat{\theta}_d$ is the average over 80 samples. To understand the content of the estimated topics, we save the most probable stems under predictive topic-specific vocabulary distributions $\hat{\beta}_k$, obtained at the 8000th iteration. In Table 3.3, we list the estimated topics that are further discussed in Section 3.4. For instance, the first topic in this table, identified as T3, corresponds to the topic-specific vocabulary distribution $\hat{\beta}_4$ (our topic IDs start from T0). Under this topic, the most probable stems include “aquabouti”, “gm”, “commerci”, and “fda”, as well as the collocation “salmon_produc”. This shows our approach where the vocabulary integrates not only individual words like “aquabouty”, which is stemmed to “aquabouti”, but also important collocations like “salmon_producer”, which is stemmed to “salmon_produc”. The topic T3 was subsequently labeled as “Aquabouty Land-based & GM”, illustrating that the estimated distributions are highly interpretable, with the most probable stems clearly reflecting the associated themes. The predictive document-specific topic distribution $\hat{\theta}_d$ is the average over 80 samples.

Finally, to estimate daily topic frequencies, we collapse all the articles for one specific day into one document. Following Hansen et al. (2018), we re-sample topic assignments $z_{n,d}$ for day-specific articles. The topics in this re-sampling step are kept fixed at the values of their predictive distributions. Therefore, we use only 20 iterations of the Gibbs sampler in this step. After re-sampling, we obtain the predictive day-specific topic distribution.

3.3.1.3 Topic model example

The result of this modelling and estimation procedure is illustrated with the example of one topic and an article in Figure 3.6. The word vector β_k representing the topic is visualised as a word cloud with font sizes that are proportional to word weights in the topic (Panel 3.6a). This particular visual representation allows one to understand the subject of the topic at a glance, in this case the Covid pandemic, particularly in the context of foodservices and retail.

Figure 3.5: Perplexity of the 100-topic LDA model estimated on the training data



Notes: The figure shows perplexity values along the chain drawn for the 100-topic model, corresponding to samples ranging from 1 to 150. These sample indices align with perplexity calculations performed across specific iterations: $\{550, 600, \dots, 8000\}$. We have retained the last 80 samples, representing the point at which the chain has converged.

The related time series of the average weight of this topic throughout the articles of the same day serves as a visual validation of the approach, as the topic rapidly shoots up in early 2020 and slowly decreases thereafter. It is not exactly zero before the onset of the pandemic and shows some variation, as some of the words in this topic appeared pre-2020. Moreover, the Dirichlet prior α creates a positive bias, so that no topic can never be exactly zero.

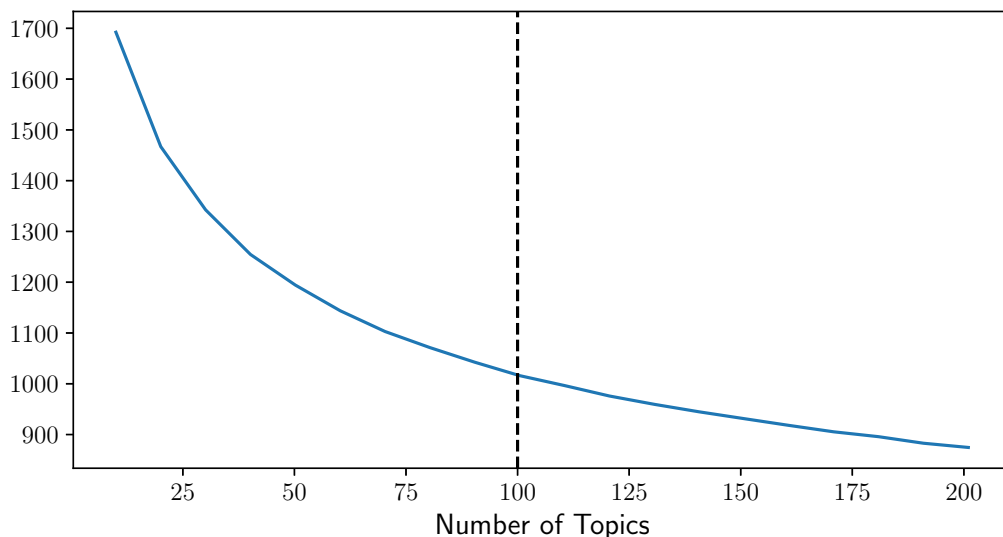
An excerpt of the topic distribution θ_d for a specific article (see Panel 3.6c for a screenshot from the Intrafish website) is shown in Panel 3.6d, with the topic example indicated in bold font. This article had the strongest value for this particular topic within the entire corpus, with a probability of it appearing of almost 20%.

3.3.1.4 Cross-validation

To choose the optimal number of topics K we employ 10-fold cross-validation. In this process, we randomly split the article data set into 10 folds. Next, we estimate the LDA model on the first 9 folds that represent the training set and keep the last 80 samples where we have observed convergence of the chain (see Figure 3.5). Thereafter,

The final perplexity for the 10th fold is calculated by averaging over the 80 samples. This procedure is repeated 9 more times, each time changing the fold that serves as the test set. The final perplexity is calculated as the average across the 10 folds. We choose the optimal number of topics K based on Figure 3.7, where the horizontal axis corresponds to the number of topics and the vertical axis to the perplexity values. Even though the perplexity measure continues to decrease for numbers of topics $K > 100$, we choose it as the cut-off, since the marginal improvements due to additional topics become relatively small, and the number of topics should remain small relative to the number of documents ($D = 6082$). Also, as the number of topics becomes too large, they tend to be overly detailed and less interpretable, reducing their utility.

Figure 3.7: Perplexity for different numbers of topics



Notes: The figure shows the average perplexity of test data for different topics, calculated according to the formula presented in Eq. 3.8. These data show that as the number of topics increases, the goodness-of-fit of the model improves. Given the relatively small size of our data set, we choose 100 topics, as the improvements thereafter are non-substantial.

After determining that the optimal number of topics K is 100, based on our statistical criterion of perplexity, we thoroughly evaluated the interpretability and coherence of the estimated topics. All topics that will be discussed in Section 3.4 are presented in Table 3.3. These topics not only adhere to the statistical criterion but are also easily interpretable and closely aligned with significant events, companies, and themes pertinent to the salmon market. For instance, topics T24 and T61 concentrate on industry challenges such as escapes and algal blooms, respectively. The table also includes analyses of the salmon market (topics T28, T39, T51, and T59), discussions on stock market and business outcomes (topics T41, T53, T68, and T69), insights into companies listed on our index (topics T33, T56, and T97), their competitors (topics T3, T11, and T52), and major

events like COVID-19 (topics T36 and T83). Collectively, these topics validate our choice of $K = 100$, demonstrating both statistical robustness and practical relevance.

Once the number of topics is set, they are algorithmically estimated without any human bias. The topics selected for detailed discussion in Section 3.4 serve one of three purposes: firstly, to identify topics have a significant impact on market volatility and are well-represented in our dataset, such as topics 36 (Covid Pandemic), 83 (Covid Cases in Production Facilities), and 41 (Business Results). Secondly, to explore topics that, when adjusted for our extended sentiment, influence stock market returns, exemplified by topics 35 (Fear, Harm, and Negative Outlook) and 43 (Plans and Strategy). Lastly, to address specific limitations of the Loughran-McDonald (LM) sentiment dictionary in our application, such as topic 36 (Covid), and to underscore the importance of incorporating competition-specific issues (topic 61 about algal blooms) and domain-specific vocabulary (e.g., topic 23 on diseases).

3.3.2 Sentiment approach

We believe that both topics and sentiment might be important for explaining stock returns because they provide different types of information. Topics refer to the content of the news, or what is being discussed, while sentiment refers to the emotional tone or valence of the news, or whether it is perceived as positive, negative, or neutral.

To quantify sentiment, we have chosen to use a lexicon-based approach. A lexicon-based approach involves using a dictionary of words or phrases and their corresponding sentiment scores to determine the overall sentiment of a text. The advantages of this method are that it is relatively simple, efficient, and does not require an annotated training set. Additionally, it is more interpretable than machine learning methods, as it allows researchers to examine the specific words and phrases that contribute to the overall sentiment score.

The literature on sentiment analysis differentiates between general and domain-specific lexicons. General lexicons are designed to be applicable across a range of different contexts. In contrast, domain-specific lexicons are tailored to a particular topic, market, or industry. For instance, in the finance literature, Tetlock (2007) examined the relationship between sentiment expressed in the news and stock market movements by using the general Harvard IV dictionary. However, Loughran and McDonald (2011) argued that a domain-specific lexicon is better suited for capturing the tone in financial texts because certain words that may have a negative sentiment in a general lexicon (such as “liability”, “cost”, or “capital”) are rather neutral in the finance domain.

Our approach to sentiment computation is similar to Shapiro et al. (2022). The authors highlight that combining lexicons can improve sentiment analysis performance, particularly when using domain-specific lexicons. In our study, we combine the Loughran-

McDonald (LM) dictionary, which is specific to the finance industry, with our own domain-specific dictionary tailored to the salmon market.

We develop a daily sentiment index by following a systematic approach. Firstly, we create a list of seed sentiment words and phrases based on our industry expertise and initial analysis of the news articles. For instance, we consider “algal bloom” as a negative phrase since the rapid growth of algae in the water can have detrimental effects on the health and survival of salmon, while “recapture” is a positive word since it implies that the salmon that have escaped from farms have been successfully brought back. We include all grammatical forms of seed words to expand our list. Secondly, we estimate a word2vec model on our corpus to identify data-driven synonyms for the seed words. Thirdly, we manually select relevant synonyms and their grammatical forms to extend our dictionary. Fourthly, we combine our dictionary with the LM dictionary.

We then use this combined dictionary to estimate sentiment scores for each news article, where sentiment is defined as the number of positive words minus the number of negative words divided by the total number of words in the text. Numbers are not included in the total word count. To account for negation, we multiply the sentiment scores of words by -1 if the word is preceded by a negation term (such as, but not exclusively “neither”, “never”, “not”, or “no”) within a three-word window. For instance, consider the sentence, “It was not a bad day for the salmon market”. Although the word ‘bad’ conveys a negative sentiment, the presence of the negation ‘not’ just before ‘bad’ reverses its sentiment impact. Thus, ‘bad’ is adjusted to reflect a positive sentiment, accurately capturing the intended positive nuance of the statement. Once we have the sentiment score for each article, we compute the daily sentiment index by taking the average of the sentiment scores across all articles on that day.

3.4 Empirical results

3.4.1 The impact of industry-specific topics

To investigate the drivers of behaviour in financial markets, we conducted Principal Component Analysis (PCA) on a set of estimated topics, in line with Larsen (2021). By extracting principal components, we effectively reduced the data’s dimensionality, which enabled us to concentrate on major “themes”, i.e. uncorrelated linear combinations of original topics, that might impact returns. For the insights from this analysis to hold value, it is essential that these themes are highly correlated with the topics that are related to each other in a meaningful way. In order to understand and appropriately label the themes, we pinpointed the five topics exhibiting the highest absolute correlation coefficients with each component.

The importance of dimensionality reduction in our study lies in the nature of the topics. As demonstrated in Table 3.3, certain topics tend to cluster around the same types of information, leading to high correlation among them. For example, topics like T36 (Covid Pandemic) and T83 (Covid Cases in Production Facilities) both focus on the impacts of the Covid pandemic; similarly, T76 (Chilean Producers) and T52 (Chilean Salmon Farming) discuss related aspects of the Chilean salmon industry, while T41 (Business Results) and T68 (Business Data) both concentrate on business performance metrics. By employing PCA, we linearly combine these correlated topics into components that are more likely to have a pronounced effect on market behavior.

The inclusion of absolute returns in this stage of analysis is necessary because topics can be framed positively or negatively. While some topics may be considered inherently “good” or “bad” news, influencing the markets correspondingly, other topics may be directional in nature, impacting markets positively or negatively based on their portrayal. To isolate the impact of specific themes on returns, we account for these effects by combining the topics with sentiment.

However, at this point, our focus is primarily on the topics themselves, and less so on the sentiment surrounding them. We are therefore concentrating on the absolute value of returns rather than simply returns. Our aim is to uncover which topics hold relevance to the market, as this would manifest through market volatility. We use standard bivariate vector autoregressions (VAR) to identify the dynamic responses of absolute log returns to surprise increases in the identified principal components. In these VAR models, our text measures are ordered first, and the number of lags is determined based on the Akaike Information Criterion (AIC). Furthermore, we generate cumulated Impulse Response Functions (IRFs) for the subsequent 20 working days to examine the continued impact of these surprise increases.

Our analysis reveals several components that exhibit a significant impact on absolute returns, indicating a pronounced response from investors. The accompanying VAR results of a few chosen ones are displayed in Figure 3.8. Moreover, other components are interesting due to the particular combination of topics that these components are most highly correlated with. Table 3.4 in the Appendix 3.A provides more details on some of these components, including their respective shares of total variation in topics. It also lists the five topics that exhibit the highest absolute correlation with each component. Detailed discussions of all analysed components and their impact on market volatility are presented in the subsequent Subsections 3.4.1.1 and 3.4.1.2.

3.4.1.1 PCA results

The first component, which accounts for approximately 4% of total variation in topic values, is labelled “Business figures”. This label has been assigned due to the component’s

high positive correlation with topics related to results (topics 53 and 41), as well as changes in numbers (topic 67), and its negative correlation with topics 19 (public and social) and 12 (future challenges). The names of the topics listed in Table 3.4 are assigned based on the most probable words under each topic according to the estimated LDA model. For example, topic 53 is labelled “quarterly results” because the words with the highest probability under this topic are ‘quarter’, ‘earnings’, ‘revenue’, ‘tax’, ‘ebitda’, and the like. The fact that this component explains the highest percentage of variance in topic values shows that news on business aspects is covered substantially. This matters for our purposes of analysing the effect of these articles on returns of the companies and is hence a reassuring result.

Given the time frame of our data set, it is not surprising that the Covid pandemic had significant effects on stock markets. Components five and six are specifically related to the pandemic, with the former exhibiting a positive correlation with topic 36 (Covid news) and topic 83 (Corona infections in production facilities such as farms and processors). Additionally, component five displays negative correlations with topics 37, 28, and 10, which pertain to kilo prices for Norwegian salmon, salmon harvesting results, and outlooks on prices, respectively.

Components five and six both relate to Covid, but they differ in their specific focus. While component five combines news about Covid and production, the sixth component contains information on the role of Covid in salmon markets. This is indicated by the high and positive correlation (36%, see Table 3.4) of this component with topic 13, which is frequently featured in articles discussing business challenges, risk, and uncertainty in the markets. Moreover, the sixth component exhibits a negative correlation with topics on salmon commodity and wholesale prices (topics 59 and 39, respectively), as well as mergers and acquisitions.

Another component of interest is component nine, which focuses on salmon production in Chile. This is most notably indicated by its high and positive correlations with topics 76 and 52 (see Table 3.4), relating to Chilean producers and salmon farming in Chile, respectively. Interestingly, topic 61, the topic on algal blooms, correlates positively with the component, too. This suggests that algal blooms, deadly to salmon, are a concern that is more relevant (albeit not exclusively) to production in Chile. This does not come as a surprise as at the beginning of 2016, high and persistent harmful algal blooms (HABs) took place in the marine ecosystems of southern Chile. A major mortality event of about 27 million salmon (i.e. 39,000 tonnes) was caused by blooms in the Los Lagos Region (León-Muñoz et al., 2018; Montes et al., 2018). Indeed, the articles in our data set with the highest values for shares of topic 61, algal blooms, cover events taking place in Chile almost exclusively. Topic 33, which discusses SalMar’s offshore farming - a firm in our index - correlates negatively with component nine, indicating once again that this

component focuses more on the competitors.

On the other hand, Component 12 is most highly and positively correlated with topics that are relevant to our index: SalMar offshore farming, NTS acquiring Norwegian Royal Salmon (NRS), and project licenses in Norway. Again, it is worthwhile to investigate the topics that correlate negatively as well: here, the ones worth mentioning are topics 85 and 3, which concentrate on Atlantic Sapphire land-based farming, and Aquabounty land-based and genetically-modified (GM) farming (see Table 3.4). These two topics cover not only competitor companies, but also competitive technologies. The Norwegian and Faeroese (i.e. Bakkafrøst) companies in our index produce mostly in pens in the sea. Their competitive advantage lies in access to high-quality locations in natural waters. However, these locations are typically remote and far removed from consumer markets. Hence, land-based farming, which in the news reporting here is opposed to Norwegian sea-based farming, may become a significant threat to the companies in our index, in case it can be scaled to produce substantial amounts of high-quality salmon.

Finally, Component 17 is highly and positively correlated with topics 41 (business results), 48 (management and boards, i.e., high-level personnel of producers), and 69 (stock market news). This component covers various aspects related to business, which are expected to significantly influence investors' decisions.

3.4.1.2 VAR results

The effect of the first component on absolute returns is small, and only significant one day after a surprise increase in the component, as illustrated in Figure 3.8. While one could argue that the reported figures are results from past business activity, which may be partially priced in already, and that the future prospects may be more relevant, direct news about the state of the business would still be expected to matter for expectations on the next rounds of dividends, i.e. cash flow to stock holders. Moreover, it indicates whether a company is on a profitable course. As we will explain in more detail below, we rather interpret this as a need to provide more structure to our analysis, as business results could be reported for any company in the salmon industry, not only those reflected in our index. Especially the competitive relationships between salmon producers are shown to matter in the latter course of this analysis. Hence, all articles cannot be treated equally.

Among the components analysed, components five and six, related to the Covid pandemic, and component 17, concerning corporate and stock matters, seem to drive market volatility the most. As seen from Figure 3.8, following a surprise increase in component five, which we label as 'Covid and production', absolute log returns increase by 0.0013 (or 0.13 percentage points, given that log returns are expressed in percentages) four days post-shock, maintain that level for another 6 days, and then revert to the original level 10 days subsequent to the shock. An explanation for the link between Covid and volatility

may be attributed to the uncertainty surrounding the pandemic’s duration and severity, as well as its potential impact on the global economy. The negative correlations observed between component 5 and topics 37, 28, and 10 suggest that these topics have a calming effect on the markets. Although it may seem counter-intuitive that articles about prices and production results are related to lower volatility, it can be explained by the fact that such articles appear regularly and occupy significant shares of daily news only during periods of relative market stability. In times of turmoil such as the beginning of the Covid pandemic, these topics are overshadowed and occupy relatively little space in salmon-related reporting.

Given that component six specifically addresses Covid in salmon markets, it is not surprising that it has a stronger effect on absolute returns than component 5, as seen in Figure 3.8. Specifically, absolute returns increase immediately one day after a surprise increase in component six (labelled “Covid and markets”), and this rise continues for the next 20 days (excluding days 8 and 9), peaking at an increase of about 0.0025. Overall, this response is more persistent and prolonged. The aforementioned negative correlation of the sixth component with topics 59, 39, and 82 can be attributed to these topics being less prominent in discussions during turbulent times.

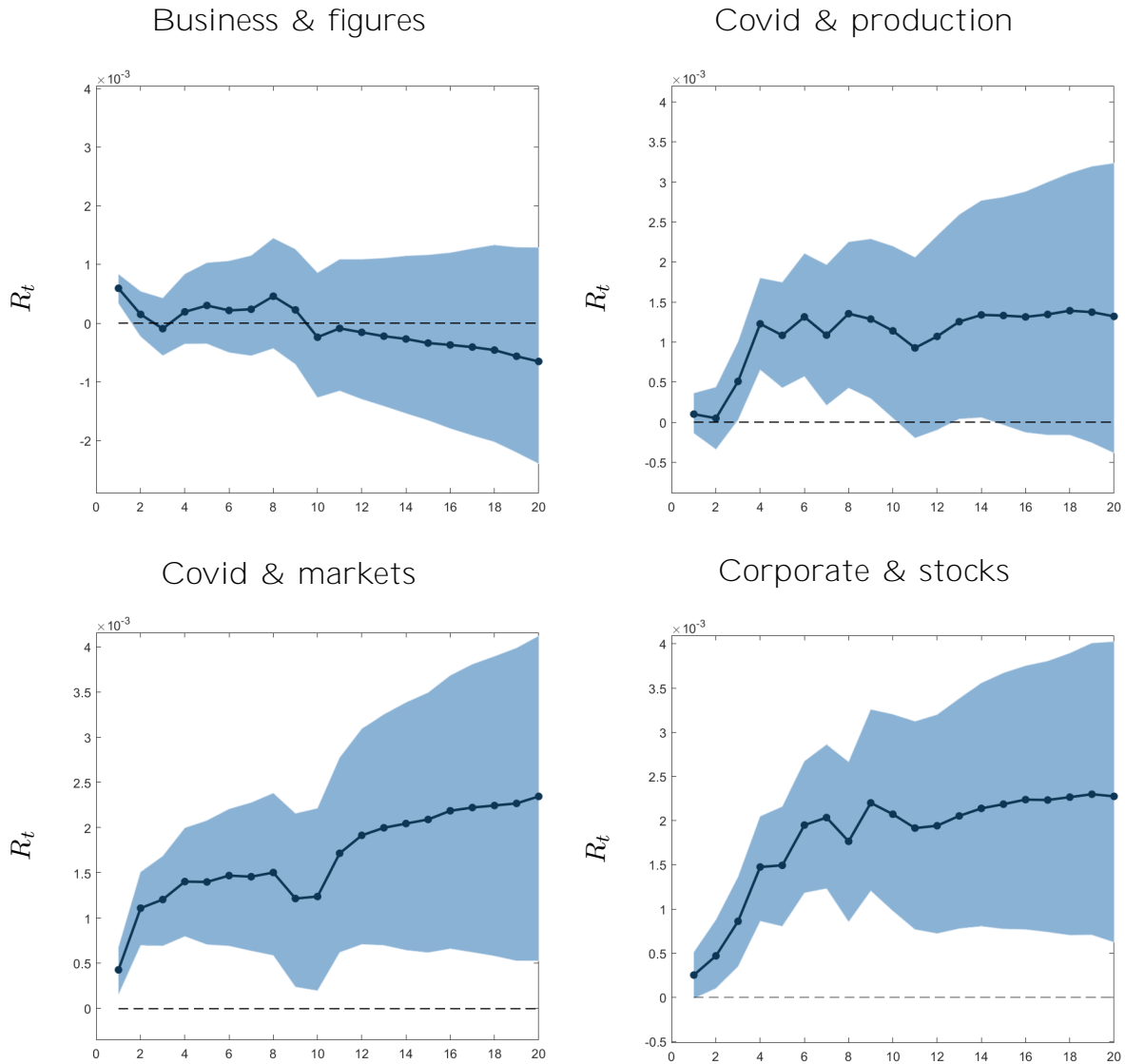
Shifting from the Covid-related components, we now turn our focus to another key driver of market volatility: Component 17, labelled “Corporate & Stocks”. This component exhibits a slowly building and prolonged influence on absolute returns, as depicted in the bottom-right panel of Figure 3.8. The peak positive response of absolute log returns occurs 9 days after the shock, and there is no evidence of a rebound effect.

Contrasting effects are observed with Component 9 and Component 12. In response to component 9 the absolute returns of the firms within our index are seen to experience a brief, significant decline for three days, before a swift recovery (IRF omitted). Hence, when news reporting is more focused, in relative terms, on issues in Chile, stocks of the firms in our index exhibit relatively less volatility. Conversely, Component 12, which concentrates on the companies from our index, has a positive and short-lived effect on market volatility (IRF omitted). These opposite effects underscore the importance of considering competitive dynamics within the salmon market in our analysis.

3.4.2 Incorporating lexicon-based sentiment

In the following, we multiply the daily topic values with the daily sentiment, specifically calculated using the LM dictionary often used in financial literature, see e.g., (Li et al., 2020). Our goal is to differentiate between positive and negative news concerning topics that inherently may not possess any sentiment. Given this distinction, we can proceed and analyse the effect of news on returns, including their direction. Prior to this step, human judgement was only involved in the choice of methods, as the procedures used were

Figure 3.8: Cumulated impulse responses for selected components



Notes: The figure shows cumulated impulse responses of SPI absolute returns to one standard deviation innovations in selected components. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

mathematical and algorithmic and did not require external guidance. However, the use of a dictionary approach for sentiment extraction introduces an element of subjectivity, as the choice of words is based on the expertise of the creators of the dictionary. In the subsequent Subsections 3.4.2.1, 3.4.2.2, and 3.4.2.3, we discuss the components extracted from sentiment-adjusted topics, explore their impact on log returns, and use the estimated topics to highlight limitations of the LM dictionary in our specific application.

3.4.2.1 PCA results

The PCA results are strikingly different to the analysis above. The topics that most highly correlate with extracted components, when multiplied with sentiment, have changed. Furthermore, the first component now explains 26.5% of total variation in the values of topics multiplied with sentiment, compared to 4% in the analysis without sentiment. We label this first component “business expectations”, due to its co-movement with topics 43 (plans & strategy), 35 (fear, harm, negative outlook), 68 (business data), 12 (future challenges), and 47 (contracts & agreements). The correlations between key components and the sentiment-weighted topics are provided in Table 3.5 in the Appendix 3.A.

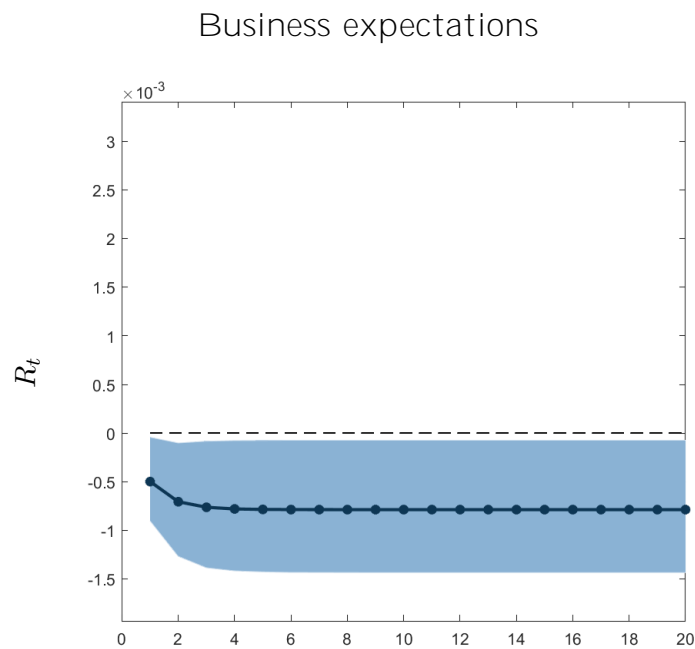
The other two components that significantly impact log returns are labeled “Investments Norway” and “Investments rest of the world (ROW)”, for components 3 and 4, respectively. Component 3 is closely associated with Norwegian companies, as it positively correlates with topics such as T33 Salmar Offshore Farming and T96 NTS acquires NRS, alongside T82 M&A and T70 Smolt Production Facilities Investment, emphasizing its focus on investments within Norway. Conversely, Component 4 broadens its scope to global markets, indicated by positive correlations with T14 Processor Acquisitions (UK) and T78 US Fund Raising, as well as T55 Strategy, Investments, Opportunities. It also shows a negative correlation with T97 (Mowi Business and Management), a Norwegian company, highlighting its international investment focus. These two components once again underscore the importance of competition in the salmon market.

The observation that some sentiment-weighted topics lack correlation with the components driving the market, as estimated in this section, is important. Specifically, topics related to Covid, which had shown to be particularly relevant in the analysis without sentiment, and biological aspects of salmon farming, such as algal blooms, do not seem to strongly correlate with the first, third, and fourth components when their values are multiplied with the LM sentiment. While it is anticipated that business and finance-related news would greatly influence returns, the lack of prominent correlation of these other topics with the components driving the market can be largely attributed to the nature of the sentiment dictionary used. The dictionary was initially designed to identify sentiment in business reports from a wide range of companies, and thus it effectively identifies sentiment in a business and finance-related context. Articles in which other topics are discussed are consequently evaluated to have neutral sentiment. As a result, variation in sentiment-weighted topics is driven by the topics that the sentiment dictionary is able to recognise, rather than topics that may be relevant to salmon producers specifically, such as those related to production processes.

3.4.2.2 VAR results

In this subsection, we analyse the impact of the extracted components on returns, demonstrating significant differences that arise when sentiment adjustments are considered. As shown in Figure 3.9, the first component has a marginally significant, small, and prolonged effect on logarithmic returns. The effect is negative, which, given that the topic values are multiplied with sentiment, stands in contrast to the expected positive effect. With sentiment included in the analysis, one would anticipate that a surprise increase in sentiment about business expectations would lead to an increase in logarithmic returns. However, we observe the opposite, i.e. a negative relationship. As this component does not differentiate between different producers, or clusters of producers, the negative correlation is mostly driven by the fact that there are only five firms in our index, and many more competitors, and hence there are also more news articles about other firms, which hence dominate the effect.

Figure 3.9: Cumulated impulse response of stock returns to innovation in component 1

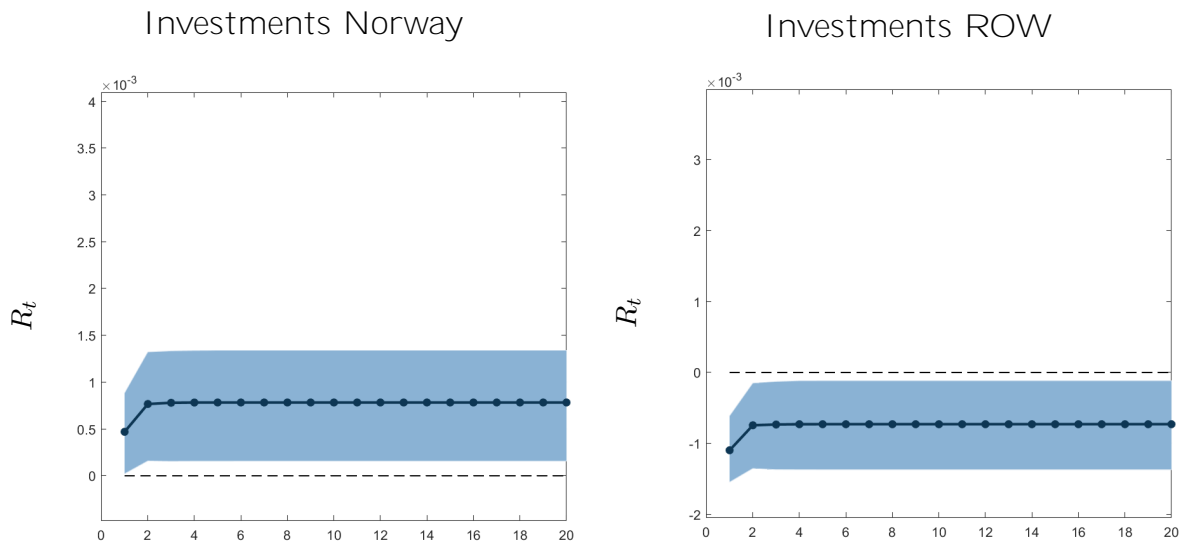


Notes: The figure shows the cumulated impulse response of logarithmic stock returns to one standard deviation innovation in component 1 labelled “Business expectations”. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

Components 3 and 4, and their respective effects on log returns, reveal one reason why sentiment does not function as expected. As shown in Figure 3.10, a surprise increase in

sentiment about investments in Norway is followed by an increase in returns. Conversely, an unanticipated increase in sentiment regarding investments in the ROW has the opposite effect. This difference in market reaction hints to the prevailing effects of competition amongst firms in salmon markets. Increased sentiment regarding investments into the productive capacities of firms reflected in our index raises the expectations of future cash flows, and hence returns increase, and vice versa for a decrease in sentiment about investment projects. Investments in the rest of the world, however, increase competitive pressures on the companies in our index. Hence, a rise in sentiment about those projects relates to declines in returns of our index, and vice versa.

Figure 3.10: Cumulated impulse responses of stock returns to innovations in components 3 and 4



Notes: The figure shows the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in components 3, labelled "Investments Norway", and 4, labelled "Investments rest of the world (ROW)". Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

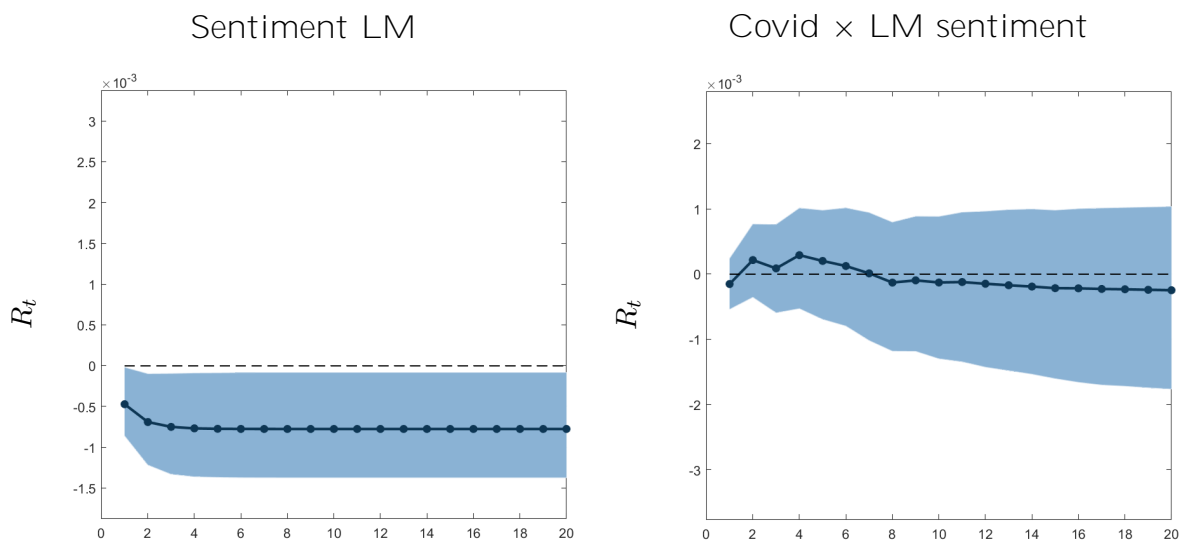
3.4.2.3 Critical assessment of the LM dictionary through estimated topics

The weaknesses of LM dictionary are clearly shown in Figure 3.11, which portrays the effect of LM sentiment (left panel), as well as of topic 36 (Covid) multiplied with LM sentiment (right panel) on logarithmic returns. It is generally anticipated that an increase in sentiment would result in a corresponding rise in logarithmic stock returns. However, an unexpected increase in LM sentiment is observed to precede a marginally significant and prolonged decrease in these returns (Figure 3.11, left panel). Moreover, an unexpected

surge in sentiment about topic 36, Covid, does not have a statistically significant influence on logarithmic returns (Figure 3.11, right panel). The unusual negative correlation between returns and LM sentiment could potentially be explained by the dictionary's inability to account for competition dynamics in the salmon market. As discussed earlier, sentiment is calculated based on articles not only about the firms in our index but also their numerous competitors. At the same time, the insignificant effect observed from the sentiment-weighted Covid topic could suggest the LM dictionary's limitations in accurately capturing the sentiment associated with non-financial news.

Moreover, we observe a striking similarity between the impulse response of stock returns to LM sentiment and to the first component. This is likely due to every daily topic value being scaled by the LM sentiment value for that same day, resulting in the variation in sentiment-weighted topics being dominated by the variation in LM sentiment. Consequently, the topics that correlate strongly with the first component serve as a reliable indicator of the type of news that the sentiment index captures - primarily, financial and business-related news in this case.

Figure 3.11: Impulse responses for LM sentiment and topic 36 (Covid) multiplied with LM sentiment



Notes: The left panel shows the cumulated impulse response of logarithmic stock returns to one standard deviation innovation in LM sentiment. The right panel shows the cumulated impulse response of logarithmic stock returns to one standard deviation innovation in topic 36 (Covid) multiplied with LM sentiment. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

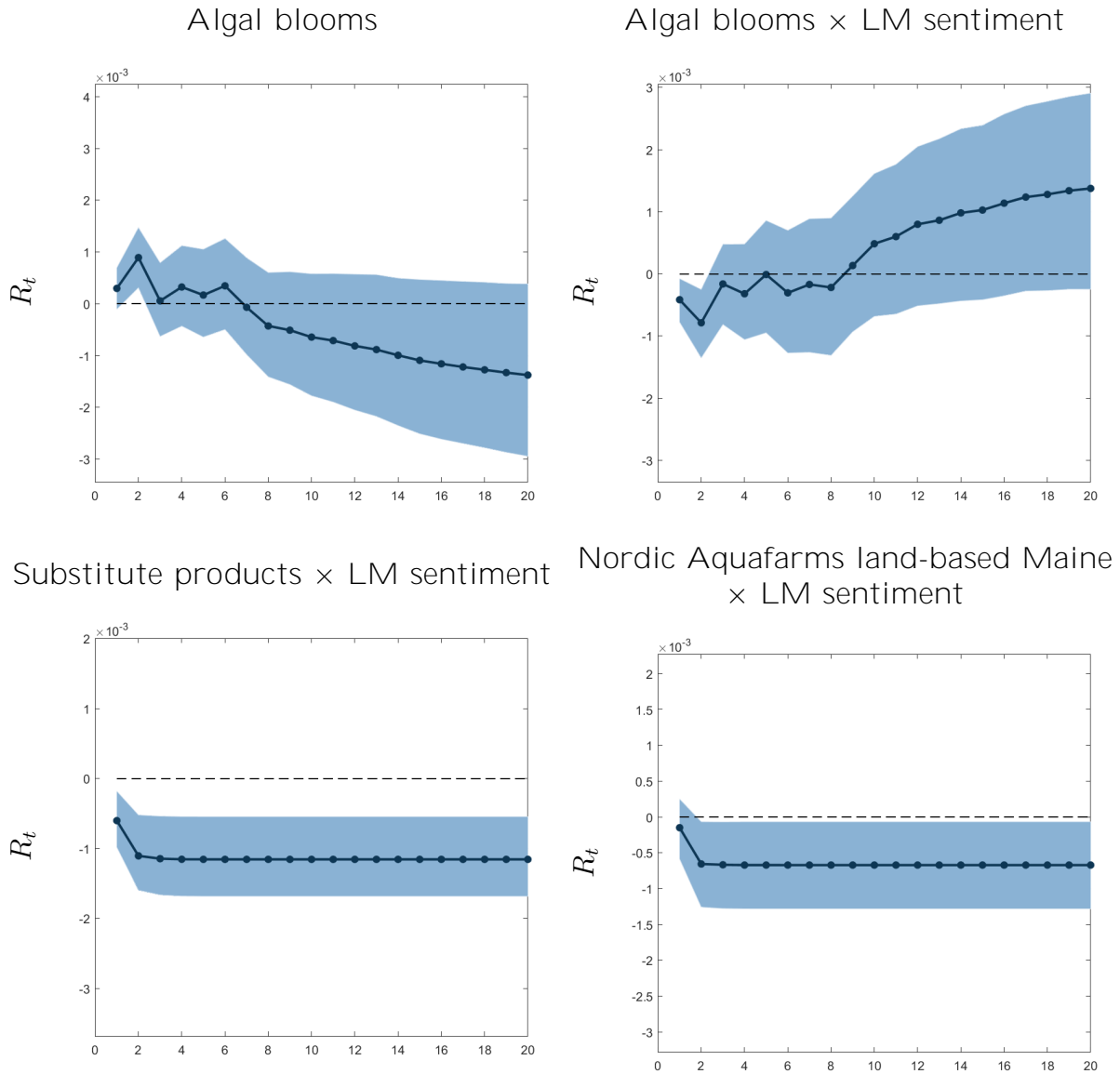
Another example highlighting the importance of considering competition in the salmon market when computing sentiment is topic 61, algal blooms. This topic correlates strongly

and positively with component 9, which pertains to salmon production in Chile and is extracted based on topics only. This strong correlation suggests that during the period our data set covers, algal blooms were a more prominent issue in that part of the world. This is confirmed by examining the articles that exhibit a high proportion of the topic about algal blooms. These articles are almost exclusively related to aquaculture in Chile. The positive correlation of this topic (not multiplied with sentiment) with the logarithmic returns of our index (top left panel in Figure 3.12) reveals the competition effect in the data. Since news about algal blooms is reported more frequently in relation to Chilean farms, we observe a positive correlation with log returns of our index. Some algal blooms that had been reported in this period led to devastating losses in Chilean farms, hence creating substantial contractions in total supply. For firms that did not suffer from these losses, the supply shock could only be experienced as increased prices on global salmon markets, thus increasing profits. Hence, this example vividly demonstrates the competition effect: the logarithmic return of our index reacts positively to news about deadly algal blooms, but the sign of the correlation changes once we account for sentiment (top right panel in Figure 3.12). This means that a surprise increase in sentiment about algal blooms was followed by a decrease in log returns of our index. It is worth noting that the LM dictionary effectively captures this topic, given the presence of words such as “loss”, “incidence”, and “suffer” that are denoted as negative in the dictionary and also bear high probabilities in the distribution of the topic on algal blooms.

Technological competition and competitive products are two specific kinds of competition effects that might partially explain the negative correlation between the logarithmic returns of our companies and the sentiment index calculated with the LM dictionary. For instance, topic 25 focuses on Nordic Aquafarms’ land-based salmon farming project in Maine. When multiplied with LM sentiment, it exhibits a negative correlation with the stock returns of our index (see the bottom right panel of Figure 3.12). Land-based production is a new and innovative technology that could potentially threaten the competitive advantage of traditional salmon producers in remote parts of the world with good access to high-quality water. If land-based salmon farming can compete at relevant scales, remote production sites could potentially become a liability for traditional producers, as production could then move closer to consumption markets. Thus, successful investments by competitors in this technology can be seen as bad news for the companies represented in our index. Similarly, the sentiment-weighted topic 42 on substitute products, such as plant-based alternatives and meat, also demonstrates a negative correlation with the logarithmic returns of salmon producers (see the bottom left panel of Figure 3.12). Positive news about products that consumers may replace their salmon consumption with can be bad news for salmon producers as well.

It is common practice to use sentiment dictionaries to identify the effects of news on

Figure 3.12: Impulse responses for topic interactions with LM sentiment



Notes: The figure shows the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in topic 61 (algal blooms), topic 61 multiplied with LM sentiment, topic 42 (substitute products, such as meat and plant-based alternatives) multiplied with LM sentiment, and topic 25 (Nordic Aquafarms' land-based project in Maine, US) multiplied with LM sentiment. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

stock prices. However, we find that using a sentiment dictionary that is not tailored to the specific industry can result in inaccurate analyses. The Loughran-McDonald dictionary, for example, was not designed for the salmon aquafarming industry and thus misses important market structure and industry-specific vocabulary. As a result, the version

of the LM dictionary that we applied does not capture the effects of natural disasters, such as algal blooms, storms or diseases that impact salmon production. Additionally, the LM dictionary does not account for the impact of the Covid-19 pandemic, which has been a major driver of market volatility in recent times. To address these limitations, we suggest modifying the dictionary by adding domain-specific vocabulary and differentiating between news about competitors and the firms in our index. Such modifications could improve the accuracy of sentiment analysis and help investors make more informed decisions based on news about the salmon aquafarming industry.

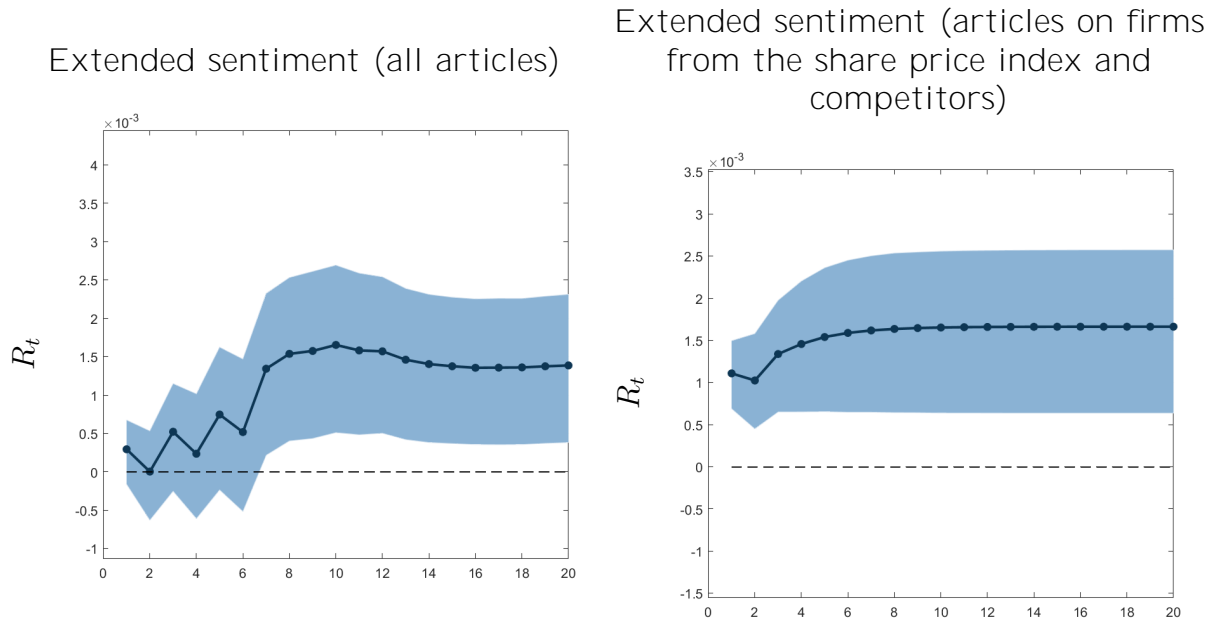
3.4.3 Resolving limitations of the Loughran-McDonald dictionary

3.4.3.1 Implementing competitive and sector-specific modifications in sentiment analysis

As highlighted earlier, the direct application of the Loughran-McDonald dictionary to our specific data suffers from two limitations: firstly, the presence of a competitive market structure, where news about our index firms elicits an opposite reaction to news on their competitors, renders the approach inappropriate. To overcome this issue, we modify the sentiment of articles mentioning at least one competitor but none of our index firms by multiplying the sentiment score by -1 . Secondly, the dictionary's lack of sensitivity to industry-specific language necessitates its expansion to include terms relevant to salmon production. These terms encompass technology, diseases and natural disasters, market-specific expressions, and general sentiment-related words that we deemed relevant but were not included in LM dictionary. The modified dictionary contains 187 additional positive words and 336 negative words, thereby supplementing LM's 347 positive and 2345 negative words. We deliberately excluded words that pertain exclusively to the Covid crisis, such as "Corona" or "pandemic", since their relevance was only apparent ex post and would have been unpredictable ex ante. However, we included "virus" since viral infections pose a significant challenge to salmon production. Nonetheless, as we shall demonstrate below, the inclusion of "virus" is insufficient to elicit substantial variation in articles related to Covid.

In Figure 3.13, we present the impulse response of logarithmic returns following a shock to the extended sentiment index. Remarkably, this small increment of words (amounting to only 16% of the words in the extended dictionary), in combination with competition effects, leads to a significant positive (0.0015, or 0.15 percentage points, given that log returns are expressed in percentages) effect of sentiment on returns seven days after the shock. Notably, our individual modifications to the dictionary could not yield such a substantial outcome, as evidenced by our detailed results (omitted for brevity). Indeed, the combination of both extensions is crucial to obtain the desired results. In the right

Figure 3.13: Impulse responses for extended sentiment measures



Notes: The figure illustrates the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in sentiment, as measured with the extended dictionary. Left panel: all articles; right panel: only articles that discuss firms from the share price index or competitors. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

panel, we display the impulse response for articles that solely mention either the companies in our index or their competitors, indicating the potential of filtering the data for pertinent articles. However, this unsophisticated analysis only provides a preliminary indication of the scope of data filtering, since it potentially discards relevant articles that do not mention firms directly, but yet carry important information for salmon markets. Developing more efficient and refined methods of data filtering remains an open research question, outside the scope of this paper.

3.4.3.2 Topics illustrating advantages of the modified methodology

Having shown that our solutions of extending the dictionary to include domain-specific vocabulary, and imposing market structure on the data analysis has the desired effect such that sentiment has a positive impact on markets, we can now proceed to present the results of a few chosen topics combined with the extended sentiment index (Figure 3.14).

Algal blooms topic (topic 61, top left panel) multiplied with extended sentiment now has the desired positive correlation with returns for three days after the shock. This is likely

to be driven mainly by the explicit introduction of a competitive market structure to our analysis, since algal blooms are more likely to be associated with Chilean aquaculture in our data set, and therefore are more likely to affect competitors. Word inclusions related to this well-known problem in aquafarming may however also have had an improving effect on the results.

Covid (topic 36, top right panel) on the other hand is still not adequately accounted for. We consciously decided against the inclusion of words that are straightforwardly linked to the pandemic, as it could not have been foreseen before its outbreak. Hence, exogenous shocks from unexpected directions are still not possible to analyse with this approach. However, this is an expected results, as some shocks may indeed be unforeseeable and cannot be planned for.

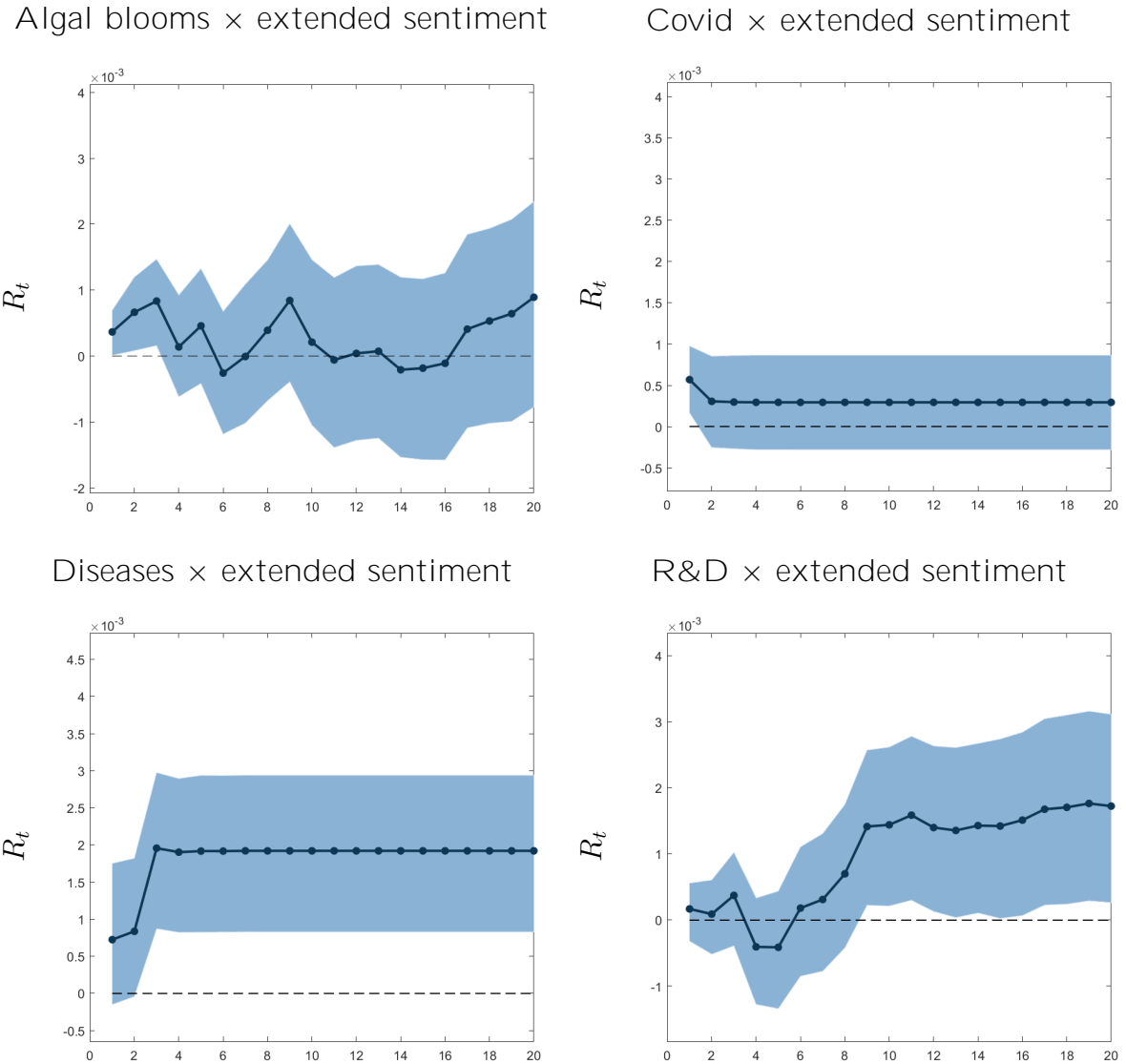
The inclusion of domain-specific vocabulary has clear benefits to this approach as well, as can be seen in the bottom panels of figure 3.14: topic 23 (bottom left), which pertains to diseases in production facilities, multiplied with extended sentiment now has a highly significant effect on returns, with the largest magnitude (0.002) observed. Naturally, the LM dictionary could not pick up news on such topics, and overall there may be good or bad news about diseases (outbreaks, or successful mitigation, for instance). By including words that relate to typical farmed salmon diseases, we could achieve this positive correlation between a sentiment-weighted topic that is highly specific, and the respective returns on company shares.

A similar effect can be observed for topic 71, which pertains to news on research and development (R&D) in salmon farming. These articles discuss technological and biological solutions to issues or difficulties in salmon production, which are also highly field-specific. With the extended sentiment, we find a positive correlation between the sentiment-weighted topic and returns in the long-run (9 days after the shock). Other examples (omitted) are the effects of topics 24 (escapes), 29 (Scottish salmon, Brexit), 57 (risk measures), 73 (construction of land-based facilities), 79 (biological performance), 87 (technology), 95 (project licenses Norway), 97 (Mowi business & management), all multiplied with extended sentiment.

3.4.3.3 PCA results

In this subsection, we analyse components derived from topics multiplied by the extended sentiment that impact log returns. The sentiment-weighted topics that exhibit strong correlation with the first component (see Table 3.6) mirror the sentiment-weighted topics that have substantial correlation with the first component in the LM-sentiment analysis (see Table 3.5 in the Appendix 3.A). The top three sentiment-adjusted topics that share the strongest correlation are the same, albeit in a different order. The fourth and fifth sentiment-adjusted topics, which demonstrate a notable correlation with the first compo-

Figure 3.14: Impulse responses for topic-sentiment interactions using the extended dictionary



Notes: The figure illustrates the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in topics 61 (algal blooms), 36 (Covid), 23 (diseases), and 71 (research & development), all multiplied with the extended sentiment index. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

ment - namely, “People, Projects & Perspectives” and “Green Bonds” - integrate seamlessly within the overarching theme of the component. This similarity can be ascribed to the relatively small addition of words to the sentiment dictionary. The first component explains a substantial 28.9% of the total variation in sentiment-weighted topic values, indicating its significant capture of sentiment-driven discourse.

We omit the second component as it primarily pertains to business-related news, similarly to the first component. However, since components are orthogonal by construction (being eigenvectors of the covariance matrix), it appears to capture rather spurious aspects in the behaviour of topics.

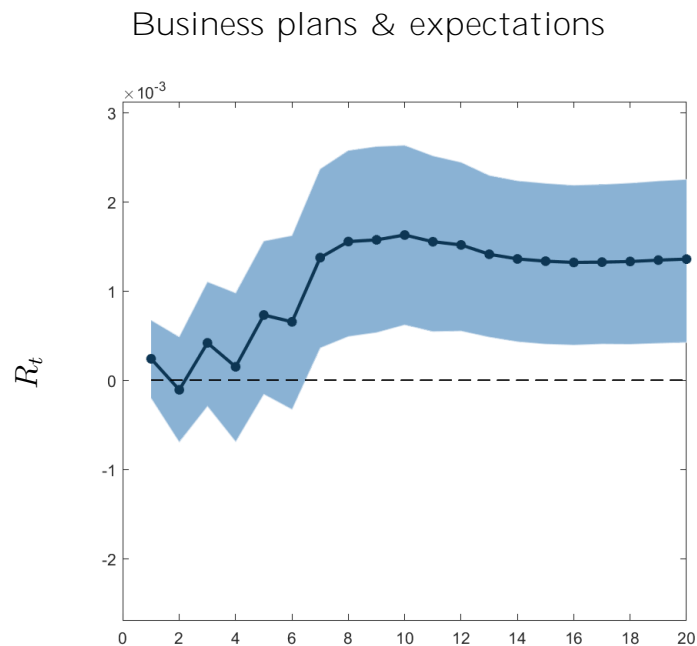
Component 4, labelled “Chile & algal blooms”, demonstrates a positive association with topic 61, concerning algal blooms, as well as topics 52 and 76, both of which address salmon aquaculture in Chile (see Table 3.6). More precisely, Component 4 suggests that, considering the increased frequency of algal blooms in Chile as opposed to Norway or the Faroe Islands, calamities occurring within the Chilean aquaculture industry may contribute to a rise in our constructed index. This verifies that the salmon market is subjected to competition. Furthermore, Component 7, broadly labelled “Salmon Industry”, relates to various topics specifically addressing aspects that are specific to salmon markets and production, such as topic 37 “Norwegian Salmon Prices”, topic 11, which covers articles about a collaboration between Salmon Evolution and Dongwon to establish a land-based production facility in Korea, and topic 28, “Salmon Harvesting Results”.

3.4.3.4 VAR results

Our analysis is supported by the VAR findings. As illustrated in Figure 3.15, after a short lag period, the impulse response of logarithmic stock returns to one standard deviation innovation in the first component turns significantly positive. Notably, the response to this innovation in the first component is nearly indistinguishable from the impulse response to an innovation in sentiment alone (see Figure 3.13), as seen in the case of topics multiplied with LM-sentiment (Figures 3.9, and 3.11, left panel). This again suggests that a substantial portion of sentiment is effectively captured by the first component. Moreover, it is promising that such a minor adjustment to the financially-oriented LM dictionary can nudge the previously negative response towards the expected direction and improve the results substantially.

Next, we draw attention to two additional components that, when subjected to a one standard deviation shock, produce statistically significant responses in log returns (Figure 3.16). Shocks to either Component 4 or Component 7 generate positive impulse responses, as anticipated due to the multiplication of topics with sentiment values. Most importantly, however, our findings indicate that the added terms in the dictionary enable us to uncover salmon industry-specific topics that are prominent in the data, as evidenced by their strong correlation with components based solely on topics. Thus, the addition of domain-specific vocabulary not only improves impulse responses to sentiment and its combination with topics but also enhances our ability to analyse a specific market with its unique characteristics.

Figure 3.15: Impulse response for component 1 based on topics multiplied with extended sentiment

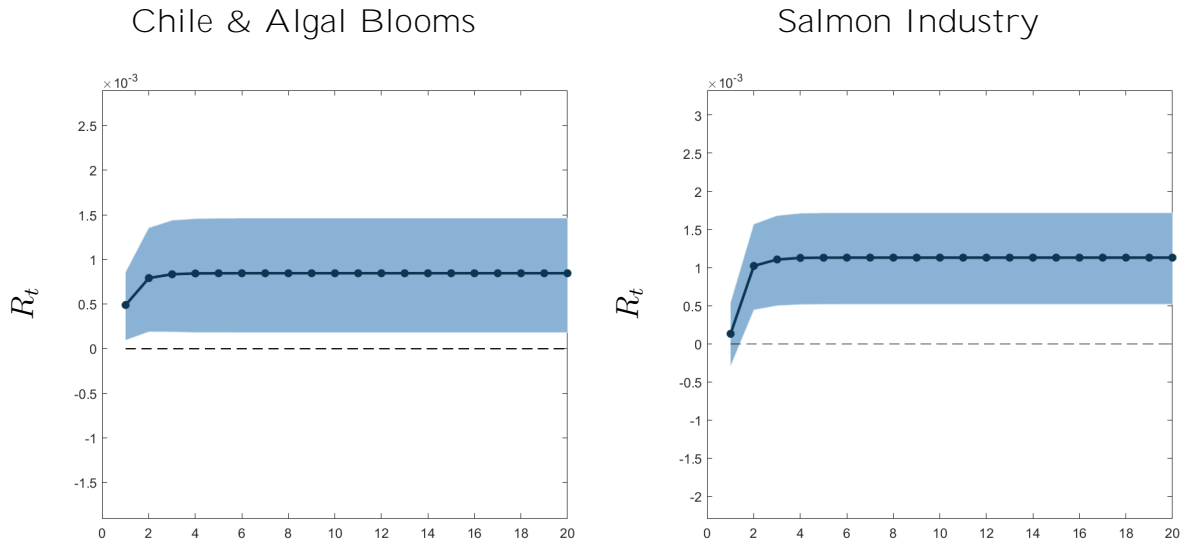


Notes: The figure illustrates the cumulated impulse response of logarithmic stock returns to one standard deviation innovation in component 1 based on topics multiplied with extended sentiment. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

3.4.3.5 Robustness check

We further strengthen our results by estimating a VAR model that incorporates three variables: one of the components based on topics multiplied with extended sentiment, the log return of the SPI, and additionally, the log return of the Oslo Stock Exchange (OSE). This approach allows us to assess whether our findings remain consistent when we include the market's systematic risk. In our VAR model, we position the component first to capture its direct effects, followed by the return of the OSE to consider the potential overarching market impact, and place the log return of the SPI third. This ordering reflects our hypothesis that market-wide factors might drive the returns of individual sectors or companies. For comparability, we maintain the same lag structure as used in the two-variable VAR models. Subsequently, we examine the impulse responses of the log return of the SPI to one standard deviation shocks in Components 1, 4, and 7 (see Figure 3.17). The consistency of these responses between the two- and three-variable models suggests that our results are robust to the inclusion of broader market dynamics, indicating that the effects of these components on SPI's log returns operate independently

Figure 3.16: Impulse responses for components 4 and 7 based on topics multiplied with extended sentiment



Notes: The figure illustrates the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in components 4 (“Chile & algal blooms”) and 7 (“Salmon Industry”) based on topics multiplied with extended sentiment. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

of overall market movements.

3.4.4 Out-of-sample forecasting exercise

While the main purpose of the study is to understand what type of information drives returns of salmon producer stocks, we validate the findings through an out-of-sample forecasting exercise. This will provide a financial significance check, determining whether these results could underpin profitable trading strategies.

Following the methodology established by (Li et al., 2020), we attempt to predict stock price movements. Specifically, we sort the close-to-close stock price returns in ascending order to label each stock. We use the 25th and 75th percentiles of the close-to-close stock price returns as thresholds for this label determination. Hence, if the return is in the bottom 25%, the label is defined as “fall” (class 0); if it is in the top 25%, the label is “rise” (class 2); for the middle 50%, it is labeled as “horizontal” (class 1).

Our dataset is partitioned into a training set (spanning 13th January 2016 to 11th November 2021) and a test set, constituting 10% of the data (12th November 2021 to 11th July 2022).

The Support Vector Machine (SVM) model serves as our fundamental analytical tool, often used as the benchmark in literature predicting stock market movements. Despite the potential for better performance from models like Long Short-Term Memory (LSTM), our focus is not on finding the most sophisticated model, but rather understanding if our proposed analysis could enhance out-of-sample forecasting.

Model parameters are fine-tuned using 10-fold cross-validation, where we select penalty C and polynomial degree d . We consider a grid of values for C ($10^0, 10^{0.25}, \dots, 10^3$) and for d (2, 3, 4, 5), aligned with (Li et al., 2020). Given the class imbalance in our data (40, 74, and 49 observations in classes 0, 1, and 2, respectively), the weighted average F1-score serves as our cross-validation metric.

Our SVM uses a polynomial kernel and adopts a “one versus one” strategy for this multi-classification task. To examine if components based on topics multiplied with extended sentiment can improve market movement prediction beyond price data alone, we estimate SVM models with varying feature sets.

The baseline model includes only the close price $Close_{t-1}$ and volume $Volume_{t-1}$, with a lag of one period, following the most popular practices (Li et al., 2020). Both variables are equally weighted for the five companies in our index. We experimented with using returns instead of close prices, but the performance on the test set was comparable, so we chose to retain the close prices.

Our second model incorporates sentiment estimated using the LM dictionary in addition to the price features, again with a lag of one period. The third model specification brings into play prices along with components derived from topics multiplied by the LM sentiment. In this instance, we factor in the first lag of the first component owing to its substantial role in explaining the variation. We also incorporate five other variables that represent lags ranging from 1 to 10 of the ten components most strongly correlated with log returns in the training dataset.

Our fourth model combines prices and extended sentiment from the period $t-1$. Lastly, our fifth model encompasses prices along with components based on topics multiplied by extended sentiment. Just as in the third model, the first lag of the first principal component is included due to the high percent of variance it explains. Additionally, five other variables are incorporated, which represent lags ranging from 1 to 10 of the ten components that are most highly correlated with log returns in the training data. Initially, we attempted to include the principal components at period t , however, the resulting weighted F1 metric was lower, and hence, in the final model, we only considered the lags of the components. Our feature selection process has been carefully designed to include the most meaningful variables for forecasting.

Model performance on the test set is evaluated using the weighted average F1-score. For models that incorporate text features, we compute the performance improvement

attributable to the news information using the following metric:

$$\Delta_{\text{news}} = \frac{F1_{p,\text{text}} - F1_p}{F1_p} \quad (3.9)$$

Here, $F1_{p,\text{text}}$ represents the weighted average $F1$ score when utilizing both information sources—text and prices—while $F1_p$ stands for the $F1$ score with only price information.

Table 3.2 presents the $F1$ -scores and Δ_{news} values for each model.

Table 3.2: Forecasting exercise results

Model	Weighted Average F1	Δ_{news}
Prices	0.28	-
Prices + LM Sentiment	0.28	0
Prices + LM Components	0.31	0.10
Prices + Extended Sentiment	0.28	0
Prices + Extended Components	0.38	0.32

Notes: The table demonstrates the forecasting improvements due to the inclusion of sentiment and topic components (news).

These results highlight that incorporating only sentiment (either LM or extended) does not enhance performance relative to the baseline prices-only model. However, including components based on topics multiplied with sentiment (LM or extended) significantly increases the $F1$ score. Specifically, the increase is by 10% with LM sentiment and 32% with extended sentiment.

In conclusion, this study demonstrates the potential of our text information extraction approach in enhancing out-of-sample forecasting. While it remains a limited experiment using a single model, it serves its purpose: to verify that text matters.

3.5 Discussion

In this study, we conducted a comprehensive examination of over 6000 news articles covering salmon production and markets, intending to assess the influence of news on the stock returns of the largest salmon-producing companies listed on the Oslo Stock Exchange. To derive meaning from this unstructured data, we employed Latent Dirichlet Allocation (LDA) to generate topics, as well as a dictionary approach to analyse sentiment.

Initially, we explored the impact of topics on markets by consolidating them into components and utilizing Vector Autoregression (VAR) analyses on these components and logarithmic stock returns. Given that news coverage can encompass both positive and negative aspects, the effect's direction remains undefined, thus constraining our analysis to absolute returns. Owing to the specific time frame examined, news concerning the Covid-19 pandemic dominated the topics and their market repercussions. Nevertheless,

we identified significant market reactions driven by the component related to corporate news and stocks, too.

Upon evaluating topics combined with sentiment, we found that sentiment dictionaries are not sufficiently adaptable to domains other than those they were originally designed for. Specifically, a surprise increase in sentiment constructed using the Loughran-McDonald dictionary, which was tailored for financial data, resulted in a marginally significant effect on logarithmic returns with an incorrect sign. One rationale for this outcome is that the dictionary was designed to detect sentimental expressions in the financial reporting of companies, neglecting industry-specific news.

Moreover, our stock index solely comprised the largest companies on the Oslo Stock Exchange, influencing the results due to the competitive market structure of salmon markets. Although concentrating only on news directly and specifically related to the underlying firms could solve this issue, it would simultaneously disregard vital general news about the salmon industry, such as research not conducted by the producers, or news about competitors that could impact other firms, either directly or indirectly. Alternatively, we addressed this issue by inverting sentiment in articles concerning competitor firms.

To tackle these challenges, we augmented the LM sentiment dictionary by incorporating industry-specific terms and considering the market structure, thereby constructing a sentiment index that has the desired impact on stock returns. This methodological contribution holds general relevance, beyond the specific domain of salmon markets, since language is highly dependent on context. Employing the extended dictionary and explicitly considering competition between producers, we achieved the anticipated positive correlation between sentiment and returns. Additionally, we observed the effects of industry-specific topics on returns, including algal blooms, diseases, and R&D.

However, we discovered that the extended dictionary and explicit competition effects could not account for the Covid topic, representing an archetypal unanticipated exogenous shock to the market that could not have been represented in the dictionary *ex ante*. Although space constraints in this paper did not permit an in-depth discussion of numerous additional topics, we contend that the enhanced sentiment index we devised will prove beneficial for future investigations of news effects on financial markets.

In our out-of-sample experiment, we aimed to demonstrate the value of incorporating sentiment and topics into financial forecasting models. We found that integrating components based on topics multiplied with sentiment, particularly our domain-specific sentiment index, significantly improved the performance of the SVM model in predicting stock market price movements, as evidenced by the higher weighted average F1 score. This improvement underlines that news information can provide substantial predictive power beyond what is captured by price data alone.

One constraint of our study was the limited number of news articles and the relatively

brief time series under examination. Future studies could consider broadening the time series horizon and incorporating articles from additional news sources to address these limitations. We believe that our extended dictionary holds potential for application in similar studies within aquaculture economics beyond the salmon industry. Moreover, future research focusing exclusively on salmon markets could contemplate further expanding the dictionary to encompass competitive seafood markets, such as shrimp, tuna, and others, thereby enhancing the scope and applicability of the sentiment analysis.

Additionally, while the dictionary approach to sentiment analysis offers high interpretability and ease of implementation, a static lexicon like ours often struggles to capture the full range of sentiment nuances. The next step could involve creating a high-quality training dataset specific to the salmon market. This dataset would require manually annotating articles as negative, neutral, or positive, and then using these annotations to train a machine learning model. Although developing such a dataset is outside the scope of this research due to the significant resources required, our methodology offers a robust foundation for researchers interested in pursuing this avenue. For instance, it might be advantageous to include news articles covering the specific topics we have identified as market drivers in the training set. Furthermore, considering the competitive structure of the market is important when annotating each article.

The exploration of enhanced data filtering techniques that preserve general market news and competitor information, while still removing unrelated noise, could be considered for future research. In particular, our study emphasised the importance of accounting for competition and market structure, but a trade-off emerged between focusing on articles directly related to the firms in question (thereby increasing significance), and retaining crucial news about the overall market. Moreover, further examinations on how sustainability-related news impact market behavior and provide significant implications for regulatory bodies and industry practices could be worthwhile for future research. However, as this falls outside the scope of the current study we encourage the exploration of such topics in future studies.

Appendix 3

3.A Estimated topics, component tables and robustness check

Table 3.3: Estimated topics with most probable words

ID	Label	Most Probable Words
T3	Aquabounty Land-based & GM	aquabounti, gm, commerci, fda, approv, genet, facil, modifi, stock, aquadvantag, indiana, salmon_produc, site
T10	Outlook and Opinions on Prices	think, much, good, big, money, littl, seem, reason, surpris, enough, lose, worri, interest, expens
T11	Land-based Facility by Salmon Evolution & Dongwon	salmon_evolut, land, south, hofseth, dongwon, aqua, facil, invest, phase, project, start, korea, construct, joint_ventur, berg, korean, establish, intern, build, biocar, capit, smart, announc, rais, sign, indr, recent, fresh, largest, haroy, enter, partner
T12	Future Challenges	need, differ, look, import, find, interest, think, term, possibl, hope, tri, moment, best, option, attract, gap, perspect, idea, stress, problem
T13	Challenges, Market risk	impact, term, challeng, situat, signific, short, neg, posit, affect, financi, effect, face, difficult, uncertainti, pressur, futur, warn
T14	Processor Acquisitions (UK)	brand, uk, retail, launch, smoke, busi, smoked_salmon, supplier, young, label, processor, line, united_kingdom, hilton, grimsbi, categori, united_st, macknight, sainsburi, tesco, store, sold, expand, salmon_product, england
T19	Public & Social	public, social, comun, respons, media, point, question, concern
T23	Diseases	site, isa, author, confirm, diseas, outbreak, viru, detect, sampl, infect, infectious_salmon_anemia, area, case, suspect, food_safeti, cage, affect, test, found, inspect
T24	Escapes	escap, site, net, cage, damag, incid, storm, pen, investig, caus, weather, fire, occur, structur, recaptur
T25	Nordic Aquafarms' land-based project in Maine, US	main, project, land, nordic_aquafarm, permit, state, nordic, site, whole_ocean, local, american, construct, belfast, citi, facil, california, build, aquafarm, move, properti, approv
T28	Salmon harvesting results	size, larg, larger, big, much, small, smaller, kilogram, differ, farmer, player, littl, trend
T29	Scottish salmon, Brexit	scotland, uk, scottish, export, scottish_salmon, brexit, govern, loch, sspo, scott, sector, scottish_salmon_farm, trade, duart, eu, problem, united_kingdom

Continued on next page

Table 3.3 continued from previous page

ID	Label	Most Probable Words
T33	SalMar Offshore Farming	salmar, offshor, ocean, fish_farm, central, aker, norwegian_salmon_farm, sea, area, project, coastal, invest, northern_norway, establish, rokk, technolog, coast
T35	Fear, Harm, Negative Outlook	sever, major, execut, unclear, fear, turn, wait, direct, none, devast, rush, lack, enorm, specul, lose, unsur, expos, sizeabl, loom
T36	Covid Pandemic	covid, pandem, foodservic, coronaviru, demand, retail, lockdown, 19_pandem, crisi, restrict, disrupt, global
T37	Norwegian Salmon Prices (kg)	export, kg, kilo, week, nok, next_week, kilogram, buyer, norwegian_salmon_pric
T39	Salmon Wholesale Prices	week, brazil, united_st, wholesal, coho, remain, size, frozen, kilo, chilean_salmon, pound, averag, chilean_salmon_pric, atlantic_salmon, rose, fell
T41	Business Results	profit, loss, revenu, end, net, earn, amount, operating_profit, turnov
T42	Substitute Products	food, protein, plant, meat, altern, consum, grow, sustain, giant, beef, anim, pork, poulttri, launch, byproduct, health, sourc, natur, seafood_industri, agricultur, option, chicken, biomega, vegan
T43	Plans & Strategy	target, strategi, aim, grow, step, effort, hope, set, key, futur, launch, goal, boost, program, ambiti, road, aggress, win
T47	Contracts & Agreements	agreement, contract, sign, deal, suppli, firm, final, start, part, supplier, extend, announc, agre, cover, negoti, reach
T48	Management & Boards	board, invest, chairman, member, hold, owner, coast, founder, largest, firm, kverva, famili, stake, broodstock, fiskeoppdrett, billionaire
T51	Salmon Market Analysis	analyst, estim, buy, target, recommend, stock, believ, salmon_pric, nordea, johannessen, sparebank, hold, share, bank, share_pric, pareto_secur, risk
T52	Chilean Salmon Farming	chile, chilean, camanchaca, region, sernapesca, nova_austr, servic, chilean_salmon_farm, aysen, salmones_camanchaca
T53	Quarterly Results	quarter, earn, revenue, tax, ebitda, earnings_before_interest, first_quarter, last_year, second_quarter, third_quarter
T55	Strategy, Investments, Opportunities	strategi, invest, focu, opportun, valu, integr, ad, focus, expand, strateg, term, look, global, establish, grow, value_chain, presenc, creat, team, strong, innov

Continued on next page

Table 3.3 continued from previous page

ID	Label	Most Probable Words
T56	Bakkafrost Acquires SSC	bakkafrost, ssc, scottish_salmon_compani, faroes, jacobsen, faroe_island, scotland, northern, link, salmon_produc, vap, full, sekkingstad, sell, releas, sharehold, final
T57	Risk measures	risk, measur, implement, reduc, number, limit, ensur, control, allow, minim, consid, river, mitig, maintain, assess, avoid, polici, environ, evalu
T59	Salmon Commodity Prices	week, kilo, compar, kilogram, averag, average_pric, nasdaq, categori, size, weight_class, salmon_pric, kilo_on_averag, market_shar, farmed_salmon_pric
T60	People, Projects & Perspectives	peopl, job, support, help, busi, commun, creat, econom, employ, famili, team, person, leader, economi, gener, cultur, popul, role, critic, staff, employe, career, modern, uniku
T61	Algal Blooms	mortal, loss, algal_bloom, site, caus, incid, die, lost, event, bloom, affect, alga, mass, biomass, oxygen_level
T63	Retail	retail, fresh, frozen, sell, custom, fillet, consum, packag, line, chain, ad, supplier, promot, store, supermarket
T65	Seafood Markets & Consumption	consum, restaur, eat, consumpt, food, onlin, peopl, buy, home, trend, purchas, healthi, prepar, spend, good, shop, chain, meal, sushi, wild, menu, chef, dine, salmon_market
T66	Green Bonds	issu, green, qualiti, bond, date, announc, significantli, convert, type, success, superior, matur, extend, regist, rate, clean, access, joint, unsecur, contribut, incent, instrument
T67	Drop, Fall, Decline	drop, fall, declin, fell, saw, hit, lower
T68	Business Data	figur, show, number, estim, last_year, total, releas, account, farmed_salmon, metric_ton, data, annual, half, doubl, declin, reach, publish, level, calcul
T69	Stock Markets	day, close, monday, share_pric, stock, announc, share, trade, valu, news, open, updat, lost, lose, gain, hour, sever, fallen, indic, stock_exchang
T70	Smolt Production Facilities Investment	smolt, facil, sea, capac, invest, hatcheri, plant, site, build, grow, expans, expand, transfer, production_capac, locat, gram, futur, processing_pl, freshwat, nova, weight, construct, releas, salmon_product
T71	Research & Development	research, studi, institut, univers, found, marin, farmed_salmon, find, scientist, water, test, speci, flavor, scienc, project, center, natur, trial, commerci, wild_salmon, conduct

Continued on next page

Table 3.3 continued from previous page

ID	Label	Most Probable Words
T73	Construction of land-based facilities	construct, facil, phase, land, build, project, complet, start, ton, proximar, expans, initi, andfjord_salmon, metric, annual, full, capac, begin, locat
T76	Chilean Producers	aquachil, chile, blumar, multiexport, chilean, agrosup, chilean_salmon_farm, region, magallan, ventisquero, fiordo, chilean_salmon_compani, invermar
T78	US Fund Raising	rais, farmer, capit, sell, seek, stolt, return, expand, fund, investor, new_york, goal, nielsen, target, halibut, fundrais
T79	Biological performance	improv, perform, biolog, strong, good, challeng, posit, better, achiev, contribut, condit, period, effici, record, solid, strengthen, optim, profit, capac
T82	Mergers & Acquisitions	acquisit, acquir, deal, hold, stake, announc, purchas, transact, owner, control, ownership
T83	Covid Cases in Production Facilities	worker, employe, case, test, coronaviru, posit, health, covid, plant, quarantin, outbreak
T85	Atlantic Sapphire Land-based	atlantic_sapphir, land, miami, facil, danish, denmark, andreassen, fire, phase, florida, based_salmon, based_salmon_farm, plant, system, bluehous, united_st, share_pric
T87	Technology	system, technolog, water, use, feed, energi, data, wast, emiss, solut, sustain, fish_farm, reduc, carbon, digit, environment, instal, tech, optim, improv, monitor, ga, power, equip, aquaculture_industri, innov
T95	Project Licenses Norway	licens, project, cermaq, cage, concept, permit, close, applic, appli, receiv, grant, reject, innov, director, directorate_of_fisleri
T96	NTS Acquisition of NRS	nt, nr, norway_royal_salmon, bid, sharehold, share, offer, salmonor, salmar, merger, announc
T97	Mowi business & Management	mowi, marine_harvest, scotland, divis, canada, aarskog, giant_mowi, vindheim, helg, record, chile, america, ceo_ivan_vindheim, faro
T98	Global Salmon Markets	suppli, demand, global, strong, rise, grow, outlook, limit, salmon_market, predict, forecast, trend, strong_demand, volatil, shock, us_market, north_america, salmon_product

Chapter 3 Salmon stock returns around market news

Table 3.4: Topic components and top-correlated topics

Component number	Component label	variation explained	Component's Top 5 Topics	Correlation
1	Business Figures	4%	T53 Quarterly Results	57%
			T19 Public & Social	-49%
			T67 Drop, Fall, Decline	48%
			T12 Future Challenges	-47%
			T41 Business Results	44%
5	Covid & Production	2.2%	T37 Norwegian Salmon Prices (kg)	-50%
			T28 Salmon harvesting results	-48%
			T36 Covid Pandemic	37%
			T83 Covid Cases in Production Facilities	30%
			T10 Outlook and Opinions on Prices	-29%
6	Covid & Markets	2.1%	T59 Salmon Commodity Prices	-41%
			T36 Covid Pandemic	37%
			T39 Salmon Wholesale Prices	-36%
			T13 Challenges, Market risk & Uncertainty	36%
			T82 Mergers & Acquisitions	-32%
9	Chile	1.7%	T76 Chilean Producers	52%
			T52 Chilean Salmon Farming	48%
			T47 Contracts & Agreements	32%
			T61 Algal Blooms	27%
			T33 SalMar O shore Farming	-25%
12	Norwegian Farming	1.5%	T33 SalMar O shore Farming	36%
			T96 NTS Acquisition of NRS	33%
			T85 Atlantic Sapphire Land-based	-30%
			T95 Project Licenses Norway	24%
			T3 Aquabounty Land-based & GM	-24%
17	Corporate & Stocks	1.4%	T41 Business Results	32%
			T98 Global Salmon Markets	-29%
			T48 Management & Boards	29%
			T69 Stock Markets	29%
			T51 Salmon Market Analysis	-29%

Notes: This table reports topic components and the five topics that exhibit the highest absolute correlation with each component.

Table 3.5: Components based on LM sentiment-weighted topics and top-correlated topics

Component number	Component label	variation explained	Component's Top 5 Topics	Correlation
1	Business expectations	26.5%	T43 Plans & Strategy T35 Fear, Harm, Negative Outlook T68 Business Data T12 Future Challenges T47 Contracts & Agreements	82% 81% 74% 74% 73%
3	Investments Norway	2.1%	T33 SalMar O shore Farming T96 NTS Acquires NRS T10 Outlook and Opinions on Prices T82 Mergers& Acquisition T70 Smolt Production Facilities Investment	39% 38% -33% 30% 31%
4	Investments ROW	2.1%	T14 Processor Acquisitions (UK) T78 US Fund Raising T97 Mowi business & Management T63 Retail T55 Strategy, Investments, Opportunities	38% 36% -36% 36% 28%

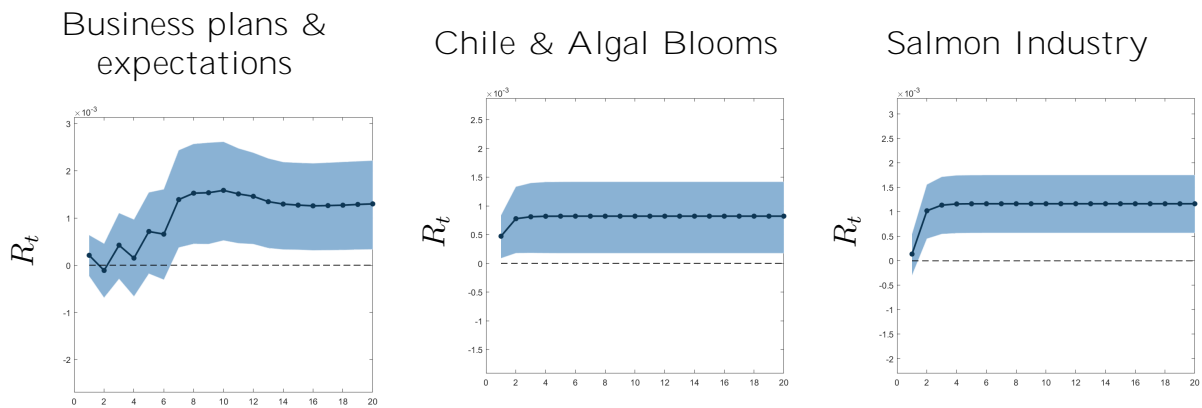
Notes: The table shows LM sentiment-weighted topic components and the five LM sentiment-weighted topics that exhibit the highest absolute correlation with each component.

Table 3.6: Components based on topics adjusted with extended sentiment and top-correlated topics

Component number	Component label	variation explained	Component's Top 5 Topics	Correlation
1	Business Plans & Expectations	28.9%	T35 Fear, Harm, Negative Outlook T43 Plans & Strategy T68 Business Data T60 People, Projects & Perspectives T66 Green Bonds	85% 83% 78% 76% 75%
4	Chile & Algal Blooms	2.0%	T61 Algal blooms T76 Chilean Producers T10 Outlook and Opinions on Prices T65 Seafood Markets & Consumption T52 Chilean Salmon Farming	45% 38% -37% -35% 33%
7	Salmon Industry	1.6%	T37 Norwegian Salmon Prices T82 Mergers & Acquisitions T56 Bakkafrøst Acquires SSC T11 Land-based Facility by Salmon Evolution & Dongwon T28 Salmon Harvesting Results	36% -36% -28% 26% 26%

Notes: This table shows components of topics multiplied with the extended sentiment and the five topics multiplied with the extended sentiment that exhibit the highest absolute correlation with each component.

Figure 3.17: Robustness check: IRFs for components based on topics multiplied with extended sentiment



Notes: The figure shows the cumulated impulse responses of logarithmic stock returns to one standard deviation innovations in components 1 (“Business plans & expectations”), 4 (“Chile & algal blooms”), and 7 (“Salmon Industry”) based on topics multiplied with extended sentiment. These responses are derived from a VAR model that includes one of these components, as well as the log returns of the OSE and the SPI. Shaded regions are 68 percent confidence bands, computed with bootstrap standard errors, using 1000 replications. The horizontal axis represents the number of lagged days in the Impulse Response Functions (IRFs).

Bibliography

- Alexopoulos, M., & Cohen, J. (2015). The power of print: Uncertainty shocks, markets, and the economy. *International Review of Economics & Finance*, 40, 8–28.
- 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.
- Andersen, B. P., & de Lange, P. E. (2021). Efficiency in the Atlantic salmon futures market. *Journal of Futures Markets*, 41(6), 949–984.
- 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.
- Ardia, D., Bluteau, K., & Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, 35(4), 1370–1386.
- Asche, F. (2008). Farming the sea. *Marine Resource Economics*, 23(4), 527–547.
- Asche, F., Misund, B., & Oglend, A. (2019). The case and cause of salmon price volatility. *Marine Resource Economics*, 34(1), 23–38.
- 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.
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, 131(4), 1593–1636.
- Bañbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting. *European Central Bank, Working Paper 1275*.

Bibliography

- Bañbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting with daily data. *European Central Bank, Working Paper*.
- Bañbura, 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.
- Barbaglia, L., Consoli, S., Manzan, S., Tiozzo Pezzoli, L., & Tosetti, E. (2025). Sentiment analysis of economic text: A lexicon-based approach. *Economic Inquiry*, 63(1), 125–143.
- Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 759–766.
- Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.
- Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10(71), 71–89.
- Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3), 623–685.
- Bloznelis, D. (2016). Salmon price volatility: A weight-class-specific multivariate approach. *Aquaculture economics & management*, 20(1), 24–53.
- Bloznelis, D. (2018). Hedging salmon price risk. *Aquaculture Economics & Management*, 22(2), 168–191.
- Bok, B., Caratelli, D., Giannone, D., Sbordone, A., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10, 615–643.
- Borup, D., Hansen, J. W., Liengard, B. D., & Montes Schuette, E. C. (2023). Quantifying investor narratives and their role during COVID-19. *Journal of Applied Econometrics*, 38(4), 512–532.
- 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.

Bibliography

- Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation. *The Journal of Business*, 78(2), 405–440.
- 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>
- Carstensen, K., Heinrich, M., Reif, M., & Wolters, M. H. (2020). Predicting ordinary and severe recessions with a three-state Markov-switching dynamic factor model: An application to the German business cycle. *International Journal of Forecasting*, 36(3), 829–850.
- 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.
- Correa, R., Garud, K., Londono, J. M., & Mislant, N. (2017). Constructing a dictionary for financial stability. *Board of Governors of the Federal Reserve System*, 6(7), 9.
- Dahl, R. E., & Oglend, A. (2014). Fish price volatility. *Marine Resource Economics*, 29(4), 305–322.
- Dahl, R. E., Oglend, A., & Yahya, M. (2021). Salmon stock market prices revealing salmon price information. *Marine Resource Economics*, 36(2), 173–190.
- Dahl, R. E., & Yahya, M. (2019). Price volatility dynamics in aquaculture fish markets. *Aquaculture Economics & Management*, 23(3), 321–340.
- 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*.
- De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy*, 98(4), 703–738.
- 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.
- Di Caro, L., Guerzoni, M., Nuccio, M., & Siragusa, G. (2017). A bimodal network approach to model topic dynamics. <http://arxiv.org/abs/1709.09373>
- 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.

Bibliography

- Ewald, C.-O., Haugom, E., Kanthan, L., Lien, G., Salehi, P., & Størdal, S. (2022). Salmon futures and the Fish Pool market in the context of the CAPM and a three-factor model. *Aquaculture Economics & Management*, 26(2), 171–191.
- Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, 38(1), 34–105.
- Fama, E. F. (1970). Efficient capital markets. *Journal of Finance*, 25(2), 383–417.
- FAO. (2018). *The state of world fisheries and aquaculture 2018*.
- FAO. (2020). *The state of world fisheries and aquaculture*.
- Fronzi, 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.
- Garlock, T., Asche, F., Anderson, J., Bjørndal, T., Kumar, G., Lorenzen, K., Ropicki, A., Smith, M. D., & Tveterås, R. (2020). A global blue revolution: Aquaculture growth across regions, species, and countries. *Reviews in Fisheries Science & Aquaculture*, 28(1), 107–116.
- 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.
- Gusev, M., Kroujiline, D., Govorkov, B., Sharov, S. V., Ushanov, D., & Zhilyaev, M. (2015). Predictable markets? A news-driven model of the stock market. *Algorithmic Finance*, 4(1-2), 5–51.
- Guttormsen, A. G. (1999). Forecasting weekly salmon prices: Risk management in fish farming. *Aquaculture Economics & Management*, 3(2), 159–166.
- Hanna, A. J., Turner, J. D., & Walker, C. B. (2020). News media and investor sentiment during bull and bear markets. *The European Journal of Finance*, 26(14), 1377–1395.
- 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*.
- Hersoug, B. (2021). Why and how to regulate Norwegian salmon production?—The history of Maximum Allowable Biomass (MAB). *Aquaculture*, 545, 737144.

Bibliography

- 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.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 168–177.*
- 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.*
- Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural language engineering, 1*(1), 9–27.
- 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.*
- Karalevicius, V., Degrande, N., & De Weerd, J. (2018). Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance, 19*(1), 56–75.
- Khedr, A. E., Yaseen, N., et al. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications, 9*(7), 22.
- Kling, G., & Gao, L. (2008). Chinese institutional investors' sentiment. *Journal of International Financial Markets, Institutions and Money, 18*(4), 374–387.
- Kräussl, R., & Mirgorodskaya, E. (2017). Media, sentiment and market performance in the long run. *The European Journal of Finance, 23*(11), 1059–1082.
- Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114, 128–147.*

Bibliography

- 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.
- Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of econometrics*, 210(1), 203–218.
- Larsen, V. H. (2021). Components of uncertainty. *International Economic Review*, 62(2), 769–788.
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.
- León-Muñoz, J., Urbina, M. A., Garreaud, R., & Iriarte, J. L. (2018). Hydroclimatic conditions trigger record harmful algal bloom in western Patagonia (summer 2016). *Scientific reports*, 8(1), 1330.
- Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, 57(5), 102212.
- Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69, 14–23.
- 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.
- Montes, R. M., Rojas, X., Artacho, P., Tello, A., & Quiñones, R. A. (2018). Quantifying harmful algal bloom thresholds for farmed salmon in southern Chile. *Harmful Algae*, 77, 55–65.
- 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.

Bibliography

- Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), 9603–9611.
- Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint: 1103.2903*.
- Oglend, A. (2013). Recent trends in salmon price volatility. *Aquaculture Economics & Management*, 17(3), 281–299.
- Okuneva, M., Hauber, P., Carstensen, K., & Bär, J. (2024). Nowcasting German GDP with Text Data. *CESifo Working Paper Series 11587*.
- 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.
- Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS – a publicly available German-language resource for sentiment analysis. *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*.
- 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.
- Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of finance*, 52(1), 35–55.
- Shrub, Y., Rieger, J., Müller, H., & Jentsch, C. (2022). Text data rule - don't they? a study on the (additional) information of Handelsblatt data for nowcasting German GDP in comparison to established economic indicators. *Ruhr Economic Papers, No. 964*.
- 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.

Bibliography

- 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>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62, 1139–1168.
- 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 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*.
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2579–2605.
- Verma, R., Baklaci, H., & Soydemir, G. (2008). The impact of rational and irrational sentiments of individual and institutional investors on DJIA and S&P500 index returns. *Applied Financial Economics*, 18(16), 1303–1317.
- Zitti, M. (2024). Forecasting salmon market volatility using long short-term memory (LSTM). *Aquaculture Economics & Management*, 28(1), 143–175.

Declaration of co-authorship

1. Doctoral candidate:

Mariia Okuneva

2. This declaration applies to the following article:

Nowcasting German GDP with text data

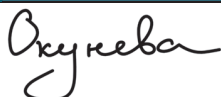
I declare that the aforementioned research paper, included in the doctoral dissertation of Ms. Okuneva and titled “Macroeconomic Forecasting and Market Analysis with Newspaper Articles”, has been developed as a collaborative work by Mariia Okuneva, Philipp Hauber, Kai Carstensen, and Jasper Bär.

I affirm that the contributions of each co-author have been appropriately recognized and that all co-authors have reviewed and approved the final version of the paper. The contributions include the research concept and ideas, data collection and analysis, methodology design, model estimation, result analysis, critical review, and co-writing of the paper. Furthermore, I confirm that all co-authors agree with the inclusion of this research paper as part of the doctoral dissertation of Ms. Okuneva and that this arrangement does not infringe upon the intellectual property rights of any party involved.

3. Signatures of all co-authors:

Date	Name	Signature
28.11.2025	Philipp Hauber	
01.12.2025	Kai Carstensen	
01.12.2025	Jasper Bär	

4. Signature of the doctoral candidate:



Declaration of co-authorship

1. Doctoral candidate:

Mariia Okuneva

2. This declaration applies to the following article:

Salmon Stock Returns around Market News

I declare that the aforementioned research paper, included in the doctoral dissertation of Ms. Okuneva and titled “Macroeconomic Forecasting and Market Analysis with Newspaper Articles”, has been developed as a collaborative work by Clemens Knoppe, Mariia Okuneva, and Mikaella Zitti.

I affirm that the contributions of each co-author have been appropriately recognized and that all co-authors have reviewed and approved the final version of the paper. The contributions include the research concept and ideas, data collection and analysis, methodology design, model estimation, result analysis, critical review, and co-writing of the paper. Furthermore, I confirm that all co-authors agree with the inclusion of this research paper as part of the doctoral dissertation of Ms. Okuneva and that this arrangement does not infringe upon the intellectual property rights of any party involved.

3. Signatures of all co-authors:

Date	Name	Signature
29.11.2025	Clemens Knoppe	
29.11.2025	Mikaella Zitti	

4. Signature of the doctoral candidate:

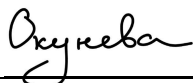


Declaration of independent work and AI use

I hereby declare that I have written my doctoral dissertation entitled “Macroeconomic Forecasting and Market Analysis with Newspaper Articles” independently and have used no aids other than those indicated. I further confirm that, as a co-author, I made substantial contributions to each of the individual scientific articles included in this dissertation.

I also confirm that I have used ChatGPT (GPT-5.1) exclusively to improve the readability and language of the dissertation. All content generated with the assistance of this AI tool has been carefully reviewed and edited by me. I take full responsibility for the accuracy, correctness, and integrity of the final version of the dissertation.

02.12.2025

A handwritten signature in black ink, appearing to read 'Byreba', written over a horizontal line.

Date and Signature