Doctoral Thesis

Justus-Liebig-University Giessen

# Text-as-Data: Methodological Advances and Applications in Economics


Author:

Viktoriia Naboka-Krell

1. Supervisor:

Prof. Dr. Peter Winker

2. Supervisor:

Prof. Dr. Nicolas Pröllochs


*Submitted in fulfillment of the requirements*
*for the degree of Doctor rerum politicarum*

*in the*


Department of Statistics and Econometrics

Faculty of Economics and Business Studies


December 9, 2024

## Affidavit

I hereby declare that I completed the papers submitted and listed hereafter independently and only with those forms of support mentioned in the relevant paper. When working with the authors listed, I contributed no less than a proportional share of the work. In the analysis that I have conducted and to which I refer in the papers, I have followed the principles of good academic practice, as stated in the Statute of Justus Liebig University Giessen for ensuring good scientific practice.

Viktoriia Naboka-Krell,
Giessen, December 9, 2024

## Submitted Articles

- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024). Analysing the Impact of Removing Infrequent Terms on Topic Quality in Latent Dirichlet Allocation Models. *arXiv, abs/2311.14505*. https://doi.org/10.48550/arXiv.2311.14505

- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024). Sampling Uncertainty of Topic Modelling. *Unpublished. Working Paper.*

- Bystrov, V., Naboka, V., Staszewska-Bystrova, A., & Winker, P. (2022). Cross-Corpora Comparisons of Topics and Topic Trends. *Jahrbücher für Nationalökonomie und Statistik, 242(4)*, 433-469. https://doi.org/10.1515/jbnst-2022-0024

- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024). Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria. *Journal of Machine Learning Research, 25(79)*, 1-30. https://www.jmlr.org/papers/v25/23-0188.html

- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024). Comparing Links between Topic Trends and Economic Indicators in the German and Polish Academic Literature. *Comparative Economic Research. Central and Eastern Europe, 27(2)*, 7-28. https://doi.org/10.18778/1508-2008.27.10

- Naboka-Krell, V. (2024). Construction and analysis of uncertainty indices based on multilingual text representations. *Economics Letters, 237*, 111653. https://doi.org/10.1016/j.econlet.2024.111653

- Latifi, A., Naboka-Krell, V., Tillmann, P., & Winker, P. (2024). Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects. *European Economic Review, 168*, 104827. https://doi.org/10.1016/j.euroecorev.2024.104827

- Latifi, A., Naboka-Krell, V., Tillmann, P., & Winker, P. (2024). Disagreement about Fiscal Policy. *Unpublished. Working Paper.*

## Acknowledgements

This thesis has been written with the support of many people and institutions to whom I would like to express my sincere gratitude. I would like to thank the University of Giessen for making my journey possible in the first place. I would also like to thank the German Research Foundation (DFG) and the National Science Centre (NCN) for their financial support of the projects that I was involved in (TEXTMOD, MaFiText). In addition, this thesis would not have been possible without the support of many others whom I would like to thank.

First of all, I would especially like to thank my supervisors, Peter Winker and Nicolas Pröllochs, who have been very supportive during this time. I am very grateful to Peter Winker, who encouraged me to start my PhD in the first place and who supported me with continuous feedback and advice over the past years. This thesis includes joint papers written as a part of two major research projects. I would like to thank Anna Staszewska-Bystrova and Victor Bystrov from the University of Lodz and Peter Winker for the great cooperation on the TEXTMOD project. That was a great start for me because I had the opportunity to learn from them how a research project works, where to start and how to structure and pursue different ideas. I would also like to thank Peter Tillmann, Albina Latifi and Peter Winker for collaborating on the MaFiText project. It was a great opportunity not only to apply methodological approaches that I am familiar with, but also to deepen my knowledge about different aspects of (fiscal) policy making in Germany.

A special thanks also goes to my dear colleagues Elena Tönjes, Jenny Bethäuser, Albina Latifi and Maykol Rodriguez who have contributed to a pleasant and joyful atmosphere. Many thanks to Carmen Hersener for always having an open ear and constant administrative support.

Last but not least, I would like to thank my family and friends who have supported me throughout this time. I would especially like to thank my husband for his constant support, patience and for not letting me doubt myself.

Without any of you, this thesis would not have been possible. Thank you very much!

# Contents

Affidavit . . . iii
Submitted Articles . . . v
Acknowledgements . . . viii

**I Introduction 1**

**1 Introduction 3**

**II Methodological Advances in Topic Modelling 7**

**2 The Impact of Text Preprocessing on Topic Quality 9**
2.1 Introduction . . . 11
2.2 Preprocessing of Text Data . . . 13
2.3 Monte Carlo Study Design . . . 14
2.3.1 Corpora Generation . . . 14
2.3.2 Removal of Infrequent Terms . . . 15
2.3.3 Evaluation . . . 17
2.4 Results . . . 18
2.5 Conclusions . . . 21

**Appendix A to Chapter 2 23**

**3 Sampling Uncertainty of Topic Modelling 29**
3.1 Introduction . . . 31
3.2 Methods . . . 32
3.2.1 Bootstrapping . . . 32
3.2.2 Structural Topic Modelling . . . 33
3.2.3 Implementation and Evaluation . . . 33
3.2.3.1 Measures . . . 34
3.2.3.2 Word Cloud Uncertainty . . . 35
3.2.3.3 Confidence Bands for Topic Weight Time Series . . . 36

3.3 Application to Scientific Abstracts . . . 37
3.4 Bootstrap Results . . . 39
3.4.1 Measures . . . 39
3.4.2 Word Cloud Uncertainty . . . 41
3.4.3 Confidence Bands for Topic Weight Time Series . . . 44
3.5 Conclusions . . . 45

**4 Cross-Corpora Comparisons of Topics and Topic Trends 47**
4.1 Introduction . . . 49
4.2 Data and methods . . . 50
4.2.1 Topic modelling and corpora comparisons . . . 50
4.2.2 Textual data for Germany and Poland . . . 52
4.3 Methodological advances . . . 54
4.3.1 Topic number selection based on singular Bayesian information criterion . . . 54
4.3.2 Topic matching . . . 56
4.3.3 Embedding based matching . . . 58
4.3.4 Topic trends comparison . . . 60
4.4 Results . . . 60
4.4.1 Number of topics . . . 60
4.4.2 Topics . . . 62
4.4.3 Matching of topics . . . 64
4.4.4 Embedding based matching of topics . . . 65
4.4.5 Time series of topic weights . . . 66
4.5 Conclusions and outlook . . . 67

**Appendix B to Chapter 4 71**

**5 Choosing the Number of Topics in LDA Models 111**
5.1 Introduction . . . 113
5.2 Model Selection Criteria for LDA . . . 114
5.2.1 Topic Similarity . . . 114
5.2.2 Topic Coherence . . . 115
5.2.3 OpTop Criterion . . . 115
5.2.4 Singular Bayesian Information Criterion . . . 116
5.3 Monte Carlo Simulations . . . 118
5.3.1 Procedure . . . 118

5.3.2 Data Generating Processes . . . . . 120
5.3.3 Details of Implementation . . . . . 121
5.4 Results . . . . . 122
5.4.1 Number of Topics . . . . . 122
5.4.2 Structure and Content of Topics . . . . . 126
5.5 Conclusions and Outlook . . . . . 129

**Appendix C to Chapter 5** **131**

**III Text-as-Data Applications in Economics** **141**

**6 Topic Trends in Academic Literature and Economic Indicators** **143**
6.1 Introduction . . . . . 145
6.2 Textual and Economic Data . . . . . 146
6.2.1 Transforming Text to Time Series . . . . . 146
6.2.2 Selecting Economic Indicators . . . . . 148
6.3 Methods . . . . . 149
6.4 Results . . . . . 150
6.5 Conclusions . . . . . 160

**Appendix D to Chapter 6** **163**

**7 Uncertainty Indices based on Multilingual Text Representations** **171**
7.1 Introduction . . . . . 173
7.2 Text Representation Techniques . . . . . 173
7.3 Data . . . . . 174
7.4 Construction of Uncertainty Indices . . . . . 174
7.5 Results . . . . . 176
7.5.1 EPU Indices . . . . . 176
7.5.2 VAR Models . . . . . 176
7.6 Conclusions . . . . . 179

**Appendix E to Chapter 7** **181**

**8 Fiscal Policy in the Bundestag** **199**
8.1 Introduction . . . . . 201
8.2 Text data . . . . . 203
8.2.1 Bundestag speeches as a novel data source . . . . . 204

8.2.2 Corpus and text preprocessing . . . . . 205
8.3 A text-based fiscal sentiment indicator . . . . . 206
8.3.1 Compiling a dictionary on fiscal policy . . . . . 207
8.3.2 Doc2Vec approach . . . . . 208
8.3.3 Baseline sentiment . . . . . 210
8.3.4 Exogenous vs endogenous fiscal sentiment . . . . . 211
8.4 Estimating the macroeconomic effects . . . . . 215
8.4.1 VAR model . . . . . 215
8.4.2 Data . . . . . 216
8.4.3 Identification . . . . . 217
8.4.4 Results . . . . . 218
8.4.5 Robustness . . . . . 223
8.4.6 Exogenous and endogenous fiscal sentiment . . . . . 225
8.5 Fiscal foresight revisited . . . . . 226
8.6 Conclusions . . . . . 229

**Appendix F to Chapter 8** **231**

**9 Disagreement about Fiscal Policy** **251**
9.1 Introduction . . . . . 253
9.2 Constructing indicators of disagreement from textual data . . . . . 256
9.3 Some stylized facts on fiscal policy disagreement . . . . . 258
9.4 The effects of fiscal disagreement . . . . . 262
9.4.1 A VAR model . . . . . 262
9.4.2 Results . . . . . 263
9.4.3 The role of the majority in the Bundesrat majority . . . . . 264
9.5 Alternative Specifications of VAR model . . . . . 268
9.6 Conclusions . . . . . 273

**Appendix G to Chapter 9** **275**

**IV Conclusion** **281**

**10 Conclusion** **283**

**Bibliography** **285**

# List of Figures

1.1 Structure of the thesis . . . . . 4

2.1 Average vocabulary size depending on the relative cut-off value . . 16
2.2 Best Matching: example . . . . . 18
2.3 Evaluation of document frequency-based vocabulary pruning for DGP1 19
2.4 Evaluation of document frequency-based vocabulary pruning for DGP2 20
A.1.1 Evaluation of absolute term frequency based vocabulary pruning for DGP1 . . . . . 23
A.1.2 Evaluation of absolute term frequency based vocabulary pruning for DGP2 . . . . . 24
A.1.3 Evaluation of TF-IDF based vocabulary pruning for DGP1 . . . . . 25
A.1.4 Evaluation of TF-IDF based vocabulary pruning for DGP2 . . . . . 25
A.2.5 Recall values and additional statistics for DGP1 . . . . . 26
A.2.6 Recall values and additional statistics for DGP2 . . . . . 26
A.2.7 Recall values for DGP1 . . . . . 26
A.2.8 Recall values for DGP2 . . . . . 27

3.1 sBIC values for abstracts corpus . . . . . 37
3.2 Wordclouds of selected original topics . . . . . 38
3.3 Relative topic weights . . . . . 39
3.4 Model fit and the 5% percentile over 500 replications . . . . . 40
3.5 Cosine distances for selected topics . . . . . 40
3.6 Defined cut-off values and corresponding recall values (bin-width=0.05) 41
3.7 Word Overlay Clouds exhibiting uncertainty of top-word weights . 43
3.8 Cross-sample Frequency Clouds exhibiting uncertainty about top words 44
3.9 Modified percentile bootstrap confidence intervals for topic weight time series . . . . . 45

4.1 Outline of the methods pipeline for corpora comparison . . . . . 50
4.2 Language of the articles in the German text corpus. . . . . 53
4.3 Distribution of the sBIC values for the Polish corpus . . . . . 61

4.4 Distribution of the sBIC values for the German corpus . . . . . 61
4.5 Evaluations metrics from *tm toolkit* . . . . . 62
4.6 $PL^{ENG}$ Topics . . . . . 63
4.7 German Topics . . . . . 63
4.8 Distribution of cosine similarities between all possible matches . . . 64
4.9 Topic Match "International economic relationships" . . . . . 65
4.10 Topic Match "Business cycle" . . . . . 65
4.11 Topic Match "Unemployment" . . . . . 66
4.12 Topic Match "Capital growth" . . . . . 66
4.13 Topic Match "International economic relationships" . . . . . 67
4.14 Topic Match "Business cycle" . . . . . 68
B.3.1 Distribution of the sBIC values for the joint German corpus . . . . 77
B.3.2 Topic Matching using Machine Translation . . . . . 78
B.5.3 Distribution of the Jensen-Shannon distances . . . . . 79
B.5.4 Differences in the assignment . . . . . 80

5.1 Generic procedure for Monte Carlo simulations with a given selection criterion . . . . . 118
5.2 Comparison of evaluation metrics for DGP1 . . . . . 123
5.3 Comparison of evaluation metrics for DGP2 . . . . . 123
5.4 Comparison of evaluation metrics for DGP3 . . . . . 123
C.2.1 Distribution of the pairwise cosine similarity values. . . . . 133
C.2.2 Similar topics in DGP 1 . . . . . 134
C.2.3 Similar topics in DGP 2 . . . . . 134
C.2.4 Similar topics in DGP 3 . . . . . 134
C.3.5 Precision and recall for DGP1 . . . . . 135
C.3.6 Precision and recall for DGP2 . . . . . 135
C.3.7 Precision and recall for DGP3 . . . . . 136
C.4.8 Precision and recall based on cosine similarity for DGP1 . . . . . 139
C.4.9 Precision and recall based on cosine similarity for DGP2 . . . . . 139
C.4.10 Precision and recall based on cosine similarity for DGP3 . . . . . 140

6.1 Relative topic importance for Poland . . . . . 150
6.2 Relative topic importance for Germany . . . . . 151
6.3 Word clouds for the topic "Capital and growth" . . . . . 153
6.4 Word clouds for the topic "International economics" . . . . . 153
6.5 Word clouds for the topic "Foreign trade" . . . . . 154

6.6 Word clouds for the topic “Monetary policy” . . . . . 155
6.7 Word clouds for the topic “Business cycle” . . . . . 155
6.8 Word clouds for the topic “Crude oil market” . . . . . 156
6.9 Word clouds for the topic “Banking and credit” . . . . . 157
6.10 Word clouds for the topic “Stock market” . . . . . 158
6.11 Word clouds for the topic “Labour market” . . . . . 159
6.12 Word clouds for the topic “Energy sector” . . . . . 159
D.2.1 Economic indicators and topic weights . . . . . 166
D.2.1 (cont.) Economic indicators and topic weights . . . . . 167
D.2.1 (cont.) Economic indicators and topic weights . . . . . 168
D.2.1 (cont.) Economic indicators and topic weights . . . . . 169

7.1 Cosine Similarity Values between *policy* and all German words . . . 175
7.2 *dic*1 and *dic*2 EPU Indices . . . . . 177
7.3 *art_emd* EPU Indices . . . . . 178
7.4 *art_sbert* EPU Indices . . . . . 178
7.5 MCTM with 40 Topics: EPU Related Topics . . . . . 179
7.6 IRFs: *dic*2 EPU Index →Industrial Production Index . . . . . 180
E.1.1 Example of Multilingual Word Embeddings provided by fastText library, dimensions reduced using Principal Component Analysis . . 181
E.1.2 The Transformer’s Structure. Adapted from: Alammar (2018) . . . 182
E.3.3 Time Series of Selected EPU Topics in Germany . . . . . 190
E.3.4 Time Series of Selected EPU Topics in Russia . . . . . 191
E.3.5 Time Series of Selected EPU Topics in Ukraine . . . . . 192
E.4.6 MCTM: “Science” topic . . . . . 195
E.5.7 BBD and *dic*2 EPU Indices . . . . . 197

8.1 Number of speeches . . . . . 206
8.2 Length of speeches before and after preprocessing . . . . . 207
8.3 Rolling-window Doc2Vec . . . . . 209
8.4 Fiscal policy sentiment . . . . . 211
8.5 Average fiscal sentiment per election period . . . . . 212
8.6 Word clouds of topics “Bundeswehr" (topic 47) and “Budget, Debt, Investment" (topic 79) . . . . . 213
8.7 Exogenous and endogenous fiscal sentiment . . . . . 214
8.8 Cross-correlation of sentiment with macroeconomic variables . . . . 215
8.9 Fiscal sentiment in selected episodes . . . . . 216

8.10 Response to fiscal sentiment (entire Bundestag) . . . . . 219
8.11 Response to fiscal sentiment (government) . . . . . 219
8.12 Response to fiscal sentiment (opposition) . . . . . 220
8.13 Response to fiscal sentiment (Bundestag): extended VAR . . . . . 221
8.14 Response to fiscal sentiment (Bundestag): extended VAR . . . . . 222
8.15 Response to fiscal sentiment (Bundestag): extended VAR . . . . . 222
8.16 Response to fiscal sentiment: open-economy VAR . . . . . 223
8.17 Response to fiscal sentiment: detrending . . . . . 224
8.18 Response to fiscal sentiment: detrending . . . . . 225
8.19 Cumulative fiscal multipliers . . . . . 226
8.20 Response to exogenous fiscal sentiment . . . . . 227
8.21 Response to endogenous fiscal sentiment . . . . . 227
8.22 Response of government expenditure shock to sentiment . . . . . 229
F.1.1 Visualization of trained text vectors for 2020Q1 . . . . . 232
F.2.2 Comparing fiscal policy sentiment in the Bundestag across methods 234
F.2.3 Response to fiscal sentiment (dictionary I) . . . . . 235
F.2.4 Response to fiscal sentiment (dictionary II) . . . . . 236
F.2.5 Response to exogenous fiscal sentiment (dictionary I) . . . . . 236
F.2.6 Response to endogenous fiscal sentiment (dictionary I) . . . . . 237
F.2.7 Response to exogenous fiscal sentiment (dictionary II) . . . . . 237
F.2.8 Response to endogenous fiscal sentiment (dictionary II) . . . . . 238
F.6.9 Response of fiscal sentiment to tax shocks . . . . . 246
F.7.10 Response to fiscal sentiment (Bundestag): inflation and monetary policy . . . . . 247
F.8.11 Response to fiscal sentiment (entire Bundestag): lag order . . . . . 248
F.9.12 Response to government spending . . . . . 249

9.1 Disagreement in the Bundestag . . . . . 259
9.2 Disagreement in each election period . . . . . 260
9.3 Average disagreement across all election periods . . . . . 261
9.4 Disagreement during recessions . . . . . 262
9.5 Response to disagreement between the government and the opposition 264
9.6 Response to disagreement within the government . . . . . 265
9.7 Disagreement between the government and the opposition and the votes in the Bundesrat . . . . . 266

9.8 Response to disagreement between the government and the opposition: controlling for Bundesrat votes . . . . . 267
9.9 Response to disagreement within the government (controlling for Bundesrat votes) . . . . . 267
9.10 Response to disagreement between the government and the opposition (alternative ordering) . . . . . 269
9.11 Response to disagreement within the government (alternative ordering) 269
9.12 Response to disagreement between the government and the opposition (including election period-specific time trends) . . . . . 271
9.13 Response to disagreement within the government (including election period-specific time trends) . . . . . 271
9.14 Response to disagreement between the government and the opposition (including the budged balance) . . . . . 272
9.15 Response to disagreement within the government (including the budget balance) . . . . . 272
9.16 Response to disagreement . . . . . 273
G.2.1 Disagreement in the Bundestag: non-standardized data . . . . . 276
G.2.2 Disagreement in the Bundestag: maximum vs minimum sentiment 277
G.2.3 Response to maxmimum/minimum disagreement . . . . . 278
G.4.4 Response to disagreement between the government and the opposition (correction outlier) . . . . . 280
G.4.5 Response to disagreement within the government (correcting outlier) 280

# List of Tables


2.1 Characteristics of DGPs . . . . . 15

3.1 Reproduced share of selected topics . . . . . 41

B.1.1 Number of Articles published in Journal of Economics and Statistics 72

B.1.2 Proceedings from Macromodels International Conference and joint meetings . . . . . 73

B.1.3 Articles published in CEJEME . . . . . 74

B.4.4 Sklearn vs Gensim: model evaluation . . . . . 78

B.6.5 Comparison of filtered weight time series . . . . . 80

5.1 Evaluation of different criteria . . . . . 124

5.2 Percentages of the estimated number of topics, $K_{\text{metric}}$, falling within intervals around the true number of topics, $K_{\text{true}}$ . . . . . 125

5.3 Descriptive statistics of recall, precision, and F1 scores . . . . . 128

C.4.1 Descriptive statistics of recall, precision, and F1 scores based on cosine similarity . . . . . 138

6.1 Summary of Common Topics and Selected Economic Indicators . . 148

6.2 Mean topics weights . . . . . 151

6.3 Granger causality tests results . . . . . 152

D.1.1 Summary of common topics and selected economic indicators . . . 164

D.2.2 Data . . . . . 165

D.2.3 Data for the model of oil shocks . . . . . 165

D.2.4 Data transformations . . . . . 169

D.2.5 Model details . . . . . 170

E.2.1 EPU Terms by Baker et al. (2016) . . . . . 184

E.2.2 EPU Terms for Germany by Baker et al. (2016) . . . . . 184

E.2.3 EPU Terms for Russia by Baker et al. (2016) . . . . . 184

E.2.4 EPU Terms and their Cosine Similarity Values for Germany based on *dic*1 approach . . . . . 185

E.2.5 EPU Terms and their Cosine Similarity Values for Russia based *dic*1 approach . . . . . 186
E.2.6 EPU Terms and their Cosine Similarity Values for Ukraine based on *dic*1 approach . . . . . 187
E.2.7 EPU Terms and their Cosine Similarity Values for Germany, Russia, and Ukraine based on *dic*2 approach . . . . . 188
E.6.8 Granger Causality Test Results (p-values) . . . . . 198

F.1.1 Term usage over time . . . . . 233
F.5.2 Endogenous fiscal policy-related topics in Bundestag speeches . . . 245
F.5.3 Exogenous fiscal policy-related topics in Bundestag speeches . . . . 245

G.1.1 Determinants of disagreement . . . . . 275

# Part I

# Introduction

# Chapter 1

# Introduction

This PhD thesis focuses on text in various formats as a primary data source. Text data have become increasingly important in a wide range of domains over the last few decades. Also in economics, textual data from news articles (Thorsrud, 2020; Kalamara et al., 2020; Mamaysky, 2023; Ellingsen et al., 2022; Adämmer et al., 2025), communications on social media platforms (Lüdering & Tillmann, 2020), government press releases (Debnath & Bardhan, 2020), companies' websites (Dörr et al., 2022) etc. have been extensively used to enhance new perspectives and solution approaches to new as well as existing problems and questions. "Text-as-Data" topic already sparked my interest at the end of my Bachelor's studies. During my Master's studies I had already gained some experience in this area, and I wanted to deepen my knowledge as part of a PhD. This concerns both the methods commonly used in text-as-data applications as well as specific applications in the field of economics, i.e. the development of text-based indicators and their integration into macroeconomic analysis. I was lucky enough to be able to do this during my PhD, while working on several projects between 2021 and 2024.

This thesis is divided into four parts. The first part contains a general introduction, the current chapter. This thesis is a collection of eight papers which are grouped into two main parts of this thesis, the second part is entitled "Methodological Advanced in Topic Modelling" and the third part is entitled "Text-as-Data Applications in Economics". Each chapter in the second and third parts corresponds to a specific paper. At the beginning of each chapter, the co-authors are listed with their respective contributions and the current status of the papers. In order to improve the joint presentation format in this thesis, the individual papers have been slightly adapted in a way that does not affect the content. Figure 1.1 visualises the structure of the thesis.

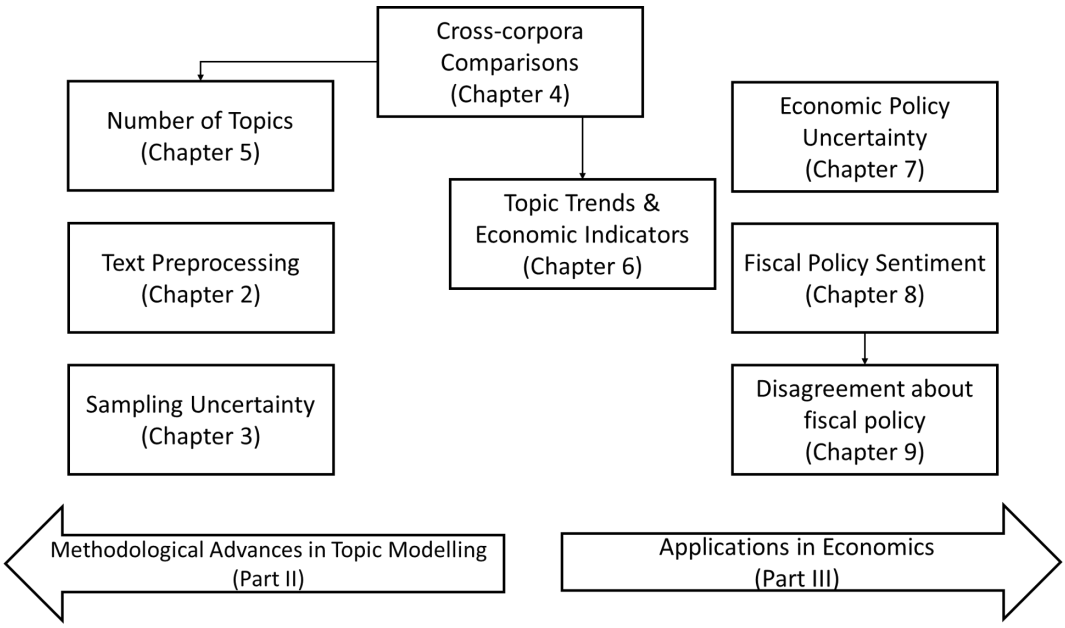



**Figure 1.1:** Structure of the thesis

The second part focuses on topic modelling as one of the well-established unsupervised techniques that is widely used to obtain first insights into the latent structures of a text collection. Latent Dirichlet Allocation (LDA) introduced by Blei et al. in 2003 is one of the most frequently used topic modelling approaches. The papers in this section deal with different aspects of this approach that might have an impact on the results. The first paper focuses on the pre-estimation phase of topic modelling, namely text preprocessing (chapter 2). The paper evaluates the impact of removing infrequent terms prior to estimation on topic quality. To provide generalizable results, a comprehensive Monte Carlo study considering two different data generation processes is conducted. Some practical recommendation are drawn on how to set the critical thresholds for removal of infrequent terms. The second paper is dedicated to analysis of sampling uncertainty in LDA applications (chapter 3). A bootstrap-based method for investigating sampling uncertainty is proposed, as well as approaches for measuring and visualising this uncertainty. The third paper, which is arranged in the middle of the two main directions in Figure 1.1, is the origin of two other papers and focuses on possible techniques for comparing the results of topic modelling of two different text collections (chapter 4). This procedure is referred to as topic matching. In addition, a new metric, singular Bayesian information criterion (sBIC), is used to identify an optimal number of topics. Both the sBIC metric and the proposed topic matching techniques are applied to collections of scientific publications in Germany and Poland to demonstrate their functionality. The fourth paper builds on the results of the third paper, namely the potential of the newly implemented

sBIC for identifying the optimal number of topics in the LDA context (chapter 5). To test its performance, when applied to different corpora and in comparison to other existing metrics, a Monte Carlo (MC) study is conducted. The results of this comprehensive MC study indicate that sBIC outperforms other criteria in identifying the true number of topics across different data generating processes (DGPs).

The third part of this thesis consists of four papers that deal with text-based indicators in an economic context. The first paper in this part considers scientific trends in publications in Germany and Poland and analyses their possible relationships with real economic indicators (chapter 6). This paper builds on the results of the paper in chapter 4, namely the uncovered topics and the identified topic matches for these text collections (see Figure 1.1). The analysis indicated significant links between scientific literature and real developments for 12 out of 13 identified topic-indicator pairs. The second paper in this part is dedicated to measurement of economic policy uncertainty in different countries based on vector representations of news articles (chapter 7). The results show the ability of the constructed indices as high-frequency indicators of economic activity. The third paper focuses on parliamentary debates in the Bundestag in order to extract relevant signals on the sentiment towards fiscal policy in Germany (chapter 8). To do so, a specific dictionary is created that contains revelant terms related to fiscal policy measures. Based on vector representations of the speeches given in the German Bundestag, a final fiscal policy sentiment is constructed. Further analysis using vector autoregressive (VAR) models shows that the constructed index has real macroeconomic effects. These results motivated a follow-up study, which is presented in the last paper of this part (chapter 9). This paper analyses fiscal policy disagreement in the German Bundestag and estimates its macroeconomic effects. Two types of disagreement are considered: disagreement between the government and the opposition, and disagreement within the coalition government parties. One of the main findings is that an increase in the latter has a contractionary effect on the economy. Disagreement between the government and the opposition, on the other hand, does not seem to affect the business cycle.

The fourth and final part contains concluding remarks on all the papers. It summarises the main findings and suggests possible lines of research.

# Part II

# Methodological Advances in Topic Modelling

# Chapter 2

# Analysing the Impact of Removing Infrequent Terms on Topic Quality in Latent Dirichlet Allocation Models

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Analysing the Impact of Removing Infrequent Terms on Topic Quality in Latent Dirichlet Allocation Models |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 40%),<br>Victor Bystrov (contribution: 20%),<br>Anna Staszewska-Bystrova (contribution: 20%),<br>Peter Winker (contribution: 20%) |
| **Status:** | *Working Paper*; submitted to *Central European Journal of Economic Modelling and Econometrics* |
| **Available from:** | `https://doi.org/10.48550/arXiv.2311.14505` (earlier version) |

Earlier versions of this paper were presented at:

- BERD@NFDI Research Symposium, Mannheim/Germany, 2024
- 26th International Conference on Computation Statistics (COMPSTAT), Giessen/Germany, 2024

# Analysing the Impact of Removing Infrequent Terms on Topic Quality in Latent Dirichlet Allocation Models*

Victor Bystrov† Viktoriia Naboka-Krell‡
Anna Staszewska-Bystrova†,‖ Peter Winker‡

**Abstract.**

An initial procedure in text-as-data applications is text preprocessing. One of the typical steps, which can substantially facilitate computations, consists in removing infrequent terms believed to provide limited information about the corpus. Despite popularity of vocabulary pruning, not many guidelines on how to implement it are available in the literature. The aim is to fill this gap by examining the effects of removing infrequent terms for the quality of topics estimated using Latent Dirichlet Allocation. The analysis is based on Monte Carlo experiments taking into account different criteria for infrequent terms removal and various evaluation metrics. The results indicate that pruning is often beneficial and that the share of vocabulary which might be eliminated can be quite considerable.

*Key Words: Topic models, text analysis, latent Dirichlet allocation, Monte Carlo simulation, text generation, text preprocessing*

*JEL classification: C49*

* Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

† University of Lodz, Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

‡ Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

‖ Corresponding author: anna.bystrova@uni.lodz.pl

## 2.1 Introduction

The use of topic modelling techniques, especially Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003), is growing fast. The methods enable the analysis of large collections of texts in an unsupervised manner by uncovering latent structures (topics) behind the data. They find application in a broad variety of domains, including economics and econometrics (see e.g. Edison & Carcel, 2021; Bystrov et al., 2024b; Adämmer et al., 2025).

Given this increasing use of LDA as a standard tool for empirical analysis, the interest in details of the method, and, in particular, in parameter settings for its implementation is also rising. Thus, since the introduction of the LDA approach, different methodological components have been studied in more detail as, for example, the choice of the number of topics (Cao et al., 2009; Mimno et al., 2011; Lewis & Grossetti, 2022; Bystrov et al., 2024a), hyper-parameter settings (Wallach et al., 2009), model design (e.g. hierarchical structure as proposed by Teh et al. (2006)), and inference methods (Griffiths & Steyvers, 2004).

However, not only the setting of technical parameters of the LDA model and the estimation algorithms are crucial for the results obtained, e.g. the identified topics. As the algorithm behind LDA "learns" from data based on co-occurrences of terms within texts, these have to be prepared in an appropriate way. LDA requires the text data to be structured in a document-term matrix (DTM), where each row corresponds to a document and each column to a specific term used throughout all documents. Then, the entry in a cell of the matrix provides the frequency of the term within a certain document. To obtain this matrix, the documents in a text corpus are usually cleaned and each document is represented as a bag-of-words (BoW), i.e. the algorithm neglects the semantic relationships between words and sentences. Together, these steps are referred to as *preprocessing*. By removing irrelevant terms and merging very similar terms (e.g. singular and plural forms of the same noun), preprocessing helps to reduce both the dimension of the DTM and its sparsity, which affect the performance of the algorithms used to estimate the LDA model (Maier et al., 2018; Stoltenberg et al., 2020).

Even though text preprocessing is an inherent component of any LDA analysis, there appear to be no common standards on how to perform it. In their illustrative application, Blei et al. (2003), for example, mention removing a standard list of stop words and all terms with an absolute frequency of one, i.e. showing up only once in the full corpus. In fact, such a step is usually performed in the majority of text-as-data applications with different lists of stop words and alternative rules for removing low- and – sometimes also – high-frequency terms. However, only a few attempts have been made so far to analyze the impact of text preprocessing on uncovered topics.

Denny & Spirling (2018) address this question by examining the impact of different combinations of text preprocessing steps on the outcomes of unsupervised techniques, including LDA. They show sensitivity of the results to preprocessing decisions; however, since the analysis is done using real datasets, they cannot draw more general conclusions. The authors

highlight the importance of careful data preparation for unsupervised techniques, like LDA, because, unlike for supervised methods, the results cannot be evaluated in a well-defined procedure (e.g. through accuracy measures as in text classification tasks). Lu et al. (2017) focus on removing terms that occur only once and those that are frequently used in three different real-world datasets (medical abstracts, articles published in biomedical journals, bibliographic records and abstracts from Elsevier Arts & Humanities journals). They measure the impact of removing (in)frequent terms by means of four different metrics. All in all, the authors come to the conclusion that removing singly occurring terms (i.e. the reduction of the vocabulary size by 30% to 40% depending on the underlying dataset) does not impact the topic modelling outcome substantially. Schofield et al. (2017) conduct some experiments to test the effect of removing common terms on topic quality using two datasets, the United States State of the Union Addresses and the annotated corpus of the New York Times. The authors conclude that removing stop words prior to model estimation does not impact topic inference.

Tang et al. (2014) analyze the properties of the data that affect the inferential performance of LDA models. They conduct small-scale Monte Carlo experiments using an LDA generative process with varying parameter configurations. In each experiment Tang et al. (2014) generate 30 corpora and compare true and estimated topics. Although they do not study the effects of text preprocessing, the results of their analysis elucidate the deterioration effect of data sparsity on the performance of LDA models.

A growing number of studies examine the consequences of text preprocessing on the results of supervised techniques (see e.g. Alam & Yao, 2019; Barushka & Hajek, 2020; Reimann & Dakota, 2021; HaCohen-Kerner et al., 2020; Al Sharou et al., 2021). These studies show that preprocessing can improve the performance of machine learning classifiers. They also highlight that each preprocessing procedure and each combination of preprocessing steps may matter for the final results and indicate the need for further systematic studies of initial text preparation.

In this contribution, we focus on the impact of removing terms with low frequency on the results of LDA modelling. Usually, low-frequency terms make up a large proportion of unique terms occurring in a corpus. This feature common to many, if not all languages can be approximated by Zipf's law, stating in its simplest version that term frequency is proportional to the inverse of the term frequency rank. A slightly more complex model has been proposed and estimated by Mandelbrot (1953). However, terms occurring only with low frequency are believed to be too specific to contribute to the meaning of the resulting topics when applying the LDA algorithm. Additionally, removing those terms substantially decreases the vocabulary size and, consequently, accelerates model estimation.

To the best of our knowledge, little research has been done so far to analyze the impact of removing infrequent terms on LDA estimation results. Stoltenberg et al. (2020) focus on the consequences for topic quality of removing both frequent and infrequent words. They conduct their experiments on three different real-world datasets and conclude that vocabulary pruning does not qualitatively impact the resulting topics. To contribute to this line of

research, we conduct a Monte Carlo simulation study. First, we define the characteristics of the data generating processes (DGPs) and following the generative model described by Blei et al. (2003) create true document-topic and topic-term distributions. For each of the DGPs, we generate a total of 1 000 pseudo-corpora. Finally, we apply different techniques, which have been proposed in the literature, to define and remove infrequent terms. Afterwards, LDA models are estimated based on the preprocessed corpora. Eventually, we can analyze the impact of different settings on the estimation results.

The remainder of this paper is structured as follows. Section 2.2 introduces the steps that are usually performed for text data under the heading of text preprocessing. Focusing on removing infrequent terms, Section 2.3 describes the design of our Monte Carlo study. Next, in Section 2.4, we present and discuss the results of the experiments. Section 2.5 concludes.

## 2.2 Preprocessing of Text Data

Since texts are considered a very unstructured data source, text preprocessing usually precedes all other steps in text-as-data applications, regardless of the field of use. In general, these preprocessing steps can be divided into standard preprocessing steps and corpus or domain-specific preprocessing steps. In our description of changes applied to the vocabulary, we refer to "terms" as unique tokens included in the vocabulary and "words" as non-unique tokens in documents.

The standard preprocessing steps include the following: removing punctuation, special characters, and numbers; lowercasing; removing language specific stop words; lemmatizing or stemming. This list can be adjusted or extended by the so-called domain-specific preprocessing steps. For example, the character "#" falls into the category of special characters, but keeping it can be useful when working with Twitter data. In addition, the removal of extremely frequent and rare terms (relative pruning) could facilitate topic modeling.

Very frequent terms, also called corpus-specific stop words, occur in the majority of all documents and are often considered to be insufficiently specific to be useful for topic identification. Therefore, Grimmer & Stewart (2013) and Maier et al. (2018) remove all terms that appear in more than 99% of all documents.

Denny & Spirling (2018) provide two rationales for removing very rare terms: First, these terms contribute little information for topics retrieval, and, second, their removal reduces the size of the vocabulary substantially and, consequently, speeds up computations. A common rule of thumb, mentioned in Denny & Spirling (2018), is to discard terms that appear in less than 0.5-1% of the documents. Denny & Spirling (2018) notice, however, that there has been no systematic study of the effects this preprocessing choice has on the modeling of the topics.

Infrequent terms can be removed using one of the following criteria:

- Document frequency: remove terms with the frequency of showing up across the documents in the corpus below the defined threshold (absolute/relative).
- Term frequency: remove terms with frequency in the corpus below the defined threshold

(absolute/relative).

- Term Frequency-Inverse Document Frequency (TF-IDF) values describing relative importance of terms for specific documents: remove terms with low TF-IDF values (Blei & Lafferty, 2009)).

There are no obvious rules for setting the required thresholds. Grimmer & Stewart (2013) notice that the choice of thresholds for removing common and rare terms from a corpus should be contingent on the diversity of the vocabulary, the average length of documents and the size of the corpus. However, this is a heuristic observation that is not based on a systematic analysis.

## 2.3 Monte Carlo Study Design

To analyze the impact of removing infrequent terms in the context of LDA in a systematic way, we conduct a Monte Carlo simulation study. The purpose of the analysis is to provide insight into the effects of vocabulary pruning on topic quality in estimated LDA models. Given that the actual topics are known in the experiments, we focus in particular on the difference between the estimated and true topics. Obviously, this difference is driven only to some extent by the specific preprocessing used, but depends also on the sampling error, which we have to take into account when summarizing our findings.

In this section, we first describe the setup of simulation experiments. Then, we present the features of the DGPs and details of the procedure of corpora generation (subsection 2.3.1). Afterwards, we define and describe the rules for the removal of infrequent terms to be applied in the Monte Carlo study (subsection 2.3.2). Finally, we discuss different quality measures used to evaluate the results (subsection 2.3.3).[1]

### 2.3.1 Corpora Generation

We start by presenting two DGPs to be considered in the Monte Carlo study. Table 2.1 summarizes the main characteristics of these DGPs. The first one contains a relatively small number of long documents covering a moderately large number of topics. These characteristics are derived from some real-world datasets such as scientific publications, reports, or speeches (e.g. Hartmann & Smets (2018)). DGP2 has the characteristics of corpora containing a large number of short texts discussing a relatively small number of topics. They are typical for collections of conference abstracts, social media, microblogs etc.

[1] Code details for data generation, model estimation, and evaluation is available on Github at `https://github.com/VikaNa/removing-infrequent-words-lda`.

| | #documents | # words per document | # unique terms | # topics, $K$ |
|---|---|---|---|---|
| DGP1 | 1,000 | 3,000 | 30,000 | 50 |
| DGP2 | 10,000 | 150 | 20,000 | 15 |

**Table 2.1:** Characteristics of DGPs

Given these features of the DGPs, we follow the generative model described by Blei et al. (2003). For each DGP, the matrix of topic-term probabilities $\boldsymbol{\beta}$ is drawn from the Dirichlet distribution using a single concentration parameter $\eta = 1/K$. Algorithm 2.1 describes how each document $\mathbf{w}$ in a corpus $D$ is generated. The length of the document $N$ is defined by drawing from a Poisson distribution where the parameter $\xi$ is equal to the expected number of words in a document, namely 3,000 for DGP1 and 150 for DGP2. For each document $\mathbf{w}$, the vector of topic probabilities $\theta$ is drawn from the Dirichlet distribution using a concentration parameter $\alpha = 1/K$. For each word in a document, a topic $z_n$ is first drawn from the multinomial distribution parametrized by vector $\theta$ and then a term $w_n$ is drawn from the multinomial distribution given the topic $z_n$ and the matrix of topic-term probabilities $\boldsymbol{\beta}$. The choice of flat priors $1/K$ for the parameters $\alpha$ and $\eta$ is in accordance with the default parameter setting in software implementations, e.g. in Python's `scikit-learn` library (Pedregosa et al., 2011). These default values are used in many text-as-data applications.

**Algorithm 2.1** Generative probabilistic model by Blei et al. (2003)

Choose $\boldsymbol{\beta} \sim Dir(\eta)$

**for** document $\mathbf{w}$ in corpus $D$ **do**

  Choose $N \sim Poisson(\xi)$

  Choose $\theta \sim Dir(\alpha)$

  **for** word $w_n = 1, 2, \ldots, N$ **do**

    (a) Choose a topic $z_n \sim Multinomial(\theta)$

    (b) Choose a term $w_n$ from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial probability distribution conditioned on the topic $z_n$

  **end for**

**end for**

We use Algorithm 2.1 to generate 1 000 different pseudo-corpora for each DGP.

### 2.3.2 Removal of Infrequent Terms

A popular approach to vocabulary pruning is to remove all terms that appear in a small number of documents in the corpus. As indicated in Section 2.2, this criterion can be based on the absolute number of documents (e.g., remove all terms that occur in no more than

one document) or the relative number of documents (e.g., remove all terms that occur in no more than in 1 percent of all documents in the corpus). In the Monte Carlo experiments we consider different values of the relative cut-off for removing terms on the basis of relative document frequency.[2]

Before fixing the range of cut-off values, we consider the resulting distribution of the vocabulary size for each DGP: Figure 2.1 shows the average vocabulary size over 1 000 corpora as a function of the relative cut-off value (relative document based frequency) for each DGP. For the cut-off value of 1% that is often used in empirical applications, the vocabulary size decreases by 9.3% and 74.1% for DGP1 and DGP2, respectively. Given these differences in the relative distributions of vocabulary sizes for selected DGPs, in the simulations, we use different ranges of cut-off values. For DGP1, we proceed in steps of 0.5% within the interval $[0.0\%; 9.5\%]$. For DGP2, we reduce the step size to 0.25% up to the cut-off value of 2.5% and set the maximum cut-off value to 4% because higher thresholds would result in an empty vocabulary.

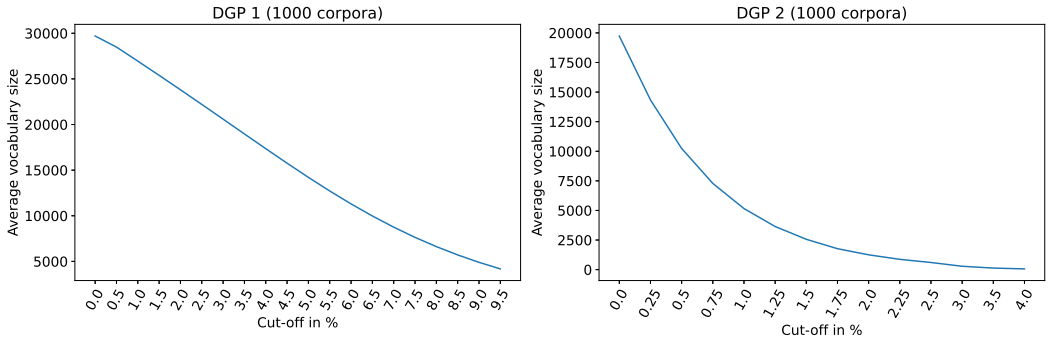



**Figure 2.1:** Average vocabulary size depending on the relative cut-off value

For every corpus generated from DGP1 we build 20 different sub-samples, according to the defined cut-off values. We use each subsample to estimate an LDA model. For every corpus obtained from DGP2, 14 different sub-samples are constructed and corresponding LDA models are estimated. For model estimation, we use Python's library `scikit-learn` (version 0.24.2). We leave all parameters at their default values except the number of topics, which is set according to the DGP characteristics defined in Table 2.1 and the maximum number of iterations, which is set to 100 due to computational constraints. In addition, we fix a random state for the reproducibility of the results.

As described in Section 2.2, there are two additional criteria for vocabulary pruning that are frequently used in applications: term frequency and TF-IDF frequency. We use them to perform robustness checks for the results presented in our study. The results of this robustness analysis are presented in Appendix A.1.

[2] We use Python's `scikit-learn` (version 0.24.2) library and its *CountVectorizer()* class to exectute this task.

### 2.3.3 Evaluation

Throughout each Monte Carlo scenario, we keep all the parameters constant, except for the document-term matrix required as input for the estimation of the LDA model. As described in the previous subsection, different variations of one corpus are created by applying various cut-off values to remove infrequent terms. As a result, we obtain 20 and 14 LDA models for DGP1 and DGP2, respectively.

Different evaluation techniques have been developed to access topic modelling quality. Some of them became standard in different text-as-data applications, e.g. topic coherence (Mimno et al., 2011) or topic similarity (Cao et al., 2009). The measure of Mimno et al. (2011) was designed to correspond with the judgement of the consistency of topics by humans. It is based on maximization of the average semantic coherence across a range of topics. The method of Cao et al. (2009) associates good topic quality with sharp topic distinction or lack of overlap. The proposed measure is computed by minimizing the average cosine similarity between each pair of topics.

Another popular measure used to evaluate the model's predictive performance on an unseen (or held-out) sample is perplexity. It is defined as the inverse of the geometric mean per-word likelihood. Blei et al. (2003) show that perplexity is monotonically decreasing in the likelihood of the test data with increasing number of topics. Reducing the size of the vocabulary while keeping the number of topics constant leads qualitatively to the same effects. For this reason, in the current study, we do not consider perplexity as an evaluation metric.

Instead, we also compute recall (or the share of reproduced topics) as proposed by Bystrov et al. (2024a) and model fit to evaluate the impact of removing infrequent terms on topic quality in LDA models.

First, using the *recall* metric, we aim to measure how the true structure of topics changes (by comparing *true* and *estimated* topic-term distributions). In the current work, we follow a similar approach to the one proposed by Bystrov et al. (2022) and apply the so-called *best matching*:

1. Compare true and estimated topic-term distributions based on the union of two vocabularies. For terms not contained in the estimated topic-term distribution, assign probability of zero. An example of this procedure is presented in Figure 2.2.

2. For each of the estimated topics, calculate *similarity/distance* to each of the true topics. Then, assign the true topic with the highest (lowest) similarity (distance).

3. Define and apply a cut-off value to keep good quality matches only. Calculate the *recall* metric as the share of correctly reproduced topics.

In their empirical application, Bystrov et al. (2022) use cosine similarity in step 2 and automatically determine a data-based cut-off as the 95% percentile of all pairwise similarity scores in step 3. Stoltenberg et al. (2020), who also studied the impact of removing infrequent terms on topic quality, perform topic matching based on top 20 topic terms following the

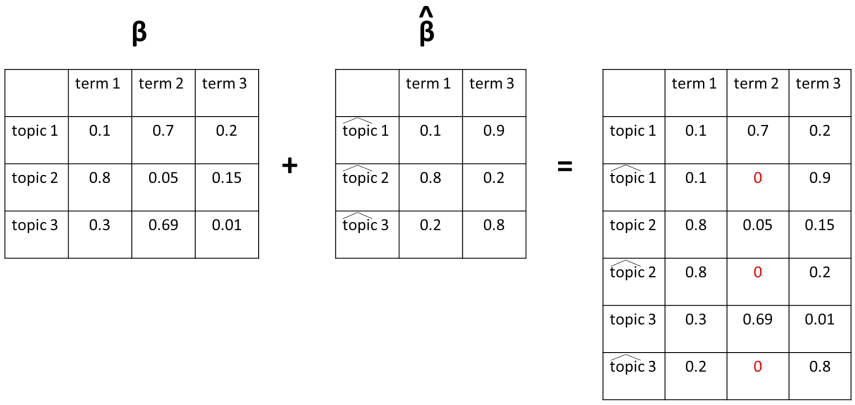



**Figure 2.2:** Best Matching: example

approach proposed by Niekler & Jähnichen (2012). The authors calculate pairwise cosine distances and apply a cut-off value of 0.5 to obtain the share of reproduced topics.

In the current application, we use different metrics to measure the similarity between true and estimated topics:

**a)** Cosine similarity: takes values between -1 (two vectors point in opposite directions) and 1 (two vectors point in the same direction).

**b)** Jensen-Shannon divergence/distance: ranges between 0 (two distributions are the same) and 1 (two distributions are completely different).

**c)** Rank-Biased Overlap (RBO) proposed by Webber et al. (2010) to compare ranked lists: ranges from 0 (ranked lists are disjoint) to 1 (ranked lists are exactly the same).

Since the true topics appear to be very distinct from each other in the current Monte Carlo study, we decided to use a cut-off value of 0.8 for the similarity metrics (cosine similarity and RBO) and 0.2 for the distance metric (Jensen-Shannon).

Alternatively, one can use *one-to-one matching* as described by Bystrov et al. (2022). The resulting measure is called *model fit*. Thereby, all of the topics have to be matched using the Hungarian algorithm and a defined distance metric. Matches are assigned to minimize the overall cost of assignment. Thus, the mean distances between the identified matches can be considered to measure the quality of the fit of the model.[3]

## 2.4 Results

In this section, we summarize the main findings of the Monte Carlo analysis. Thereby, we focus on the removal of infrequent terms according to their document frequency in the corpus as described in Section 2.3.2. The results presented here are based on 1 000 replications.

[3] In the case of simulated data, the model fit metric based on one-to-one matching offers a different focus on the defined problem. Since the true number of topics is known, it is of special interest to see how the true and estimated topics are assigned to each other when none of the topics is left out.

We also perform robustness checks using 100 replications for the alternative criteria for vocabulary pruning, namely absolute term frequency and TD-IDF values, and present the results in Appendix A.1.

Figures 2.3 and 2.4 present the metrics values obtained after document frequency pruning. The cutoff values exhibited on the $x$ axis in Figures 2.3 and 2.4 describe the minimum share of documents in which a term must be included to not be removed from the corpus. Thus, a cut-off value of 0.0% corresponds to keeping all terms (30K for DGP1 and almost 20K for DGP2), while 9.5% in Figure 2.3 refers to the removal of all terms which do not show up in at least 9.5% of all documents leaving only about 4K terms in the corpus. Consequently, in Figure 2.4, the value of 4.0% corresponds to keeping only those terms, which appear in at least 4.0% of all documents reducing the size of the vocabulary to 60 terms.

On the ordinate, Figures 2.3 and 2.4 show, as solid lines, the means of the evaluation metrics obtained over 1 000 replications for DGP1 and DGP2, respectively. The dashed lines in the first three subplots provide the 20% and 80% quantiles of the distributions of these metrics. The corresponding bands for the measures from the last panel (*recall*) are shown in Figures A.2.7 and A.2.8 in Appendix A.2. The metrics considered include: *model fit* (Bystrov et al., 2024a) (to be minimized), topic *similarity* (Cao et al., 2009) (to be minimized), topic *coherence* (Mimno et al., 2011) (to be maximized), and *recall* (to be maximized). In empirical applications, the true DGPs and corresponding topics are unknown. Thus, the *recall* criteria cannot be applied. The observed collapse of *recall* for higher cut-off values indicates that the remaining vocabulary is no longer sufficient to identify the true topics.

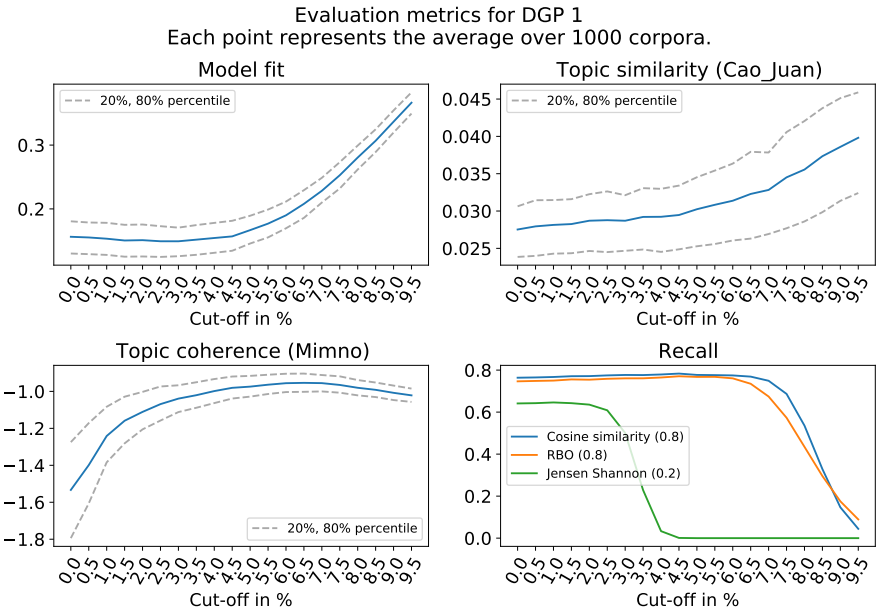



**Figure 2.3:** Evaluation of document frequency-based vocabulary pruning for DGP1

It becomes obvious from Figures 2.3 and 2.4 that removing infrequent terms has consequences for the results of the LDA estimation. As a general pattern, we conclude that the application of pruning is beneficial for low cut-off values. This might be attributed to two effects. First, terms appearing only in a few documents do not contain much informa-

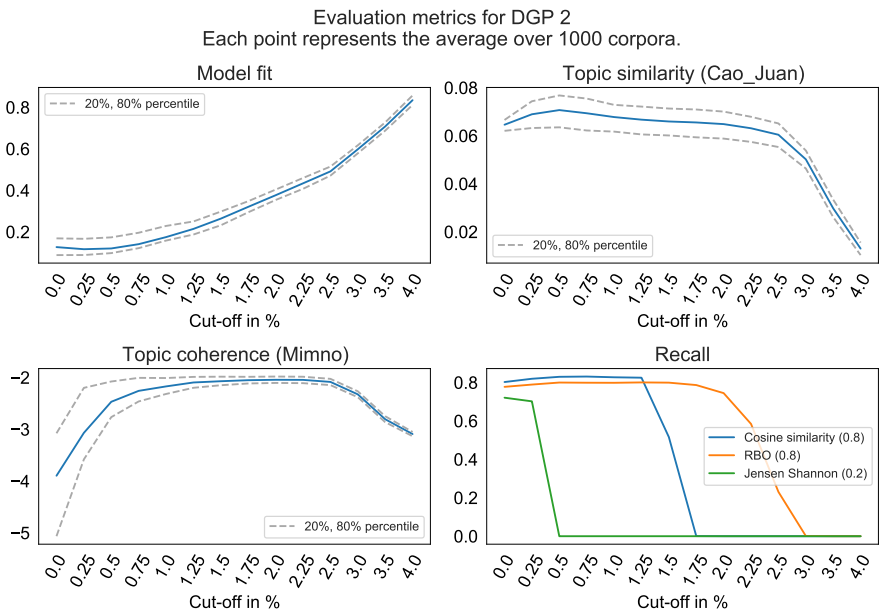



**Figure 2.4:** Evaluation of document frequency-based vocabulary pruning for DGP2

tion about more general topics. Second, removing these terms substantially reduces the dimensionality of the estimation problem, which increases the efficiency of the estimators. However, beyond a certain point the increasing loss of information resulting from the removal of more and more infrequently used terms dominates the gains due to reduced dimensionality. Comparing the findings from Figures 2.3 and 2.4, it appears that gains and losses from decreasing vocabulary size by eliminating rare terms are weighted somewhat differently by alternative evaluation criteria.

For DGP1 (Figure 2.3), the lowest average distances between the true and estimated topic sets as measured by *model fit* correspond to cut-off values from 3% to 4.5%. Further removal of infrequent terms leads to increased distortions in estimated topics. The best values of *coherence* are obtained for thresholds of 3%-6.5%. The metric is quite sensitive to keeping too many infrequent terms in the texts, showing significantly smaller values for initial thresholds. In the case of *similarity*, thresholds up to 4.5% lead to similar metric values. Eventually, alternative versions of *recall* measures indicate that the maximum threshold that might be considered is about 3% (metric based on Jensen-Shannon distance) or 6.5% (cosine similarity and RBO-based metrics). Altogether, if all metrics are considered jointly, the best threshold is about 3%. A similar conclusion is reached if the TF-IDF or absolute term frequency-based vocabulary pruning is performed instead of the document frequency pruning (see Appendix A.1).

A similar analysis for DGP2 (Figure 2.4) suggests the following cut-off values. According to *model fit* the interval from 0.25% to 0.75% could be considered, while the *coherence* metric indicates the range 0.5%-2.5%. Topic *similarity* is quite similar for cut-off values up to 0.5% and *recall* metrics suggest stopping at 0.25%, 1.25% or 2% starting from the most restrictive measure. Thus, in general, a threshold of about 0.25%-0.5% might be selected. This finding is again quite robust with respect to the criterion used for the removal of infrequent terms

(see Appendix A.1).

For a better understanding of the results from Figures 2.3 and 2.4, the selected thresholds were juxtaposed with the corresponding shares of terms removed from the vocabularies (see Figures A.2.5 and A.2.6 in Appendix A.2). The cut-off value of 3% for DGP1 corresponds to reducing the size of the vocabulary by 30% and cut-offs of 0.25-0.5% for DGP2 imply removing 27-48% of all terms. Thus, in both cases, it could be concluded that the reduction in vocabulary size, which could accelerate the estimation process without affecting the results qualitatively, was considerable and amounted to about 30% of all terms. These results show that guidelines focusing on removing infrequent terms up to a certain share of all terms might be worth following up. Our conclusions are also in line with the findings of Lu et al. (2017).

## 2.5 Conclusions

The focus of this paper was on preprocessing of text data in the context of LDA analysis. Although text preprocessing is an essential part of data preparation in text-as-data applications and some rules-of-thumb of text preprocessing sequences exist and are often followed, there is only little evidence on how particular text preprocessing decisions might affect the final results. In the specific setting considered in this paper, the outcome of interest were the estimated topics and the analyzed preprocessing step was the removal of infrequent terms in a text corpus.

To allow for a systematic evaluation of the impact of different techniques for reducing vocabulary size and generalizable conclusions, we conducted a Monte Carlo simulation study. We first generated data from scratch based on two pre-defined DGPs following the probabilistic model proposed by Blei et al. (2003). For each of the defined DGPs, we then applied different techniques to remove rare terms from the texts and estimated multiple LDA models varying the text input only. Finally, we evaluated the results using some well established metrics such as *coherence* and *topic similarity* that focus on the estimated set of topics as well as *model fit* and *recall* metrics that are based on the comparison between the true and estimated set of topics.

Our results indicate that appropriate removal of infrequent terms can improve the LDA estimation results. This is caused by the reduction of dimensions and the sparsity of the document-term matrix paired with a limited loss of information about the content of the topics.

The results have at least two practical implications. First, we show that across the DGPs considered, about 30% of terms can be removed without qualitative losses in the resulting topics. This is a valuable insight for the scientists who work with substantial sets of data containing long texts on average. Most real-world data sets have large or even very large vocabularies. In such cases, removing 30% of terms could lead to a considerable decrease in computing time and an increase in efficiency. Second, we demonstrate the robustness of these conclusions with respect to the application of different techniques to reduce the size of vocabularies. This implies that in practice, vocabulary pruning can be based on either of the

popular criteria.

Our results suggest that future research could follow the ideas of Denny & Spirling (2018) and focus on an evaluation of different combinations of text preprocessing steps. However, performing this analysis in a systematic manner by means of Monte Carlo experiments would require substantially more computational resources. For example, it might be worthwhile to consider the combined impact of stemming/lemmatizing and vocabulary pruning.

# Appendix A

## A.1 Robustness Checks

### Term frequency

This approach to vocabulary pruning is based on the absolute frequency of terms in the considered corpus. The rule was applied e.g. by Blei et al. (2003) who removed all terms that occurred only once in the corpus used in their illustrative example. To make the results based on term frequency comparable to the results based on document frequency, we consider specific sequences of cut-off values for each DGP. These thresholds are such that the vocabulary sizes were comparable to vocabulary sizes implied by document frequency cut-off values. To compute them we identify vocabulary sizes corresponding to the relative cut-offs applied in document frequency based pruning. Then, we identify minimum absolute term frequencies corresponding to the considered relative cut-offs. Figures A.1.1 and A.1.2 show the results for DGP1 and DGP2, respectively.

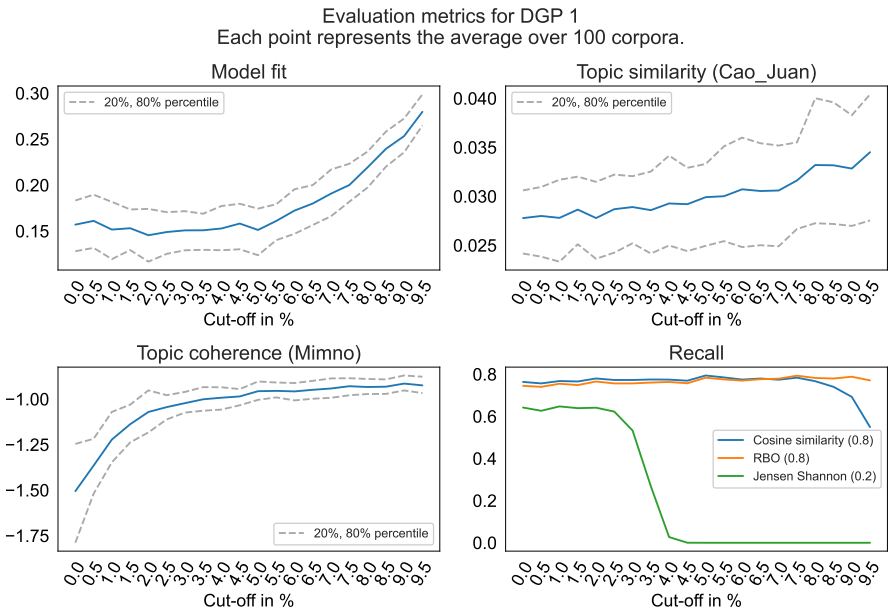



**Figure A.1.1:** Evaluation of absolute term frequency based vocabulary pruning for DGP1

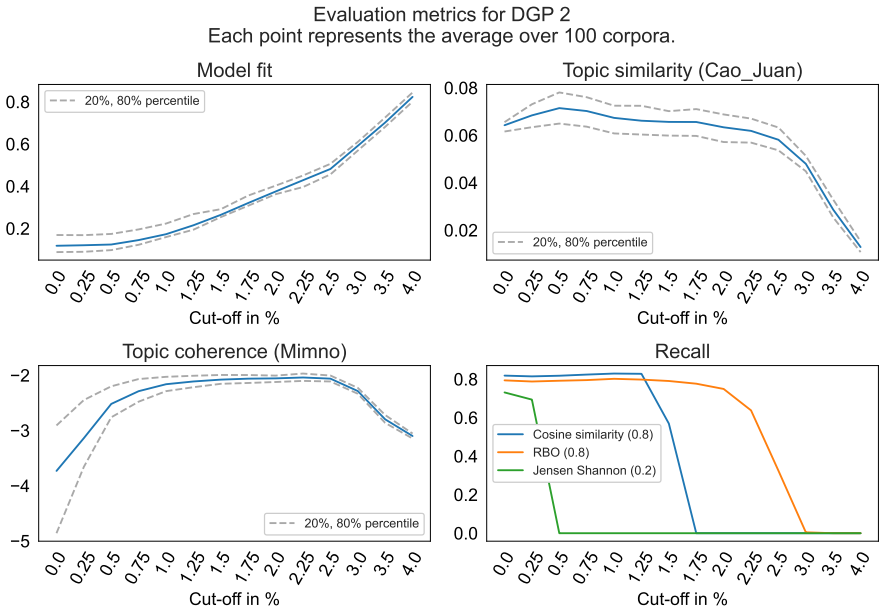



**Figure A.1.2:** Evaluation of absolute term frequency based vocabulary pruning for DGP2

### TF-IDF

Blei & Lafferty (2009) propose to use TF-IDF to prune the vocabulary. In their experiments, they consider the top 10,000 terms with highest TF-IDF values. TF-IDF is a weighted measure that is used to determine the importance of a term for a given corpus and consists of two parts, namely term frequency (TF) and inverse document frequency (IDF):

$$\text{Term Frequency}_{w,D} = \frac{\text{Number of times term } w \text{ appears in document } D}{\text{Total number of term } w \text{ in document } D} \tag{A.1.1}$$

$$\text{Inverse Document Frequency}_{w} = log\frac{\text{Total number of documents}}{\text{Number of documents with term } w} \tag{A.1.2}$$

The IDF part accounts for terms that occur in the majority of documents (e.g. stop words) and scales down their importance. Finally, TF-IDF score is calculated by multiplying TF and IDF:

$$\text{TF-IDF}_{w,D} = \text{Term Frequency}_{w,D} * \text{Inverse Document Frequency}_{w} \tag{A.1.3}$$

For each of the corpora generated from DGP1, we build 20 different sub-samples considering the top $V$ terms with the highest TF-IDF values. To make the results comparable, we choose $V$ equal to the vocabulary size that results when document frequency-based rules are applied (see Figure 2.1). For example, if applying a document frequency cut-off value of 6 percent results in a vocabulary size of about 10,000 terms for corpus $x$, we consider only 10,000 terms with the highest TF-IDF values for this corpus. Figures A.1.3 and A.1.4 present the results based on TF-IDF vocabulary pruning.

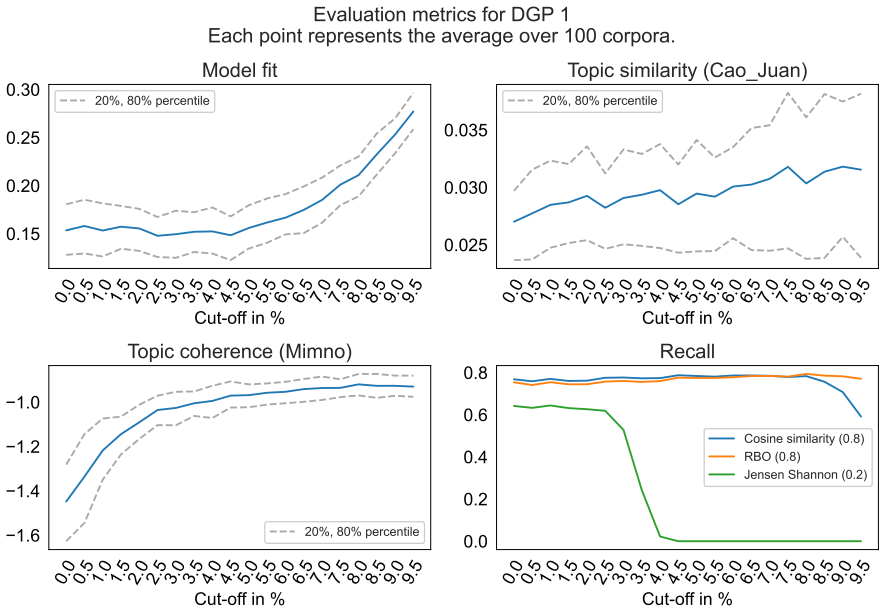



**Figure A.1.3:** Evaluation of TF-IDF based vocabulary pruning for DGP1

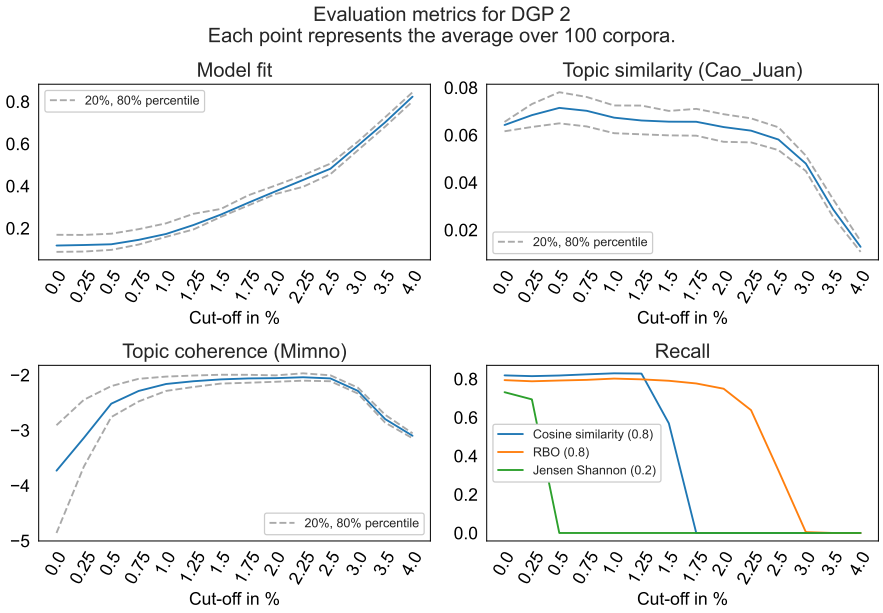



**Figure A.1.4:** Evaluation of TF-IDF based vocabulary pruning for DGP2

## A.2 Additional Visualizations

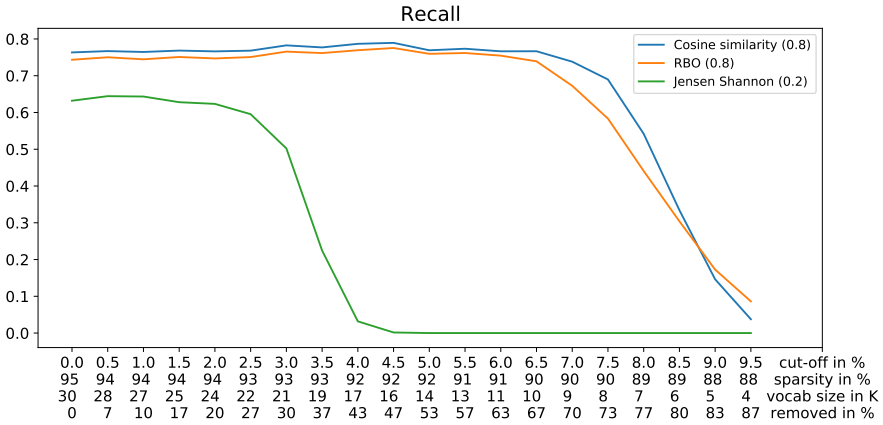



**Figure A.2.5:** Recall values and additional statistics for DGP1

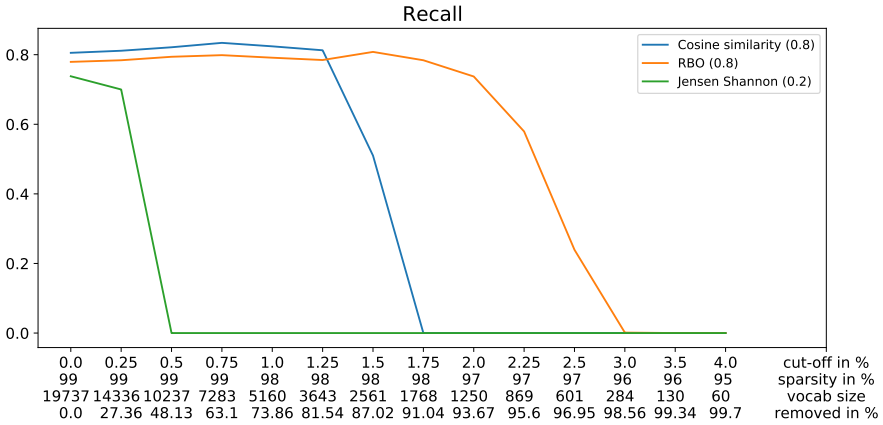



**Figure A.2.6:** Recall values and additional statistics for DGP2

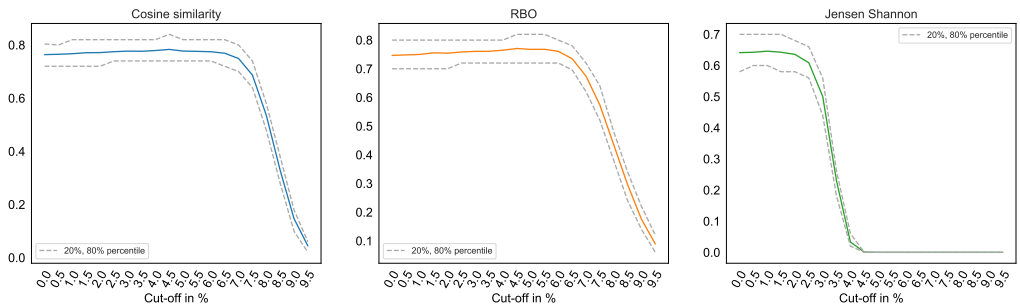



**Figure A.2.7:** Recall values for DGP1

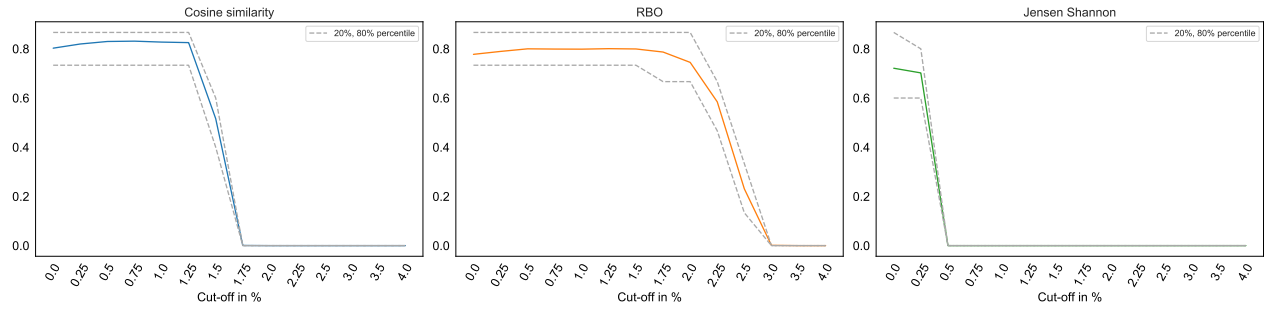



**Figure A.2.8:** Recall values for DGP2

# Chapter 3

# Sampling Uncertainty of Topic Modelling

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Sampling Uncertainty of Topic Modelling |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 40%),<br>Victor Bystrov (contribution: 20%),<br>Anna Staszewska-Bystrova (contribution: 20%),<br>Peter Winker (contribution: 20%) |
| **Status:** | *Working Paper*; will be submitted to *Scientometrics* |

# Sampling Uncertainty of Topic Modelling*

Victor Bystrov† Viktoriia Naboka-Krell‡
Anna Staszewska-Bystrova† Peter Winker‡

**Abstract.**

Topic modelling estimation results, related to topic-term and document-topic probabilities, are typically reported with no indication of sampling uncertainty. The lack of additional information on sampling uncertainty might result in misleading conclusions. We propose to measure sampling variability using the bootstrap method and describe how it can be captured by novel types of word clouds reporting topic-term probability estimates and by confidence intervals or bands designed for reporting time series estimates of topic weights. The application of the new measures and methods is illustrated with an empirical example involving conference abstracts.

*Key Words: Topic models, text analysis, latent Dirichlet allocation, bootstrapping, uncertainty*

*JEL classification: C49*

* Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from financial support from HiTEc Cost Action CA 21163.

† University of Lodz, Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

‡ Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

## 3.1 Introduction

Topic modelling, combining natural language processing (NLP) and machine learning methods, is frequently used to extract main themes from large collections of textual data and to analyze weights or aggregated weights of these topics for selected documents, e.g., to describe their dynamics in time.

Some recent applications of topic models include the paper by Savin et al. (2022) who obtain topics from textual robotic patent data and track their importance over time or Bystrov et al. (2024b) examining topics from academic economic articles and analyzing the relationship between topic trends and the corresponding economic indicators. Adämmer et al. (2025) use economically related newspaper articles to obtain topics and aggregated topic proportions which are then applied in macroeconomic forecasting.

The outcomes of topic modelling depend on many factors including a specific sample of documents, pre-processing of textual data, model selection and initial configuration of the estimation algorithm. The estimation uncertainty is, however, seldom measured and reported when presenting the final results of text analyses.

The aims of this paper is to introduce methods for investigating sampling uncertainty associated with topic modelling, to suggest alternative ways of measuring and presenting it, and to conclude that such results should be reported on a regular basis when estimating a topic model. Besides presenting summary measures of topic uncertainty, We first propose to modify word clouds (see, e.g., Heimerl et al. (2014)) as they are typically used to visualize the estimated topics. The purpose of the novel types of world clouds is to exhibit how content and structure of estimated topics might vary across samples. We also investigate the uncertainty associated with time series of topic weights and present it using confidence bands.

Our approach towards analyzing estimation uncertainty due to sampling of documents is based on the bootstrap method of Efron (1979). To focus on sampling variability only, we fix all model parameter settings throughout the bootstrap replications. Sensitivity of the variational Bayesian estimator to initial configurations is accounted for by employing a deterministic initialization algorithm implemented by Roberts et al. (2019). Non-parametric bootstrap is applied, because the variational Bayesian estimator may not converge to the true topic distribution (see, e.g., Chen et al. (2018)).

Papers related to our study include the analysis by Winker (2023) who proposed a method for visualising estimation uncertainty in topic models and illustrated the method by showing the effects of changing the initial conditions. Sensitivity analysis of estimation results to the choice of the modelled number of topics can be found, e.g., in Bystrov et al. (2024a). The impact of pre-processing corpora by removing terms with low frequency is studied by Bystrov et al. (2023).

The usage of the bootstrap for analyzing sampling uncertainty in the context of topic models was proposed in a number of studies. Rieger et al. (2020) assessed the so-called

aleatoric uncertainty of the text generating process focusing on a large dataset containing news articles from a German newspaper, Süddeutsche Zeitung. They used the bootstrap method to construct confidence intervals for topic proportions in single texts and text corpora over time. Tang et al. (2014) conducted a systematic analysis of data properties and model settings that affect the inferential performance of latent Dirichlet allocation (LDA). They evaluated effects of various characteristics including the number of documents, the number of topics, and the Dirichlet hyper-parameters on synthetic and real data. In experiments on real datasets, the authors resampled words in documents or sampled documents from the corpus. Bootstrapping was also used for topic modelling, e.g., by Kontoghiorghes & Colubi (2023) who compared topic prevalence in sub-sets of documents.

Investigating sampling uncertainty of topic modelling by the application of methods proposed in this paper should make it possible to demonstrate the degree of robustness of estimation results and to identify those outcomes which could be considered as artefacts of a specific sample of textual data. Additionally, the construction of confidence bands for topic trends could be helpful in formulating conclusions related to the significance of changes in topic weights.

The structure of the paper is as follows. In section 3.2, we describe the methods, including the bootstrap algorithm used to study sampling uncertainty of parameter estimates and procedures for evaluating and reporting it. The textual data providing the base for the empirical analysis are presented in section 3.3, while the results are described in section 3.4. Section 3.5 provides conclusions and discusses possible extensions of our study.

## 3.2 Methods

### 3.2.1 Bootstrapping

The bootstrap method consists in resampling from the original set of data in order to obtain the distribution of an estimator or a test statistic. Most commonly, bootstrapping is applied to quantitative data for which various algorithms have been developed (see, e.g., Herwartz & Lange (2020) or MacKinnon (2009)), however there are also studies using the method for the analysis of textual information (for an overview of some applications see Egbert & Plonsky (2020)). The approaches proposed for bootstrapping text corpora are summarized below:

1. resampling documents (Tang et al., 2014),
2. treating document structure as given and resampling words in each text (BWord method in Rieger et al. (2020), Tang et al. (2014)),
3. resampling sentences within each document (BSentence method in Rieger et al. (2020)),
4. resampling sentences and words for each document (BSentenceWord method in Rieger et al. (2020)).

The rationales behind these different approaches are as follows. Since a text corpus can be viewed as a finite set of representative documents for a particular generative process, resampling these documents captures uncertainty associated with the possibility that a different collection of texts might have been obtained.

Resampling words from each document uses information on the observed frequencies of terms in documents. Under the assumption of a topic model, these frequencies are a superposition of document-topic and topic-term frequencies. This type of resampling is associated with the uncertainty of describing document-specific topics in a particular way. If the structural unit is a sentence then resampling sentences or a combined approach of drawing sentences and words should be used instead of resampling words.

Rieger et al. (2020) conclude that resampling words might lead to underestimating uncertainty. Thus, in this study we construct bootstrap samples by drawing with replacement documents from the original sample.

### 3.2.2 Structural Topic Modelling

The evaluation of sampling uncertainty is carried out in the framework of the structural topic model which was developed by Roberts et al. (2016) and is implemented in the R package `stm` (Roberts et al., 2019). This framework allows us to control for algorithmic uncertainty caused by the stochastic initialization of parameters in mixed-membership topic models.

In the structural topic model the initial matrix of topic-term probabilities is uniquely identified and estimated by the method of moments (Arora et al., 2013). The identification is attained by using a combinatorial algorithm to select an anchor term for each topic, i.e., a term that appears only in that specific topic. The restrictions, imposed by selecting anchor terms, are used to recover initial parameters by matching observed counts of term co-occurrences and counts implied by the topic model. This deterministic initialization procedure produces initial document-topic and topic-term probability matrices.

Once the initial probability matrices are selected the structural topic model is estimated by the variational expectation maximization algorithm which optimizes variational posteriors for document-topic and topic-term distributions.

By controlling for algorithmic uncertainty we can evaluate the sampling uncertainty of estimated topic-term probabilities and topic prevalence using bootstrap procedures.

### 3.2.3 Implementation and Evaluation

To evaluate sampling uncertainty we implement document-based bootstrapping described in Section 3.2.1. For this purpose we proceed as follows:

1. For a given text collection, estimate an initial LDA model using a pre-defined number of topics $K$ that is considered true.

2. Create 1,000 bootstrapped samples by repeated drawing from the original corpus.

3. For each of the bootstrapped corpora, estimate an LDA model with $K$ topics.

4. Evaluate sampling uncertainty.

First, we estimate an initial model with $K$ that we consider to be the true number of topics. Thereby, we use the estimation procedure implemented in R package `stm` (Roberts et al., 2019). We opt for deterministic (spectral) initialization and estimate a structural topic model without metadata covariates. Second, we create new text corpora by resampling documents (see Algorithm 3.1). Third, for each of the generated corpora, we estimate a topic model using `stm` algorithm with $K$ topics. Finally, we analyze sampling uncertainty using quantitative metrics and visualization. The impact of sampling on the content of the estimated topics is measured by the *model fit* and *recall* metrics proposed by Bystrov et al. (2024a). These metrics are described in more detail in subsection 3.2.3.1. The visualization of sampling uncertainty is implemented in the context of word clouds (see subsection 3.2.3.2) and time series of topic weights (see subsection 3.2.3.3). Word clouds incorporating information on uncertainty allow evaluating effects of sampling on topic structure and confidence bands constructed for time series of topic weights provide insights into the evolution of topic weights over time.

**Algorithm 3.1** Resampling documents

1: Define corpus size (equal to the original corpus size)

2: Draw documents with replacement from the original corpus

3: Estimate an LDA model using $K$ topics

#### 3.2.3.1 Measures

To quantify changes in content due to sampling uncertainty, we propose to use two metrics, namely *model fit* and *recall*. *Model fit* considers topic-term distributions of two topic sets and reports the average over the distances between these topic sets. We refer to this process as *one-to-one* matching, i.e. each topic from one topic set has to be assigned to a topic from another topic set. In particular, the Hungarian algorithm is implemented to find matches of the original topic-term and estimated/bootstrapped topic-term distributions. This procedure is aimed to minimize the overall costs of assignment between these topic sets. The resulting metric, the average over the distances between the identified matches, varies between 0 and 1. The lower the value of *model fit*, the better the estimated topic set reproduces the original one.

In standard classification tasks, *recall* describes the share of relevant items retrieved. In the context of topic modelling, the *recall* metric corresponds to the share of correctly reproduced topics. First, we perform the so-called *best matching* between the initial topic-term distribution and the estimated ones. For each topic in the initial topic set cosine similarities to the topics in a given estimated topic set are calculated. If this cosine similarity score is above a certain pre-defined value (also referred to as cut-off value), the topic is considered to be correctly uncovered. It means that, in contrast to the *one-to-one* matching

described above, some of the original topics might not find a match. To identify a critical value, all the similarity values between two topic sets are calculated and the 95% percentile of this distribution is considered to be a cut-off value. Finally, the number of correctly reproduced topics is divided by the number of topics $K$. Therefore, recall values can vary between 0 and 1. A value of one means that all original topics are reproduced correctly by the estimated/bootstrapped topic set, i.e. cosine similarity scores of all the best matches are above the cut-off value.

#### 3.2.3.2 Word Cloud Uncertainty

A further approach for evaluating the sampling uncertainty of topics focuses on word clouds, a standard tool for presenting topics by means of the words with highest probability in the respective topic. To allow for a fast screening of word clouds by humans, often the number of highest probability words is restricted, in the present application to 25.

Typically, word size corresponds to the actual probability of words. Often, the scaling is chosen such that probability is directly proportional to the area covered by the word. The bootstrapping procedure delivers a large number of alternative estimates, which – after the one-to-one matching described above – correspond to a large number of word clouds. Low sampling uncertainty would imply that all these word clouds are quite similar, while larger sampling uncertainty would result in major differences. Furthermore, the higher uncertainty is, the more likely it is that the set of the top 25 words will not coincide across replications.

Obviously, these features render it quite difficult to exhibit the uncertainty surrounding a word cloud in a way which is both informative and easy to grasp by a human. Therefore, we propose two alternative approaches. The first one, proposed by Winker (2023), allows to visualize the uncertainty by presenting the topic top words of the original estimates with their corresponding "confidence bands" based on bootstrap replications. The idea of confidence intervals is implemented by adding copies of each word on top of each other in different colours. Thereby, the size of the copies depends on quantiles of the distribution of corresponding probabilities of the word across bootstrap replications, e.g., the 10, 20, 50, 80 and 90% quantiles. This novel type of word clouds is labelled as Word Overlay Clouds (WOC) in the following.

If all these superimposed words are of similar size, the distribution is very concentrated around the median, i.e., uncertainty is low. If, however, the sizes differ substantially, uncertainty is high. In particular, the values of the lower quantiles might become so small, that the word size becomes too little to be recognized in print. Thus, not only substantial differences in size, but also missing colours corresponding to lower quantiles indicate large uncertainty around a word with respect to a specific topic.

The second approach focuses less on word probabilities, but instead on the frequency of words showing up among the top words for a topic across bootstrap replications. Therefore, we label the resulting word clouds as Cross-sample Frequency Clouds (CSFC). Thereby, word size corresponds to the probability of the estimates based on the original corpus. Now,

colours change according to the frequency of words belonging to the top words for the same topic in the bootstrap replications. This way, the colours provide a direct insight about the stability of top words within the topic with regard to sampling uncertainty. Obviously, whenever a top word does not show up among the top words in a bootstrap replication, some other word might appear. Therefore, we also count the frequency of these new words and add the most frequent ones with a different colour scheme to the word cloud. Thereby, the size of these new words still corresponds to their weight in the topic estimated on the original corpus. Given that these novel ways of presenting uncertainty in topic modelling might require some habituation, we will describe them in the results section in more detail.

#### 3.2.3.3 Confidence Bands for Topic Weight Time Series

While often the interest in topic modelling is focused on word clouds, representing weights of words in topics, the temporal dynamics of topic prevalence is also frequently investigated. As a typical corpus comprises documents from different time periods, a natural question is whether topic prevalence is changing over time. For a selected topic, the time series of topic weights is easily calculated from the estimated model by aggregating document-topic weights over all documents belonging to a specific time frame. Consequently, conducting a bootstrap analysis also allows constructing a large number of such time series corresponding to the same original topic, using the one-to-one matching approach described above.

To construct bootstrapped time series we retain time label for each resampled document. For each bootstrap replication, after estimating the topic model for the generated corpus and matching the bootstrapped topic to the original, the resampled documents are reordered chronologically and the document-topic weights are aggregated over all resampled documents belonging to the same period. This way we construct bootstrapped time series of topic weights matched to the original series of topic weights.

In this setting, low uncertainty would be depicted by the bootstrapped time series that are close to the original series of topic weights, while large uncertainty would be represented by substantially different levels and dynamics of bootstrapped time series. A graphical presentation of this uncertainty is straightforward following the idea of confidence bands. As a simplest approach we use pointwise percentile bands described in Lütkepohl et al. (2015).[1] For each time period we compute lower and upper quantiles of all bootstrapped time series. The smaller these bands are, the more accurate the estimate of the topic weight series with regard to sampling uncertainty. Given the constraints for all topic weights to be positive and to sum up to one, symmetric bands, e.g., 5% and 95% quantiles, might be too large. Therefore, we will apply pointwise modified bootstrap bands by choosing the smallest interval containing, e.g., 90% of all replications.

[1] The use of methods for constructing joint confidence bands using some of the methods described in Lütkepohl et al. (2020) would also be feasible.

## 3.3 Application to Scientific Abstracts

The collection of texts we consider in our study comprises abstracts submitted to the European Research Consortium for Informatics and Mathematics (ERCIM) and Computational and Financial Econometrics (CFE) conferences in the period from 2007 to 2023. The corpus contains 19,059 documents (1121.12 submitted abstracts per year on average) of a mean length of 158 words (including title). We perform standard preprocessing steps such as removing punctuation, numbers, and special characters, lowercasing, lemmatization.[2] We also remove 0.5% of the rarest terms (see Bystrov et al. (2023) for the discussion of the impact of removing infrequent words on the estimation outcomes). This results in a vocabulary size of 1,844 unique terms.

To follow the procedure described in 3.2.3, we first identify an optimal number of topics in the original data set. To do so, we apply the singular Bayesian Information Criterion (sBIC) recently proposed by Bystrov et al. (2022). The authors show in a comprehensive simulation study (see Bystrov et al. (2024a)) that sBIC delivers a sensible number of topics across different corpora by reflecting a trade-off between goodness-of-fit and model complexity. The sBIC values obtained for the range between 10 and 40 topics are shown in Figure 3.1. According to sBIC, the number of topics that should be considered for this corpus is 24.

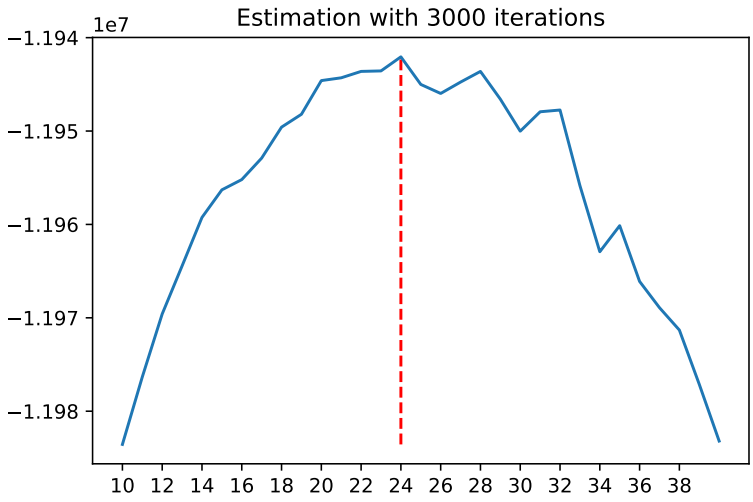



**Figure 3.1:** sBIC values for abstracts corpus

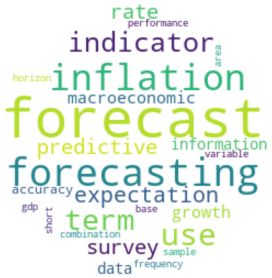
Figure 3.2 shows some examples of topics uncovered in the original corpus focusing on `portfolio management`, `risk management`, `forecasting`, `energy market`, `signal extraction` and `economic indicators`. These topics correspond well to the aims and scope of the conferences. Figure 3.3 shows the corresponding relative topic weights over time. To construct such topic time series, individual document-topic weights are averaged on a yearly basis. For example, the importance of the topic on `portfolio management` seems to

[2] To lemmatize texts, spaCy library in Python and the model "en_core_web_lg" were used. We adjusted the lemmas for "datum" and "data" to result in a single form "data" to avoid multiple occurrences of this word

decrease overtime, while the importance of the `forecasting` topic appears to remain more or less the same. The interest in the topic on `signal extraction` increases over time.


**(a)** Portfolio Management


**(b)** Risk Management




**(c)** Forecasting


**(d)** Energy Market


**(e)** Signal Extraction


**(f)** Economic Indicators

**Figure 3.2:** Wordclouds of selected original topics

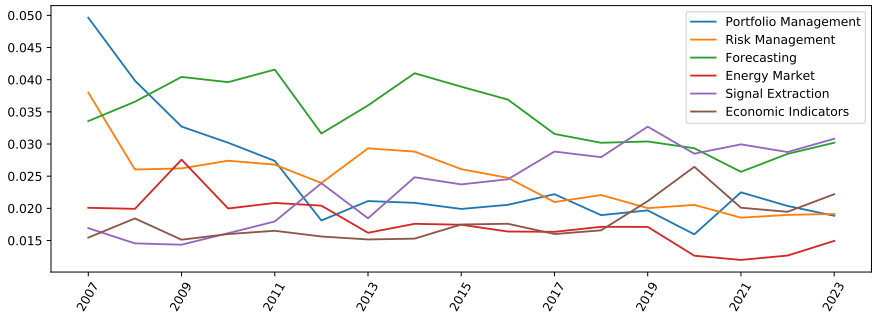



**Figure 3.3:** Relative topic weights

## 3.4 Bootstrap Results

### 3.4.1 Measures

In this subsection, we report the results obtained using the *model fit* and *recall* metrics as described in section 3.2.3.

#### Model Fit

This metric describes the average distance between two topic sets. In our setting, we apply it to measure mean distances between the true/original topic set and the topic collections from bootstrapped samples. Cosine distance is chosen to define the costs of assignment. Figure 3.4 shows the distribution of model fit values over 500 replications (in blue). The lower the value, the better the assignment between the true and estimated topic-term distributions. To allow for a more complete analysis, for each replication we also calculate the 5% percentile of all possible cosine distances (576 values). Finally, we obtain a distribution of the 5% percentile over 500 replications (shown in grey). As can be seen in Figure 3.4 both distributions do not overlap which means that in all of the replications the average distance between the original topic set and the bootstrapped one is considerably low (lower than the 5% percentile).

Alternatively, we can take a look at single topics and the distances between the true topic-term distribution and the assigned topics across bootstrapped samples. For example, Figure 3.5 shows the distributions of distances for the topics `portfolio management`, `risk management`, `forecasting`, `energy market`, `signal extraction` and `economic indicators`. The topics `portfolio management` and `risk management` seem to be very persistent across bootstrapped samples. Although the `forecasting` topic shows a greater variance, the majority of the distances are below 0.5. In contrast, the topics presented in the second row do not appear to be clearly identifiable, as the distance values vary substantially across the replications.

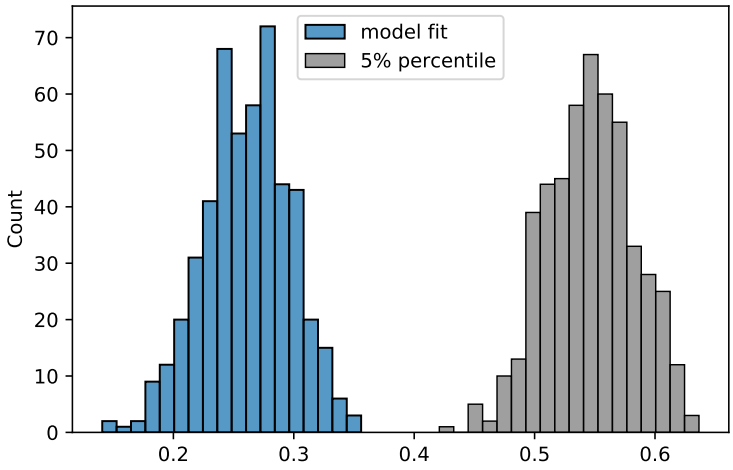
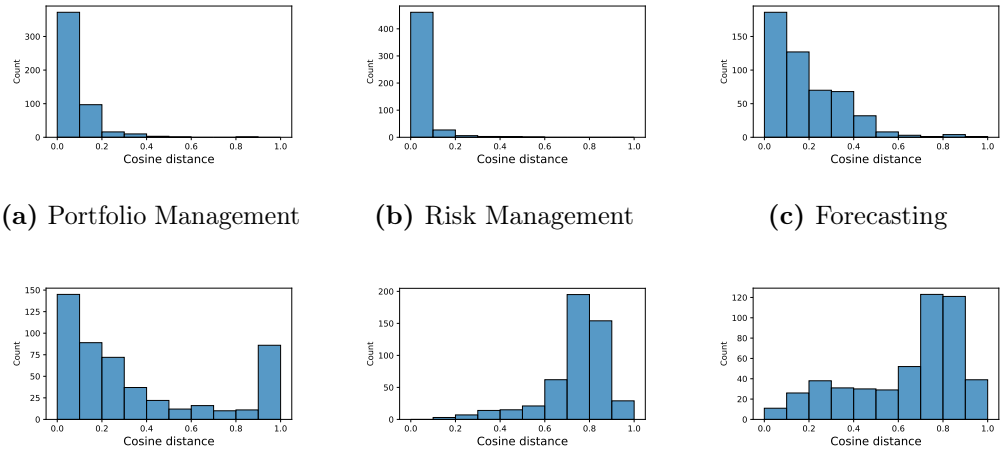



**Figure 3.4:** Model fit and the 5% percentile over 500 replications




**(a)** Portfolio Management **(b)** Risk Management **(c)** Forecasting

**(d)** Energy Market **(e)** Signal Extraction **(f)** Economic Indicators

**Figure 3.5:** Cosine distances for selected topics

### Recall

To calculate the recall metric, first a cut-off value is defined automatically by calculating all pairwise cosine similarities and taking the 95% percentile. The distribution over the obtained cut-off values is shown on the left-hand side in Figure 3.6. We use then this cut-off value to filter out matches considered to be too poor. The corresponding recall values are shown on right-hand side in Figure 3.6.

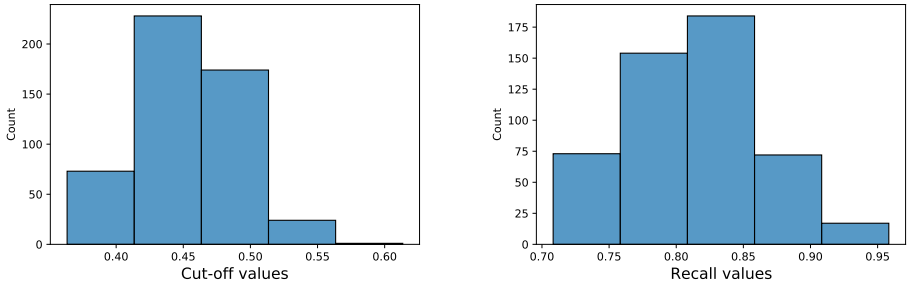



**Figure 3.6:** Defined cut-off values and corresponding recall values (bin-width=0.05)

Further, the cosine similarity distribution for selected topics could be of interest and the corresponding frequency with which a topic was correctly identified. Table 3.1 reports the reproduced share of selected topics. In line with the analysis of the *model fit* metric presented above, it can be concluded from the table that the topics `portfolio management`, `risk management` and `forecasting` find sensible matches almost across all bootstrapped samples (above 95%), while the topics on `energy market`, and, in particular, `signal extraction` and `economic indicators` often fail to find a sensible match.

| Topic | Reproduced Share in % |
|---|---|
| `portfolio management` | 99.4 |
| `risk management` | 99.6 |
| `forecasting` | 96.0 |
| `energy market` | 68.2 |
| `signal extraction` | 8.2 |
| `economic indicators` | 28.6 |

**Table 3.1:** Reproduced share of selected topics

## 3.4.2 Word Cloud Uncertainty

For the selected topics shown in Figure 3.3, we start with the presentation of sampling uncertainty for each of the 25 top words by means of Word Overlay Clouds. Thereby, the five colors reflect the 10, 20, 50, 80 and 90% quantiles of the bootstrap distribution. For

example, for the first topic shown in Figure 3.7, panel (a), the word "portfolio" shows up in very similar size from the 10 up to the 90% quantile. We may interpret this as a tight confidence band indicating a low sampling uncertainty for the weight of this word within the topic. In contrast, when focusing on the word "factor", we see substantially different word sizes corresponding to the 50, 80 and 90% quantiles, while the relative topic weight for the 10 and 20% quantiles of the bootstrap distribution seem to be too small for visual perception. Only when zooming in with a factor of 10, one may recognize tiny yellow and red letters in the middle of the box corresponding to that word. This reflects a rather broad confidence band for the weight of this word within the topic. However, for most top words of this topic, the distribution of weights across bootstrap replications seems to be quite concentrated, indicating a topic with low sampling uncertainty. The only exceptions seem to be the words "predictability", "pricing" and – as already mentioned – "factor".

The word clouds for the topics `risk management` (b) and `forecasting` (c) exhibit a similar quality. For the `risk management` topic only the words "extreme", "quantile" and "contagion" show larger bands indicating that these top words might be subject to a larger sampling uncertainty as compared to terms such as "risk" or "measure". In the topic labeled `forecasting`, a few more words have low weights for the topic when considering the lower quantiles. For example, "inflation" and "rate" might be subject to a larger sampling uncertainty as only some of the abstracts dealing with forecasting also use the example of inflation rates.

The picture changes substantially, when considering the other three topics shown in the lower panel of Figure 3.7. In these cases, only for few of the top words the 10 and 20% quantiles are still visible, while in particular for the last two topics (e) and (f) even for the most important words "process" and "index" already the estimated weights for the median of the bootstrap replications (green colour) is very small. Therefore, the estimation of these topics has to be considered as not very robust with regard to sampling uncertainty.

As a first tentative summary, we might conclude that from the six selected topics, the first three are rather robust to sampling uncertainty, while the other three are less well founded in the corpus, but depend to a larger extent on specific documents being in or out of the bootstrap samples. It should be noted that the examples have been selected for purpose to include both topics exhibiting small and large variations across bootstrap replications.

As an alternative way of presenting sampling uncertainty in word clouds, we introduced in Subsection 3.2.3.2 the Cross-sample Frequency Clouds. This approach focuses on the frequency of words showing up among the 25 top words of a topic across bootstrap replications. In Figure 3.8, the top words of the topic estimated on the original corpus are shown in colours ranging from bright orange to bright grey (upper row of legend). Thereby, words coloured in bright orange show up among the top words of the topic in less than 20% of all bootstrap replications, while the bright grey ones show up in at least 80% of all bootstrap replications. The size of words corresponds to their weights in the topic estimated on the original corpus.

In addition to the frequency of top words present in the bootstrap replications, the word clouds also contain 10 new words in colours ranging from bright turquoise to red (second row of legend). Thereby, the colours correspond to the frequency that the word showed up


(a) Portfolio Management

(b) Risk Management

(c) Forecasting

(d) Energy Market

(e) Signal Extraction

(f) Economic Indicators

**Figure 3.7:** Word Overlay Clouds exhibiting uncertainty of top-word weights

among the 25 top words across the bootstrap replications. The included words are those with highest frequency over all bootstrap replications. The size of these new words still corresponds to their weights in the original estimate.


**(a)** Portfolio Management **(b)** Risk Management **(c)** Forecasting

**(d)** Energy Market **(e)** Signal Extraction **(f)** Economic Indicators

**Figure 3.8:** Cross-sample Frequency Clouds exhibiting uncertainty about top words

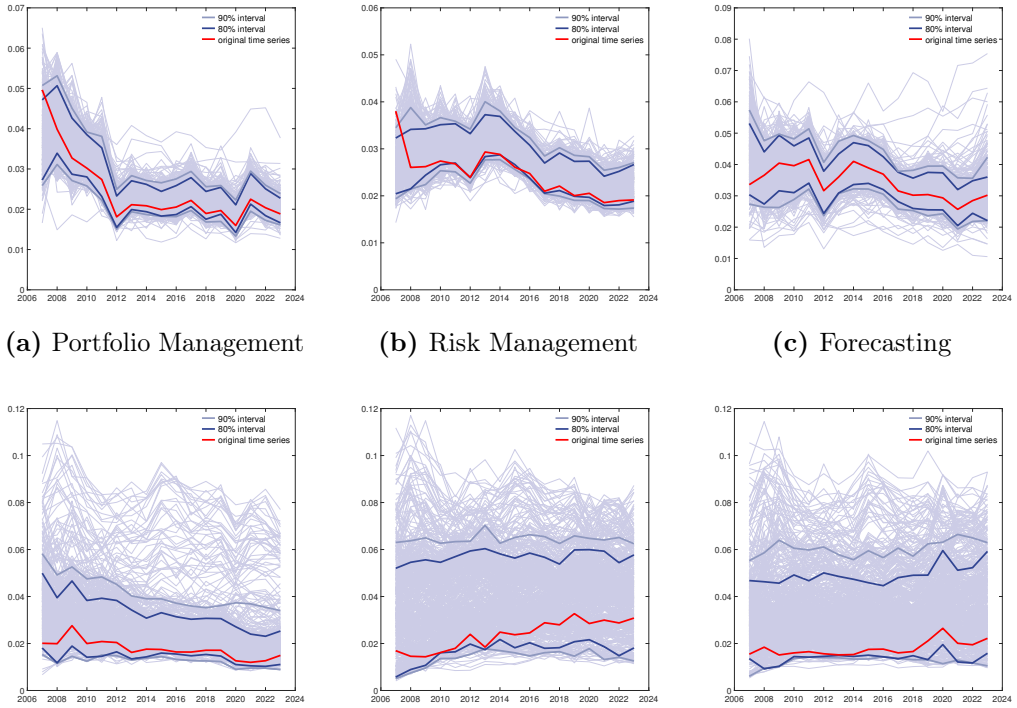
The CSFCs confirm the findings derived based on the distribution of weights above and reflected by the WOCs. For the first three topic shown, it can be seen from the shading that many words, in particular the most important ones in terms of weight, show up among the top words in a large share of all bootstrap replications, often exceeding the 80% threshold. The new words entering often fit also well with the topic at hand. Again, the impression is different when considering the second row, in particular sub-figures (e) and (f). Here, hardly any of the top words of the original topic shows up at high frequency, indicating that the identification and estimation precision of these topics under sampling uncertainty is rather low.

Summarizing the findings, both approaches for the visualization of sampling uncertainty provide a consistent view, while delivering different details with a focus on either importance (weights by the WOCs) or frequency (among top words by the CSFCs).

### 3.4.3 Confidence Bands for Topic Weight Time Series

Figure 3.9 shows the time series of topic weights for the original corpus as red lines. They are obtained by aggregating the topic weights of all abstracts in each year. Corresponding time series for the bootstrapped samples are shown as bright grey lines. Using the modified

percentile method (as in Lütkepohl et al. (2015)), for each year, smallest confidence intervals covering 80% (dark blue) and 90% (light blue) of the bootstrapped values are calculated resulting in the presented point-wise confidence bands.




**(a)** Portfolio Management **(b)** Risk Management **(c)** Forecasting

**(d)** Energy Market **(e)** Signal Extraction **(f)** Economic Indicators

**Figure 3.9:** Modified percentile bootstrap confidence intervals for topic weight time series

The results allow for the conclusion that those topics, which turned out to be more robust when focusing on their top words, also result in topic weight time series with less uncertainty as indicated by the smaller confidence bands in the upper part of Figure 3.9. The topics exhibiting more uncertainty when analyzing their word distribution correspond to time series exhibiting a lower level, i.e., the relevance of these topics for the corpus turns out to be smaller. Furthermore, the uncertainty about these topic weights is substantially higher as reflected by the larger confidence bands.

Overall the bootstrap approach allows obtaining measures of uncertainty not only of the topic themselves, but also of time series generated from the topics by clustering observations over time periods. For the actual application, we might state that the topics related to `portfolio management` and `risk management` lost importance over time, while forecasting remained rather stable. For the less precisely estimated topics, the large confidence bands do not allow for a reliable conclusion.

## 3.5 Conclusions

The paper introduced new methods of measuring and presenting sampling uncertainty for topic models. In particular, we focused on the problem of evaluating sampling uncertainty

of the estimated topic-term and document-topic probabilities which are used to describe the structure of uncovered latent topics and their weights in a text corpus. To evaluate the uncertainty we employed a non-parametric bootstrap method. Since it is most common to report the estimation results using word clouds for the largest topic-term weights and trajectories for document-topic probability estimates aggregated over time, we proposed to incorporate the information on uncertainty into these plots.

We show, that the two novel types of word clouds (WOC and CSFC) and topic trend confidence bands described in this study provide much better grounds for understanding the results of estimating topic models than bare point parameter estimates. As illustrated by our empirical example, topic-term probabilities for different terms within the same topic can be estimated with varying precision. This implies that, at least for some of the topics, standard word clouds might be misleading because different clouds could be similarly or even more likely for the same topic. Such greater or smaller robustness of estimates with respect to sampling uncertainty should be taken into account when formulating the final conclusions regarding contents and focus of a text corpus. We proposed two attractive ways of reporting this uncertainty graphically, in the form of extended word clouds: first, by showing words from the original cloud together with bootstrap "confidence bands" reflecting changing weights of the words in resampled corpora (WOC) and second, by presenting the frequency of terms belonging to the top words for a topic across bootstrap replications (CSFC).

Similar considerations apply when the main goal of the analysis is to investigate the importance of topic weights over time. Our study revealed that the uncertainty associated with estimated topic weights can also be quite substantial. In such cases interpreting point estimates of topic weights for individual periods and comparing these point estimates for different periods is not a good idea. We demonstrate that estimation precision of a topic trend is related to the precision of estimating prevalence of terms for this topic, i.e., there is more certainty about estimated importance over time for topics for which contents were estimated more accurately. The proposed confidence bands make it possible to present the uncertainty of estimated topic trends and to analyze whether there were significant changes in the weights over time.

When measuring estimation uncertainty for topic models we focused on bootstrap distributions of individual topic-term and aggregated document-topic probability estimators. This approach ignored correlations between prevalence of different terms or topics. Another possibility might be to consider joint distributions of probability estimators for a number of weights and to report their sampling variability. This extension, consisting in constructing joint confidence regions, is left for future work. Further follow up research could consider different resampling schemes of textual data within the bootstrap method, e.g., to construct bootstrap samples by drawing sentences and words. An additional idea, so far neglected in the literature would be to apply a joint approach of resampling documents and sentences and/or words.

# Chapter 4

# Cross-Corpora Comparisons of Topics and Topic Trends

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Cross-Corpora Comparisons of Topics and Topic Trends |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 35%),<br>Victor Bystrov (contribution: 25%),<br>Anna Staszewska-Bystrova (contribution: 25%),<br>Peter Winker (contribution: 15%) |
| **Status:** | Published: *Jahrbücher für Nationalökonomie und Statistik*, vol. 242, no. 4, 2022, pp. 433-469 |
| **Available from:** | `https://doi.org/10.1515/jbnst-2022-0024` |

# Cross-Corpora Comparisons of Topics and Topic Trends

Victor Bystrov[†] Viktoriia Naboka-Krell[‡]
Anna Staszewska-Bystrova[†] Peter Winker[‡,‖]

**Abstract.**

Textual data gained relevance as a novel source of information for applied economic research. When considering longer periods or international comparisons, often different text corpora have to be used and combined for the analysis. A methods pipeline is presented for identifying topics in different corpora, matching these topics across corpora and comparing the resulting time series of topic importance. The relative importance of topics over time in a text corpus is used as an additional indicator in econometric models and for forecasting as well as for identifying changing foci of economic studies. The methods pipeline is illustrated using scientific publications from Poland and Germany in English and German for the period 1984 to 2020. As methodological contributions, a novel tool for data based model selection, sBIC, is impelemented, and approaches for mapping of topics of different corpora (including different languages) are presented.

*Key Words: Topic models, text analysis, latent Dirichlet allocation, singular Bayesian information criterion, topic matching*

*JEL classification: C49*

[†] University of Lodz, Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

[‡] Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

[‖] Corresponding author: Peter.Winker@wi.jlug.de

## 4.1 Introduction

Textual data gained relevance as a novel source of information for applied economic research. Examples include text based indicators of economic uncertainty (Baker et al., 2016) and economic or political sentiment (Shapiro et al., 2020; Jentsch et al., 2020), the analysis of central bank communication (Hansen & McMahon, 2016; Lüdering & Tillmann, 2020), using textual information for now- and forecasting (Larsen & Thorsrud, 2019; Thorsrud, 2020; Kalamara et al., 2020; Ellingsen et al., 2022; Foltas, 2022) or describing the diffusion of innovations (Lenz & Winker, 2020) and innovation cluster (Krüger et al., 2020), and the link between real economic developments and scientific publications in economics (Lüdering & Winker, 2016; Wehrheim, 2019). Textual data have also been used during the Covid-19 pandemic for policy and impact analysis, e.g., by Debnath & Bardhan (2020), Mamaysky (2023), and Dörr et al. (2022).

In most of these applications, the main interest is on temporal patterns in the textual information and how these relate to other developments over time. However, also a comparison of the content of different text corpora might be of interest, e.g., when comparing sentiments in different countries or topics of economic research present in different scientific journals. When longer periods are considered, it might also become necessary to merge the information content of various textual data sources available for different sub-periods. In such settings, it is imperative to identify those topics which are common or at least similar across the corpora and their evolution over time. Obviously, there is no guarantee for finding matching topics ex post. Therefore, it is also relevant to identify those topics which are specific for certain corpora only. Although such analyses are urgently needed there is no consensus yet on which methods are most appropriate for specific applications.

In this paper, we extend the existing literature on topic modelling by proposing methods for comparing (matching) topics identified for various corpora. This approach also includes proposing a data based criterion for assessing the quality of a match, which eventually serves for identifying real matches. Additionally, we suggest to select the number of topics in each corpus based on singular Bayesian information criterion (sBIC), which appears more robust compared to other commonly used tools used for this model selection step.

The working of the methods is presented using economic text corpora in two languages. However, the proposed tools are generic and can be applied also beyond economic research. For example, the results described in this paper might be of interest to political and communication scientists who often need to consider multilingual textual sources (see e.g. Lucas et al. (2015) and Maier et al. (2022)).

The presented methods pipeline is not meant to be the only viable approach for such analyses, but rather a basic setting building mostly on established procedures. Although the approach is kept simple, it still involves a number of steps which require choosing several parameters. We will stick as far as possible to standard parameter values and discuss some central aspects in more detail, in particular those related to choosing the number of topics in

a corpus and the threshold-values for defining a meaningful match of topics across corpora.

As illustration of the application of the proposed methods pipeline, we consider two corpora of scientific publications in economics over the period 1984–2020, one published in Germany and one in Poland. These datasets and the basic methods pipeline building on latent Dirichlet allocation (LDA) (Blei et al., 2003) for topic modelling are introduced in Section 4.2. Methodological extensions are presented in Section 4.3, in particular using a singular Bayesian information criterion for selecting the number of topics in Subsection 4.3.1 and the approaches for matching topics across corpora in Subsections 4.3.2 and 4.3.3. The results of the application to scientific publications are provided in Section 4.4. Section 4.5 summarizes the findings and provides an outlook to further analysis.

## 4.2 Data and methods

### 4.2.1 Topic modelling and corpora comparisons

In this section, the text corpora are described. Figure 4.1 summarizes our research methodology.

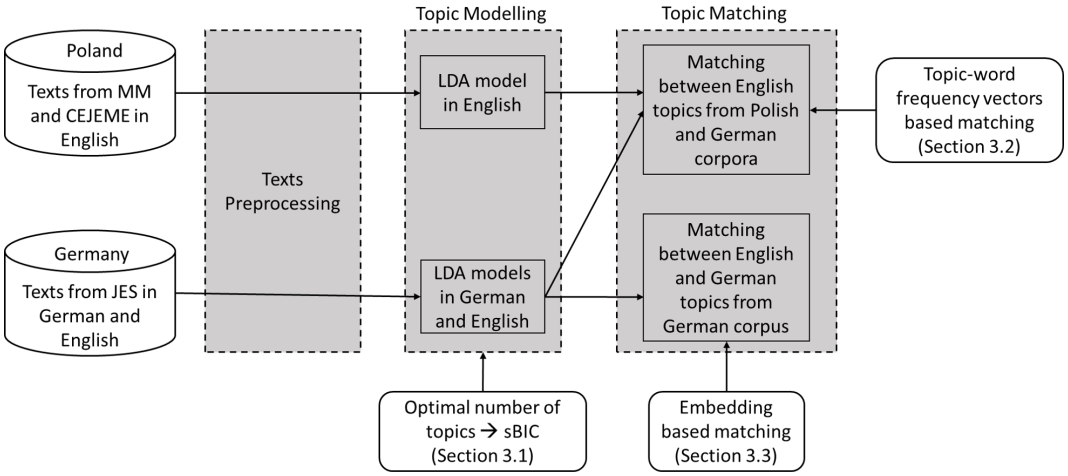



**Figure 4.1:** Outline of the methods pipeline for corpora comparison

Textual data for the current application consist of scientific articles published in Germany, in the Journal of Economics and Statistics (JES), and in Poland, in the proceedings of the Macromodels International Conference (MM) and the Central European Journal of Economic Modelling and Econometrics (CEJEME). A detailed description of the data sources is provided in Section 4.2.2. After the documents from the different sources were collected and prepared, the following common text preprocessing steps were applied:[1]

1. Removing all punctuation marks, special characters, numbers.

[1] To perform steps 5 and 6, Python's *spacy* module and the corresponding language models were used.

2. Following Lüdering & Winker (2016), who also applied topic modelling to data from JES, we also decide to remove words that contain fewer than 3 and more than 20 characters in order to capture further stop words and reduce the vocabulary size.

3. Removing English and German stop words (see Appendix B.2).

4. Removing especially rare/common words: all words with frequency over all articles in the text corpus under 2.5% and above 75% were removed. We prefer relative thresholds to absolute ones as they will depend on the size of the corpus and the vocabulary. Therefore, using relative threshold appears more appropriate for establishing a standardized pipeline.

5. Lemmatizing of texts, i.e., grouping inflected forms together as a single base form.

6. Removing certain parts-of-speech, the so called PoS tags, such as determiners, adpositions, conjunctions, pronouns to the extent that they are not contained in the usually rather short lists of stop words.

The resulting Bag-of-Words (BoW) representations of the documents were further used to train LDA models for the Polish and German data sets. LDA is one of the best-known and most widely used topic modelling approaches (Blei et al., 2003). It is based on the assumption that each document in a corpus is a distribution over some latent topics and each topic is a distribution over a fixed corpus vocabulary. Therefore, the algorithm behind the LDA approach aims to identify this hidden structure and to uncover the underlying latent topics in a corpus.

For the German corpus, two different LDA models were trained for the two subsets in English and German languages. Running LDA on the combined corpus would result in two sets of topics which are language-specific, although some of them might cover the same semantic content. Therefore, a separate modelling and matching post hoc appears preferably. As a robustness check, we also did an analysis on a joint corpus, for which all German texts have been translated to English using DeepL API Pro (see Appendix B.3, pp. 32-33).[2] This robustness check using machine translation has revealed that the most of the topics from the joint German dataset can be also found under the topics uncovered in the English subset of the data. Furthermore, we also show that for the identification of relevant matches between the two countries using the joint LDA model does not impact the results substantially. For the Polish corpus, only one LDA model was estimated. The models were trained using Python's *sklearn* module. The implementation in *sklearn* follows M. Hoffman et al. (2010) and M. D. Hoffman et al. (2013) and provides a method for estimating LDA models based on the online variational Bayes algorithm. Except for the number of topics and the number of iterations in the training process, all the parameters were kept at default values. Since there are several Python modules that implement the LDA algorithm, we also perform the analysis using *gensim*, another popular module for LDA topic modelling. In doing so, we

[2] Note that this alternative comes at additional cost for the translation. Therefore, we stick to the matching approach for the standard pipeline.

aim to account for possible differences resulting from different LDA implementations. We show that the qualitative findings of the current work do not change. The results of this robustness check are described in Appendix B.4, p.35.

The choice of an optimal number of topics in LDA models still remains a challenge in applied research. Although there are several criteria for selecting the optimal number of themes, the ultimate choice is often based on human judgement concerning interpretability of selected topics. In the current work, we aim to avoid the subjectivity of topic selection and use sBIC to determine the optimal number of topics for each of the text corpora. We further discuss interpretability of topics selected by sBIC. To our knowledge this is the first application of sBIC to LDA modelling and so we provide more methodological and practical details in Section 4.3.1.

In the topic modelling stage, we obtained three different sets of topics corresponding to two countries and two languages. To distinguish these sets we will use the following notation: $PL^{ENG}$, $DE^{ENG}$ and $DE^{GER}$ where $PL$, $DE$ indicate the country of publication of a corpus i.e. Poland or Germany, while the superscripts $ENG$, $GER$ inform about the language of publication.

We first focus on the matching between $DE^{ENG}$ and $PL^{ENG}$ topics. The matching of two topic sets of different LDA models can be done based on topic-word frequency vectors. Thereby, the distributions of topics over the vocabulary words are compared. However, since this standard approach to topic matching assumes the vocabularies are in the same language, it cannot be applied to LDA models trained on corpora in different languages. For this reason, we also propose a different embedding based approach to topic matching to compare topic sets in different languages. Both topic-word frequency vectors based matching and embedding based matching approaches are described in Sections 4.3.2 and 4.3.3, respectively.

In the final step, we qualitatively analyse the resulting topic matches and define thresholds for matches expected to be meaningful in a colloquial sense. We also construct topic time series based on the topic weights for each of the topic sets and descriptively analyse the time series trends of the matched topics.

### 4.2.2 Textual data for Germany and Poland

The illustration of the methods pipeline is based on corpora of scientific articles published in Germany and Poland. Given the interest in comparing trends of research topics over time, it is important that both corpora cover a long period. For our application, the overlap of both corpora is from 1984 to 2020. While the time span of the sample is rather long, the number of documents per year is substantially smaller than in other applications covering recent years. Furthermore, scientific articles in economics are more focused than general interest documents. Therefore, the number of distinct topics to be expected in these corpora is rather small. Nevertheless, the example might well illustrate the general procedure of cross-corpora topic and topic trends comparison outlined in Subsection 4.2.1.

### German text collection

The German textual data consist of articles published in the Journal of Economics and Statistics. The Journal has been published since 1863 containing articles that cover topics from economics with a focus on empirical economics and applied statistics. During the sample period 1984–2020 publications have been either in German or English. The distribution of the articles' languages is presented in Figure 4.2.[3]

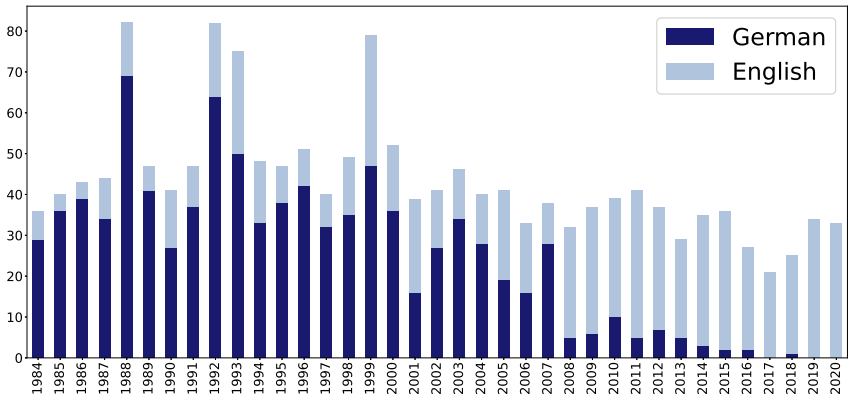



**Figure 4.2:** Language of the articles in the German text corpus.

The volumes of the journal published annually between 1984 and 2020 comprise usually 4 to 6 issues.[4] The details of data collection and further steps of preparation are described in Appendix B.1. The corpus used for the application comprises 903 articles in German and 704 articles in English.

### Polish text collection

Two sources of textual data for Poland were considered. Firstly, proceedings of the Macromodels International Conference (MM) and joint meetings were used providing textual data for the years 1984–2011. Secondly, papers published in the Central European Journal of Economic Modelling and Econometrics (CEJEME) in the period 2009–2020 were analysed.

The Macromodels International Conference has been organised in Poland every year since 1974. The printed materials analyzed in this article included also papers presented at the meetings held jointly with MM, such as Econometric Modelling and Forecasting Socialist Economies (Models & Forecasts, MF), the Multivariate Statistical Analysis (MSA) conference and the Association for Modelling and Forecasting Economies in Transition (AMFET) meetings. As indicated at the Macromodels' webpage (`www.macromodels.uni.lodz.pl`), the aim of the conference is to "bring together scientists who work in the field of econometric modelling [...]. Within the scope of interest are issues such as the problems of estimation,

[3] To automatically identify the language of an article, Python's module *langdetect* was used.

[4] In the years 1986, 1988, 1992, 1993, and 1999, there were two volumes per year.

simulation, developing econometric models and their use for policy analyses. Recently, a special attention has been given to modelling economies of new EU member countries". Conference materials printed as books are available since 1984 and continue up to the year 2011. Altogether 41 conference volumes comprising a total of 514 articles were used in the analysis. The language of the conference is English. After several preprocessing steps that are described in more detail in Appendix B.1, a structured database with bibliographic information for the articles was created in Python. The data include information on the year of publication, names of the authors, title of the paper, abstract and the main text.

The article collection from the Central European Journal of Economic Modelling and Econometrics includes 145 scientific articles which appeared in 46 issues of the journal, starting with the first issue from 2009 (1/2009) and ending with the fourth issue from the year 2020 (4/2020). As indicated by the aims and scope of CEJEME, the papers are focused on the theory and applications of mathematical and statistical models in economic sciences. All articles are in English. Detailed information on the preparation of the data from CEJEME publications can be found in Appendix B.1. The Polish data set used for the application consists of the main texts of the articles (without abstracts) from MM and CEJEME.

## 4.3 Methodological advances

This section first describes the proposed information criterion for determining the optimal number of topics in a corpus. Then, we present a general topic matching approach for LDA models trained on different corpora in the same language. As the German corpus consists of texts in English and German, we also propose a further topic comparison approach based on multilingual word representations that can be applied to LDA models trained on corpora in different languages.

### 4.3.1 Topic number selection based on singular Bayesian information criterion

Several criteria, which are often used for selecting the number of topics in LDA modelling, are based on specific semantic properties of selected topics, such as similarity and coherence (see Cao et al. (2009) and Mimno et al. (2011)). The number of topics selected by these criteria frequently differs considerably and the final choice is based on human judgement concerning interpretability of selected topics. In this paper, the number of topics is chosen using an information criterion that does not directly quantify any semantic property of topics, but balances the goodness-of-fit and model complexity. The model selection procedure based on the information criterion does not rely on topic interpretability, but chooses the optimal number of topics that can be used for inference. Nevertheless, topics selected by the information criterion are expected to be interpretable for a text corpus generated by an LDA model.

The implementation of information criteria for topic number selection in the LDA analysis

is complicated because it is based on a singular statistical model: the Fisher information matrix is not positive definite. The usual BIC cannot be implemented for evaluation of singular models as the penalty for model complexity used in the BIC is too large for singular models: too few topics would be selected in the LDA modelling if the regular BIC was used.

Drton & Plummer (2017) proposed a model selection criterion, called singular BIC (sBIC), that uses the Bayesian model averaging and a smaller penalty than the penalty used in the regular BIC. Hayashi (2021) derived the asymptotic learning coefficient for LDA that can be used for the evaluation of penalty in sBIC. In this paper the model averaging method proposed by Drton & Plummer (2017) and the asymptotic learning coefficient derived in Hayashi (2021) are combined in order to implement sBIC for the selection of number of topics in the LDA modelling. As it is a novel application of sBIC, essential theoretical and practical details of this procedure are briefly described below.

In order to present essential details of sBIC, let us consider a document corpus $\mathcal{D}$ that includes $N$ documents and uses a vocabulary of $M$ words. An LDA model is described by $N \times H$ matrix $\theta$ of document-topic frequencies and $H \times M$ matrix $\beta$ of topic-word frequencies with the dimensions of these matrices depending on the number of topics $H$. A set of candidate LDA models is thus determined by the numbers of topics in candidate models: $H \in \{H_{min}, \ldots, H_{max}\}$.

The marginal likelihood of corpus $\mathcal{D}$ given a model with $H$ topics can be written as

$$L(\mathcal{D}|H) = \int_{\theta,\beta} P(\mathcal{D}|\theta, \beta, H) dP(\theta, \beta|H),$$

where $P(\mathcal{D}|\theta, \beta, H)$ is the likelihood of $\mathcal{D}$ given matrices $\theta$ and $\beta$. The Fisher matrix for LDA models is singular, and the quadratic approximation of marginal likelihood, which is used in the derivation of the regular BIC, is not possible. But the singular BIC can be derived using the decomposition (Watanabe (2009))

$$\log L(\mathcal{D}|H) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, H) - \lambda(H) \log(n) + (m(H) - 1) \log\log(n) + O_p(1),$$

where $\hat{\theta}$ and $\hat{\beta}$ are consistent estimators of corresponding matrices, $\lambda(H)$ is a learning coefficient, measuring stochastic complexity of a model with $H$ topics, $m(H)$ is the multiplicity of the learning coefficient, and $n$ is the number of words in the document corpus. In practice, $\lambda(H)$ and $m(H)$ are not known as they depend on the true model dimension. This problem can be solved by the model averaging which is described in Drton & Plummer (2017).

An approximation of the marginal likelihood for a model with $H$ topics, based on averaging of submodels with number of topics $h \leq H$, is obtained as

$$L'(\mathcal{D}|H) = \frac{\sum_{h \leq H} L'_{Hh} L(\mathcal{D}|h) P(h)}{\sum_{h \leq H} L(\mathcal{D}|h) P(h)}, \tag{4.1}$$

where $P(h)$ is the prior of a model with $h$ topics,

$$L'_{Hh} = P(\mathcal{D}|\hat{\theta}, \hat{\beta}, H) \frac{(\log n)^{m_{Hh}-1}}{n^{\lambda_{Hh}}}$$

and the constants $\lambda_{Hh}$, $m_{Hh}$ can be computed for any $h \leq H$ using formulas from Hayashi (2021). Following Drton & Plummer (2017), we replace unknown marginal likelihoods $L(\mathcal{D}|h)$ on the right-hand side of (4.1) by their approximations, $L'(\mathcal{D}|h)$,

$$L'(\mathcal{D}|H) = \frac{\sum_{h \leq H} L'_{Hh} L'(\mathcal{D}|h) P(h)}{\sum_{h \leq H} L'(\mathcal{D}|h) P(h)},$$

and define the singular Bayesian information criterion for a model with $H$ topics as

$$sBIC(H) = \log L'(\mathcal{D}|H),$$

where $L'(\mathcal{D}|H)$ is the unique solution of the equation system

$$\sum_{h \leq H} [L'(\mathcal{D}|H) - L'_{Hh}] L'(\mathcal{D}|h) = 0 \tag{4.2}$$

that can be found recursively with $L'(\mathcal{D}|H_{min}) = L'_{H_{min}H_{min}}$ for the minimal model. The optimal number of topics maximizes sBIC over the set $\{H_{min}, \ldots, H_{max}\}$.

For an empirical implementation of sBIC in the LDA modelling approach, it is essential to use high-precision computations in case of large-scale data sets. Since the likelihood function for an LDA model is a product of a large number of word frequencies, it usually takes extremely small positive values. Correspondingly, the log-likelihood function takes negative values of extremely large modulus. To avoid exponent underflow and overflow in floating point computations, the outer limits allowable for exponents of floating-point numbers have to be sufficiently large. The computational precision (the size of fractional part of floating-point numbers) has to be sufficiently large for solving the system of quadratic equations (4.2) that represents a bottleneck of the sBIC algorithm.

The selection of an appropriate precision needed to avoid rounding errors might depend on a particular dataset. However, as compared to the estimation time of LDA models the additional time needed for high-precision computations in the sBIC algorithm is not substantial.

### 4.3.2 Topic matching

The outcome of a LDA model is a matrix containing the probabilities of occurence of each word in each topic. Therefore, a standard and intuitive way to compare two LDA models, or the hidden structures behind the data, is to compare the distributions of topics over the vocabulary words. Each topic can be represented as a vector with the length equal to the vocabulary size.

For the comparison, the topic vectors from different LDA models should have the same length. However, it is quite improbable that the vocabularies from different corpora are exactly the same. There are two possibilities to create topic-word frequency vectors of the same length. First, one of the vocabularies can be considered as the base vocabulary. If some of the words are missing in the other vocabulary, the probabilities are replaced with zeros. Alternatively, one can use only the intersection of the vocabularies of the considered corpora,

i.e., only the words that occur in both corpora. In the current work, we use the second solution as only minor differences have been observed when comparing both approaches. In general, this choice bears the risk that matched topics can still differ substantially with regard to the non overlapping parts of the vocabularies. Thus, in particular for less homogeneous corpora than considered in our application, one might also consider matching based on the union of the vocabularies.

In the next step, the similarities of the topic vectors are calculated. To this end, we consider two similarity measures:

- **Jensen-Shannon divergence (JSD)**, which is closely related to the Kullback-Leibler divergence (KLD), measures the similarity between two probability distributions or, in the current case, two word-topic distributions. The Jensen-Shannon divergence between two probability distributions $P$ and $Q$ is calculated as follows:

$$JSD(P||Q) = \frac{1}{2}Entropy(P||M) + \frac{1}{2}Entropy(Q||M),$$

  where $M = (P+Q)/2$. The square root of the Jensen-Shannon divergence is a distance metric.

- **Cosine similarity** is an alternative measure of similarity of two vectors and is often used when working with textual data. Cosine similarity is the cosine of the angle between the two vectors. For example, a cosine similarity of 1 implies that two vectors have the same orientation in the corresponding vector space.

The final step is the actual matching of the topics. Again, there are two alternative approaches to matching that can be applied to obtain topic pairs. The first one is the so-called *one-to-one matching* using the Hungarian algorithm (Kuhn & Yaw, 1955). The Hungarian algorithm is an optimization algorithm that, given a cost matrix containing the assignment costs between the topics of two LDA models, aims to find an optimal assignment of rows to columns with minimal costs. It is also possible to apply this algorithm if the number of topics of two LDA models is not the same. In this case, some of the topics of the larger LDA model remain unmatched. *One-to-one matching* can be applied, for example, when the two corpora are expected to cover the same set of topics. When implementing one-to-one matching, it is recommendable to use distance metrics such as the Jensen-Shannon divergence or the cosine distance, which is defined as $1 -$ *cosine similarity*, as the cost measure, as the Hungarian algorithm is formulated as a minimization problem.

The second option is *best matching* using the nearest neighbours approach, i.e., for each topic in a corpus, its nearest neighbour in the other corpus is chosen as its match. Hereby, the topics can be assigned multiple times. *Best matching* is a better choice when the thematic focus of the corpora to be compared is quite different and it is unclear whether each topic in one corpus can find a meaningful match in the other one.

However, given that each topic is assigned its nearest neighbour independently of the corresponding minimum distance, there is also no guarantee that all of the identified best matches actually correspond to a match according to the common understanding. For this

reason, a cut-off value has to be set in order to select only topic pairs sharing a high enough similarity. At this point, it is important to mention that the *best matching* is a non-symmetric process. For example, if for the German Topic $b$ the Polish Topic $a$ is the nearest neighbour in the Polish topic set (direction Germany $\rightarrow$ Poland), it does not necessarily imply that for the Polish Topic $a$ the German Topic $b$ is the nearest neighbour in the German topic set (direction Poland $\rightarrow$ Germany). To account for this non-symmetry, it is advisable to check the topic assignments in both directions.

In the current application, we use the cosine similarity measure to evaluate the topic similarity and perform best matching. We set the cut-off value based on the distribution of the cosine similarity values between all possible topic pairs. Subsequently, we also perform the matching using Jensen-Shannon distance as a robustness check.

### 4.3.3 Embedding based matching

The standard matching described in the previous subsection is restricted to the comparison of models in the same language. To enable multi-language analyses, we propose a further approach that uses the so-called word embeddings. These word vector representations have been attracting a lot of attention in recent years and are widely used in different applications also beyond the natural language processing field. One of the most important characteristics of such word embeddings is the interpretability of the distances between them. It means that semantically similar words tend to be close to each other in the shared vector space. For more details on how word embeddings are trained see Mikolov, Chen, et al. (2013) and Mikolov, Sutskever, et al. (2013). Recently, such word embeddings have been also used in the context of topic modelling. For example, Dieng et al. (2020) introduce the embedded topic model (ETM) where each word and each topic in a corpus are represented in the same embedding space. The authors claim that the proposed approach addresses the drawbacks of a classical LDA model, namely dealing with large vocabularies. Empirically, it is shown that the method leads to better results compared to other approaches including classical LDA as measured by the coherence criterion introduced by Mimno et al. (2011). However, it is not discussed how ETM performs in a multilingual context when a dataset consists of texts in different languages, as in the current case. Since in the proposed approach word and topic embeddings are trained based on the underlying texts and word co-occurrences, applying it to a multilingual corpus would probably result in an embedding space that contains multiple sub-spaces related to the languages contained in the corpus. Therefore, we decided to not further consider ETM for our analysis.

Bianchi et al. (2021) address exactly this problem and develop a language-agnostic approach to topic modelling – Multilingual Contextualized Topic Modelling (MCTM). The authors develop a topic modelling approach that is based on document representations from SBERT, a novel Transformer based technique to language modelling. The main advantage of the proposed approach is that a model can be trained based on one corpus and topic distributions for documents in unseen languages can be inferred just based on the multilingual

vector representations. In the current case, we could apply MCTM and train the model, for example, for the Polish dataset and infer topic distributions for the documents from the German dataset. In doing so, we would, however, restrict ourselves only to the topics in the Polish dataset. Some latent topics that are specific only for the German dataset would be missing. Therefore, we decided to stick to our embedding based matching approach that is described in more detailed in the following.

In the last few years, a lot of pre-trained word vectors have been released. For example, the fastText[5] library provides pre-trained word embeddings for over 150 different languages (Joulin et al., 2018; Grave et al., 2018). Many attempts have also been made to train multilingual word embeddings, i.e., training a shared vector space for multiple languages. For example, Conneau et al. (2018) introduce both supervised and unsupervised approaches to learning cross-lingual word embeddings. The authors provide multilingual embeddings for 30 languages based on fastText monolingual word vectors.[6] These multilingual embeddings can be used to obtain language independent topic representations. Thereby, each topic could be also represented as a vector in the shared multilingual vector space using the word embeddings of its most frequent words (see options 1 to 3 below). We consider the following options for obtaining multilingual topic vector representations:

1. Represent a topic as the sum vector of $n$ word vectors in the embedding space corresponding to its $n$ most frequent words.

2. Represent a topic vector as the weighted sum of $n$ word vectors corresponding to its $n$ most frequent words. The weights are given by the estimated LDA models and represent the probabilities of each word occurring in a certain topic. Thereby, rescale the original probabilities given by the LDA output depending on the number of words considered.

3. Represent a topic vector as the weighted sum of all the vocabulary word vectors, i.e., the word embeddings of all the vocabulary words vectors multiplied by the probabilities of occurring given by the LDA output.

4. "Translate" the words of one model into the language of the other model using word embeddings. For example, for each word in the English corpus vocabulary, search for the first nearest neighbour in German language and use the corresponding word as the translation of the English word. Afterwards, apply the standard matching approach described previously.

Further steps, i.e., calculating the similarity and applying one of the matching types, are performed analogously to the standard matching approach. In the current work, we use the first option and represent the topics as the sum vectors of 100 word vectors corresponding to their 100 most frequent words (not weighted). While preliminary analysis indicated

[5] fastText is a free library for text classification and representation learning.

[6] The vectors can be downloaded from Github under `https://github.com/facebookresearch/MUSE#multilingual-word-embeddings`.

no qualitative differences in the results for our analysis, an in-depth comparison of the performance of the different alternatives is left for future research.

### 4.3.4 Topic trends comparison

The methods described in this section aimed at identifying similar topics based on their textual content. A further aspect of interest is the development of these topics over time, namely the relative importance of the identified matched topics in their corpora at certain points in time. For this comparison we construct topic weight time series and are interested in whether the identified topic matches exhibit similar dynamics over the considered period of time.

As described above, for each document in a corpus, the estimated LDA models provide probabilities of each topic occurring is this document, e.g., each document is represented as a vector with the length equal to the number of topics selected and sums up to one. To construct topic time series, the probabilities of each topic to occur in documents of the corpus are aggregated over all documents published in a given year and averaged on an annual basis.

To construct time series for topic matches identified within the German corpus for the two different languages considered, the average of the individual topic time series was calculated. If one of the values was missing in one of the time series, this value was replaced with the value from the second time series. In doing so, we were able to provide longer time series for the $DE^{ENG}$ and $DE^{GER}$ matches, as the share of German articles in the German corpus was substantially higher until the early 2000s.

In order to describe similarity between the time series for the matched topics we perform two steps. Firstly, given that the trajectories are quite ragged due to the limited number of texts per year, to ease visual inspection we smooth the series using a two-sided filter. In the second step, we evaluate the correlation coefficient and compute the Euclidean distance for the pairs of filtered series.

## 4.4 Results

### 4.4.1 Number of topics

As described in Section 4.2, the first step of the analysis consisted in identifying the optimal number of topics for each of the text corpora. The number of topics was selected by maximizing the singular BIC with a minimal number of topics set to $H_{min} = 10$ and a maximal number of topics set to $H_{max} = 100$. These boundaries were set based on the assessment of the variety of topics in the scientific publications considered. The models with different number of topics in the predefined range were assumed to have the same priors, i.e., $P(H_{min}) = P(H_{min} + 1) = \cdots = P(H_{max})$. The values of the learning coefficients $\lambda_{Hh}$ and their multiplicities $m_{Hh}$ were obtained using the formulas provided by Hayashi (2021). High precision computations were implemented using the Python module *decimal*.

Using the model selection procedure based on the sBIC, we identified 37 topics for the Polish data set, 20 topics for the $DE^{GER}$ data set and 60 topics for the $DE^{ENG}$ data set. The sBIC values for Poland and Germany depending on the number of topics are shown in Figures 4.3 and 4.4, respectively.[7] The red dashed lines indicate the selected number of topics for each corpus. For the $DE^{ENG}$ data set, maximizing the sBIC would lead to 74 topics. However, it can be seen in Figure 4.4b that the shape of the curve of sBIC values in the interval from 55 to 75 topics is almost flat. For this reason and due to a rather small data set consisting of 704 articles, we decided to consider 60 topics for this corpus.

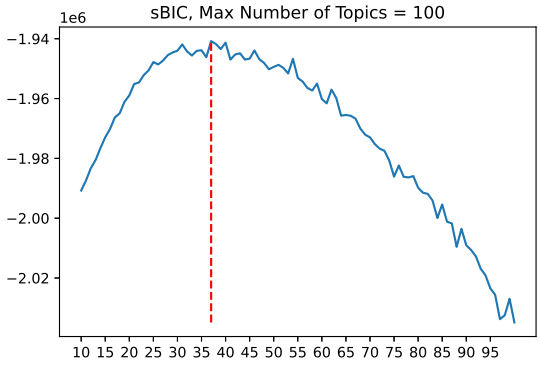
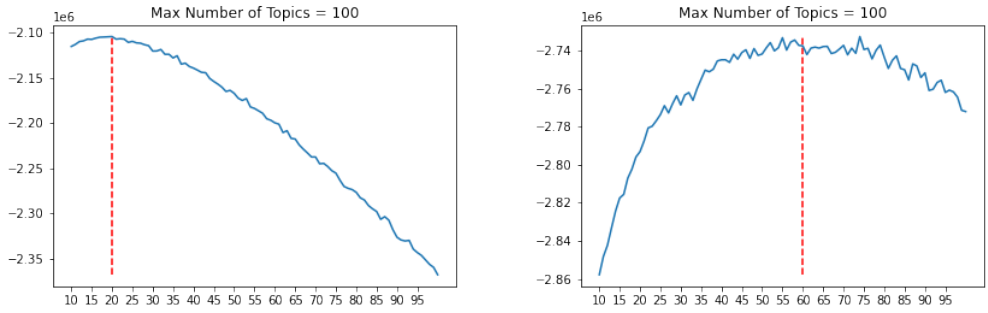



**Figure 4.3:** Distribution of the sBIC values for the Polish corpus



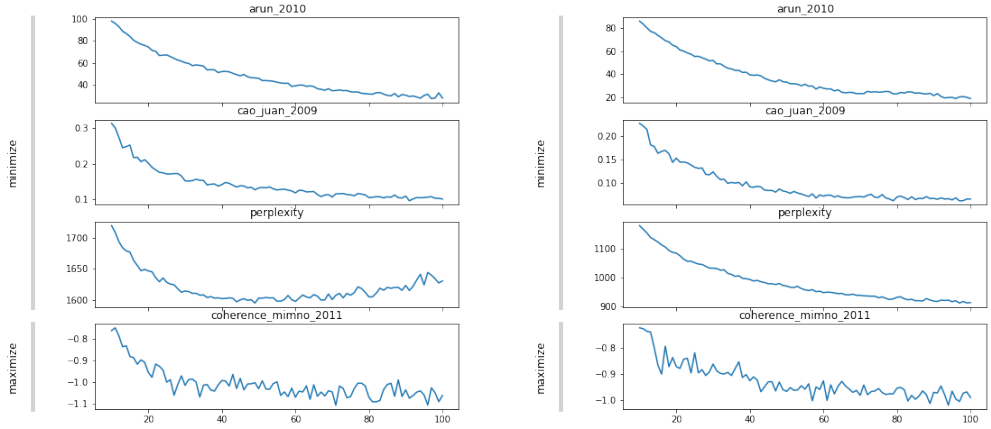

**(a)** $DE^{GER}$ Data **(b)** $DE^{ENG}$ Data

**Figure 4.4:** Distribution of the sBIC values for the German corpus

For comparison, we also applied some of the techniques commonly used in the literature to chose the optimal number of topics. We used Python's module *tm toolkit* and calculated the following evaluation metrics available in this module: arun_2010 (Arun et al., 2010), cao_juan_2009 (Cao et al., 2009), perplexity, and coherence_mimno_2011 (Mimno et al.,

[7] As the LDA estimation procedure is based on an iterative Bayesian method and, thus, contains a randomness component, the resulting sBIC values also contain a random component. Increasing the number of iterations within the LDA estimation procedure reduces this stochastic component and makes the decision about the number of clusters more robust, which goes with higher computational cost.

2011). For our application, however, none of these metrics provides a clear indication of an optimal number of topics (see Figure 4.5 for the German corpora). In fact, the first two criteria seem to suggest always the largest number of topics, while the coherence criterion appears to favour a very small number of topics. Only perplexity suggests an inner solution for the smaller German corpus. The results for the Polish corpus are also inconclusive. Therefore, we stick to the novel sBIC measure with a strong theoretical background.




**(a)** $DE^{GER}$ Data **(b)** $DE^{ENG}$ Data

**Figure 4.5:** Evaluations metrics from *tm toolkit*

All topics identified in the LDA model selected by sBIC both for German and Polish corpora are interpretable, i.e. by a visual inspection of word clouds composed of the 50 most common words for each topic (see Section 4.4.2 and the supplementary material B.7 to B.9) we are able to link the topics to relevant economic issues. Thus, although sBIC does not directly measure any semantic quality of topics, the outcome of the model selection procedure using sBIC is a set of interpretable topics. If another criterion was used, then the selected number of topics would be very large or very small as compared to the number of topics selected by sBIC (see discussion above). It would imply obtaining either a small model, in which some interpretable topics were omitted, or a large model, in which some topics might be meaningless.

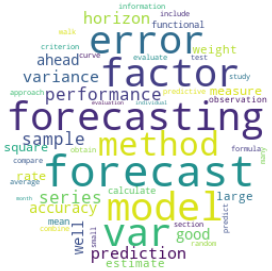
### 4.4.2 Topics

Figure 4.6 shows some topics from the Polish corpus. The uncovered topics deal with different aspects of econometric models (Topics 3 and 36), forecasting (Topic 16), and modelling of macroeconomic indicators (Topics 9, 10, 17). Figure 4.7 presents some $DE^{GER}$ topics discussing unemployment (Topic 0), consumption&income (Topic 14), government spending (Topic 15) as well as some $DE^{ENG}$ topics discussing business indicators (Topic 2), wage (Topic 6), and stock market (Topic 14). The font size of the words in the presented word clouds corresponds to the relative importance of the words in a topic. The full set of

all topics obtained for all corpora can be found in the supplementary material B.7 to B.9.

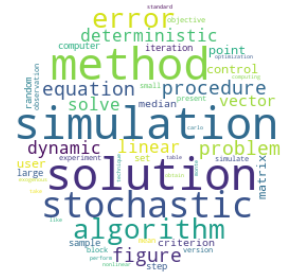

Topic 9


Topic 10

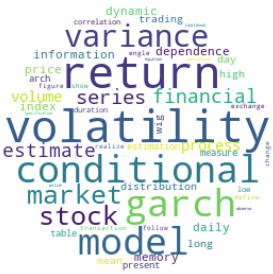



Topic 16


Topic 17




Topic3




Topic 36

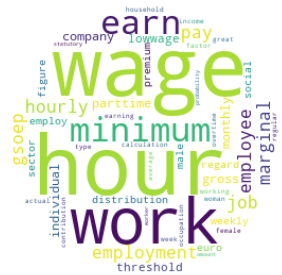
**Figure 4.6:** $PL^{ENG}$ Topics


Topic 0


Topic 14


Topic 15


Topic 2




Topic 6


Topic 14

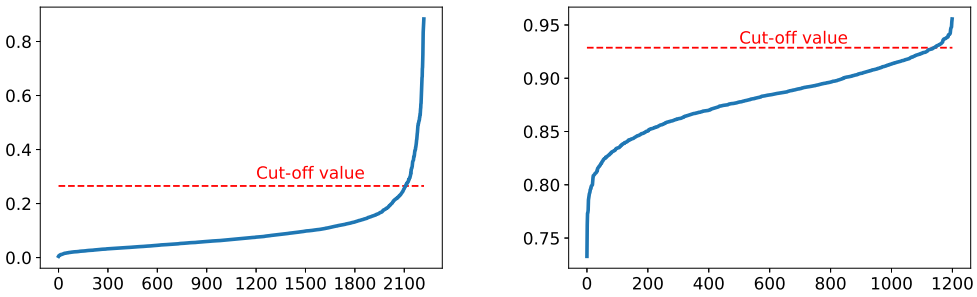
**Figure 4.7:** German Topics

### 4.4.3 Matching of topics

In the matching stage, we first performed topic matching between $DE^{ENG}$ and $PL^{ENG}$ topics based on the topic-word vectors and the intersection of the two vocabularies (2523 words). We identified best matches based on the cosine similarity values.

Given that the topic matching procedure provides a match for all topics in the corpus considered, we have to differentiate between "sensible" matches, i.e., pairs of topics with high similarity, and best matches, which pair quite different topics. To this end, we propose to determine a cut-off value based on the distribution of the cosine similarity values between all possible topic pairs, which should provide an approximation of the values we might expect for random matches. Figure 4.8a presents this distribution of the cosine similarity values which exhibits an "elbow" around a value of 0.2.

We decided to use the 95% percentile (0.265) of the empirical distribution as the cut-off value. An alternative approach for determining this cut-off value could be based on Monte Carlo simulations for corpora with common and different topics. The computational resources required for such Monte Carlo simulations would be very high, and the setup would have to take into account how similar the topics within each corpus are, i.e., results could be used only for a very specific setting. Therefore, we leave such an analysis to future work. Apart from defining a cut-off value, we also checked systematically whether there are some multiple assignments, i.e., topics matched with the same topic in the targeted corpus. In this case, we only kept the pairs with the highest cosine similarity value. At the same time, we took the non-symmetry of the cosine similarity measure into account and checked whether the topics in the selected matches are the nearest neighbours of each other also when reversing the direction of matching.




**(a)** Topic-word frequencies based Matching **(b)** Embedding based Matching

**Figure 4.8:** Distribution of cosine similarities between all possible matches

Using this approach, a total number of 24 topic pairs were identified. Figures 4.9 and 4.10 show two of them. The matched topics seem to be quite similar as can be concluded from the word clouds comprising the 50 most frequent words. While the first one deals with international economic links, the second one is about business cycle analysis. Further matches deal with topics such as loan debt, hypothesis testing, forecasting methods, labour market and (un)employment, capital growth, oil shocks, inflation, income, trade etc. (see


$DE^{ENG}$ Topic 33


$PL^{ENG}$ Topic 14

**Figure 4.9:** Topic Match "International economic relationships"


$DE^{ENG}$ Topic 2


$PL^{ENG}$ Topic 23

**Figure 4.10:** Topic Match "Business cycle"

supplementary material B.10). Results of a robustness check using Jensen-Shannon distance as the similarity measure instead of cosine similarity are provided in Appendix B.5.

### 4.4.4 Embedding based matching of topics

For the multilingual corpus, we applied the proposed embedding based approach to match the topics between the $DE^{ENG}$ and $DE^{GER}$ data subsets. To this end, each topic was represented as the sum vector of 100 word vectors corresponding to its 100 most frequent words. Cosine similarity values were calculated between the topic pairs and for each topic in one language its nearest neighbour in the other language was chosen as its match. Analogously to the topic-word based matching, we used the 95% percentile of the cosine values between all possible topic pairs, 0.93, as the cut-off value to identify "sensible" matches (see Figure 4.8b). This approach resulted in 16 topic pairs within the German data set (see supplementary material B.11). Finally, we made use of the English part of these matches to obtain overall matches between both German topic sets and the $PL^{ENG}$ topics.

Figures 4.11 and 4.12 show examples of these multilingual matches of the two corpora. In the first example it becomes obvious that both German topics and the corresponding Polish topic deal with the labor market and unemployment. However, not all of the obtained $DE^{ENG}$ and $DE^{GER}$ topic pairs appear meaningful to the same extent. The second example

exhibits that the $DE^{GER}$ topic deals with private consumption and income and might be related to the analysis of the lifetime cycle of private households. The matched $DE^{ENG}$ topic is about investment and capital growth as is the one in $PL^{ENG}$. This unsatisfactory outcome might be due to the specific multilingual embedding that was used for the matching. Therefore, further research is required for selecting or generating appropriate embeddings in order to improve the proposed approach to multilingual topic matching.


$DE^{GER}$ $DE^{ENG}$ $PL^{ENG}$

**Figure 4.11:** Topic Match "Unemployment"


$DE^{GER}$ $DE^{ENG}$ $PL^{ENG}$

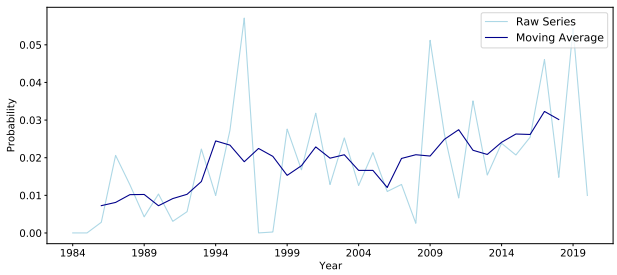
**Figure 4.12:** Topic Match "Capital growth"

### 4.4.5 Time series of topic weights

To enable comparison of patterns in the series of weights, the time series for the $PL^{ENG}$ and $DE^{ENG}$ text corpora were filtered with the centered equally-weighted moving average computed using 5 observations, MA(5). In the next step, the Euclidean distance and correlation coefficients were evaluated for the smoothed series. The values of these measures as well as cosine similarity scores are reported in Table B.6.5 in the Appendix B.6. Below we discuss relations between the weight series of two topic pairs.

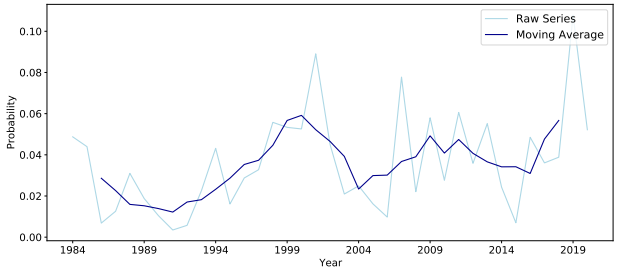
Figure 4.13 and Figure 4.14 present word clouds and weight series (both raw and filtered) for two selected pairs of topics from $PL^{ENG}$ and $DE^{ENG}$ corpora. Analogous figures for all topic pairs are provided in supplementary material B.10. Figure 4.13 shows the interest in time in the topics on international economic links. This match was characterised by

a high cosine similarity score of 0.86146. The Euclidean distance between filtered series amounted to 0.10798 and the coefficient of correlation had a value of 0.63348. Filtered series for topics identified in both text corpora show a mild upward trend: an increasing interest in international links might be associated with an increasing openness and integration in the European Union economies. High positive correlation can be additionally attributed to common patterns in the dynamics which are synchronised in time.





$DE^{ENG}$ Topic 33





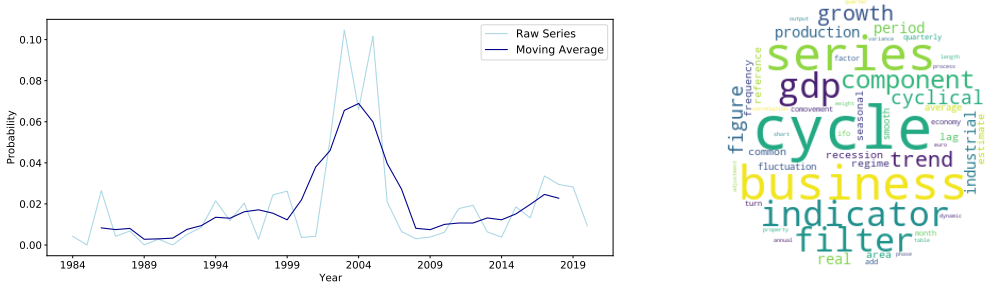
$PL^{ENG}$ Topic 14

**Figure 4.13:** Topic Match "International economic relationships"

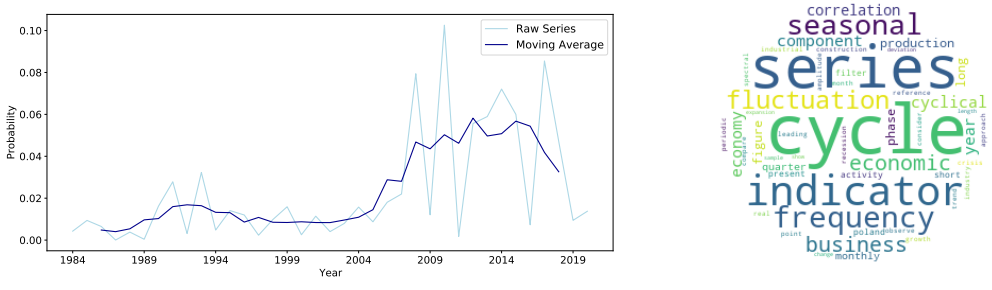
Figure 4.14 shows words clouds and plots of weight series for pair of topics concerning the business cycle. Although cosine similarity for this pair of topics is also high (0.83445) the Euclidean distance is larger (0.16063) and a negative correlation coefficient (-0.20020) indicates weaker co-movement. The negative correlation can be explained by the misalignment of interest in these topics in time due to different economic circumstances of Germany and Poland. The creation of the euro area brought about an increased interest in business cycle studies in Germany. This can be explained by the need to better understand economic fluctuations in the common currency area. The importance of a similar topic for Poland grew later – after joining the common market.

## 4.5 Conclusions and outlook

The present work considered scientific publications from Germany and Poland. The primary aim was to uncover main topics in the corpora using LDA modelling and to compare them with each other on the basis of the proposed matching approaches. The results of the current paper are a valuable contribution to the growing body of literature on text-as-data




$DE^{ENG}$ Topic 2




$PL^{ENG}$ Topic 23

**Figure 4.14:** Topic Match "Business cycle"

applications for several reasons.

First, we address one of the great challenges in the topic modelling, namely the choice of the optimal number of topics. We suggest to select the number of topics based on sBIC, a Bayesian information criterion adapted to the singularity of LDA models. Our analysis shows that the proposed information criterion leads to coherent topics in the considered text corpora. Second, we propose a topic matching approach that allows to compare the topic-word distributions of two different LDA models and to identify suitable topic pairs across text corpora. This matching approach made it possible to find meaningful topic pairs describing similar concepts in Polish and German corpora. Third, we suggest a data based procedure for identifying potentially meaningful matches of topics across corpora. Using a data based cut-off value for the minimum value of cosine similarity for matched topics, we were able to separate sensible matches from the remaining ones which do not correspond to similar topics in the colloquial sense. Fourth, we address the problem of topic matching between two corpora trained in different languages by proposing a language agnostic topic matching approach using multilingual word embeddings.

The work could be extended along the following lines. Further research is required to examine the randomness component of the sBIC criterion, e.g., by conducting a simulation study for sBIC. Additional work is also needed to improve the proposed embedding based matching approach as well as to examine further possibilities for topic matching in a multilingual context. This is recommended since not all topics matched across different languages in the current study were fully convincing. Furthermore, some of the identified topic matches seem to be closely related to themes concerning real macroeconomic activities,

e.g. inflation, unemployment, income etc. Further research will examine more closely the links between the corresponding topic time series and real macroeconomic variables with a focus on potential differences across countries.

# Appendix B

## B.1 Data Preparation

### German Data

In the first step, data were downloaded from the De Gruyter website. Table B.1.1 summarizes number of articles published each year. Up to 2000, the volumes were available as scanned pdf files. The Optical Character Recognition (OCR) was used to transform existing pdf files into text files. The text files were then copied into Microsoft Word and saved again with other coding (Unicode UTF-8). After that, the new text files were again copied into Word and the following preparation steps were taken:

1. Mark each issue number with heading 1.
2. Mark each article title with heading 2.
3. Remove the following non-textual elements:
    - Table of contents,
    - Author names and article numbers,
    - Formulas and special characters,
    - Bibliographies,
    - Tables and appendices.

After these preparation steps, the data could be imported into Python and be further preprocessed and analysed.

### Polish Data

The texts from the two data sources for Poland had different forms. The conference proceedings were available as hard copies of the volumes, while CEJEME articles were digital and had the format of LaTeX or pdf files.

The available conference volumes (including more than 9000 pages) were scanned and saved as pdf files. The description of the MM data is provided in Table B.1.2. Altogether, the data included 514 full length papers (with or without an abstract) and 231 abstracts (without the main text). In the next step, OCR was performed and the texts were saved as

| Year | Volume | Number of Articles |
|---|---|---|
| 1984 | 199 | 42 |
| 1985 | 200 | 46 |
| 1986 | 201&202 | 49 |
| 1987 | 203 | 49 |
| 1988 | 204&205 | 93 |
| 1989 | 206 | 52 |
| 1990 | 207 | 46 |
| 1991 | 208 | 53 |
| 1992 | 209&2010 | 88 |
| 1993 | 211&2012 | 81 |
| 1994 | 213 | 53 |
| 1995 | 214 | 53 |
| 1996 | 2015 | 57 |
| 1997 | 216 | 45 |
| 1998 | 217 | 53 |
| 1999 | 218&219 | 84 |
| 2000 | 220 | 52 |
| 2001 | 221 | 39 |
| 2002 | 222 | 41 |
| 2003 | 223 | 46 |
| 2004 | 224 | 40 |
| 2005 | 225 | 45 |
| 2006 | 226 | 35 |
| 2007 | 227 | 42 |
| 2008 | 228 | 34 |
| 2009 | 229 | 41 |
| 2010 | 230 | 44 |
| 2011 | 231 | 46 |
| 2012 | 232 | 43 |
| 2013 | 233 | 35 |
| 2014 | 234 | 41 |
| 2015 | 235 | 39 |
| 2016 | 236 | 33 |
| 2017 | 237 | 26 |
| 2018 | 238 | 27 |
| 2019 | 239 | 36 |
| 2020 | 240 | 35 |

**Table B.1.1:** Number of Articles published in Journal of Economics and Statistics

docx files. Over the years, the volumes were published by various publishing houses, using alternative typesetting styles and techniques. Thus, also the resulting source files differed considerably and extensive manual labor was needed to clean the texts. This preparatory step involved removing front and back matter, running heads and feet, tables, footnotes, figures, equations and other mathematical expressions as well as references. The beginning of each article was also manually marked. In addition, within each paper, information on the title, authors, affiliations, abstract (if present) and main body of the text were uniformly organized so that they could be easily identified by the code.

| Year of conference | No. of volumes | Meetings | Contents |
|---|---|---|---|
| 1984 | 1 | MM and MF | 17 full papers |
| 1985 | 2 | MM | 26 full papers |
| 1986 and 1987 | 1 | MM | 17 full papers |
| 1988 and 1989 | 1 | MM | 18 full papers |
| 1990 | 1 | MM | 11 full papers |
| 1991 | 1 | MM | 12 full papers |
| 1992 | 1 | MM | 16 full papers |
| 1993 | 2 | MM | 24 full papers |
| 1994 | 1 | MM | 11 full papers |
| 1995 | 2 | MM and MSA | 29 full papers |
| 1996 | 3 | MM and MSA | 46 full papers |
| 1997 | 2 | MM and AMFET | 20 full papers |
| 1998 | 1 | MM and AMFET | 8 full papers |
| 1999 | 2 | MM and AMFET | 40 full papers |
| 2000 | 2 | MM and AMFET | 14 full papers |
| 2001 | 2 | MM and AMFET | 33 full papers |
| 2002 | 2 | MM and AMFET | 25 full papers |
| 2003 | 2 | MM and AMFET | 20 full papers |
| 2004 | 2 | MM and AMFET | 23 full papers |
| 2005 | 2 | MM and AMFET | 27 full papers |
| 2006 | 2 | MM and AMFET | 25 full papers |
| 2007 | 2 | MM and AMFET | 25 full papers |
| 2008 | 1 | MM and AMFET | 10 full papers |
| 2009 | 1 | MM and AMFET | 6 full papers |
| 2010 | 1 | MM and AMFET | 6 full papers |
| 2011 | 1 | MM and AMFET | 5 full papers |

**Table B.1.2:** Proceedings from Macromodels International Conference and joint meetings

The input files from CEJEME used for modelling had LaTeX format[8]. Detailed information on the numbers of papers published in each volume and issue of CEJEME is presented in Table B.1.3. All papers had an abstract.

[8] The only exception was the second paper from 2012, issue 1. For this article only the pdf file was available. A LaTeX file was created and manually cleaned of mathematical expressions, tables etc.

Initially, a structured database on the documents was created using Matlab. This step consisted in extracting from LaTeX files information on the publication year, names of authors, title of each paper, key words, JEL codes and abstracts. Abstracts were cleaned of all mathematical expressions and LaTeX formatting commands. Gathering this information was facilitated by a relatively stable LaTeX template used in the publication process.

| year | volume | issue | number of papers | year | volume | issue | number of papers |
|---|---|---|---|---|---|---|---|
| 2009 | 1 | 1 | 5 | 2015 | 7 | 1 | 3 |
| | 1 | 2 | 4 | | 7 | 2 | 3 |
| | 1 | 3 | 4 | | 7 | 3 | 3 |
| | 1 | 4 | 4 | | 7 | 4 | 3 |
| 2010 | 2 | 1 | 4 | 2016 | 8 | 1 | 3 |
| | 2 | 2 | 3 | | 8 | 2 | 3 |
| | 2 | 3 | 3 | | 8 | 3 | 3 |
| | 2 | 4 | 3 | | 8 | 4 | 3 |
| 2011 | 3 | 1 | 3 | 2017 | 9 | 1 | 3 |
| | 3 | 2 | 3 | | 9 | 2 | 3 |
| | 3 | 3 | 3 | | 9 | 3 | 3 |
| | 3 | 4 | 3 | | 9 | 4 | 3 |
| 2012 | 4 | 1 | 3 | 2018 | 10 | 1 | 3 |
| | 4 | 2 | 3 | | 10 | 2 | 3 |
| | 4 | 3 | 3 | | 10 | 3 | 3 |
| | 4 | 4 | 3 | | 10 | 4 | 3 |
| 2013 | 5 | 1 | 3 | 2019 | 11 | 1 | 3 |
| | 5 | 2 | 3 | | 11 | 2 | 3 |
| | 5 | 3 | 3 | | 11 | 3 | 3 |
| | 5 | 4 | 3 | | 11 | 4 | 3 |
| 2014 | 6 | 1 | 3 | 2020 | 12 | 1 | 3 |
| | 6 | 2 | 3 | | 12 | 2 | 4 |
| | 6 | 3 | 3 | | 12 | 3 | 4 |
| | 6 | 4 | 3 | | 12 | 4 | 4 |

**Table B.1.3:** Articles published in CEJEME

In the next step, to form the text corpus(es) appropriate for further probabilistic analysis, the text files had to be suitably prepared. Initial editing was done in two steps. In the first step, the original files with .tex extension were modified to obtain the main body of the text by removing the following elements:

1. Initial article information including: the author name(s), affiliation(s), e-mail address(es), dates of submitting and accepting the article,
2. The abstract, keywords and JEL classification codes,
3. Text appearing in running head (the author name(s) and short title of the paper) and running foot (the author name(s) and information on the volume and issue) of the journal,
4. Figures and tables,

5. Formulas, mathematical symbols and Greek letters,
6. References,
7. Selected LaTeX commands which prevented compilation after the above alternations of the texts were done, e.g. those introducing change of line.

In the second step, PDF files were obtained on the basis of the filtered LaTeX files. The pdfs were transformed to a text format.

## B.2 Stopwords

**English stopwords** removed from article texts using the **R** package *tm*:

I, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, would, should, could, ought, I'm, you're, he's, she's, it's, we're, they're, I've, you've, we've, they've, I'd, you'd, he'd, she'd, we'd, they'd, I'll, you'll, he'll, she'll, we'll, they'll, isn't, aren't, wasn't, weren't, hasn't, haven't, hadn't, doesn't, don't, didn't, won't, wouldn't, shan't, shouldn't, can't, cannot, couldn't, mustn't, let's, that's, who's, what's, here's, there's, when's, where's, why's, how's, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very

**Additional stopwords** removed from the texts of articles:

appendix, acknowledgements, introduction

**German stopwords**[9] removed from article texts:

a, ab, aber, ach, acht, achte, achten, achter, achtes, ag, alle, allein, allem, allen, aller, allerdings, alles, allgemeinen, als, also, am, an, ander, andere, anderem, anderen, anderer, anderes, anderm, andern, anderr, anders, au, auch, auf, aus, ausser, ausserdem, außer, außerdem, b, bald, bei, beide, beiden, beim, beispiel, bekannt, bereits, besonders, besser, besten, bin, bis, bisher, bist, c, d, d.h, da, dabei, dadurch, dafür, dagegen, daher, dahin, dahinter, damals, damit, danach, daneben, dank, dann, daran, darauf, daraus, darf, darfst, darin, darum, darunter, darüber, das, dasein, daselbst, dass, dasselbe, davon, davor, dazu, dazwischen,

[9] This comprehensive German stopwords list was downloaded from `https://github.com/stopwords-iso/stopwords-de`

daß, dein, deine, deinem, deinen, deiner, deines, dem, dementsprechend, demgegenüber, demgemäss, demgemäß, demselben, demzufolge, den, denen, denn, denselben, der, deren, derer, derjenige, derjenigen, dermassen, dermaßen, derselbe, derselben, des, deshalb, desselben, dessen, deswegen, dich, die, diejenige, diejenigen, dies, diese, dieselbe, dieselben, diesem, diesen, dieser, dieses, dir, doch, dort, drei, drin, dritte, dritten, dritter, drittes, du, durch, durchaus, durfte, durften, dürfen, dürft, e, eben, ebenso, ehrlich, ei, ei„, eigen, eigene, eigenen, eigener, eigenes, ein, einander, eine, einem, einen, einer, eines, einig, einige, einigem, einigen, einiger, einiges, einmal, eins, elf, en, ende, endlich, entweder, er, ernst, erst, erste, ersten, erster, erstes, es, etwa, etwas, euch, euer, eure, eurem, euren, eurer, eures, f, folgende, früher, fünf, fünfte, fünften, fünfter, fünftes, für, g, gab, ganz, ganze, ganzen, ganzer, ganzes, gar, gedurft, gegen, gegenüber, gehabt, gehen, geht, gekannt, gekonnt, gemacht, gemocht, gemusst, genug, gerade, gern, gesagt, geschweige, gewesen, gewollt, geworden, gibt, ging, gleich, gott, gross, grosse, grossen, grosser, grosses, groß, große, großen, großer, großes, gut, gute, guter, gutes, h, hab, habe, haben, habt, hast, hat, hatte, hatten, hattest, hattet, heisst, her, heute, hier, hin, hinter, hoch, hätte, hätten, i, ich, ihm, ihn, ihnen, ihr, ihre, ihrem, ihren, ihrer, ihres, im, immer, in, indem, infolgedessen, ins, irgend, ist, j, ja, jahr, jahre, jahren, je, jede, jedem, jeden, jeder, jedermann, jedermanns, jedes, jedoch, jemand, jemandem, jemanden, jene, jenem, jenen, jener, jenes, jetzt, k, kam, kann, kannst, kaum, kein, keine, keinem, keinen, keiner, keines, kleine, kleinen, kleiner, kleines, kommen, kommt, konnte, konnten, kurz, können, könnt, könnte, l, lang, lange, leicht, leide, lieber, los, m, machen, macht, machte, mag, magst, mahn, mal, man, manche, manchem, manchen, mancher, manches, mann, mehr, mein, meine, meinem, meinen, meiner, meines, mensch, menschen, mich, mir, mit, mittel, mochte, mochten, morgen, muss, musst, musste, mussten, muß, mußt, möchte, mögen, möglich, mögt, müssen, müsst, müßt, n, na, nach, nachdem, nahm, natürlich, neben, nein, neue, neuen, neun, neunte, neunten, neunter, neuntes, nicht, nichts, nie, niemand, niemandem, niemanden, noch, nun, nur, o, ob, oben, oder, offen, oft, ohne, ordnung, p, q, r, recht, rechte, rechten, rechter, rechtes, richtig, rund, s, sa, sache, sagt, sagte, sah, satt, schlecht, schluss, schon, sechs, sechste, sechsten, sechster, sechstes, sehr, sei, seid, seien, sein, seine, seinem, seinen, seiner, seines, seit, seitdem, selbst, sich, sie, sieben, siebente, siebenten, siebenter, siebentes, sind, so, solang, solche, solchem, solchen, solcher, solches, soll, sollen, sollst, sollt, sollte, sollten, sondern, sonst, soweit, sowie, später, startseite, statt, steht, suche, t, tag, tage, tagen, tat, teil, tel, tritt, trotzdem, tun, u, uhr, um, und, uns, unse, unsem, unsen, unser, unsere, unserer, unses, unter, v, vergangenen, viel, viele, vielem, vielen, vielleicht, vier, vierte, vierten, vierter, viertes, vom, von, vor, w, wahr, wann, war, waren, warst, wart, warum, was, weg, wegen, weil, weit, weiter, weitere, weiteren, weiteres, welche, welchem, welchen, welcher, welches, wem, wen, wenig, wenige, weniger, weniges, wenigstens, wenn, wer, werde, werden, werdet, weshalb, wessen, wie, wieder, wieso, will, willst, wir, wird, wirklich, wirst, wissen, wo, woher, wohin, wohl, wollen, wollt, wollte, wollten, worden, wurde, wurden, während, währenddem, währenddessen, wäre, würde, würden, x, y, z, z.b, zehn, zehnte, zehnten, zehnter, zehntes, zeit, zu, zuerst, zugleich, zum, zunächst, zur, zurück, zusammen, zwanzig, zwar, zwei, zweite, zweiten, zweiter, zweites, zwischen, zwölf, über,

überhaupt, übrigens

## B.3 Robustness Check: Translation

As a robustness check, we translated the German texts from JES into English using DeepL API. The vocabulary of the joint dataset is almost identical (about 85%) to the vocabulary of the English subset of the data. For the joint dataset $DE^{JOINT}$, we identified the optimal number of topics using the proposed sBIC measure (see Figure B.3.1).

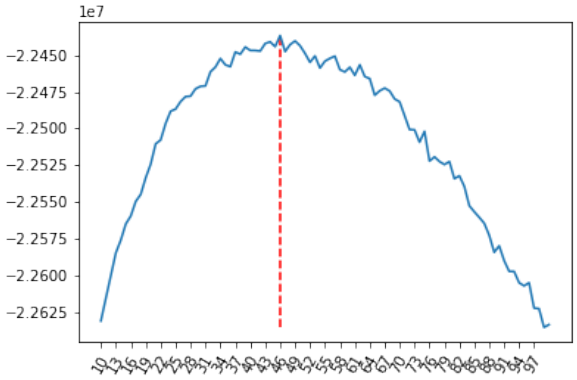



**Figure B.3.1:** Distribution of the sBIC values for the joint German corpus

In the next step, we performed standard topic matching in both directions $DE^{JOINT} \rightarrow DE^{ENG}$ as well as $DE^{ENG} \rightarrow DE^{JOINT}$ to account for the non-symmetry of the matching procedure. 34 out of the 46 identified topic pairs are the best matches of each other.

Using the proposed standard matching approach we identified 24 sensible matches between the corpora for both countries reported in Table B.6.5 in Appendix B.6. Next, we could analyse whether matching $PL^{ENG} \rightarrow DE^{JOINT}$ and then $DE^{JOINT} \rightarrow DE^{ENG}$ would result in the same topic pairs as compared to directly matching $PL^{ENG} \rightarrow DE^{ENG}$. An example of this analysis is shown in Figure B.3.2. The $PL^{ENG}$ Topic 10 was initially assigned to $DE^{ENG}$ Topic 52, both dealing with inflation and monetary policy. Using the $DE^{JOINT}$ LDA model, we find a similar topic that is the best match to both $PL^{ENG}$ Topic 10 and $DE^{ENG}$ Topic 52. For 17 out of the 24 relevant topic pairs, a similar result is obtained.

Therefore, in general, machine translation might be considered a good alternative when dealing with multilingual corpora. However, the additional costs, the quality of translation for certain corpora, and the "black box" character of machine translations must be taken into account.

## B.4 Robustness Check: Sklearn vs Gensim

To account for possible differences in topic distributions resulting from different LDA implementations, we considered a further Python module *gensim*. For all the considered datasets, namely $PL^{ENG}$, $DE^{ENG}$, $DE^{GER}$, we estimated LDA models using gensim with

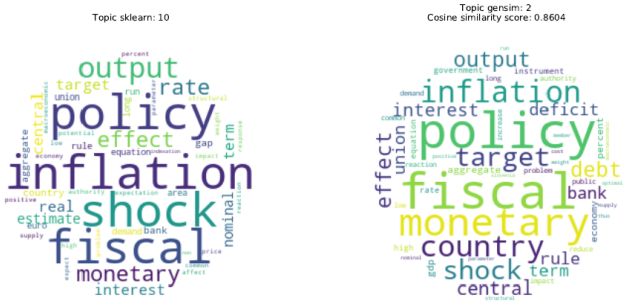

**Figure B.3.2:** Topic Matching using Machine Translation

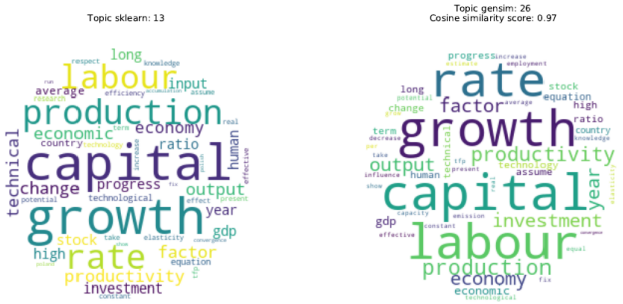
the number of topics according to sBIC. For each dataset, we calculated the following evaluation metrics: perplexity, average topic similarity Cao et al. (2009) and average topic coherence Mimno et al. (2011). The results are summarized in Table B.4.4. According to the perplexity and average topic similarity measures, *sklearn* seems to perform better. As for average coherence of the resulting topics, the scores seem to be quite similar.

| | $PL^{ENG}$ | | $DE^{ENG}$ | | $DE^{GER}$ | |
|---|---|---|---|---|---|---|
| | gensim | sklearn | gensim | sklearn | gensim | sklearn |
| Perplexity | 873.4 | 842.8 | 998.02 | 949.4 | 1697.62 | 1641.45 |
| Cao et al. (2009) | 0.18 | 0.14 | 0.11 | 0.08 | 0.28 | 0.2 |
| Mimno et al. (2011) | -0.76 | -0.79 | -1.04 | -0.93 | -0.92 | -0.95 |

**Table B.4.4:** Sklearn vs Gensim: model evaluation

As we are most interested in topics, for each dataset we compared the topic-word distributions using the proposed standard matching approach. In doing so, we aimed to find out whether topics uncovered using the two different LDA implementations overlap to a large extent or not. We found that most of the topics that are later identified as meaningful matches can be found by means of both implementations (see examples below).






## B.5 Robustness Check: Similarity Measure

We also performed topic matching using a different similarity measure, Jensen-Shannon distance, to see whether the main results and the resulting topic pairs change considerably. Analogously to the process presented in the main part of this paper, we first calculated the JS distances between all possible topic matches to derive a suitable cut-off value. Figure B.5.3 presents the distribution of the JS distance values. The lower the distance between two topic vectors, the more similar they are to each other. We took 0.05% (0.64) percentile as a cut-off value. We then removed multiple assignments keeping just the topic pair with the lowest distance. It resulted in 23 topic pairs. Four out of 24 assignments were different as compared to the results when using cosine similarity. One possible reason for this is that the $DE^{ENG}$ topic set is larger and contains some quite similar topics, i.e. one $PL^{ENG}$ topic could be a suitable pair to more than one of the $DE^{ENG}$ topics. An example is shown in Figure B.5.4. Both $DE^{ENG}$ topics seem to be related to the $PL^{ENG}$ topic. Overall, it could be observed that the use of a different similarity measure does not impact the results significantly.

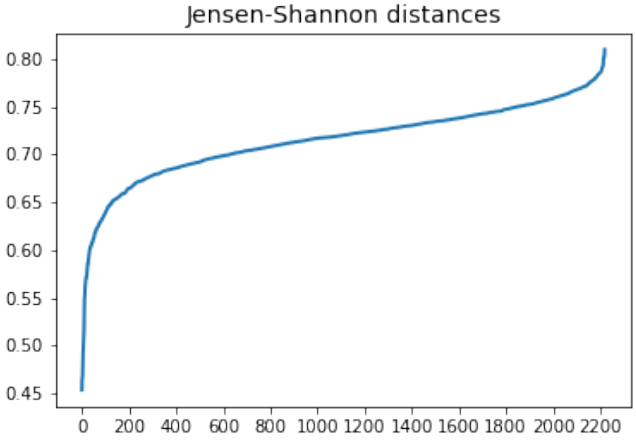
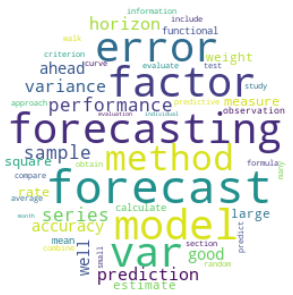



**Figure B.5.3:** Distribution of the Jensen-Shannon distances




$PL^{ENG}$ Topic 16 $DE^{ENG}$ Topic 1 $DE^{ENG}$ Topic 25

**Figure B.5.4:** Differences in the assignment

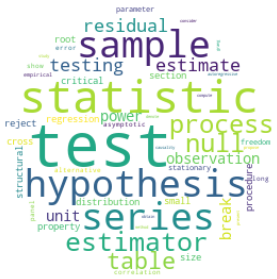
## B.6 Time Series Analysis

| Topic number in $PL^{ENG}$ | $DE^{ENG}$ | Cosine similarity* score | Euclidean distance | Correlation coefficient |
|---|---|---|---|---|
| 0 | 4 | 0.84617 | 0.15223 | -0.11350 |
| 1 | 15 | 0.45003 | 0.22556 | -0.47935 |
| 2 | 47 | 0.58684 | 0.24628 | 0.58412 |
| 3 | 13 | 0.50440 | 0.07083 | 0.42358 |
| 4 | 12 | 0.75001 | 0.24791 | -0.02885 |
| 6 | 36 | 0.48663 | 0.07005 | -0.39230 |
| 10 | 52 | 0.69914 | 0.08590 | -0.40108 |
| 11 | 20 | 0.72227 | 0.07436 | 0.60665 |
| 13 | 21 | 0.80610 | 0.20421 | 0.05272 |
| 14 | 33 | 0.86146 | 0.10798 | 0.63348 |
| 15 | 54 | 0.56860 | 0.14301 | 0.85228 |
| 16 | 1 | 0.88338 | 0.09773 | 0.03375 |
| 17 | 57 | 0.49535 | 0.13781 | 0.51459 |
| 21 | 43 | 0.63251 | 0.12119 | -0.58288 |
| 22 | 46 | 0.54288 | 0.12549 | -0.81556 |
| 23 | 2 | 0.83445 | 0.16063 | -0.20020 |
| 25 | 39 | 0.50876 | 0.09650 | -0.61320 |
| 26 | 37 | 0.65153 | 0.11825 | -0.24760 |
| 27 | 26 | 0.85116 | 0.09879 | 0.00184 |
| 29 | 30 | 0.66695 | 0.30690 | -0.44140 |
| 31 | 38 | 0.52991 | 0.06821 | 0.07816 |
| 32 | 53 | 0.70459 | 0.14300 | -0.55365 |
| 34 | 31 | 0.52545 | 0.08092 | 0.41222 |
| 35 | 28 | 0.34393 | 0.10423 | -0.59937 |

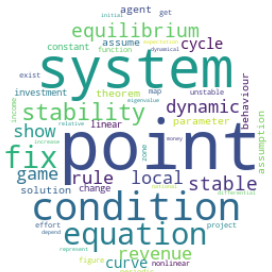
*: These values refer to the **topics content** and were calculated between word distributions of the topics, whereas Euclidean distance and correlation were calculated using the resulting topic time series.

**Table B.6.5:** Comparison of filtered weight time series

## B.7 LDA Topic for Poland

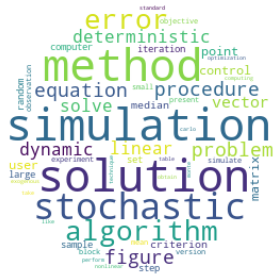



Topic 0

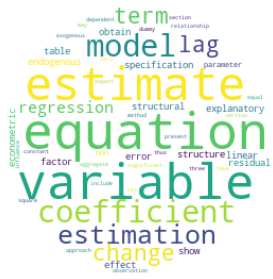



Topic 1


Topic 2




Topic 3




Topic 4


Topic 5


Topic 6


Topic 7


Topic 8


Topic 9


Topic 10

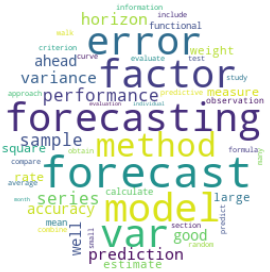

Topic 11


Topic 12


Topic 13

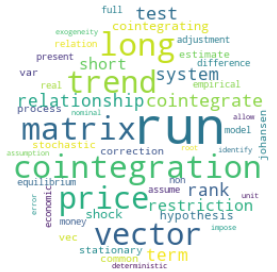

Topic 14


Topic 15

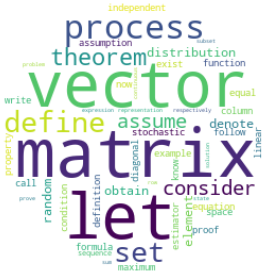



Topic 16

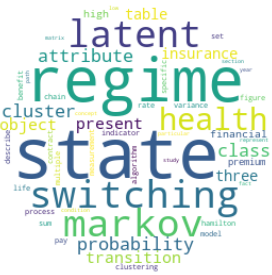

Topic 17



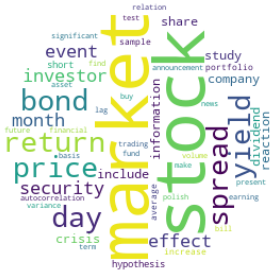

Topic 18




Topic 19




Topic 20

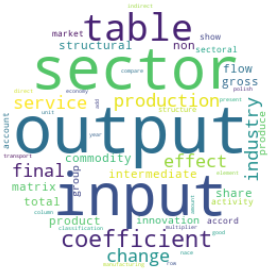



Topic 21

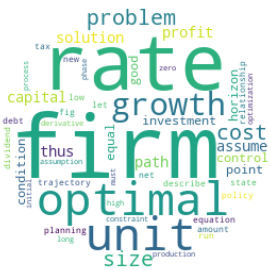

Topic 22


Topic 23


Topic 24

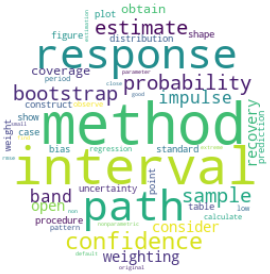



Topic 25




Topic 26

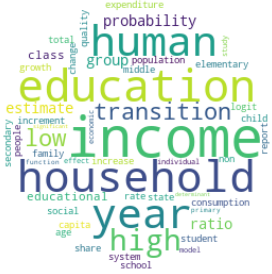

Topic 27



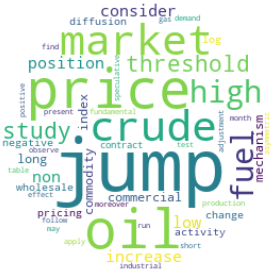

Topic 28


Topic 29

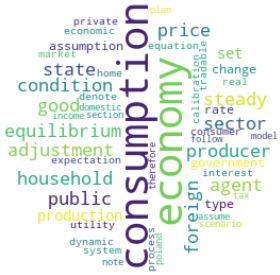

Topic 30




Topic 31



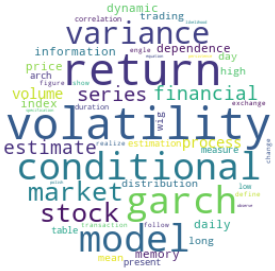

Topic 32




Topic 33


Topic 34

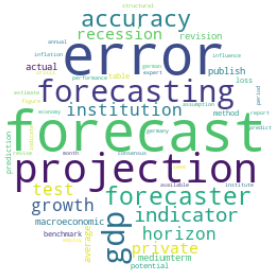

Topic 35




Topic 36

## B.8 LDA Topics for Germany (English)

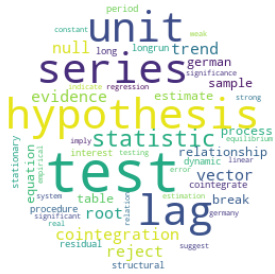

Topic 0




Topic 1

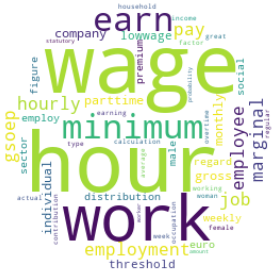

Topic 2


Topic 3

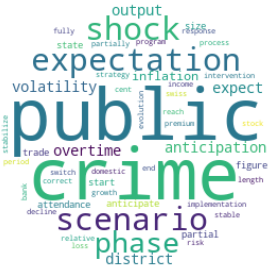



Topic 4


Topic 5




Topic 6



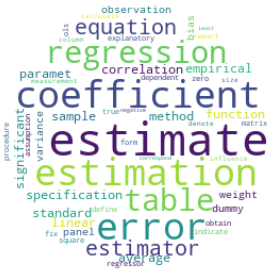

Topic 7


Topic 8


Topic 9


Topic 10

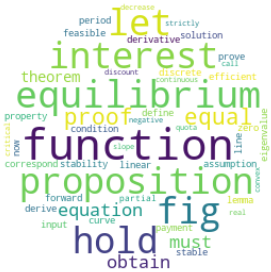

Topic 11



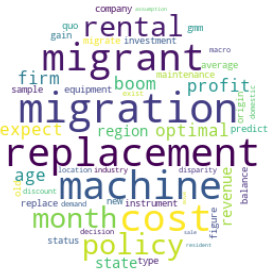

Topic 12


Topic 13


Topic 14




Topic 15




Topic 16


Topic 17


Topic 18


Topic 19


Topic 20


Topic 21


Topic 22

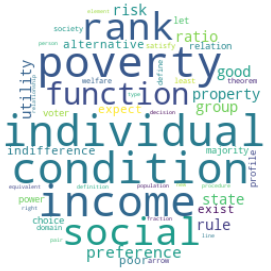

Topic 23


Topic 24


Topic 25

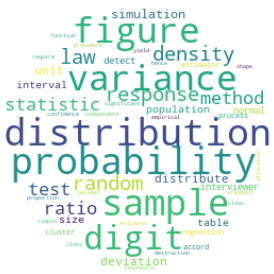

Topic 26


Topic 27




Topic 28


Topic 29




Topic 30

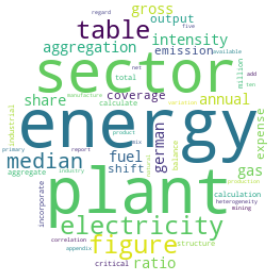

Topic 31


Topic 32


Topic 33


Topic 34

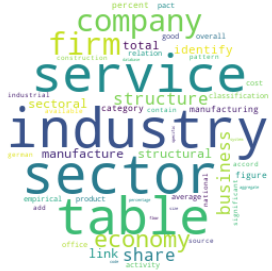

Topic 35




Topic 36

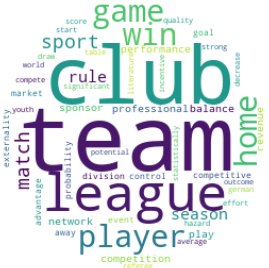

Topic 37


Topic 38




Topic 39


Topic 40

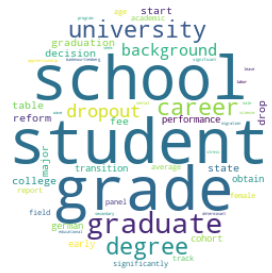



Topic 41


Topic 42


Topic 43


Topic 44

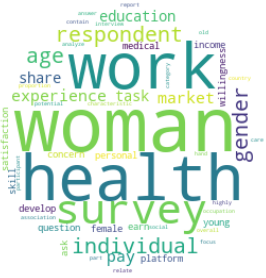



Topic 45


Topic 46


Topic 47


Topic 48



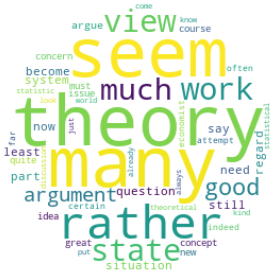

Topic 49


Topic 50


Topic 51


Topic 52


Topic 53




Topic 54


Topic 55


Topic 56


Topic 57


Topic 58


Topic 59

## B.9 LDA Topics for Germany (German)


Topic 0


Topic 1


Topic 2


Topic 3

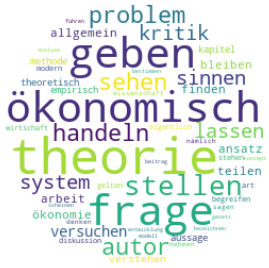

Topic 4


Topic 5


Topic 6

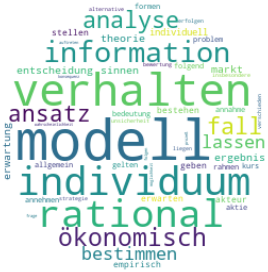

Topic 7


Topic 8




Topic 9


Topic 10




Topic 11


Topic 12


Topic 13


Topic 14


Topic 15

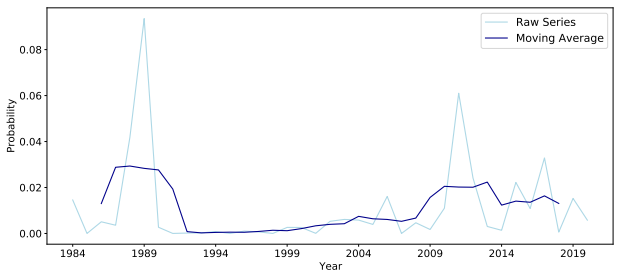

Topic 16

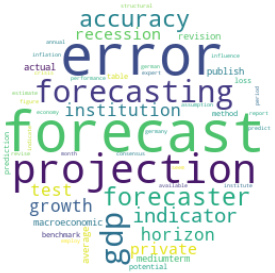

Topic 17


Topic 18

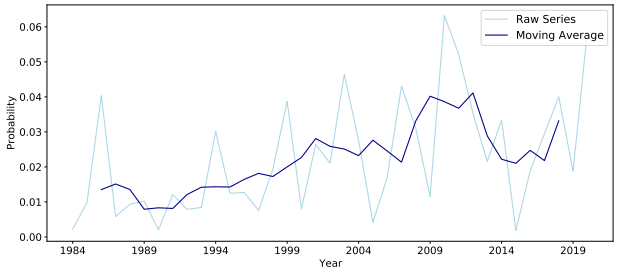

Topic 19

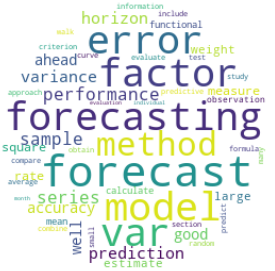
## B.10 $DE^{ENG}$ and $PL^{ENG}$ topic matches



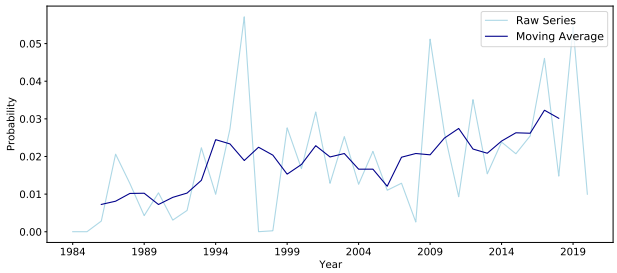




$DE^{ENG}$ Topic 1



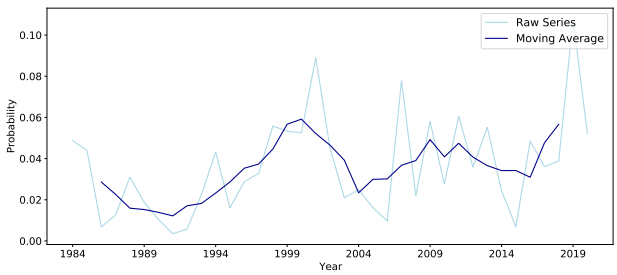




$PL^{ENG}$ Topic 16



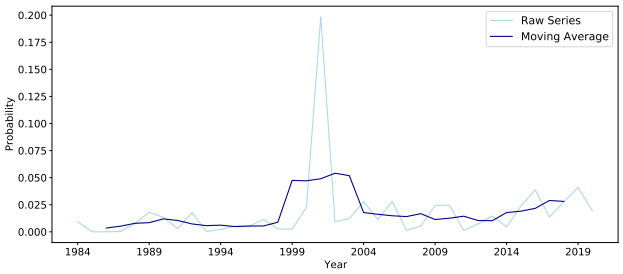


$DE^{ENG}$ Topic 33



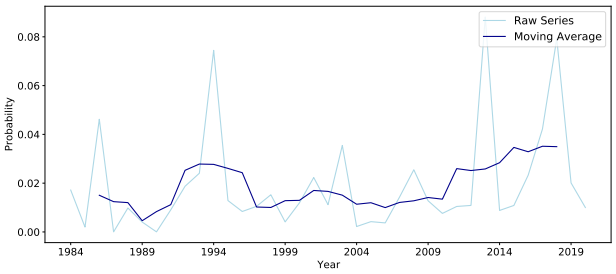


$PL^{ENG}$ Topic 14



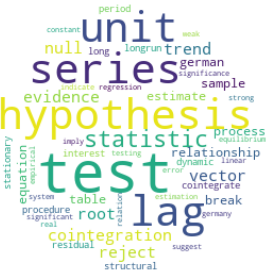


$DE^{ENG}$ Topic 26

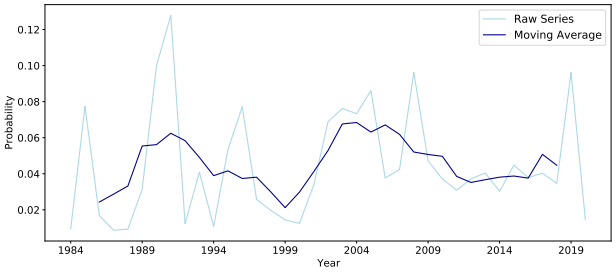


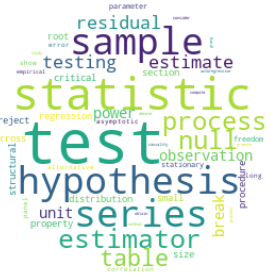


$PL^{ENG}$ Topic 27

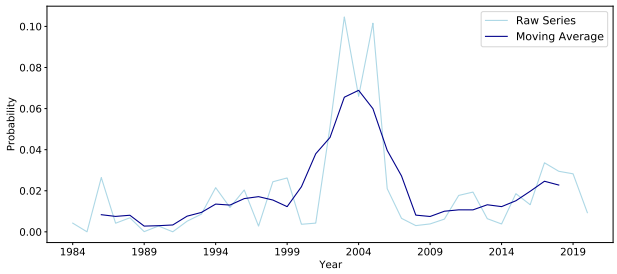




$DE^{ENG}$ Topic 4



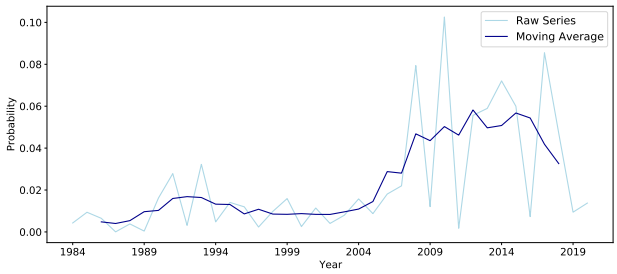




$PL^{ENG}$ Topic 0



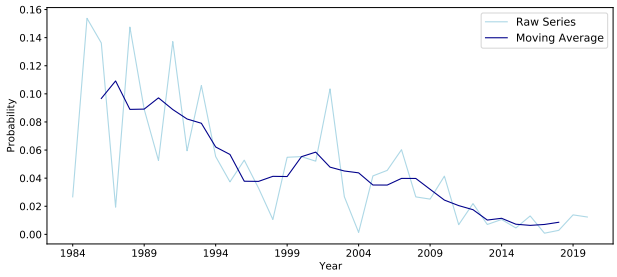


$DE^{ENG}$ Topic 2



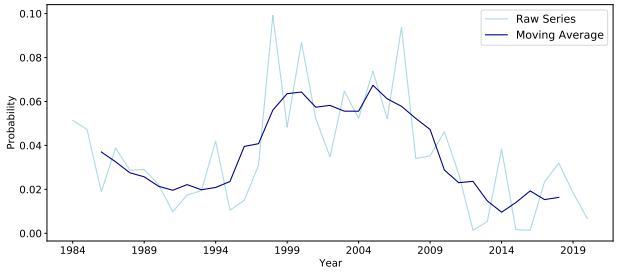


$PL^{ENG}$ Topic 23



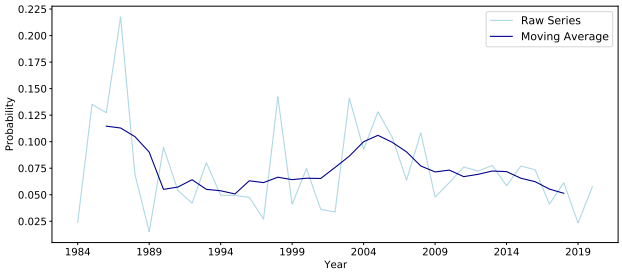


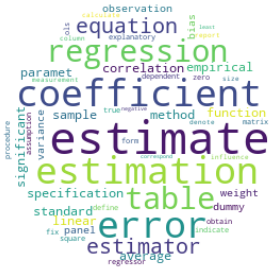
$DE^{ENG}$ Topic 21



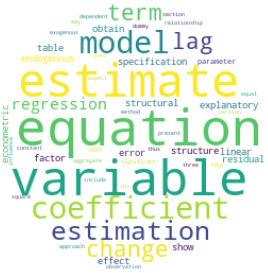


$PL^{ENG}$ Topic 13

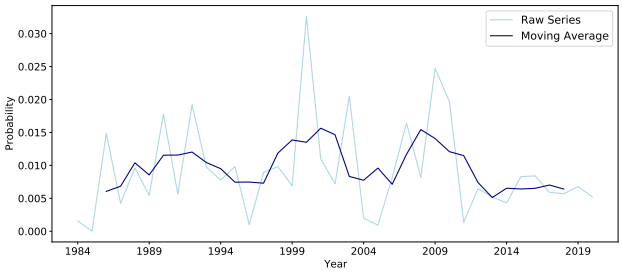



$DE^{ENG}$ Topic 12

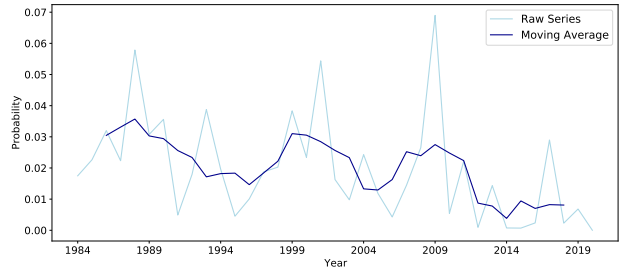




$PL^{ENG}$ Topic 4





$DE^{ENG}$ Topic 20

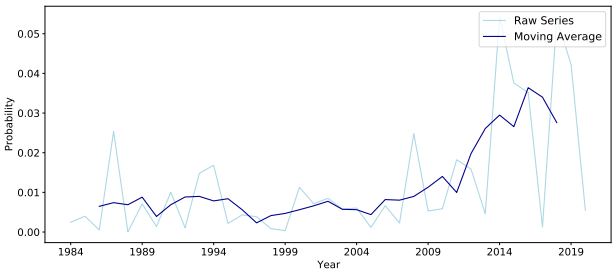


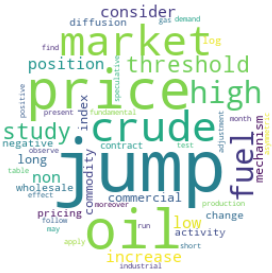


$PL^{ENG}$ Topic 11

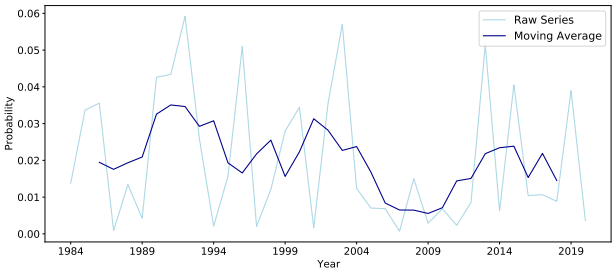


$DE^{ENG}$ Topic 53



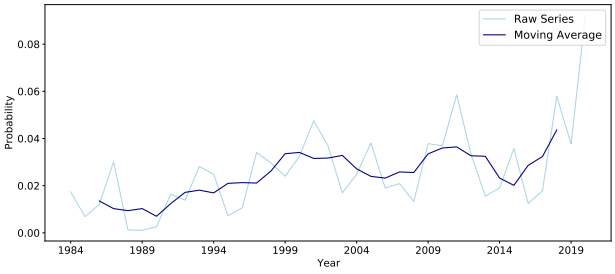




$PL^{ENG}$ Topic 32



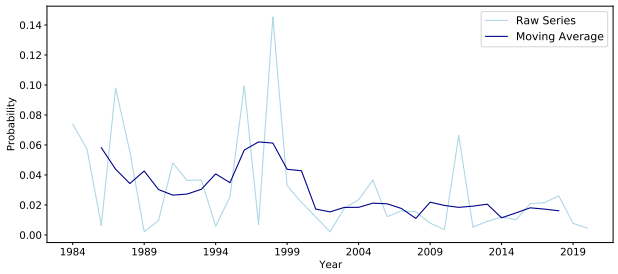


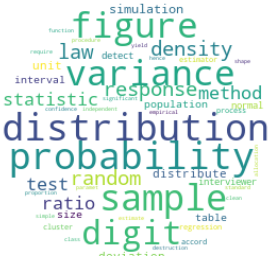
$DE^{ENG}$ Topic 52



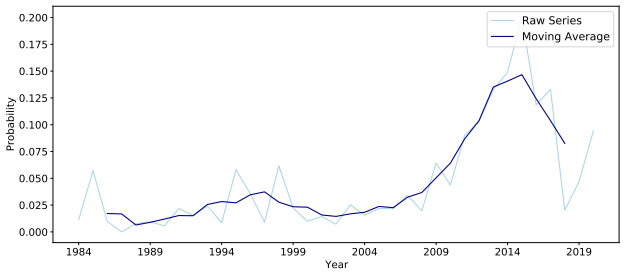


$PL^{ENG}$ Topic 10



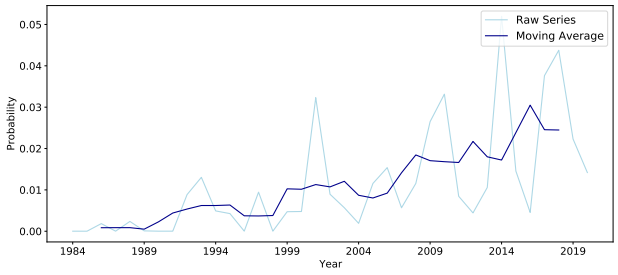




$DE^{ENG}$ Topic 30



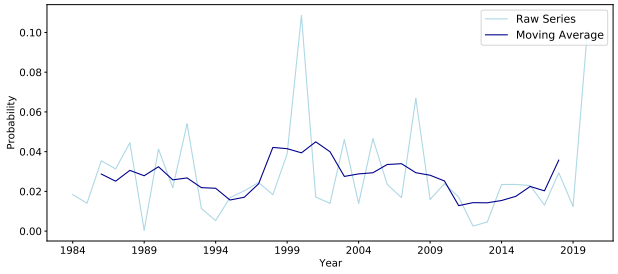


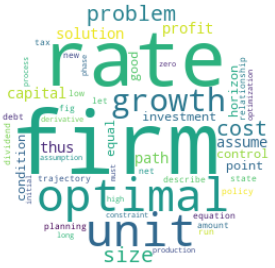
$PL^{ENG}$ Topic 29



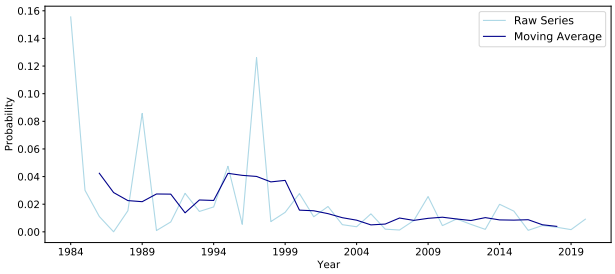


$DE^{ENG}$ Topic 37



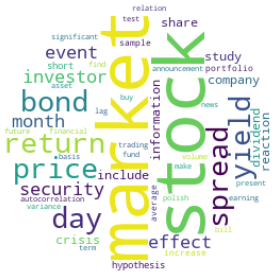




$PL^{ENG}$ Topic 26

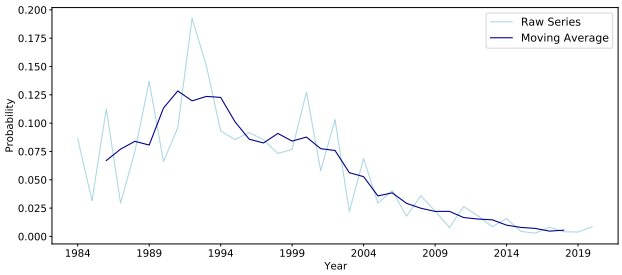




$DE^{ENG}$ Topic 43

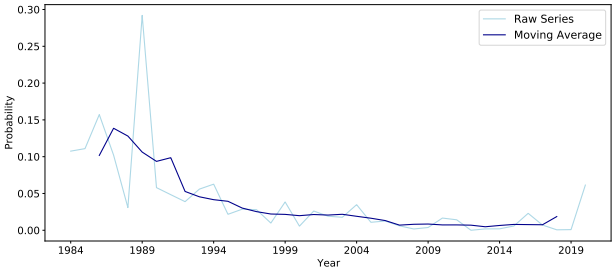




$PL^{ENG}$ Topic 21



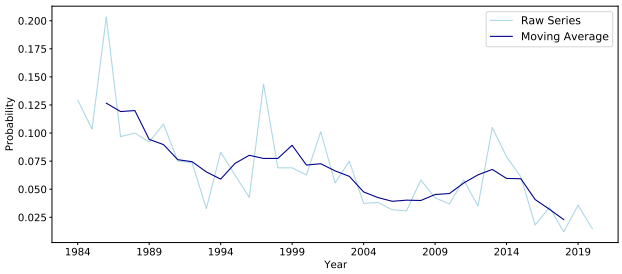


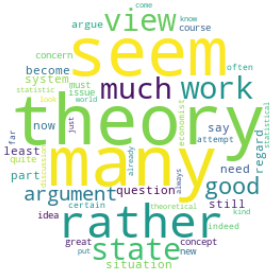
$DE^{ENG}$ Topic 47



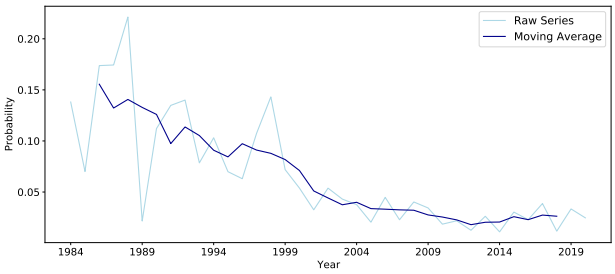


$PL^{ENG}$ Topic 2



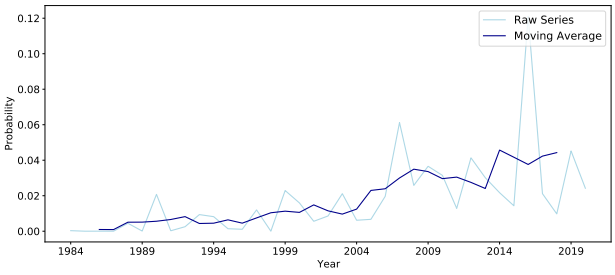




$DE^{ENG}$ Topic 54



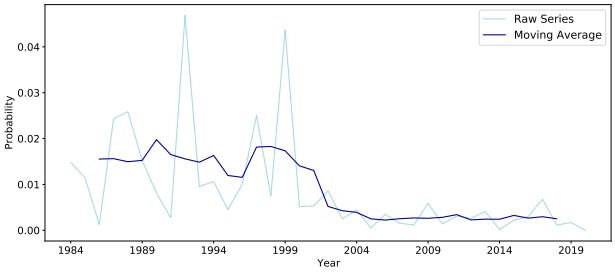


$PL^{ENG}$ Topic 15



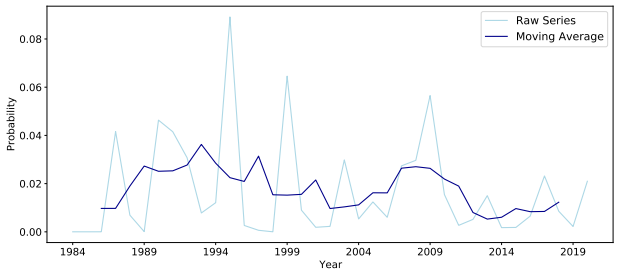


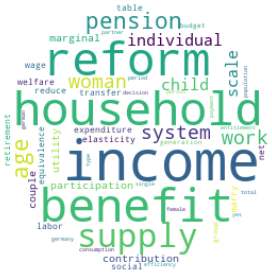
$DE^{ENG}$ Topic 46



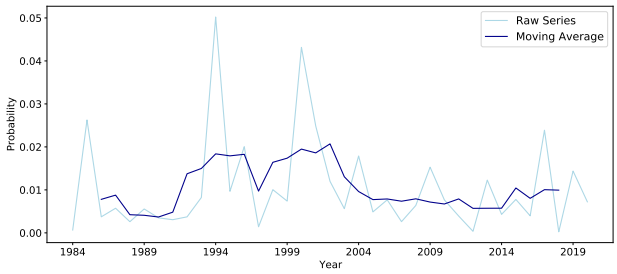


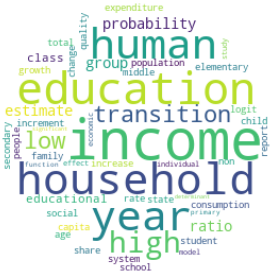
$PL^{ENG}$ Topic 22



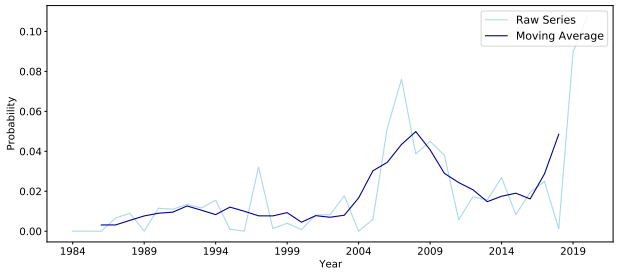




$DE^{ENG}$ Topic 38



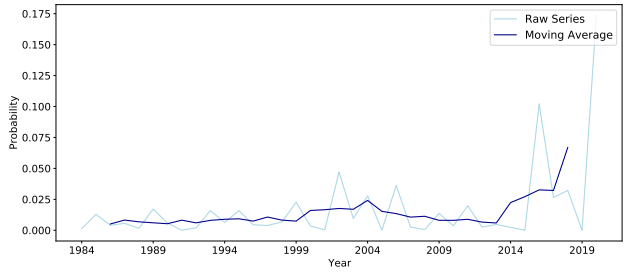




$PL^{ENG}$ Topic 31



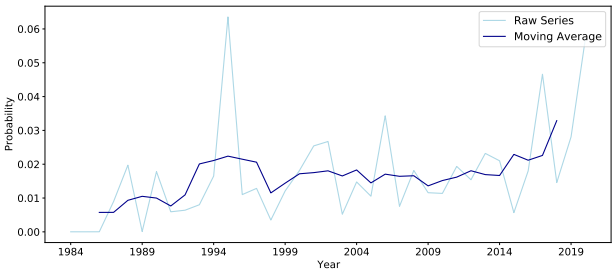


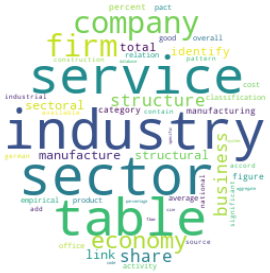
$DE^{ENG}$ Topic 31



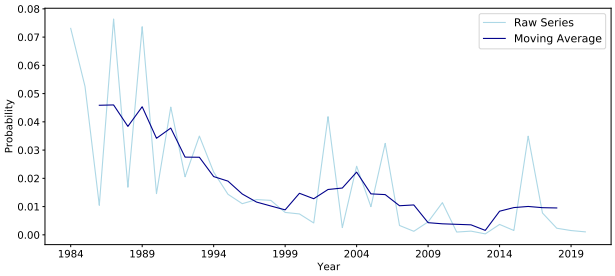


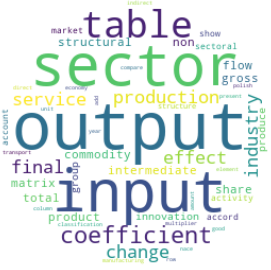
$PL^{ENG}$ Topic 34



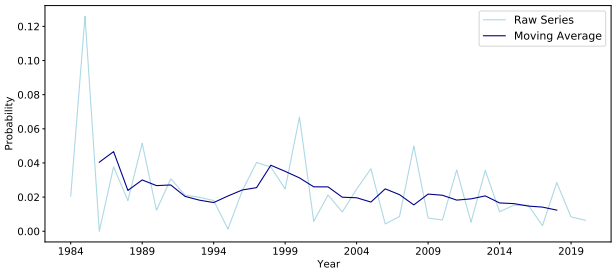




$DE^{ENG}$ Topic 39



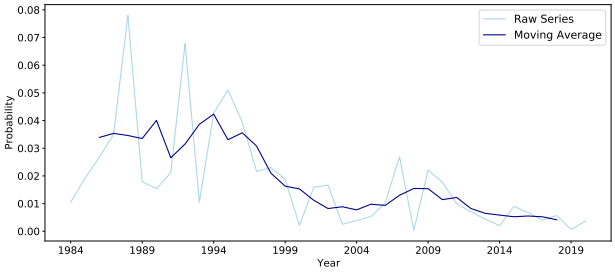



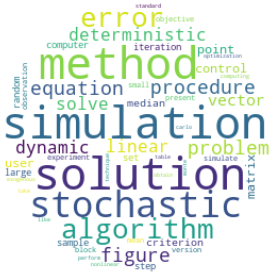

$PL^{ENG}$ Topic 25



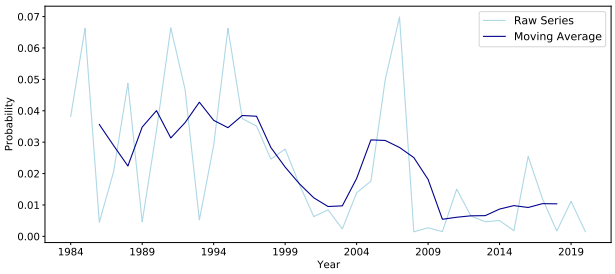


$DE^{ENG}$ Topic 13



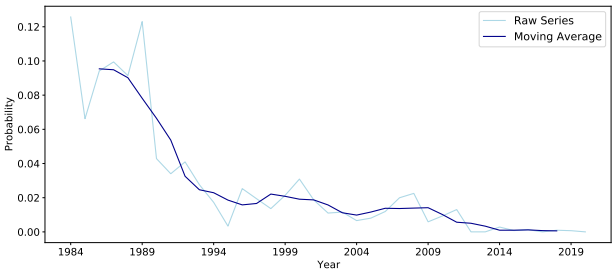




$PL^{ENG}$ Topic 3



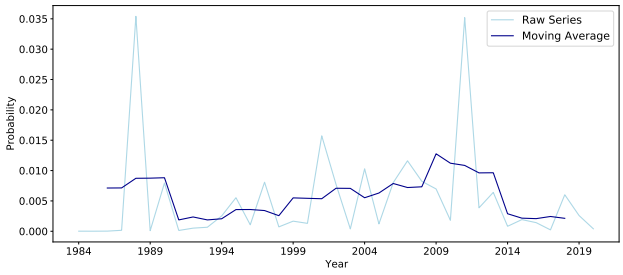


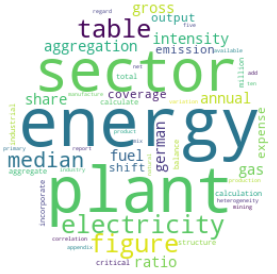
$DE^{ENG}$ Topic 57



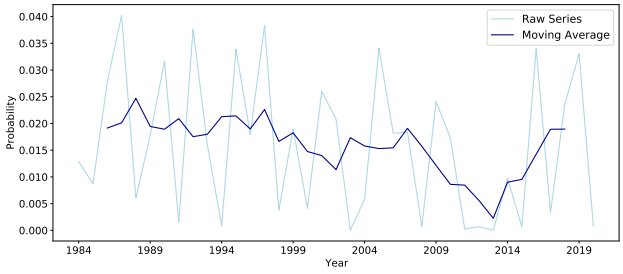


$PL^{ENG}$ Topic 17



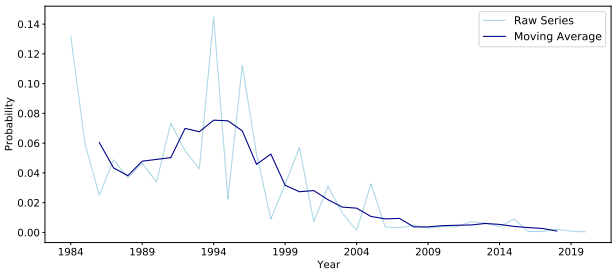



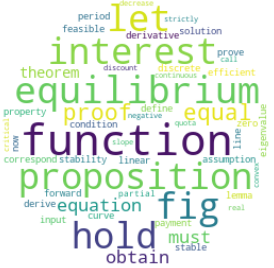

$DE^{ENG}$ Topic 36



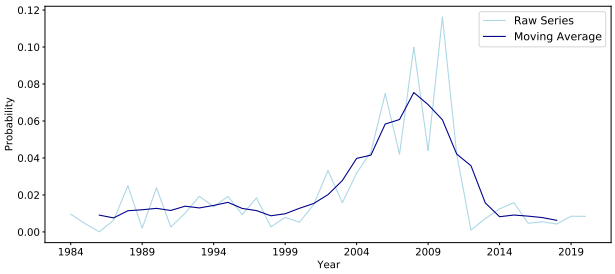


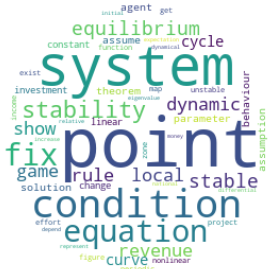
$PL^{ENG}$ Topic 6



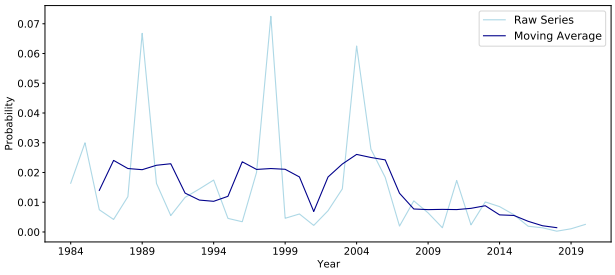



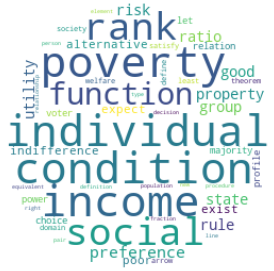

$DE^{ENG}$ Topic 15



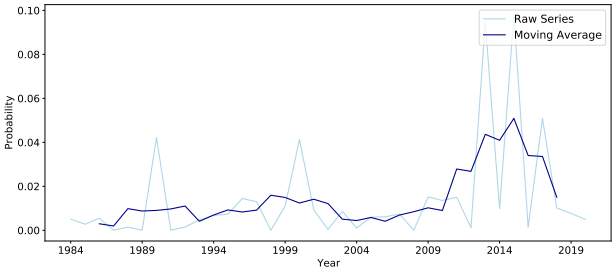




$PL^{ENG}$ Topic 1




$DE^{ENG}$ Topic 28



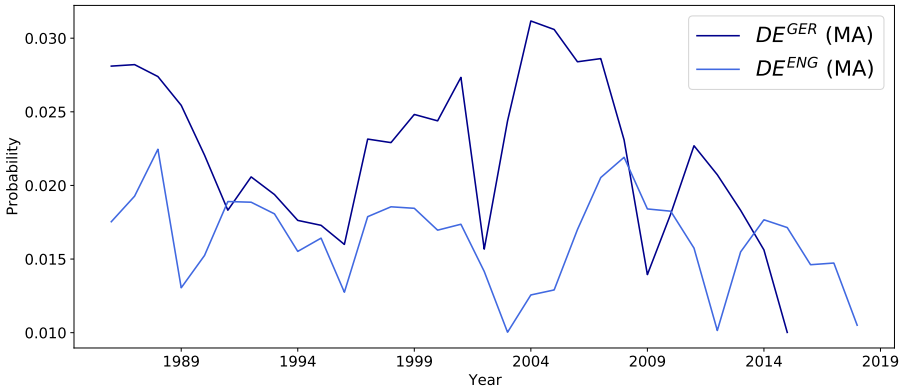


$PL^{ENG}$ Topic 35

## B.11 $DE^{ENG}$ and $DE^{GER}$ topic matches

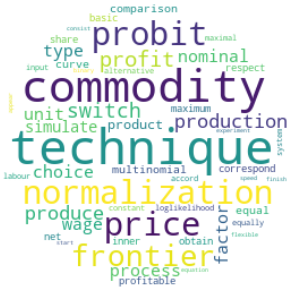

$DE^{GER}$ Topic 0


$DE^{ENG}$ Topic 19

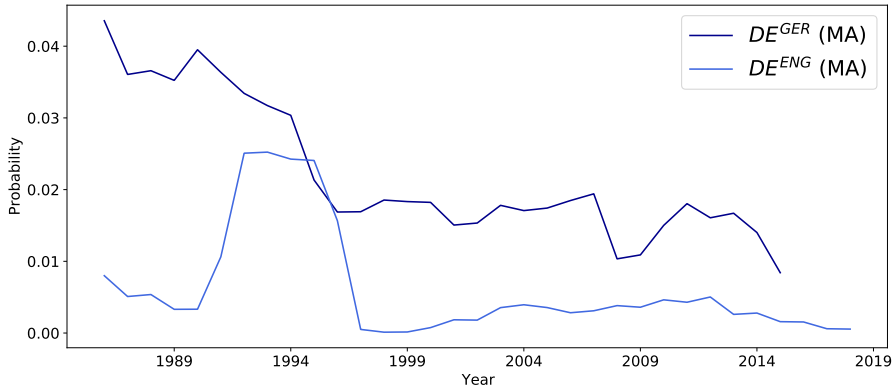




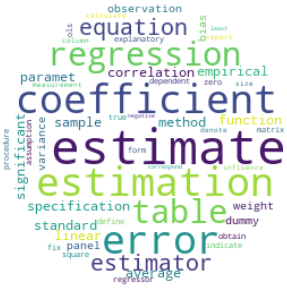
$DE^{GER}$ Topic 1



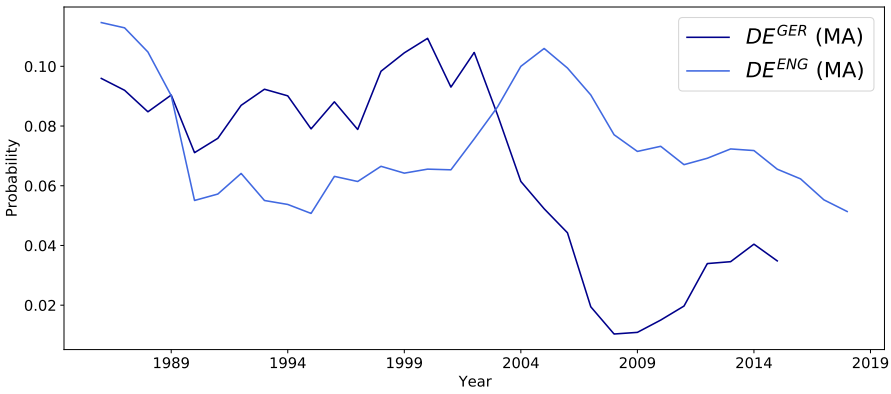

$DE^{ENG}$ Topic 55




$DE^{GER}$ Topic 2

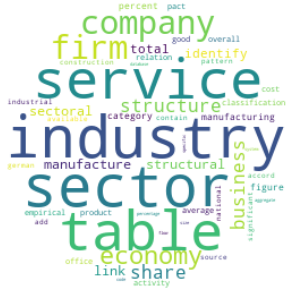



$DE^{ENG}$ Topic 12

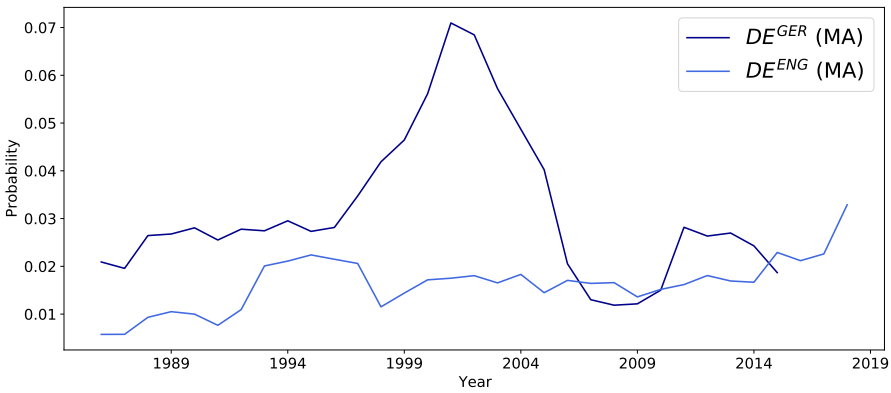




$DE^{GER}$ Topic 3

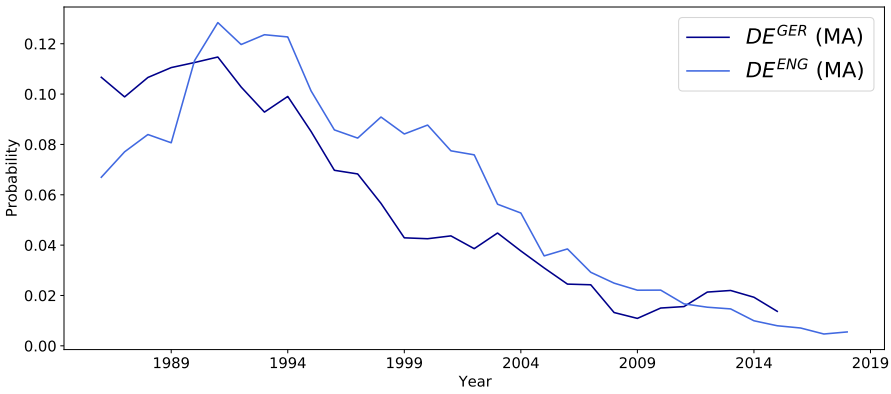



$DE^{ENG}$ Topic 39



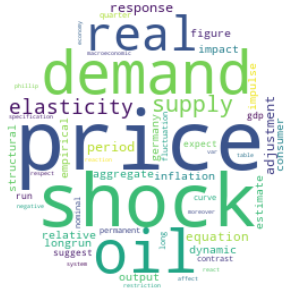

$DE^{GER}$ Topic 4


$DE^{ENG}$ Topic 47

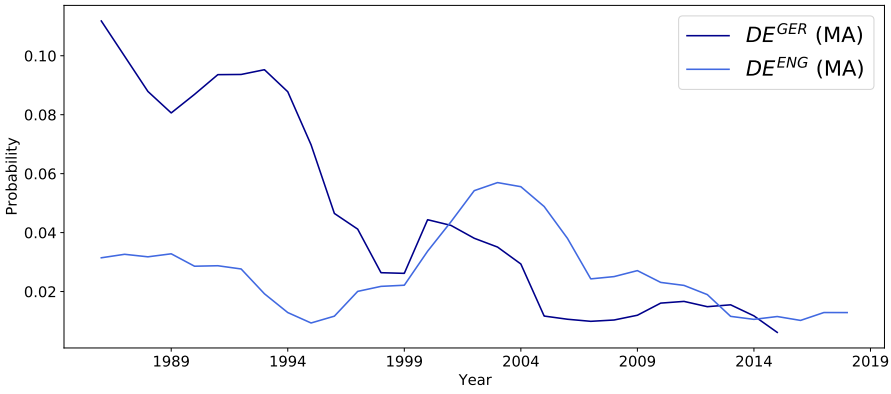




$DE^{GER}$ Topic 5

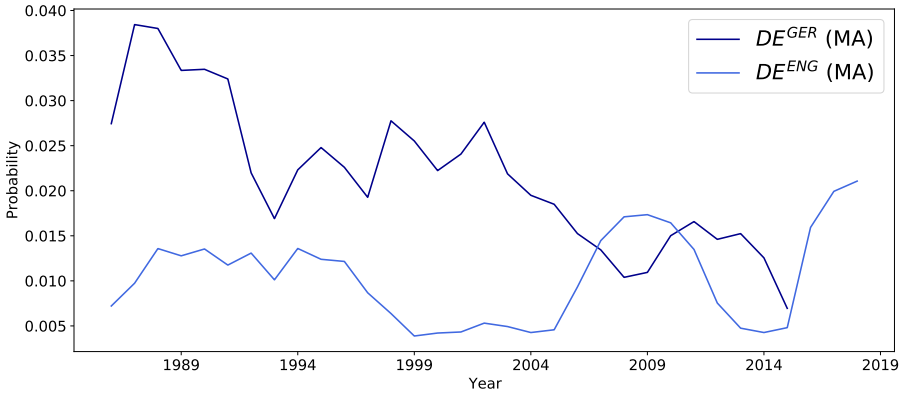



$DE^{ENG}$ Topic 53




$DE^{GER}$ Topic 6

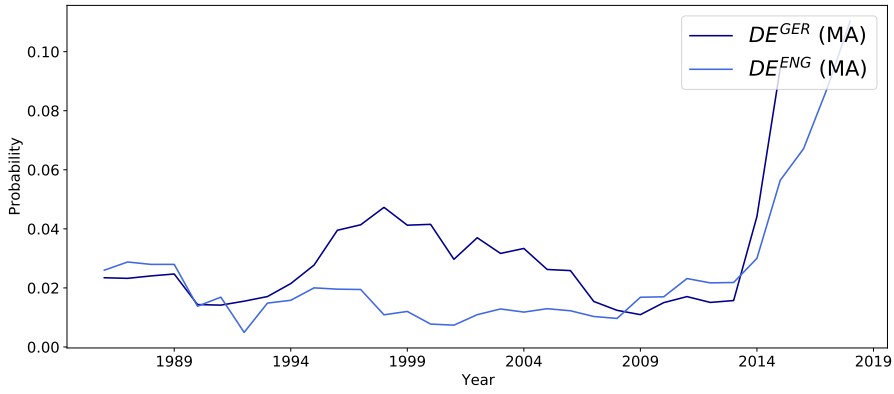

$DE^{ENG}$ Topic 35





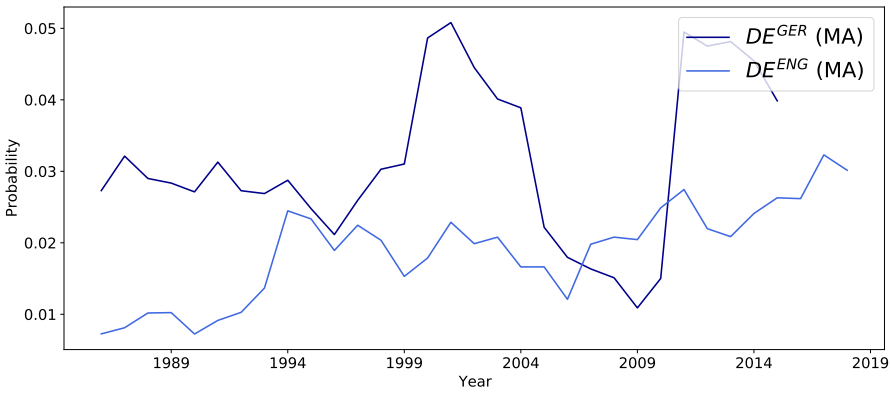
$DE^{GER}$ Topic 7

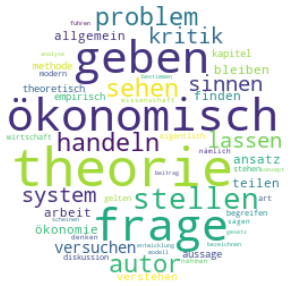

$DE^{ENG}$ Topic 10



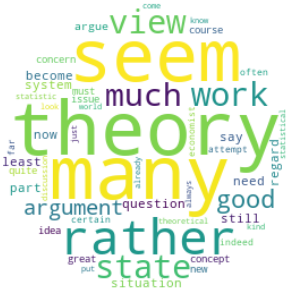

$DE^{GER}$ Topic 8


$DE^{ENG}$ Topic 33

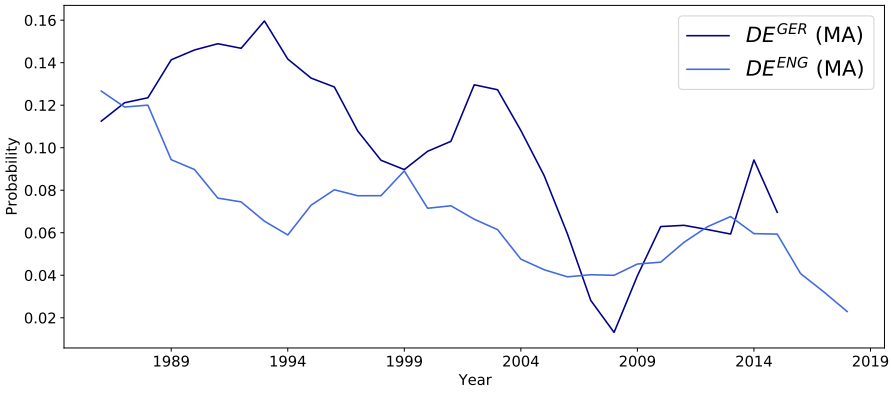



$DE^{GER}$ Topic 9

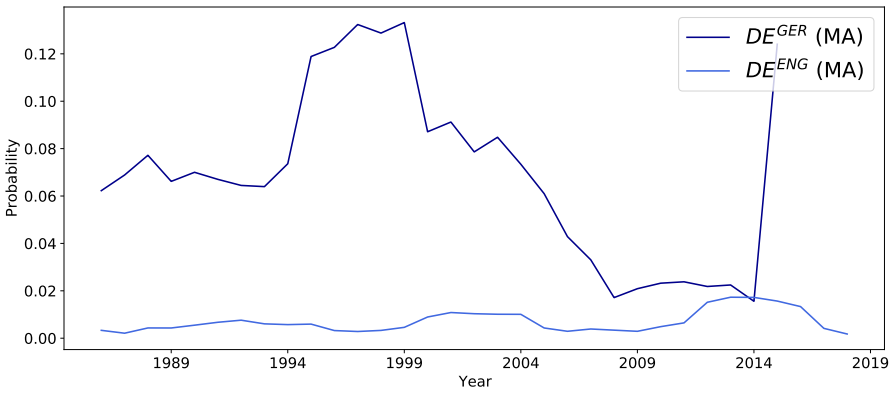


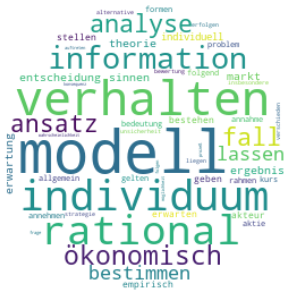

$DE^{ENG}$ Topic 54




$DE^{GER}$ Topic 10

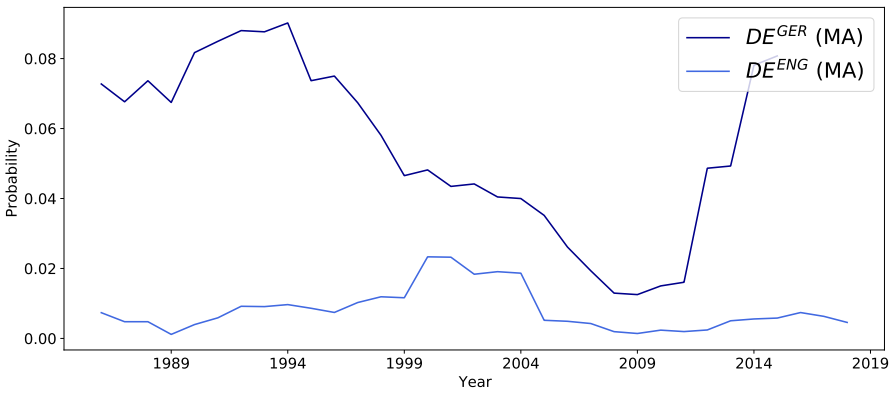

$DE^{ENG}$ Topic 8



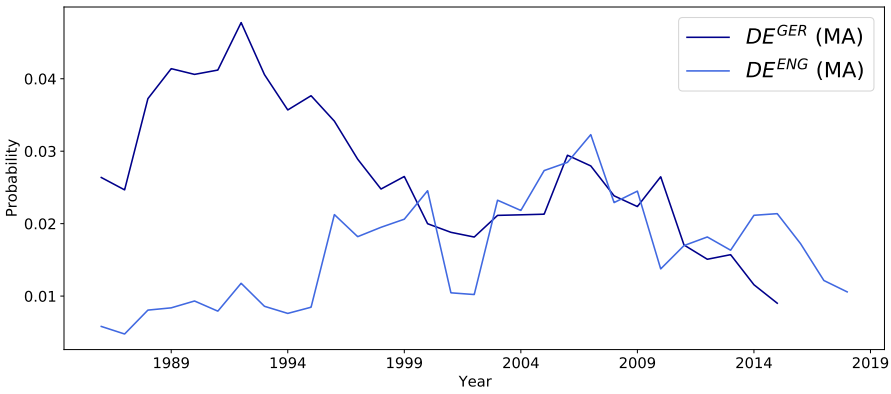

$DE^{GER}$ Topic 11


$DE^{ENG}$ Topic 51




$DE^{GER}$ Topic 13

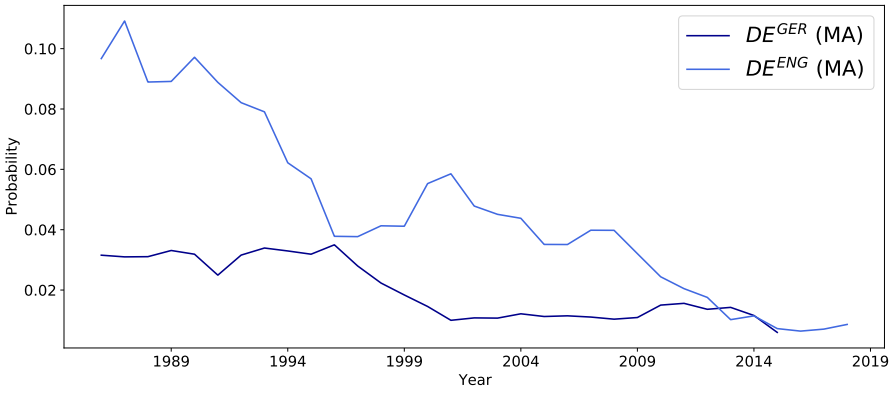

$DE^{ENG}$ Topic 29





$DE^{GER}$ Topic 14


$DE^{ENG}$ Topic 21

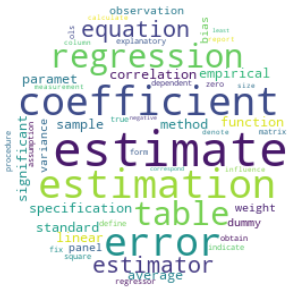



$DE^{GER}$ Topic 15

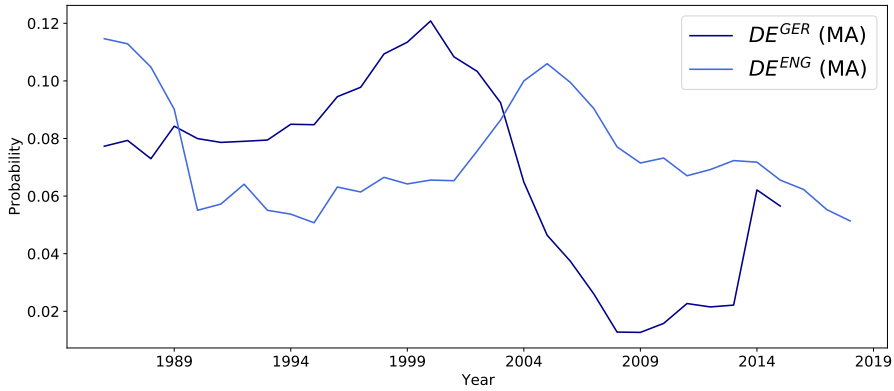

$DE^{ENG}$ Topic 50



$DE^{GER}$ Topic 17




$DE^{ENG}$ Topic 12



# Chapter 5

# Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 35%), Victor Bystrov (contribution: 30%), Anna Staszewska-Bystrova (contribution: 20%), Peter Winker (contribution: 15%) |
| **Status:** | Published: *Journal of Machine Learning Research*, vol. 25, no. 79, 2024, pp. 1-30 |
| **Available from:** | `http://jmlr.org/papers/v25/23-0188.html` |

Earlier versions of this paper were presented at:

- 24th International Conference on Computational Statistics (COMPSTAT 2022), Bologna, Italy (presented by Co-Author)
- 48th International Conference MACROMODELS, Wieliczka, Poland, 2022 (presented by Co-Author)
- 25th International Conference on Computational Statistics (COMPSTAT 2023), London, UK (presented by Co-Author)
- BERD@NFDI Research Symposium, Mannheim, Germany, 2024

# Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria*

Victor Bystrov† Viktoriia Naboka-Krell‡,‖
Anna Staszewska-Bystrova† Peter Winker‡

**Abstract.**

Selecting the number of topics in Latent Dirichlet Allocation (LDA) models is considered to be a difficult task, for which various approaches have been proposed. In this paper the performance of the recently developed singular Bayesian information criterion (sBIC) is evaluated and compared to the performance of alternative model selection criteria. The sBIC is a generalization of the standard BIC that can be applied to singular statistical models. The comparison is based on Monte Carlo simulations and carried out for several alternative settings, varying with respect to the number of topics, the number of documents and the size of documents in the corpora. Performance is measured using different criteria which take into account the correct number of topics, but also whether the relevant topics from the considered data generating processes (DGPs) are revealed. Practical recommendations for LDA model selection in applications are derived.

*Key Words: Topic models, text analysis, latent Dirichlet allocation, singular Bayesian information criterion, Monte Carlo simulation, text generation*

* Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

† University of Lodz, Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

‡ Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

‖ Corresponding author: Viktoriia.Naboka-Krell@wi.jlug.de

## 5.1 Introduction

Text data have been increasingly used in different applications lately. One of the main challenges in working with text data is to structure and to quantify these data. To this end, probabilistic topic modelling approaches are often applied, as they allow to uncover hidden structures behind text data. One of the best-known and widely used topic modelling approaches is Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). For some recent applications making use of this method, see, e.g., Lüdering & Winker (2016), Thorsrud (2020), Ellingsen et al. (2022), and Savin et al. (2022).

LDA is a generative model that builds on two basic assumptions. First, it is assumed that each document in a corpus represents a mixture of topics. The second assumption is that each topic is determined by a mixture of terms from the vocabulary. The number of these topics, or themes, is a parameter to be set by the researcher. Often this decision is based on human/expert judgment and is, therefore, rather subjective. In order to account for possible subjectivity and to allow for a more standardised estimation procedure, various evaluation metrics have been developed for identifying an optimal number of topics in LDA models. Some of them aim to minimize the similarity of different topics (Cao et al., 2009), maximize the topic coherence (Mimno et al., 2011) or maximize the goodness-of-fit between the estimated and the actual document-term frequencies (Lewis & Grossetti, 2022). These criteria, however, often result in (substantially) different numbers when applied to the same corpus. Their performance might also differ across corpora depending on the underlying data set (see examples in Section 5.2). Bystrov et al. (2022) propose to use a new measure for selecting an optimal number of topics, namely the singular Bayesian information criterion (sBIC). This information criterion reflects the trade-off between goodness-of-fit and model complexity and showed promising results in a first application.

There have been some attempts to compare selected criteria based on individual real datasets.[1] In this paper, comprehensive Monte Carlo (MC) simulations are proposed, which allow a systematic evaluation going beyond individual case reports by using a large number of datasets coming from well defined data generating processes (DGP) with known properties. Thereby, we consider three different data generating processes to reflect different types of text data commonly used in applications. In a first step, we generate corpora with a known (true) number of topics under the assumption that the underlying DGPs follow an LDA process (LDA based text generation). Afterwards, to each of the generated corpora in each of the considered DGPs, LDA models with different numbers of topics are fitted. Then, we apply several alternative criteria to select the number of topics and evaluate the performance of criteria over many MC replications. To the best of our knowledge, no such systematic and comprehensive comparison analysis of the criteria used for selecting the number of topics in LDA models has been performed yet.

[1] A notable exception including also a small scale Monte Carlo simulation is Lewis & Grossetti (2022).

The contribution of this paper is threefold. First, with the sBIC we implement a new measure for identifying the true number of topics in LDA models. Second, we perform proper Monte Carlo (MC) simulations to evaluate the proposed criterion as well as several alternatives commonly used in applications. Third, we evaluate the performance of studied criteria quantitatively and qualitatively, i.e., we consider whether the actual number of topics as well as the content and structure of the estimated topics are approximated well

The remainder of this paper is structured as follows. The considered model selection criteria are described in Section 5.2. Section 5.3 presents the design and the implementation details of the MC simulations. The results of the MC simulations for three different DGPs are presented in Section 5.4, which is divided in two subsections to address the main trade-off between *number* of topics and *coherence/structure* of the uncovered topics. The final section summarises the findings and provides recommendations for applications.

## 5.2 Model Selection Criteria for LDA

The selection of the optimal number of topics for LDA models can be based either on measures of topic quality (similarity or coherence) or on measures of goodness-of-fit and model complexity.

Let us consider an LDA model under a standard "bag-of-words" assumption. For a document corpus $\mathcal{D}$ that consists of $J$ documents, each document $j$ ($j = 1, 2, \ldots, J$) is a set of $N_j$ words, where the ordering of words is ignored. The total number of words in the corpus is equal to $N = \sum_{j=1}^{J} N_j$. The document corpus $\mathcal{D}$ can be characterized by a $J \times I$ document-term frequency matrix $X = \{x_{ji}\}_{j,i=1}^{J,I}$, where $x_{ji}$ is the frequency of term $i$ encountered in document $j$ and $I$ is the number of different terms in the vocabulary.

Under the "bag-of-words" assumption, an LDA model can be summarized by a $J \times K$ matrix $\theta$ of document-topic frequencies and a $K \times I$ matrix $\beta$ of topic-term frequencies with the dimensions of these matrices depending on the number of topics $K$. The estimated document-term matrix is a product of estimates $\hat{\theta}$ and $\hat{\beta}$: $\hat{X} = \hat{\theta} \times \hat{\beta}$. A set of candidate LDA models is determined by the numbers of topics in candidate models: $K \in \{K_{\min}, \ldots, K_{\max}\}$.

In the following, we describe two popular semantic measures of topic quality, which are often used in applications, and two recently developed goodness-of-fit measures.

### 5.2.1 Topic Similarity

Following Cao et al. (2009), the optimal number of topics is often selected by minimizing the average cosine similarity across topics:

$$\text{Cao\_Juan}(K) = \frac{\sum_{k=1}^{K} \sum_{l=k+1}^{K} \text{corr}(k, l)}{K \times (K-1)/2},$$

where

$$\text{corr}(k,l) = \frac{\sum_{i=1}^{I} \beta_{ki}\beta_{li}}{\sqrt{\sum_{i=1}^{I} \beta_{ki}^2}\sqrt{\sum_{i=1}^{I} \beta_{li}^2}},$$

and $\beta_{ki}$ is the frequency of term $i$ in topic $k$.

The average cosine similarity is extensively used for selecting the number of topics in different text-as-data applications, e.g. analyzing scientific articles to examine the evolution of research over time and identify future fields of research (Loureiro et al., 2021; Tiba et al., 2018), analyzing the speeches by Executive Board members of the European Central Bank (Hartmann & Smets, 2018), investigating news data in the context of economic reforms (Lin & Katada, 2022), analyzing and categorizing innovation projects (Dahlke et al., 2021).

### 5.2.2 Topic Coherence

Mimno et al. (2011) proposed a model selection procedure that maximizes the average semantic coherence of topics:

$$\text{Mimno}(K) = \frac{1}{K}\sum_{k=1}^{K} \text{coh}(k, \mathbf{i}^{(k)}),$$

where $\text{coh}(k, \mathbf{i}^{(k)})$ is the coherence metric for topic $k$,

$$\text{coh}(k, \mathbf{i}^{(k)}) = \frac{2}{v \times (v-1)} \sum_{m=2}^{v}\sum_{n=1}^{m-1} \log \frac{f(i_m^{(k)}, i_n^{(k)}) + \epsilon}{f(i_n^{(k)})},$$

$\mathbf{i}^{(k)} = (i_1^{(k)}, \ldots, i_v^{(k)})$ is the list of the $v$ most frequent terms in topic $k$, $f(i)$ is the document frequency of term $i$ (i.e., the number of documents with at least one token of type $i$), and $f(i, i')$ is the co-document frequency of terms $i$ and $i'$ (i.e., the number of documents containing one or more tokens of type $i$ and at least one token of type $i'$). The smoothing parameter $\epsilon$ is included to avoid taking the logarithm of zero and its default value is given by $e^{-12}$. The number of the most frequent terms, $v$, is set to the default value of 20.

The average semantic coherence is often used for selecting the number of topics in applied topic mining, e.g., for the analysis of monetary policy speeches (Ferrara et al., 2022), stock market news (Adämmer & Schüssler, 2020), tweets concerning the energy market (Polyzos & Wang, 2022), or survey responses on the consequences of the Covid-19 pandemic (Kleinberg et al., 2020).

### 5.2.3 OpTop Criterion

Lewis & Grossetti (2022) proposed to use a goodness-of-fit statistic based on the comparison of actual and estimated document-term frequencies. The frequency of term $i$ in document $j$ estimated in an LDA model with $K$ topics is

$$\hat{x}_{ji}^{(K)} = \sum_{k=1}^{K} \hat{\theta}_{jk}^{(K)} \hat{\beta}_{ki}^{(K)}.$$

Because the matrix of document-term frequencies is usually sparse, Lewis & Grossetti (2022) suggest collapsing relatively rare terms in a single frequency bin. For document $j$, they order terms from the smallest to the largest estimated frequency, $(i_1^{(j)}, i_2^{(j)}, \ldots, i_I^{(j)})$ such that $\hat{x}^{(K)}_{ji_1^{(j)}} \leq \hat{x}^{(K)}_{ji_2^{(j)}} \leq \ldots \leq \hat{x}^{(K)}_{ji_I^{(j)}}$, and select a sub-vector of relatively rare terms $(i_1^{(j)}, i_2^{(j)}, \ldots, i_p^{(j)})$. The cumulative frequency of relatively rare terms in document $j$ estimated in an LDA model with $K$ topics is

$$\hat{x}^{(K)}_{j,\min} = \sum_{i \in (i_1^{(j)}, \ldots, i_p^{(j)})} \hat{x}^{(K)}_{ji},$$

where $\hat{x}^{(K)}_{ji_p^{(j)}}$ is the largest frequency such that $\sum_{i=i_1^{(j)}}^{i_p^{(j)}} \hat{x}^{(K)}_{ji} < x_{\text{cutoff}}$, and $x_{\text{cutoff}}$ is a cumulative frequency cut-off value. Following Lewis and Grossetti (2022), we use $x_{\text{cutoff}} = 0.05$ as a baseline cut-off value. (For a robustness check, we also consider a cut-off value of 0.20). The actual cumulative frequency of relatively rare terms in document $j$ is

$$x_{j,\min} = \sum_{i \in (i_1^{(j)}, \ldots, i_p^{(j)})} x_{ji}.$$

The resulting goodness-of-fit statistic is

$$\text{OpTop}(K) = \sum_{j=1}^{J} \left[ (P_j + 1) \left( \sum_{i \in (i_{p+1}^{(j)}, \ldots, i_I^{(j)})} \frac{(\hat{x}^{(K)}_{ji} - x_{ji})^2}{\hat{x}^{(K)}_{ji}} + \frac{(\hat{x}^{(K)}_{j,\min} - x_{j,\min})^2}{\hat{x}^{(K)}_{j,\min}} \right) \right], \quad (5.1)$$

where $(i_{p+1}^{(j)}, \ldots, i_I^{(j)})$ is a sub-vector of relatively frequent terms in the $j$th document and $P_j$ is the length of this sub-vector. Lewis & Grossetti (2022) propose to select an optimal number of topics by minimizing the OpTop statistic (5.1). Unlike the criteria proposed by Cao et al. (2009) and Mimno et al. (2011), the OpTop statistic is not a semantic measure of topic quality, but a goodness-of-fit measure that can be easily computed.

### 5.2.4 Singular Bayesian Information Criterion

The last model selection criterion – the singular Bayesian information criterion – is a generalization of the Bayesian information criterion (BIC) that can be applied to singular statistical models (Drton & Plummer (2017)). The criterion was successfully used by Bystrov et al. (2022) for selecting parsimonious LDA models with coherent topics, however the properties of the criterion as applied to LDA modelling have not been studied in a simulation setup.

The standard BIC for an LDA model with $K$ topics is of the form

$$\text{BIC}(K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \frac{d_K}{2} \log(N), \quad (5.2)$$

where $P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K)$ is the value of the likelihood function for corpus $\mathcal{D}$ given the estimated matrices of document-topic and topic-term probabilities ($\hat{\theta}$ and $\hat{\beta}$), $d_K = J(K-1) + (I-1)K$

is the number of estimated parameters (model dimension), and $N$ is the total number of words in the corpus. The model dimension, $d_K$, is a linear function of the number of topics, $K$, and the term $\frac{d_K}{2}\log(N)$ in equation (5.2) is a penalty for increasing the number of parameters.

The general formula of the BIC was derived by Schwartz (1978) as a quadratic approximation for the logarithm of the marginal likelihood under the assumption of a regular (non-singular) model for which the Fisher information matrix is positive definite. For Latent Dirichlet Allocation (LDA) models the Fisher matrix is singular and the quadratic approximation of the log-marginal likelihood, which is used in the derivation of the standard BIC (5.2), is not possible.

The singular Bayesian information criterion (sBIC) can be derived using an approximation of the log-marginal likelihood described by Watanabe (2009). For an LDA model, this approximation can be written as

$$\log L(\mathcal{D}|K) \simeq \log P(\mathcal{D}|\hat{\theta},\hat{\beta},K) - \left[\lambda_{Kr}\log(N) - (m_{Kr}-1)\log\log(N)\right], \tag{5.3}$$

where $\lambda_{Kr}$ is a rational number in the interval $[0, d_K/2]$, $m_{Kr}$ is a natural number in the range $\{1, 2, \ldots, d_K\}$, and $r$ is an intrinsic value of the true distribution, $r = \text{rank}(\theta \times \beta)$, that depends on the true number of topics (see Hayashi (2021)). The term $[\lambda_{Kr}\log(N) - (m_{Kr}-1)\log\log(N)]$ is an approximation of the model complexity in LDA, which is determined by the number of non-redundant parameters in matrices of document-topic and topic-term probabilities. It is smaller than the penalty in the standard BIC and, moreover, as a sub-linear function of the number of topics, $K$, the term $[\lambda_{Kr}\log(N) - (m_{Kr}-1)\log\log(N)]$ grows slower than the penalty in the standard BIC (see Watanabe (2009) and Hayashi (2021)). Therefore, a criterion based on the approximation (5.3) selects an LDA model with more topics than the standard BIC (5.2).

The coefficients $\lambda_{Kr}$ and $m_{Kr}$ cannot be computed directly, because they depend on the true number of topics. This problem can be overcome by applying model averaging as proposed by Drton & Plummer (2017). In this approach, a feasible singular Bayesian information criterion for an LDA model with $K$ topics is defined as an approximation of the log-marginal likelihood obtained by the averaging of sub-models (models with smaller or equal number of topics). The feasible sBIC satisfies the equation

$$\text{sBIC}(K) = \log P(\mathcal{D}|\hat{\theta},\hat{\beta},K) - \log\left[\sum_{k\leq K}\omega_{Kk}N^{\lambda_{Kk}}(\log N)^{-(m_{Kk}-)}\right],$$

where the penalty term is the logarithm of the weighted average of $\left[N^{\lambda_{Kk}}(\log N)^{-(m_{Kk}-1)}\right]$ with coefficients $\lambda_{Kk}$ and $m_{Kk}$ depending on the number of topics in sub-models, $k \leq K$, and weights $\omega_{Kk}$ depending on the data. The computation of the feasible sBIC involves calculating $\lambda_{Kk}$ and $m_{Kk}$ for all $k \leq K$ as well as solving a system of quadratic equations. Therefore, full details of computing the feasible sBIC for an LDA model are provided in Appendix C.1.

## 5.3 Monte Carlo Simulations

Despite the broad usage of some metrics described in the previous section, there is not yet consensus on which metric performs best, when it comes to selecting the number of topics. Given that the data generating process (DGP) is unknown in applications, the relative performance of the metrics can only be assessed on the basis of a subjective analysis of the estimated topics. To account for this issue, we carry out a Monte Carlo (MC) simulation study, for which the data are generated by well-defined DGPs with known numbers of topics.[2] This allows us to compare the performance of alternative metrics with regard to the number of topics identified as well as to evaluate whether certain characteristics of the corpora such as number or length of documents might affect the relative performance. Furthermore, in the MC simulation study, we not only compare the selected number of topics to the number of topics in the DGP, but we also evaluate whether the estimated topics closely match the topics in the DGP.

This section provides the details of the Monte Carlo simulation setup used for the comparison of the methods described in Section 5.2. First, in Subsection 5.3.1 we present the general framework that is applied for each of three different DGPs. Second, in Subsection 5.3.2 we describe the DGPs, which are derived from actual corpora with typical characteristics of textual data used in applications. Finally, Subsection 5.3.3 provides some technical implementation details.

### 5.3.1 Procedure

The three DGPs used in the Monte Carlo simulations are designed to replicate the characteristics of a given real document corpus. Figure 5.1 presents the generic procedure which is applied to each of these DGPs.

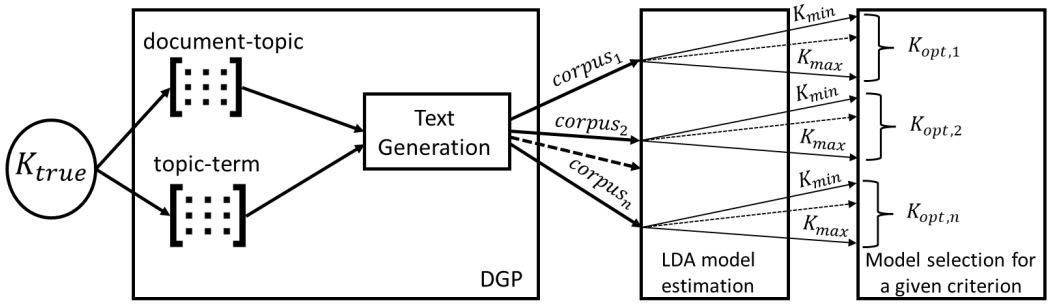



**Figure 5.1:** Generic procedure for Monte Carlo simulations with a given selection criterion

[2] The idea of using Monte Carlo simulations for obtaining well-defined text corpora has been applied recently by Wang et al. (2021) in the context of model selection for text classification tasks. The authors use the generated data to evaluate the classification performance of different topic models.

As described in Section 5.2, redLatent Dirichlet Allocation (LDA) is based on the assumption that each document in a corpus is a distribution over a given number $K_{\text{true}}$ of latent topics and each topic is a distribution over a fixed corpus vocabulary (Blei et al., 2003). Thus, an LDA model can be described by two matrices, the first containing the probabilities of occurrence of each term in each topic (topic-term distribution), and the second providing the probabilities of each topic occurring in a single document (document-topic distribution). These matrices are used to generate text corpora in the MC simulations.

In the first step of the procedure, an LDA model is estimated using a real document corpus with a number of topics that was used in previous analysis of the selected corpus. In order to make certain that only distinct topics will be used for the MC generation of text corpora in the following step, topics exhibiting a cosine similarity with other topics larger than a selected cut-off value (95% or 99% percentile) are dropped and the document-topic matrix is re-scaled to ensure that topic weights add up to one (see Appendix C.2 for more details). The data generating process based on distinct topics is intended to approximate a feature of the generative LDA model described by Blei et al. (2003) where topics are independently drawn from a Dirichlet distribution. Dealing with well-separated topics stabilizes the performance of the estimation methods developed for LDA.

In the second step, text corpora are generated using the estimated document-topic and topic-term distributions. The text generation process based on LDA is presented in Algorithm 5.1. For each document in the original corpus, a new Monte Carlo document is created with the same number of words and document-topic distribution. For each word in this document, a topic is randomly selected based on the known document-topic distribution and then a term is drawn from the vocabulary using the known topic-term distribution.

As mentioned above, Algorithm 5.1 does not exactly reproduce the generative procedure described in Blei et al. (2003) where rows of document-topic and topic-term frequency matrices are drawn from Dirichlet distributions. In applications, hyper-parameters of these distributions are not often estimated, and using exchangeable Dirichlet distributions in the generating process could result in document-topic and topic-term frequency matrices that structurally differ from frequency matrices estimated for actual text corpora. Therefore, we use a synthetic approach that, on the one hand, approximates the essential features of the generative LDA model and, on the other hand, replicates properties of frequency matrices estimated for real data.

Algorithm 5.1 is implemented in each DGP with 1 000 Monte Carlo replications, i.e., 1 000 corpora containing the same number of documents of same length as the original corpus.

In the third step of the procedure, we estimate LDA models with the number of topics ranging from $\max\{2; K_{\text{true}} - 20\}$ to $K_{\text{true}} + 20$, where $K_{\text{true}}$ is the number of topics in the data generating process. The maximum length of the range of admitted values for the number of topics is equal to 41 with the true number of topics, $K_{\text{true}}$, located in the center of the range if $K_{\text{true}} > 20$. Otherwise, the lower bound is set to 2, the lowest sensible number of topics. This limited range of admitted values for the number of topics is due to the high computational costs of model estimation. The optimal number of topics is determined for

**Algorithm 5.1** Text generation

1: **for** $document = 1, 2, \ldots, J$ **do**
2: $document\_length = original\_document\_length$
3: **for** $word = 1, 2, \ldots, document\_length$ **do**
4: Randomly select a topic from the document-topic distribution of the current document
5: Randomly select a term from the topic-term distribution
6: Append the selected term to the current document
7: **end for**
8: Append the generated document to the corpus.
9: **end for**

each of the selected criteria based on the estimated models.

In the final step, we compare the number of topics selected by different criteria using descriptive statistics such as standard deviation, mean, median, and skewness (see Subsection 5.4.1). For the visualisation of the distributions over the number of topics determined according to the considered criteria, we use histograms. Furthermore, in Subsection 5.4.2, we provide information about the extent to which the content of topics used for generating texts is revealed in the estimated LDA with the number of topics selected by the different criteria.

### 5.3.2 Data Generating Processes

The three data generating processes (DGPs) used for the Monte Carlo (MC) simulations are related to three actual text corpora:

- DGP 1 replicates the characteristics of a corpus consisting of scientific papers published in the Journal of Economics and Statistics (JES).
- DGP 2 reproduces features of the corpus consisting of abstracts submitted to European Research Consortium for Informatics and Mathematics (ERCIM) and Computational and Financial Econometrics (CFE) conferences.
- DGP 3 reproduces the properties of a corpus containing Newsticker items from heise online.

The data from JES used for DGP 1 cover the period from 1984 to 2020 and consists of 704 documents with an average text length of about 3,000 words. The size of the vocabulary for this corpus is equal to 3,911 terms. The collection focuses on scientific publications in empirical economics and applied statistics. The initial number of topics selected was equal to 60 as in Bystrov et al. (2022). After removing topics which were too similar, the final number of topics used in DGP 1 is equal to 38 ($K_{\text{true}} = 38$).

The conference abstract data used for DGP 2 cover the period from 2007 to 2019 and consists of 11,387 documents with an average text length of about 80 words. For this corpus

the vocabulary is composed of 1,796 terms. The focused nature of conference abstracts suggests a limited number of topics. The initial number of topics selected for this corpus was equal to 20. This number was reduced to 12 ($K_{\text{true}} = 12$) after removing the topics that were too similar.

The heise data used for DGP 3 cover the period from 1996 to 2021 and include 181,402 documents with an average length of about 120 words. The number of terms in the vocabulary for this corpus is equal to 4,675. The news platform discusses a significant number of topics concerning technological advances. The initial number of topics selected was equal to 120. After removing the most similar topics, the final number of topics used in DGP 3 was equal to 70 ($K_{\text{true}} = 70$). In the analysis we used only the most recent 50,000 documents from this corpus because using the whole dataset would increase the computational costs of MC simulations beyond the available capacities.

At this point, we would like to emphasize that in the described experiments, texts are generated using an LDA model with distinct topics. It means that each generated text is a "bag-of-words", where semantic and syntactic relationships between words, observed in actual texts, are neglected. However, it allows us to create a controlled setting for text generation as well as for evaluating model selection criteria. The results of applying the considered criteria to actual corpora, which do not emerge from the generative LDA model, may therefore differ from those presented in this study. Nevertheless, the results of the described experiments provide insights into the usability of the model selection criteria in settings when LDA constitutes a reasonable approximation to the actual DGP.

### 5.3.3 Details of Implementation

All Monte Carlo simulations were implemented using Python. To generate random sequences used in the text generation stage (Algorithm 5.1), the random number generator from Pythons' numpy package was used (`https://numpy.org/doc/stable/reference/random/generator.html`).

LDA models were estimated using the Gibbs sampler as implemented in the Python package "lda" (`https://pypi.org/project/lda/`). The number of iterations was set to a relatively small value of 1 000 due to computational constraints. Most other parameters of the package were used at the default values. The point estimates of document-topic and topic-term frequency matrices are computed as in Griffiths & Steyvers (2004).

For DGP 1, the numbers of topics in the estimated models ranged from 18 to 58; for DGP, the number of topics ranged from 2 to 32; and for DGP 3 - from 50 to 90.

The average topic similarity (Cao_Juan) and the average semantic coherence (Mimno) criteria were computed using the Python package "tmtoolkit" (`https://pypi.org/project/tmtoolkit/`). The Python implementations of the singular Bayesian information criterion (sBIC) and the goodness-of-fit statistic (OpTop) model selection criteria were written by the authors.

For high-precision computations in the implementation of sBIC we used the Python

module "decimal" and wrote a procedure that augments precision if it is necessary. The outer limits allowable for exponents of floating-point numbers have to be sufficiently large in order to avoid exponent underflow and overflow in the computation of sBIC which is a solution of a recursive system of quadratic equations parameterized by likelihood values in sub-models (see Appendix C.1). Compared to the estimation time of LDA models, the additional time needed for high-precision computations in the sBIC algorithm is not substantial.

Computations were performed using the high-performance-computing-cluster at Justus Liebig University Giessen (justHPC) (`https://www.hkhlr.de/de/cluster/justhpc-giessen`).[3]

## 5.4 Results

This section summarizes the results of the Monte Carlo simulations. It is divided into two subsections. Subsection 5.4.1 presents and discusses the results of estimating the optimal number of topics and subsection 5.4.2 evaluates the structure and contents of the estimated topics.

### 5.4.1 Number of Topics

The first set of results concerns the estimation of the number of topics, $K$. Figures 5.2-5.4 present histograms of the numbers of topics selected by different criteria for the three considered data generating processes (DGPs). In each of the histograms, the red vertical line depicts the true number of topics, $K_{\text{true}}$, used for generating the corpora. The shape and location of histograms shown in Figures 5.2-5.4 suggest that the sBIC is clearly the best method for selecting the number of topics for DGP 1 and DGP 2, while it performs similarly to the method of Cao et al. (2009) for DGP 3.

[3] Code details can be found in the Github repository for this paper at `https://github.com/VikaNa/sBIC`.

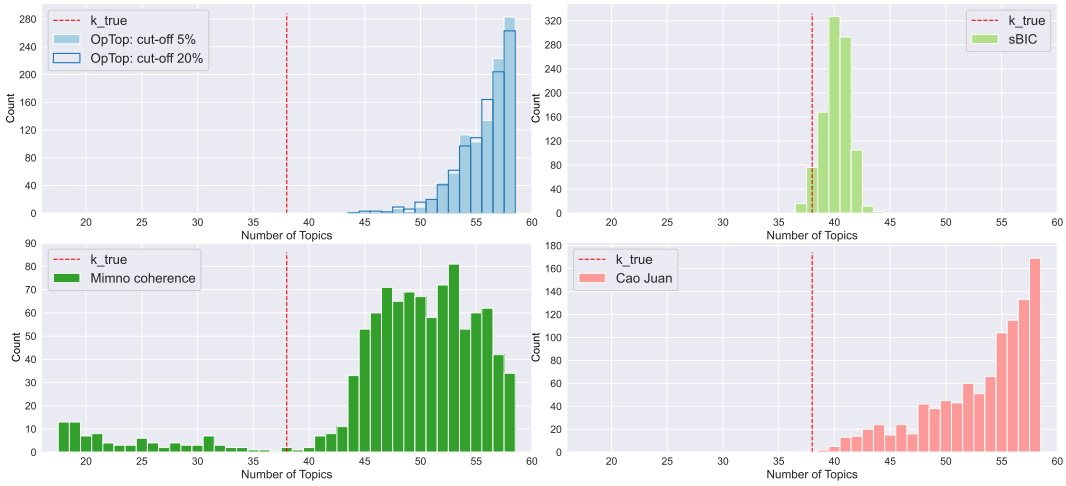



**Figure 5.2:** Comparison of evaluation metrics for DGP1

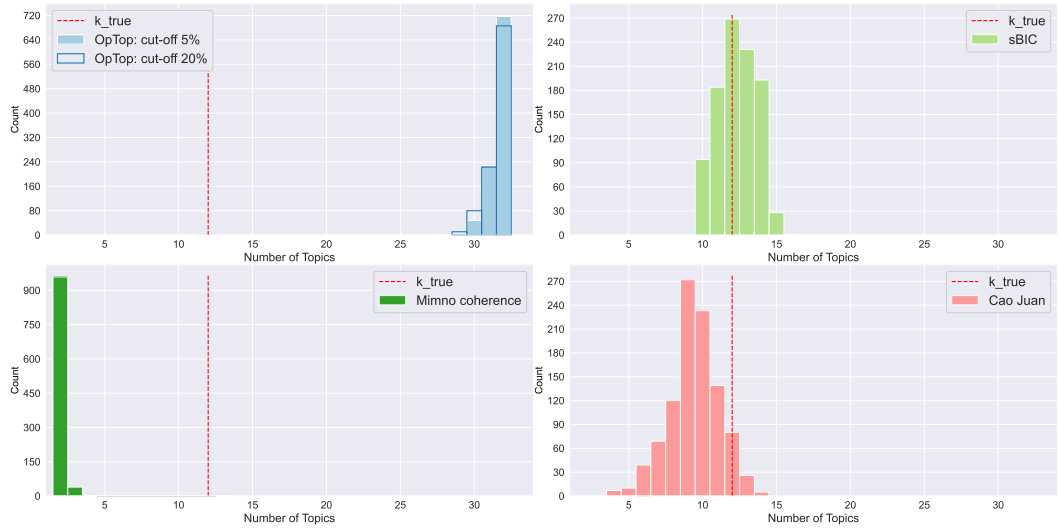



**Figure 5.3:** Comparison of evaluation metrics for DGP2

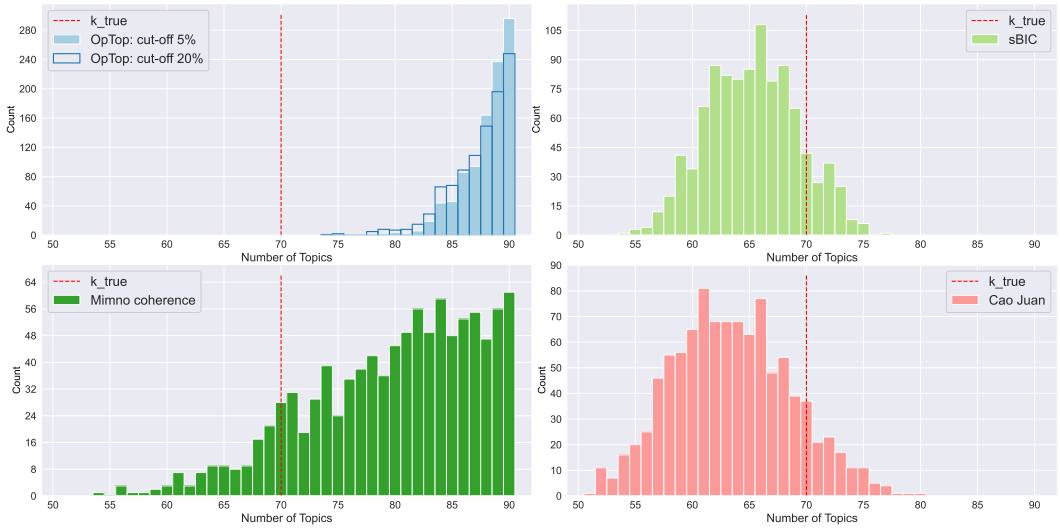



**Figure 5.4:** Comparison of evaluation metrics for DGP3

| | | DGP1 ($K_{\text{true}} = 38$) | DGP2 ($K_{\text{true}} = 12$) | DGP3 ($K_{\text{true}} = 70$) |
|---|---|---|---|---|
| sBIC | std | 1.23 | 1.35 | 4.09 |
| | mean | 40.15 | 12.28 | 65.43 |
| | median | 40.00 | 12.00 | 66.00 |
| | skewness | -0.24 | 0.00 | -0.04 |
| Cao_Juan | std | 4.54 | 1.68 | 5.36 |
| | mean | 53.40 | 9.43 | 63.53 |
| | median | 55.00 | 9.00 | 64.00 |
| | skewness | -1.16 | -0.28 | 0.12 |
| Mimno | std | 9.23 | 0.20 | 7.55 |
| | mean | 47.88 | 2.04 | 79.50 |
| | median | 50.00 | 2.00 | 81.00 |
| | skewness | -1.89 | 5.55 | -0.61 |
| OpTop 5% | std | 2.30 | 0.59 | 2.06 |
| | mean | 55.81 | 31.70 | 88.03 |
| | median | 57.00 | 32.00 | 89.00 |
| | skewness | -1.22 | -2.17 | -1.28 |
| OpTop 20% | std | 2.39 | 0.67 | 2.61 |
| | mean | 55.67 | 31.63 | 87.38 |
| | median | 56.00 | 32.00 | 88.00 |
| | skewness | -1.37 | -1.78 | -1.30 |

**Table 5.1:** Evaluation of different criteria

Table 5.1 provides descriptive statistics computed for the number of topics selected by each criterion. The results in Table 5.1 demonstrate that the mean and the median of the number of topics estimated by the sBIC is the closest to the true value for all DGPs. For DGP 2 the median of the estimates provided by the sBIC is the actual number of topics. The performance of the criterion differs for DGP 1 and DGP 3. In the first case, the sBIC tends to select too many topics and in the second case, on average, it selects too few topics. The differences between the true and the estimated values are relatively small, but both types of estimation errors have their consequences. Overestimation of the number of topics means that some spurious topics will be generated, while underestimation implies that relevant topics will be omitted. These issues are further discussed in Section 5.4.2 where the structure and content of the estimated topics is evaluated.

The performance of the OpTop statistic (goodness-of-fit) is rather poor as it has a strong

tendency to select too many topics for each DGP. This result is robust with respect to the choice of the cut-off value for low-frequency terms (5% or 20%). In each case, the mean and median values of the estimates are very close to the maximum of the range of candidates for the optimal number of topics. Such large overestimation errors mean that a substantial number of irrelevant topics would be estimated. Since, as noted by Mimno et al. (2011), there is a trade-off between obtaining many refined and meaningful topics, the quality of these additional topics found by the OpTop method might be expected to be rather low.

| | | $K_{\text{true}} - k \leq K_{\text{metric}} \leq K_{\text{true}} + k$ | | | | | |
|---|---|---|---|---|---|---|---|
| DGP | Metric | $k = 0$ | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ |
| DGP 1 | Cao_Juan | 0.1 | 0.3 | 0.8 | 2.1 | 3.5 | 5.5 |
| | Mimno | 0.2 | 0.3 | 0.6 | 1.4 | 2.4 | 3.7 |
| | OpTop 20% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | OpTop 5% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | sBIC | 7.6 | 26.0 | 58.7 | 88.0 | 98.5 | 99.7 |
| DGP 2 | Cao_Juan | 8.0 | 24.5 | 48.3 | 75.5 | 87.5 | 94.4 |
| | Mimno | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| | OpTop 20% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | OpTop 5% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | sBIC | 26.9 | 68.4 | 97.1 | 99.9 | 100.0 | 100.0 |
| DGP 3 | Cao_Juan | 3.7 | 9.7 | 17.4 | 23.9 | 32.7 | 40.1 |
| | Mimno | 2.8 | 8.0 | 11.6 | 15.4 | 20.1 | 23.4 |
| | OpTop 20% | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 |
| | OpTop 5% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | sBIC | 4.2 | 13.4 | 25.8 | 36.2 | 47.8 | 56.9 |

**Table 5.2:** Percentages of the estimated number of topics, $K_{\text{metric}}$, falling within intervals around the true number of topics, $K_{\text{true}}$

The performance of the average cosine similarity (Cao_Juan) varies depending on the DGP. The mean/median number of topics selected for DGP 1 is too large as compared to the true number of topics, while the mean/median number of topics selected for DGPs 2 and 3 is too low as compared to the true number of topics. This outcome might depend on particular features of the DGPs (e.g. DGP 1 including a relatively small number of longer documents) which could be subject to further analyses. On the whole, the estimation errors are larger than for the sBIC and smaller than in case of the OpTop criterion.

The unsystematic behaviour in terms of the tendency to over- or underestimate can be also seen for the average semantic coherence (Mimno). The mean/median number of topics

selected for DGP 1 and DGP 3 is too large as compared to the true number of topics, while there is severe underestimation problem for DGP 2. The performance of this procedure seems to be quite unstable as the estimates have the largest variance compared to the remaining methods for DGP 1 and DGP 3.

Table 5.2 reports the percentages of the number of topics, estimated by each criterion, falling within symmetric intervals centered at the true number of topics, $[K_{\text{true}} - k, K_{\text{true}} + k]$, $k \in \{0, 1, 2, 3, 4, 5\}$. The sBIC clearly outperforms other criteria for all DGPs. This result holds for all considered intervals. For example, in 88% of the cases sBIC delivers a topic number between 35 and 41 for DGP 1 ($K_{\text{true}} = 38$). In nearly 60% of cases sBIc proposes a number of topics between 65 and 75 for DGP 3 ($K_{\text{true}} = 70$) as opposed to 40% by Cao_Juan and 23% by Mimno.

### 5.4.2 Structure and Content of Topics

While the selected *number* of topics delivers first general insights on the performance of different criteria, this indicator does not contain information on the correspondence between the topics used to generate the text corpora and the topics obtained using the selected number of topics in the estimation procedure. Therefore, the structure and the *content* of topics should be also evaluated.[4] To this end, we propose to consider the problem as a classification task. This allows us to compare the results obtained using all the different selection criteria quantitatively making the use of well established performance measures, precision and recall. In standard applications, these are defined as follows:

- **Recall** describes how many relevant items are retrieved.
- **Precision** indicates how many retrieved items are relevant.

In standard classification tasks, the length of predicted and actual labels is the same. In our case it might be different, as the number of topics selected by each of the considered criteria can deviate from the true number of topics as described in the previous subsection. Thus, we define the True Positive (TP) class as those topics that were correctly identified, i.e., true topics which find their match in the set of estimated topics for the number of topics indicated by the given selection criterion. Using this definition, precision and recall can be defined and calculated as follows:

$$\text{Recall} = \frac{|\text{TP}|}{K_{\text{true}}}, \tag{5.4}$$

where $|\text{TP}|$ denotes the cardinality of the set TP and $K_{\text{true}}$ is the true number of topics in a particular DGP.

[4] In applications, sometimes the quality of topics is analyzed based on human judgment. For example, Morstatter & Liu (2018) present an approach based on existing measures of topic coherence and extending them by a measure of topic consensus by humans. Although this approach delivers some measure of interpretability by humans, the authors point out the need for automated and reproducible measures of topic quality.

$$\text{Precision} = \frac{|\text{TP}|}{K_{\text{metric}}}, \tag{5.5}$$

where $K_{\text{metric}}$ is the proposed number of topics according to the selection criterion considered.

As there might be a trade-off between recall and precision, the F1 measure is often used as a combined measure. F1 is calculated as follows:

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5.6}$$

For computing these measures, estimated topics have to be matched with true topics from the data generating process (DGP). This matching can be done using the topic matching technique proposed by Bystrov et al. (2022), the so-called *best matching*. Thereby, each topic is represented as a probability vector over the vocabulary. For each "true" topic, a match in the set of estimated topics is identified using the cosine similarity measure. Cosine similarity is often used in natural language processing to measure similarity between high-dimensional text representations. For each topic of the "true" topic set, cosine similarities to all the topics from the estimated topic set are calculated. Initially, a topic pair with the highest cosine similarity value is considered a "best match". Obviously, a "best match" does not have to be a sensible match, i.e., close to the true topic. Therefore, we apply a threshold for the cosine similarity which has to be surpassed in order to consider a match as being a sensible match. This threshold is the same as used for the topic number reduction step for each DGP described in Subsection 5.3.2 (see Appendix C.2 for further details). If one of the "true" topics finds several sensible matches, we only consider the matches with the highest cosine similarities. The number of identified matches corresponds to "true positives", i.e. the number of correctly identified topics. This number of reproduced topics is then divided either by the true number of topic $K_{\text{true}}$ (recall) or the estimated number of topics $K_{\text{metric}}$ (precision).

Table 5.3 describes the distribution of precision and recall for each DGP and each evaluation metric.[5] As mentioned before, our application differs from standard classification problems as the number of true and estimated topics might differ. Hence, the interpretation of the results is slightly different.

A precision value of 1 means that all of the estimated topics are sensible matches to some of the true topics. However, it does not imply that all of the true topics are uncovered. Consequently, this measure might overestimate the performance of a metric if it tends to underestimate the true number of topics. For example, in case of DGP 2, the average precision of topics selected by the Mimno criterion (average semantic coherence) is equal to 1, while the average recall is equal to 0.17. In the previous subsection, it was shown that the Mimno metric tends to underestimate the true number of topics for DGP 2. Thus, the high precision value only indicates that these few estimated topics are related to the true

[5] As a robustness check we also calculate the described performance metrics using cosine similarities instead of the binary indicator match/no match. The procedure is described in Appendix C.4. The results do not differ qualitatively.

| | | Recall | | Precision | | F1 | |
|---|---|---|---|---|---|---|---|
| data | metric | mean | std | mean | std | mean | std |
| DGP1 | Cao_Juan | 1.00 | 0.00 | 0.72 | 0.07 | 0.83 | 0.04 |
| | Mimno | 0.96 | 0.12 | 0.78 | 0.09 | 0.85 | 0.06 |
| | OpTop 20% | 1.00 | 0.00 | 0.68 | 0.03 | 0.81 | 0.02 |
| | OpTop 5% | 1.00 | 0.00 | 0.68 | 0.03 | 0.81 | 0.02 |
| | sBIC | 0.99 | 0.01 | 0.94 | 0.03 | 0.97 | 0.01 |
| DGP2 | Cao_Juan | 0.78 | 0.13 | 1.00 | 0.02 | 0.87 | 0.09 |
| | Mimno | 0.17 | 0.02 | 1.00 | 0.00 | 0.29 | 0.02 |
| | OpTop 20% | 1.00 | 0.00 | 0.38 | 0.01 | 0.55 | 0.01 |
| | OpTop 5% | 1.00 | 0.00 | 0.38 | 0.01 | 0.55 | 0.01 |
| | sBIC | 0.93 | 0.06 | 0.92 | 0.07 | 0.92 | 0.04 |
| DGP3 | Cao_Juan | 0.87 | 0.05 | 0.96 | 0.03 | 0.91 | 0.02 |
| | Mimno | 0.93 | 0.04 | 0.83 | 0.05 | 0.87 | 0.02 |
| | OpTop 20% | 0.98 | 0.01 | 0.79 | 0.03 | 0.87 | 0.02 |
| | OpTop 5% | 0.98 | 0.01 | 0.78 | 0.02 | 0.87 | 0.02 |
| | sBIC | 0.88 | 0.04 | 0.95 | 0.03 | 0.91 | 0.02 |

**Table 5.3:** Descriptive statistics of recall, precision, and F1 scores

topics. On the other hand, for the sBIC, which performs very well in case of DGP 2, there are relatively high values of both recall and precision (0.93 and 0.92, respectively) indicating that mostly true topics and most of the true topics are recovered.

A recall value of 1 means that all of the true topics are uncovered by the estimated topics. However, it does not imply that $K_{\text{metric}} = K_{\text{true}}$. Consequently, this measure might lead to overestimation of the performance of a metric if it tends to select too many topics. For DGP 1, for example, the Cao_Juan metric (average cosine similarity) reveals an average recall value of 1, while the average precision value of 0.72 is substantially lower. Also in this example, sBIC performs well with average recall and precision values of 0.99 and 0.94, respectively.

To take account of the trade-off described above, it seems appropriate to combine precision and recall measures. This is done by using the F1 score which is defined in equation 5.6 a the harmonic mean of precision and recall. The interpretation of F1 is straightforward: the higher the values the better the joint score of precision and recall. The results indicate that the sBIC outperforms the other evaluation metrics for DGP 1 and DGP 2. For DGP 3, according to the F1 score the sBIC is found to perform similarly to the Cao_Juan criterion, while still exhibiting some advantages compared to the other criteria.

## 5.5 Conclusions and Outlook

Estimating Latent Dirichlet Allocation (LDA) models requires making a number of decisions regarding parameter settings. This paper considered the problem of selecting the value of one of those essential parameters, viz. the number of topics discussed in the text corpus. The main aim was to analyze the performance of various model selection criteria with special focus on the recently proposed singular Bayesian information criterion. The performance of the methods was examined via Monte Carlo experiments using synthetic data generating processes (DGPs) based on empirical text corpora which differed with respect to the number and length of documents and the number of topics. This text generation process was based on the assumption that the considered DGPs actually follow an LDA process (or could be approximated by an LDA process). The generalizability of the results is therefore limited. The performance of different model selection procedures was evaluated by not only examining the accuracy of estimating the actual number of topics but also by analyzing the structure and contents of the estimated topics.

Simulation results showed that the singular Bayesian information criterion (sBIC) performed relatively well for all data generating processes considered in the experiments. It was the best method for estimating the number of topics as it was associated with the smallest estimation errors as compared to the competitors. In addition, it resulted in topics with good content and structure and performed in a relatively stable fashion for all data generating processes. Across the DGPs, the performance of the sBIC was worst for DGP 3 corresponding to a text corpus with a large number of short documents and a substantial number of topics. In this setting, sBIC exhibited a certain downward bias in the selected number of topics which might be taken into account in applied work. The reasons for this finding and possible adjustments to the method might be subject to further analyses.

The performance of the methods proposed by Cao et al. (2009) (the average cosine similarity) and Mimno et al. (2011) (the average semantic coherence) depended on the DGP. For each of these methods, the experiments revealed cases of systematic under- or overestimation of the true number of topics. The estimation errors were larger than those found for the sBIC and had some negative consequences for the structure and content of the estimated topics. Dependence on the DGP implies that reliability and stability of these methods cannot be guaranteed in applied work unless further analyses will explain the relation between features of a DGP and the model selection results. Despite these drawbacks, the method of Cao et al. (2009) was still overall the second best approach to LDA model selection in the experiments reported in this paper. It was found that the method could be particularly useful for modelling collections of many short texts related to a large range of topics.

The final set of conclusions relates to the OpTop criterion (the goodness-of-fit statistic). It was shown that the method tends to select models with an excessively large number of topics. The estimation errors were very substantial and led to small precision and F1 metric

values used for examining the content and structure of estimated topics. These results imply that using this criterion in applied work can result in obtaining some spurious topics, which do not correspond to the data generating process. It seems that poor estimation properties of the OpTop procedure could be improved by the introduction of an appropriate penalty for model complexity (which increases with the number of topics) into the test statistic formula. This adjustment constitutes a direction of future research.

# Appendix C

## C.1 Computation of the singular Bayesian Information Criterion (sBIC)

The marginal likelihood of a corpus $\mathcal{D}$ composed of $J$ documents with a vocabulary including $I$ terms given an LDA model with $K$ topics is defined as

$$L(\mathcal{D}|K) = \int_{\theta,\beta} P(\mathcal{D}|\theta,\beta,K)dP(\theta,\beta|K), \tag{C.1.1}$$

where $P(\mathcal{D}|\theta,\beta,K)$ is the value of the likelihood function for the corpus $\mathcal{D}$ given $(J \times K)$ matrix of document-topic probabilities $\theta$ and $(K \times I)$ matrix of topic-term probabilities $\beta$, and $P(\theta,\beta|K)$ is a prior distribution of matrices $\theta$ and $\beta$ in the LDA model with $K$ topics.

An approximation of the marginal likelihood (C.1.1), based on the averaging of sub-models with number of topics $k = K_{\min},\ldots,K$, is defined as (see Drton & Plummer (2017))

$$L'(\mathcal{D}|K) = \frac{\sum_{k=K_{\min}}^{K} L'_{Kk}L(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L(\mathcal{D}|k)P(k)}, \tag{C.1.2}$$

where $P(k)$ is a prior for a model with $k$ topics (assumed to be a known positive constant). The term $L'_{Kk}$ is

$$L'_{Kk} = P(\mathcal{D}|\hat{\theta},\hat{\beta},K)N^{-\lambda_{Kk}}(\log N)^{m_{Kk}-1}, \tag{C.1.3}$$

where $P(\mathcal{D}|\hat{\theta},\hat{\beta},K)$ is the value of the likelihood function given the estimated parameter matrices $\hat{\theta}$ and $\hat{\beta}$ in the LDA model with $K$ topics, and coefficients $\lambda_{Kk}$, $m_{Kk}$ are computed for $k = K_{\min},\ldots,K$ using formulas from Hayashi (2021) where the rank of the matrix product $\theta \times \beta$ in true distribution, $r$, is replaced by the number of topics in a sub-model, $k$:

1. If $J + k \leq I + K$, $I + k \leq J + K$, $K + k \leq I + J$ and
   (a) if $I + J + K + k - 1$ is odd, then
   $\lambda_{Kk} = \frac{1}{8}\left\{2(K+k)(I+J) - (I-J)^2 - (K+k)^2\right\} - \frac{1}{2}J$ and $m_{Kk} = 1$
   (b) if $I + J + K + k - 1$ is even, then
   $\lambda_{Kk} = \frac{1}{8}\left\{2(K+k)(I+J) - (I-J)^2 - (K+k)^2 + 1\right\} - \frac{1}{2}J$ and $m_{Kk} = 2$
2. Else if $I + K < J + k$, then $\lambda_{Kk} = \frac{1}{2}\{IK + Jk - Kk - J\}$, $m_{Kk} = 1$
3. Else if $J + K < I + k$, then $\lambda_{Kk} = \frac{1}{2}\{JK + Ik - Kk - J\}$, $m_{Kk} = 1$

4. Else (i.e. $I + J < K + k$), then $\lambda_{Kk} = \frac{1}{2}(IJ - J)$, $m_{Kk} = 1$.

In equation (C.1.2) the approximation of the marginal likelihood, $L'(\mathcal{D}|K)$, is expressed as a function of the actual marginal likelihoods $L(\mathcal{D}|k)$, $k = K_{\min}, \ldots, K$. Drton & Plummer (2017) resolve this problem by replacing the unknown marginal likelihoods $L(\mathcal{D}|k)$ on the right-hand side of (C.1.2) by their approximations, $L'(\mathcal{D}|k)$, and considering a system of equations

$$L'(\mathcal{D}|K) = \sum_{k=K_{\min}}^{K} \frac{L'(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L'(\mathcal{D}|k)P(k)} L'_{Kk}, K = K_{\min}, \ldots, K_{\max}, \tag{C.1.4}$$

where $L'_{Kk}$ and $P(k)$ are known constants and $L'(\mathcal{D}|K)$ are unknowns to be found. Then the singular Bayesian information criterion for a model with $K$ topics is defined as

$$\text{sBIC}(K) = \log L'(\mathcal{D}|K), \tag{C.1.5}$$

where $L'(\mathcal{D}|K)$ is the unique solution of the transformed equation system (assuming that $P(K) > 0$ for $K = K_{\min}, \ldots, K_{\max}$)

$$\sum_{k=K_{\min}}^{K} [L'(\mathcal{D}|K) - L'_{Kk}]L'(\mathcal{D}|k) = 0,\ K = K_{\min}, \ldots, K_{\max} \tag{C.1.6}$$

that can be found inductively with $L'(\mathcal{D}|K_{\min}) = L'_{K_{\min}K_{\min}} > 0$ for the minimal model. Proceeding by induction, if $L'(\mathcal{D}|k)$ have been computed for all $k = K_{\min}, \ldots, (K-1)$, then $L'(\mathcal{D}|K)$ is the unique positive solution of quadratic equation

$$L'(\mathcal{D}|K)^2 + b_K L'(\mathcal{D}|K) - c_K = 0, \tag{C.1.7}$$

with

$$b_K = -L'_{KK} + \sum_{k=K_{\min}}^{K-1} L'(\mathcal{D}|k)\frac{P(k)}{P(K)} \text{ and } c_K = \sum_{k=K_{\min}}^{K-1} L'_{Kk}L'(\mathcal{D}|k)\frac{P(k)}{P(K)}.$$

Since $c_K > 0$ by induction, the quadratic equation (C.1.7) has the unique positive solution

$$L'(\mathcal{D}|K) = \frac{1}{2}\left(-b_K + \sqrt{b_K^2 + 4c_K}\right)$$

for $K = K_{\min}+1, \ldots, K_{\max}$. Given formulas (C.1.3), (C.1.4) and (C.1.5), sBIC should satisfy

$$\text{sBIC}(K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \log\left[\sum_{k=K_{\min}}^{K} \omega_{Kk} N^{\lambda_{Kk}} (\log N)^{-(m_{Kk}-1)}\right]$$

where weights $\omega_{Kk}$ are defined as

$$\omega_{Kk} = \frac{L'(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L'(\mathcal{D}|k)P(k)},\ k = K_{\min}, \ldots, K.$$

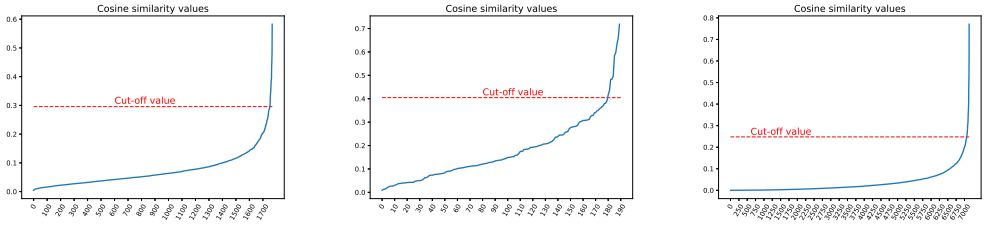
Because the coefficient $\lambda_{Kk}$ is less than half the model dimension, $\frac{1}{2}[J(K-1) + (I-1)K]$, for every $k = K_{\min}, \ldots, K$, the penalty in the singular BIC is less than in the standard BIC for an LDA model with the same number of topics. Moreover, the penalty for increasing the number of topics in the singular BIC grows slower as compared to the standard BIC.

## C.2 Topic Number Reduction

The goal of the topic number reduction step in preparing our DGPs for the Monte Carlo simulations was to use well separated topics allowing for a robust comparison of the topics estimated with the underlying DGPs. The process of topic number reduction comprises the following three steps:

1. Starting with the estimated LDA for a given corpus, for each topic the most similar other topic is identified using the standard matching proposed by Bystrov et al. (2022).

2. For deciding whether a pair of topics is "too similar", i.e., will be excluded before generating synthetic data within the Monte Carlo simulation, a threshold value has to be defined. This value is also obtained by a data driven approach. We calculate all pairwise cosine similarity scores for each DGP providing $\frac{K^2-K}{2}$ typical values. Sorting them in increasing order provides the distributions shown in Figure C.2.1. Following the approach of the "elbow" criterion, we set percentile values defining the cut-off value for each DGP. These values are shown in the figure by the red horizontal line and correspond to the 95% percentile for DGP2 and to the 99% percentile for DGPs 1 and 3, respectively.




**(a)** DGP1 **(b)** DGP2 **(c)** DGP3

**Figure C.2.1:** Distribution of the pairwise cosine similarity values.

3. All topics belonging to matched topic pairs above the cut-off value are considered as being too similar and, consequently, are removed from the model before starting the data generation within the Monte Carlo simulation. Figures C.2.2, C.2.3, and C.2.4 show examples of pairs including redundant topics in each DGP, which are eliminated by this method.

## C.3 Recall and Precision

Figures C.3.5, C.3.6, and C.3.7 exhibit the scatter plots of recall and precision values for each DGP separately. Thereby, each point corresponds to one of the simulated corpora. Consequently, there is a total of 300 points in each plot. However, the evaluation metrics considered may result in the same recall and precision scores for multiple corpora. Thus, some points may overlap.

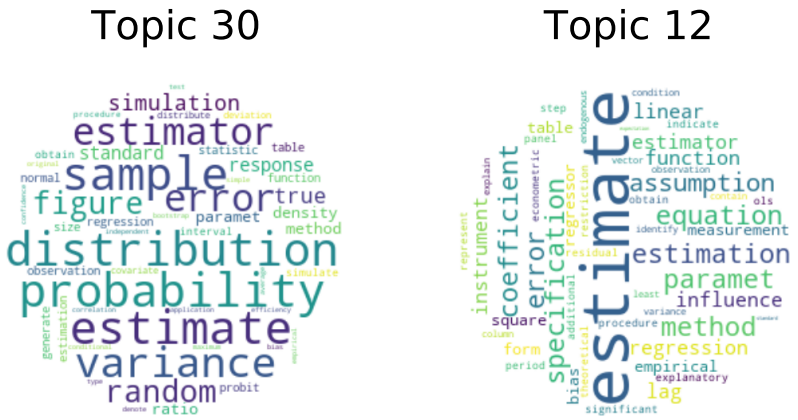



**Figure C.2.2:** Similar topics in DGP 1

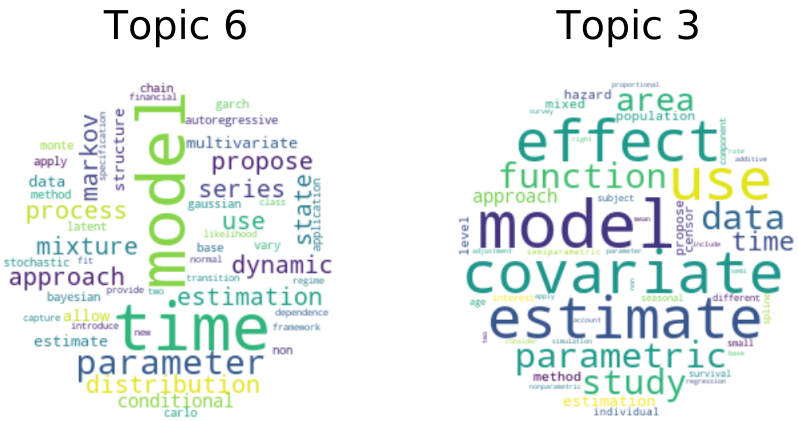



**Figure C.2.3:** Similar topics in DGP 2

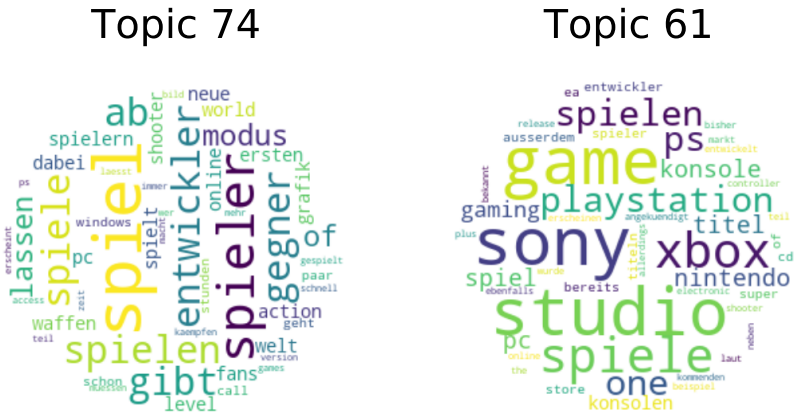



**Figure C.2.4:** Similar topics in DGP 3

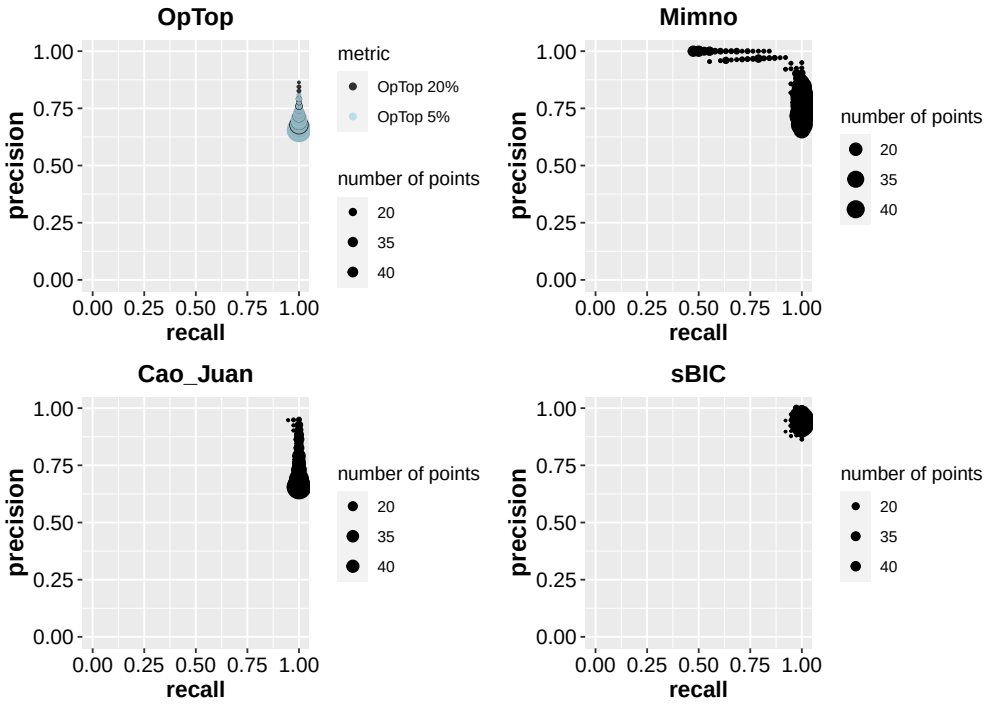



**Figure C.3.5:** Precision and recall for DGP1

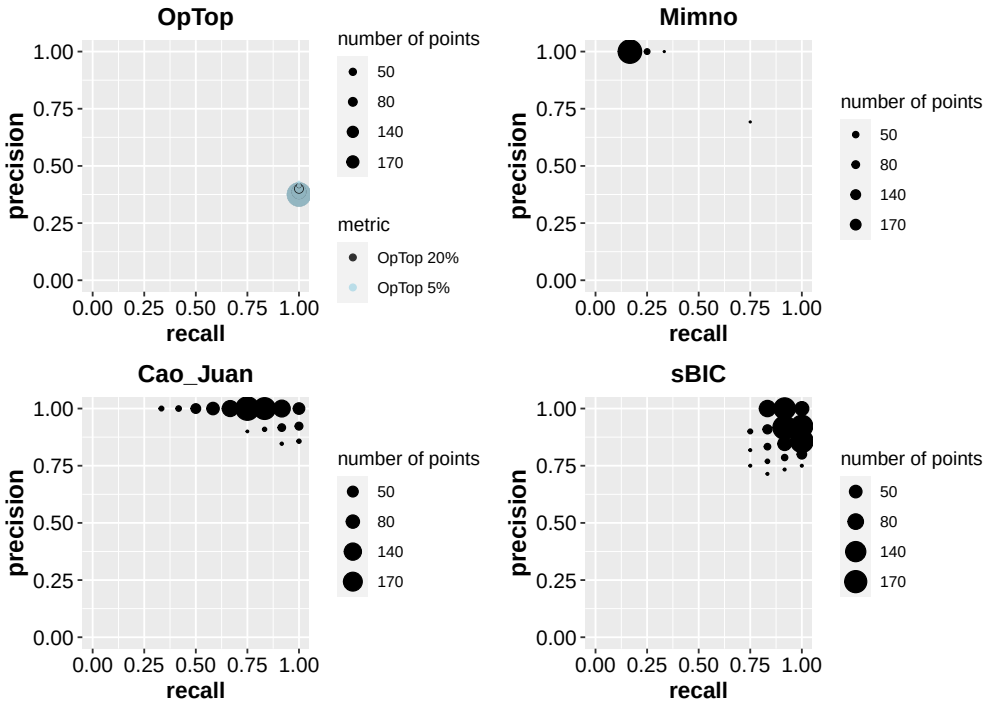



**Figure C.3.6:** Precision and recall for DGP2

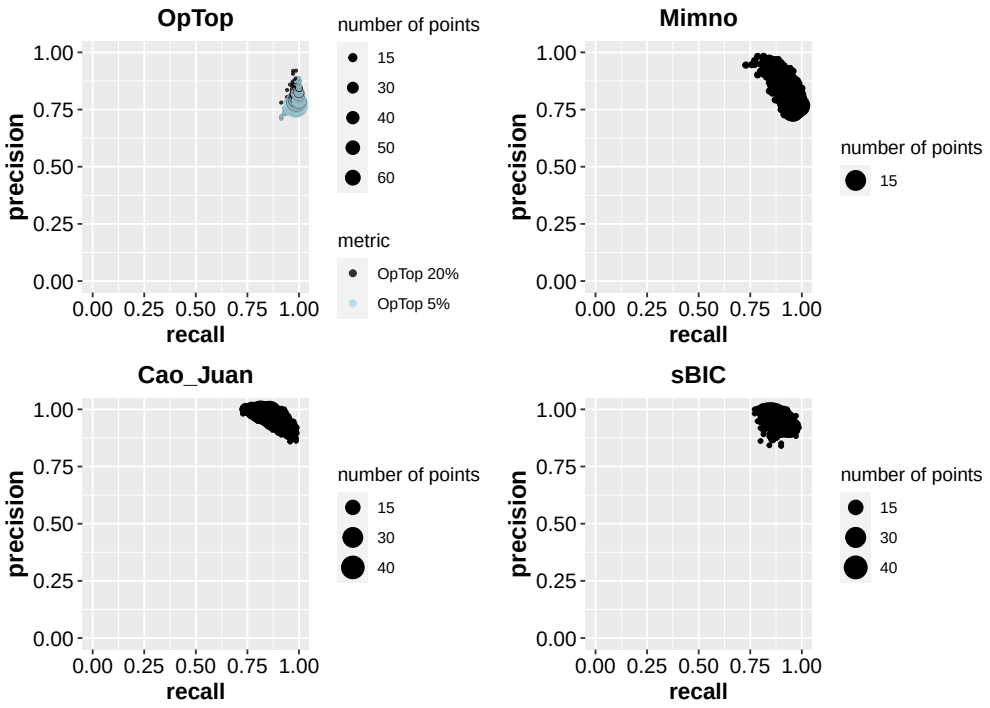



**Figure C.3.7:** Precision and recall for DGP3

## C.4 Weighted Recall & Precision

As a robustness analysis, we report results for alternative definitions of recall and precision. We identified a True Positive (TP) for the measures in Section 5.4.2, when the similarity of matched topics was above a predefined threshold. Here, we use the actual cosine similarity scores instead, which would be close to 1 for good matches. Hence, recall and precision values are calculated as follows:

$$\text{Recall} = \frac{\sum_{i=1}^{n} \text{cosine\_similarity\_score}_i}{K_{\text{true}}}, \tag{C.4.8}$$

$$\text{Precision} = \frac{\sum_{i=1}^{n} \text{cosine\_similarity\_score}_i}{K_{\text{metric}}}, \tag{C.4.9}$$

where $K_{\text{true}}$ is the true number of topics in a particular DGP. $K_{\text{metric}}$ is the proposed number of topics for the evaluation metric considered. The numerator contains the sum of cosine similarity values of all the $n$ identified matches. Therefore, recall presents the average cosine similarity value among the matches relative to the true number of topics. Precision presents the average cosine similarity value between the matches relative to the estimated number of topics.

Table C.4.1 summarizes the recall, precision, and F1 score values for this alternative definitions of recall and precision. As expected, the values are smaller than the values shown in Table 5.3 for the original definitions, but the qualitative findings about the relative performance of the different criteria remain unchanged. According to the F1 scores, sBIC performs best for DGP 1 and DGP 2, while the average F1 scores are quite similar for all the considered metrics in DGP 3, still with a minor advantage for Cao_Juan and sBIC.

While recall and precision values of our standard implementation are discrete leading to clustering of points in the scatter plots shown in Appendix C.3, the weighted recall and precision values reported in this section are continuous and each point is actually unique due to the differences in the cosine values, although these might be minor. Therefore, we do not use the type of plots from Appendix C.3 taking into account the clustering, but standard scatter plots in Figures C.4.8, C.4.9, and C.4.10.

| | | Recall | | Precision | | F1 | |
|---|---|---|---|---|---|---|---|
| data | metric | mean | std | mean | std | mean | std |
| DGP1 | Cao_Juan | 0.99 | 0.00 | 0.71 | 0.07 | 0.82 | 0.04 |
| | Mimno | 0.95 | 0.13 | 0.76 | 0.08 | 0.83 | 0.07 |
| | OpTop 20% | 0.98 | 0.00 | 0.67 | 0.03 | 0.80 | 0.02 |
| | OpTop 5% | 0.98 | 0.00 | 0.67 | 0.03 | 0.80 | 0.02 |
| | sBIC | 0.99 | 0.02 | 0.93 | 0.03 | 0.96 | 0.02 |
| DGP2 | Cao_Juan | 0.76 | 0.14 | 0.96 | 0.03 | 0.84 | 0.10 |
| | Mimno | 0.11 | 0.02 | 0.66 | 0.02 | 0.19 | 0.02 |
| | OpTop 20% | 1.00 | 0.00 | 0.38 | 0.01 | 0.55 | 0.01 |
| | OpTop 5% | 1.00 | 0.00 | 0.38 | 0.01 | 0.55 | 0.01 |
| | sBIC | 0.92 | 0.07 | 0.91 | 0.06 | 0.91 | 0.05 |
| DGP3 | Cao_Juan | 0.85 | 0.06 | 0.94 | 0.02 | 0.89 | 0.03 |
| | Mimno | 0.92 | 0.04 | 0.81 | 0.05 | 0.86 | 0.02 |
| | OpTop 20% | 0.98 | 0.02 | 0.78 | 0.03 | 0.87 | 0.02 |
| | OpTop 5% | 0.98 | 0.02 | 0.78 | 0.02 | 0.87 | 0.02 |
| | sBIC | 0.86 | 0.05 | 0.92 | 0.03 | 0.89 | 0.03 |

**Table C.4.1:** Descriptive statistics of recall, precision, and F1 scores based on cosine similarity

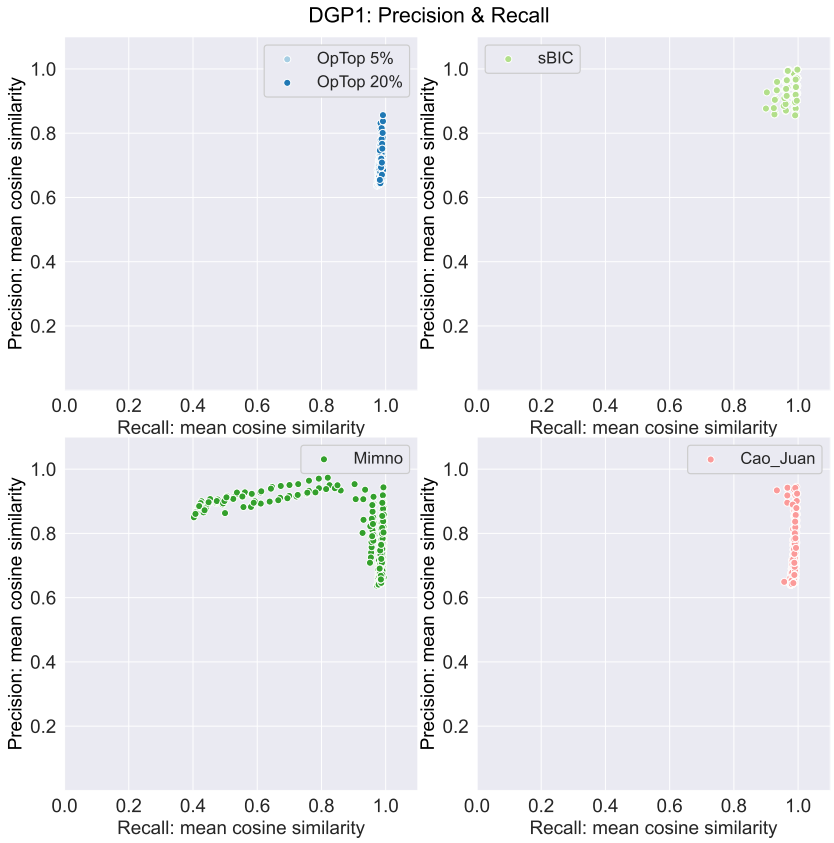



**Figure C.4.8:** Precision and recall based on cosine similarity for DGP1

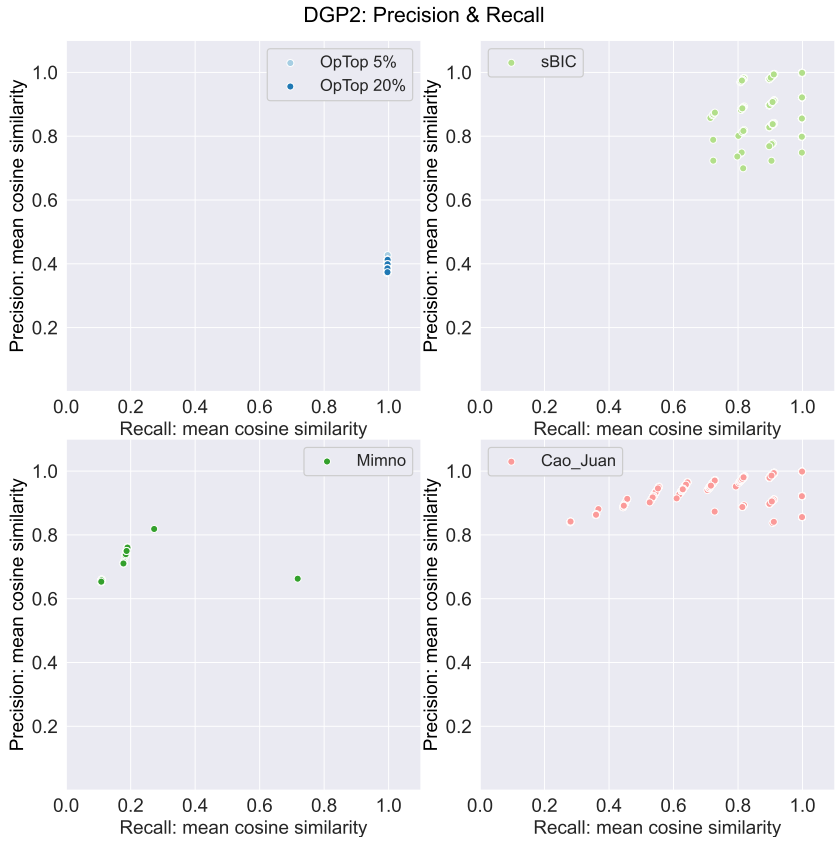



**Figure C.4.9:** Precision and recall based on cosine similarity for DGP2

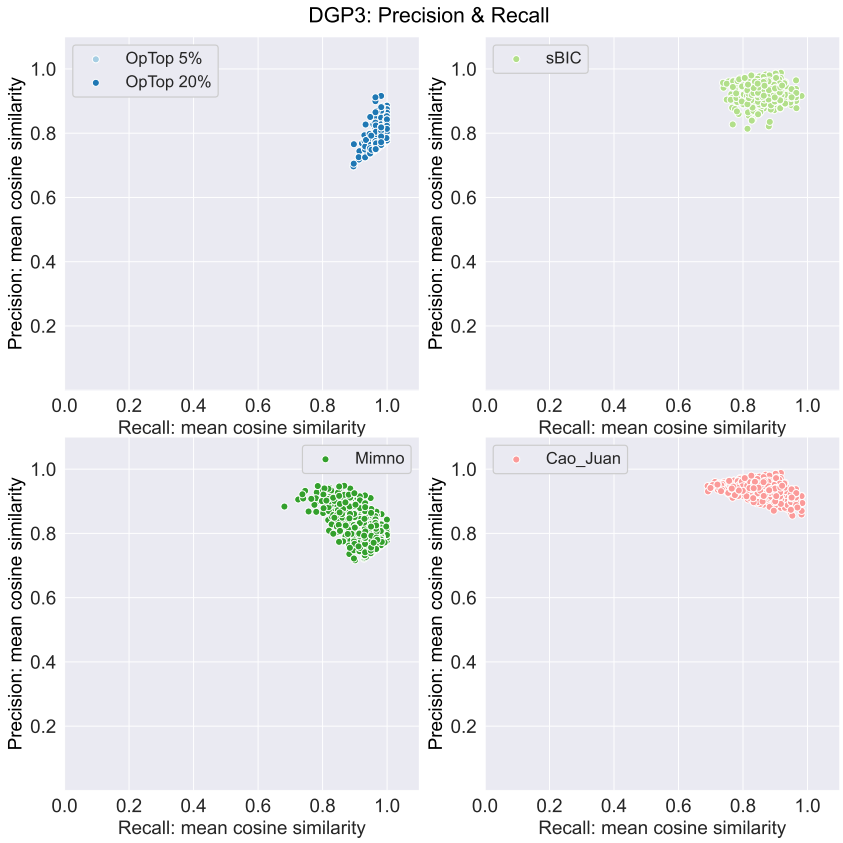



**Figure C.4.10:** Precision and recall based on cosine similarity for DGP3

# Part III

# Text-as-Data Applications in Economics

# Chapter 6

# Comparing Links between Topic Trends and Economic Indicators for German and Polish Academic Literature

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Comparing Links between Topic Trends and Economic Indicators for German and Polish Academic Literature |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 20%),<br>Victor Bystrov (contribution: 30%),<br>Anna Staszewska-Bystrova (contribution: 30%),<br>Peter Winker (contribution: 20%) |
| **Status:** | Published: *Comparative Economic Research. Central and Eastern Europe*, vol. 27, no. 2, 2024, pp. 7-28 |
| **Available from:** | `https://doi.org/10.18778/1508-2008.27.10` |

# Comparing Links between Topic Trends and Economic Indicators for German and Polish Academic Literature*

Victor Bystrov† Viktoriia Naboka-Krell‡
Anna Staszewska-Bystrova†,‖ Peter Winker‡

**Abstract**

The popularity of methods that include variables obtained from text mining in econometric models grows rapidly. One of the frequently applied approaches is to identify topics from large corpora which allows to determine trends reflecting the changing relevance of topics over time. We address the question whether such topic trends are linked to quantitative economic indicators typically used for analyzing the objects described by a topic. The analysis is based on scientific economic articles from Poland and Germany for the years 1984–2020. A specific interest is in whether relationships between topic trends and indicators are similar across national economies. The connection between topic trends and indicators is analyzed using vector autoregressive models and Granger causality tests.

*Key Words:* Topic modeling, text analysis, latent Dirichlet allocation, Granger causality, topic trends

*JEL* classification: C49

* Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

† University of Lodz, Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

‡ Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

‖ Corresponding author: anna.bystrova@uni.lodz.pl

## 6.1 Introduction

The popularity of analyses including variables obtained from texts in econometric models grows rapidly. Identification of topics in large corpora is often done as it allows to determine trends reflecting the changing relevance of topics over time. Such trends might be informative about economic developments and subsequently they can be used as (additional) indicators in economic analyses. Therefore, we address the question whether topic trends obtained from scientific text corpora from Poland and Germany might be linked to quantitative economic indicators typically used for analyzing the objects described by a topic. The textual data cover the period 1984 – 2020. Given the different developments of national economies in Poland and Germany over this period, we are particularly interested in whether relationships between topic trends and indicators are similar across national economies. Consequently, we focus on topics which are common to both text corpora.

The underlying textual data concern empirical economics and applied statistics (German text collection) as well as econometric modelling in general (Polish text collection). Given these specific characteristics of the two corpora, the uncovered topics are either associated with economic phenomena or methods used to analyze them. The relative interest over time in both categories of topics can be potentially linked to real economic processes as certain developments can directly motivate the discussion in the literature and be reflected in a changing popularity of particular methods or frameworks of doing the analysis. Since the first type of link might be more pronounced and as themes related to methods or theory found for both collections of texts were quite general, we focus on topics with macroeconomic content in this contribution.

The goal of the analysis is twofold. Firstly, we provide further evidence that results of topic modelling, i.e. uncovered latent topics and topic weight time series, can be used in the next step to study the links between real developments and the popularity of topics in a given text corpus. This type of result provides new insights for the description of a text collection. Secondly, we compare the relations between interest in topics common to Polish and German text corpora and developments in the national and global economies. Thereby, given the differences between the corpora and the economies considered, it is not obvious how many common topics can be identified. The paper contributes to the literature on the links between textual data and real indicators by considering topics which were discussed in texts published in two countries and by providing results of causality tests for both of these countries.

Establishing links between textual data and quantitative economic indicators has been subject of some analyses in the past with a fast growing interest during the last few years. Thereby, one may differentiate between keyword and sentiment based methods and approaches focusing on the identification of topics. Examples of the first group are the uncertainty index proposed by Baker et al. (2016) and, more recently, the fiscal sentiment indicator discussed by Latifi et al. (2024). Our contribution belongs to the second group, which also includes

papers on using topic modeling in the context of innovation activities by Venugopalan & Rai (2015) and Bergeaud et al. (2017). They analyzed the classification of patents based on patent applications, though without an explicit focus on the time dimension, while Lenz & Winker (2020) extracted innovation related topics from news-ticker data and generated time series of topic weights linked to fields of innovations. With a similar approach, Savin et al. (2022) provide an analysis of diffusion for different types of service robots. More closely related to our research are contributions comparing times series of topic weights with quantitative economic indicators such as Lüdering & Winker (2016) who studied the dynamic links between the relevance of topics in scientific publications and the development of corresponding macroeconomic indicators over the period from 1949 to 2010 for Germany, and Hansen et al. (2017) who analyzed the impact of increased transparency on monetary policy using publications of the Federal Open Market Committee. Dybowski & Adämmer (2018) used a topic model for analyzing fiscal policy in in the US, while Huang et al. (2018) and Larsen & Thorsrud (2022) considered the dynamic links between topic importance and financial market outcomes. Thorsrud (2020) and Ellingsen et al. (2022) used related approaches to exploit textual data from newspapers for improving GDP forecasts.

The structure of the article is as follows. The German and Polish text corpora, topics found using latent Dirichlet allocation, a topic matching procedure, and real economic indicators are presented in Section 6.2. Section 6.3 describes time series methods used to analyze the relations between topic trends and economic indicators. The results on the links of topics and real economic developments are provided in Section 6.4. A summary of the central findings and suggestions for further work are given in Section 6.5.

## 6.2 Textual and Economic Data

The analysis is based on two different types of data. On the one hand, these are textual data in the form of scientific research articles, and on the other hand, standard economic variables. While the textual data have to be transformed to quantitative indicators by means of topic modelling, the challenge regarding the economic indicators consists in selecting appropriate series corresponding to the topics identified based on the textual data. The two types of data are described in the following two subsections.

### 6.2.1 Transforming Text to Time Series

The text corpora used consist of research papers in the field of economics and econometrics published in the years 1984–2020 in Germany and in Poland, respectively. For Germany, we use all original articles published in the Journal of Economics and Statistics (JES),[1] which were published mainly in German during the first half of the sample, then with an

[1] For a previous application of topic modelling to articles published in this journal see Lüdering & Winker (2016).

increasing share in English reaching almost 100% towards the end of the sample. For Poland, we combine contributions to the proceedings of the Macromodels International Conference (MM) until 2011 and the Central European Journal of Economic Modelling and Econometrics (CEJEME) since 2009. Detailed description of both text corpora can be found in Bystrov et al. (2022).

While the scope of both scientific outlets is international, we observe that contributions discussing issues related to the German and Polish economy, respectively, constitute considerable shares of articles over the sample period. This is the rationale for contrasting results from topic modelling with some national as well as global economy indicators. Furthermore, given the low frequency of publications (6 issues per year for JES, 1 for MM and 4 for CEJEME), the following analysis will be done at an annual level. Thereby, we also alleviate the problem of publication lags inherent both in conference contributions and even more so in journal publications, but return to this issue in the discussion of the Granger causality analysis in Section 6.4.

The first step in transforming the text corpora to quantitative indicators consists in applying a latent Dirichlet allocation (LDA) model to both corpora (Blei et al., 2003). For details on pre-processing, parameter choice and the handling of the multilingual German corpus see again Bystrov et al. (2022). The LDA estimation results provide topics discussed in the text corpora and their relative weights which change over time. These weights correspond to the topic time series which are considered in the subsequent analysis.

The model selection step for the LDA indicated 37 and 60 topics for the Polish and German corpora, respectively. The larger number of different topics for the German dataset is plausible given the broader scope of JES as compared to MM and CEJEME. In further analysis, we focus on topics and their corresponding weight time series which turn out to be relevant in both corpora. Such common themes were identified using the topic matching method proposed by Bystrov et al. (2022) which is based on the comparison of distributions of topics over a joint vocabulary. Topic resemblance is evaluated using the cosine similarity measure and the matching is done by finding the nearest neighbour of a topic estimated for one corpus in the other corpus. The pairs which are considered as reasonable matches are selected on the basis of a cut-off value for the cosine similarity measure, i.e., only matches with a high enough cosine similarity are kept. For the present application, this threshold was set to 0.265 resulting in 24 topics showing up in both corpora and listed in Table D.1.1 in Appendix D.1. The labels of the topics were assigned based on the inspection of the corresponding word clouds available in Bystrov et al. (2022) and the documents with highest weights for the particular topic.

For further analysis, we restrict the set of topics to those corresponding to applied economic research which might be more closely related to specific economic developments than dynamics of topics related to purely methodological or theoretical research in economics, econometrics and statistics such as `forecasting`, `simulation methods` or `welfare economics`. Furthermore, our focus is on the national economy level. Therefore, we choose topics describing macroeconomic relationships rather than processes at the micro level such

as `firm growth` or `household income`. Table D.1.1 in Appendix D.1 provides information on this selection of topics for further analysis. Table 6.1 lists the remaining topics, the corresponding cosine similarity values and the economic indicators chosen as described in the following subsection.

| Topic label | Proposed indicator | Cosine similarity |
|---|---|---|
| International economics | Trade share | 0.86146 |
| Banking and credit | Credit-to-GDP ratio | 0.85116 |
| Business cycle | Output gap | 0.83445 |
| Capital and growth | Growth rate of investment | 0.80610 |
| Labour market | Unemployment rate | 0.72227 |
| Crude oil market | Oil market shocks | 0.70459 |
| Monetary policy | Policy rate | 0.69914 |
| Stock market | Stock market return, stock prices volatility | 0.63251 |
| Foreign trade | Net export share | 0.49535 |
| Energy sector | Total primary energy production | 0.48663 |

**Table 6.1:** Summary of Common Topics and Selected Economic Indicators

### 6.2.2 Selecting Economic Indicators

The selection of economic indicators related to the topics shown in Table 6.1 is based on the interpretation of these topics and considering related articles with the highest weights. Furthermore, this selection has to take into account data availability for both countries over the sample period. The specific sources and available samples of these economic time series are summarized in Table D.2.2 in Appendix D.2.

Given the structure of textual resources, all data are annual. This frequency of observations seems appropriate for considering relations between real developments and their discussion in journal and conference articles as publication lags or time for conducting research on a new topic might be considerable.

While for most of selected topics a suitable observable economic indicator is assigned both for the German and the Polish economy, the setting is different for the topic referring to the crude oil market. In this case, a suitable indicator would be one of shocks to the international oil market. Given that such an indicator cannot be observed directly, it has to be derived first using an auxiliary model. To this end, we use the method proposed by Kilian (2009). Then, given that global oil market shocks are considered, the same time series of shocks is used in the analysis for the Polish and German data. A more detailed discussion of these variables as listed in Table 6.1 follows in Section 6.3 together with the word clouds and the results of Granger causality testing. Furthermore, Figure D.2.1 from Appendix D.2 presents topic time series and economic indicators selected for the analysis.

## 6.3 Methods

The quantitative analysis is based on vector autoregressive models (VARs) which are a natural choice if the aim is to test for (Granger) causality when the direction of causality is not known a priori. We use bivariate VAR models, where one of the variables is the weight of a topic aggregated over all documents for each time period. We label these series as topic weight series. The second variable is the economic indicator linked to the specific model. A 2-dimensional VAR($p$) is given by (see Kilian & Lütkepohl (2017))

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + B d_t + u_t \,, \tag{6.3.1}$$

where $y_t = (topic_t,\ ind_t)'$ is the vector of topic weight and economic indicator in period $t$. The parameter matrices $A_i$ $(i = 1, \ldots, p)$ have dimensions $(2 \times 2)$, $\nu$ is a 2-dimensional intercept vector, $d_t$ includes all remaining necessary deterministic terms like dummy variables or trends with the corresponding parameters gathered in matrix $B$ and $u_t$ is a 2-dimensional vector of error terms.

Prior to estimation, variables included in the model are differenced if results of the ADF test indicate presence of a unit root. Furthermore, to take into account trending and nonlinear behaviour of topic weights which could be observed in some cases, vector $d$ includes the deterministic trend $t$ and the second power of this variable in selected models. The lag order of the VAR, $p\widehat{p}$, is chosen using the Akaike information criterion (AIC) applying a maximal lag length of 4 years. The models are tested for autocorrelation and ARCH effects of the error terms. Information on data transformations, use of deterministic variables (apart from the intercept) and the selected lag order of the VAR for all the models is provided in Tables D.2.4 and D.2.5 in Appendix D.2.

In the last step, to check the existence of dynamic relations between economic indicators and topic trends, Granger causality tests are performed. The hypothesis of instantaneous causality is also tested. If autocorrelation or ARCH errors are detected, these tests are based on the HAC estimator of the variance matrix for the OLS estimates (see Table D.2.5 in Appendix D.2 for the type of estimator used). The significance level for Granger causality tests is set to $\alpha = 0.1$. The outcomes of the tests may be to detect no causality, causality in one direction or causality in both directions. If topic weights turn out to be Granger causal for real economic indicators, this implies that changes in discussion intensity of the topic in the scientific literature preceded some relevant economic developments. The opposite outcome indicates that relevant developments in parts of the economy led to a more intensive scientific discussion of these aspects afterwards. If Granger causality is found for both directions, both channels are relevant, i.e., specific economic developments are accompanied with a change in the relevance of the topic in science both ex ante and ex post. While Granger causality focuses on the dynamic interlink between variables, it does not cover mutual influences taking place within one period, i.e., one year. Such contemporaneous effects are reflected by a significant correlation of the error terms of the VAR model across equations and are often labeled as instantaneous causality. Given low frequency of our data, instantaneous causality

might comprise the links which would be measured as Granger causality at higher frequency. This effect could be quite pronounced as scientific publications and conference proceedings were published with substantial publication lags for the period considered in our empirical application. Therefore, we also report the results of the tests of instantaneous causality.

## 6.4 Results

In this section results of Granger causality analysis are presented for the selected topics common to the JES and MM/CEJEME text corpora as described in Section 6.2. Figures 6.1 and 6.2 present the distribution of the relative topic importance for both corpora that is measured by computing the mean of topic weights over the sample period. The topics are ordered in descending order according to their relative importance for the respective corpora. The bars in orange represent "matched" topics meaning that they are common for both countries as identified by Bystrov et al. (2022) (see subsection 6.2.1). Correspondingly, the bars in blue represent "unmatched" topics meaning that these topics are specific for a single corpus. Dashed orange bars highlight topics selected for further analysis within VAR models together with corresponding economic indicators (see Table 6.1 in Section 6.2.2). These topics are also listed in Table 6.2 which additionally informs about their relative importance in both text collections. They are ordered according to their weight averaged over the two corpora, starting from the most relevant one.

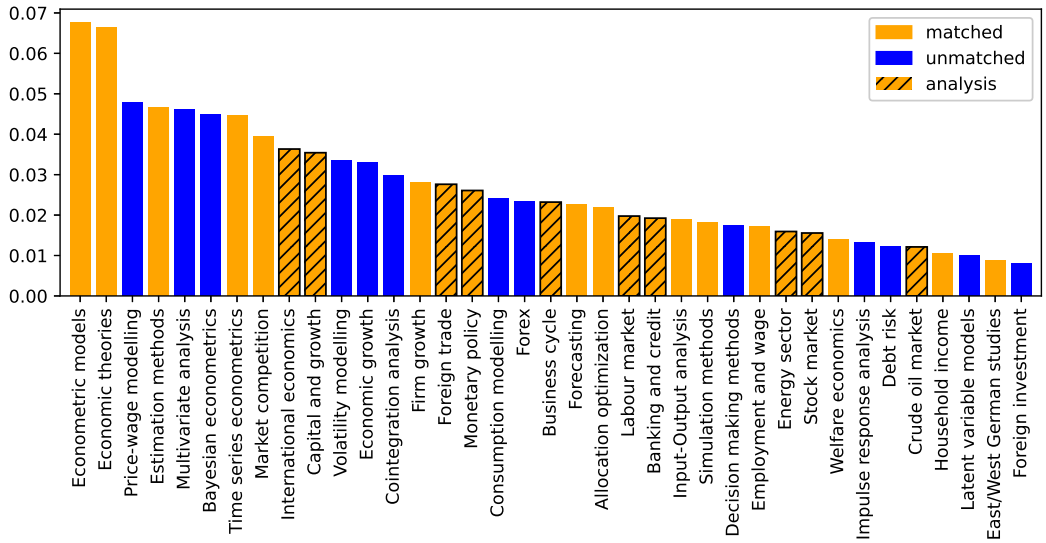



**Figure 6.1:** Relative topic importance for Poland

The outcomes of Granger causality testing are summarized in Table 6.3. These results are discussed in more detail for each topic separately in the following subsections.

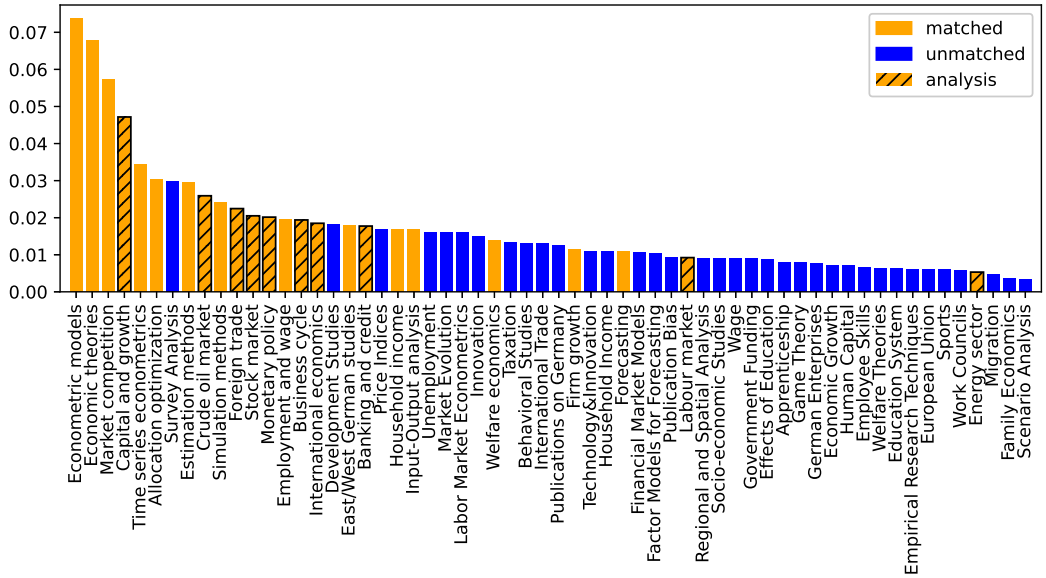



**Figure 6.2:** Relative topic importance for Germany

| | Average topic weight | |
|---|---|---|
| Topic label | Poland | Germany |
| Capital and growth | 0.0354 | 0.0472 |
| International economics | 0.0363 | 0.0185 |
| Foreign trade | 0.0276 | 0.0224 |
| Monetary policy | 0.0261 | 0.0201 |
| Business cycle | 0.0232 | 0.0194 |
| Crude oil market | 0.0121 | 0.0259 |
| Banking and credit | 0.0192 | 0.0177 |
| Stock market | 0.0156 | 0.0205 |
| Labour market | 0.0198 | 0.0092 |
| Energy sector | 0.0159 | 0.0053 |

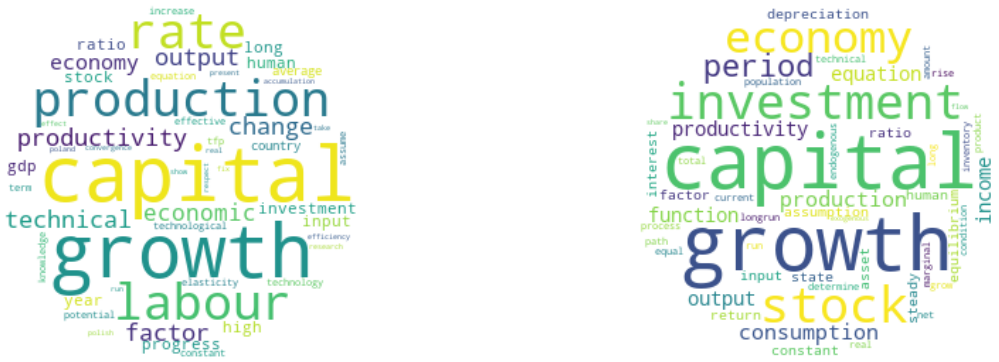
**Table 6.2:** Mean topics weights

| Topic Label | Economic Indicator | Country | topic weight is not a Granger cause for economic indicator | Hypothesis: economic indicator is not a Granger cause for topic weight | no instantaneous causality |
|---|---|---|---|---|---|
| Capital and growth | Growth rate of gross fixed capital formation | Poland | 1.5933 [0.2114] | **2.4496 [0.0763]** | 0.7304 [0.4710] |
| | | Germany | **2.3099 [0.0470]** | **4.3290 [0.0226]** | -0.5192 [0.6071] |
| International economics | Trade share | Poland | 0.2718 [0.6069] | **3.0073 [0.0951]** | -0.7904 [0.4362] |
| | | Germany | 0.0314 [0.8605] | 0.4400 [0.5122] | -1.2610 [0.2162] |
| Foreign trade | Share of net export in GDP | Poland | **9.4156 [0.0005]** | **13.790 [0.0001]** | **-3.2370 [0.0034]** |
| | | Germany | **7.3949 [0.0007]** | 1.9984 [0.1317] | -0.5633 [0.5774] |
| Monetary policy | Policy rate | Poland | 1.2599 [0.2719] | **4.2389 [0.0497]** | -0.3062 [0.7616] |
| | | Germany | 0.3366 [0.5658] | **3.7268 [0.0622]** | 0.6340 [0.5303] |
| Business cycle | Output gap | Poland | 1.6602 [0.2094] | 0.1777 [0.6770] | -0.5225 [0.6054] |
| | | Germany | 2.1504 [0.1192] | 1.1875 [0.3346] | -0.5355 [0.5960] |
| Crude oil market | Oil supply shocks | Poland | 0.0835 [0.7746] | 0.2850 [0.5972] | **3.0807 [0.0041]** |
| | | Germany | 1.3461 [0.2543] | 1.3274 [0.2576] | 0.4892 [0.6278] |
| Crude oil market | Aggregate demand shocks | Poland | 1.4327 [0.2404] | 0.0710 [0.7916] | -1.0949 [0.2813] |
| | | Germany | 1.9530 [0.1716] | **3.3648 [0.0756]** | 1.3285 [0.1929] |
| Crude oil market | Oil specific-demand shocks | Poland | 0.3835 [0.5402] | **3.7426 [0.0622]** | -0.8218 [0.4169] |
| | | Germany | 0.0748 [0.7862] | 2.0362 [0.1630] | 0.4259 [0.6728] |
| Banking and credit | Credit-to-GDP ratio | Poland | 2.1402 [0.1565] | 1.0865 [0.3076] | -1.0753 [0.2925] |
| | | Germany | 0.0670 [0.7975] | **7.0072 [0.0125]** | -0.6651 [0.5106] |
| Stock market | Stock market return | Poland | 0.3974 [0.5349] | 0.2459 [0.6249] | -0.4490 [0.6576] |
| | | Germany | **5.3744 [0.0280]** | 0.3702 [0.5478] | 1.5390 [0.1346] |
| Stock market | Stock prices volatility | Poland | **8.5297 [0.0022]** | **15.8720 [0.0002]** | -0.1086 [0.9146] |
| | | Germany | 1.0500 [0.3140] | 1.9445 [0.1738] | -0.3685 [0.7151] |
| Labour market | Unemployment rate | Poland | 0.4140 [0.5256] | 0.1316 [0.7197] | **4.8774 [0.0000]** |
| | | Germany | 2.3394 [0.1153] | **3.6196 [0.0405]** | -0.9967 [0.3264] |
| Energy sector | Primary energy production | Poland | **3.1800 [0.0862]** | 0.00017401 [0.9896] | 0.2342 [0.8166] |
| | | Germany | 1.0585 [0.4107] | **4.1260 [0.0189]** | **-4.1868 [0.0003]** |

**Table 6.3:** Granger causality tests results

## Capital and Growth

The topic label `capital and growth` was obtained based on the inspection of word clouds presented in Figure 6.3. This topic was relatively important both in the MM/CEJEME and JES corpora. Average weights were given by 0.0472 and 0.0354 which meant that it was the




Poland Germany

**Figure 6.3:** Word clouds for the topic "Capital and growth"

4th and the 10th most discussed theme in respective text collections. The time series of topic weights were modelled together with an economic indicator given by the growth rate of gross fixed capital formation. Statistical tests (see Table 6.3) indicated Granger causality from the economic indicator to topic weights for the Polish texts and two-way Granger causality for the German corpus.

### International Economics


Poland Germany

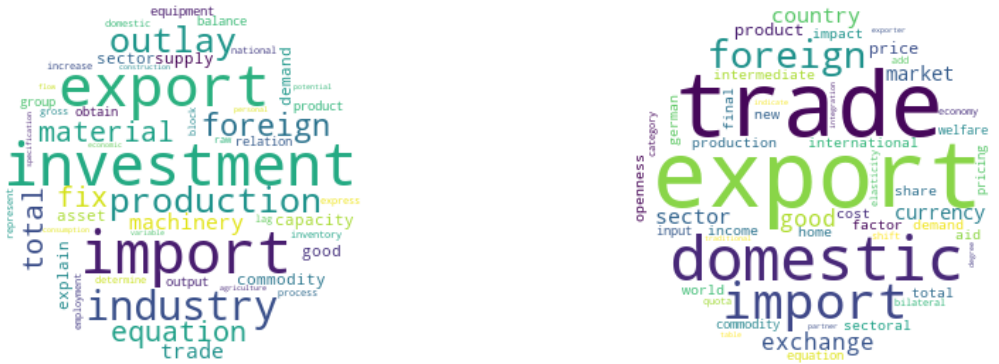
**Figure 6.4:** Word clouds for the topic "International economics"

As indicated by Table 6.2, the topic on `international economics` was relatively important for both text corpora. In the Polish text collection it was the 9th most popular theme (with an average weight 0.0363) and in the German corpus it was 16th (with an average weight of 0.0185). World clouds associated with this theme are shown in Figure 6.4. Even though these topics were paired on the basis of the topic matching algorithm, they seem to concern slightly different matters as the German topic is more related to cross country comparisons and international spillovers, while the Polish topic puts more emphasis

on international trade. This different focus complicates the selection of an economic indicator which was eventually specified as trade share defined as the sum of the nominal value of imports and exports expressed as percent of nominal GDP.

According to the results provided in Table 6.3, no Granger causality between topic weights and the indicator was found for Germany, while shocks in international trade seemed to Granger cause the topic weights for Poland. For both countries no significant instantaneous link was detected.

### Foreign Trade




Poland Germany

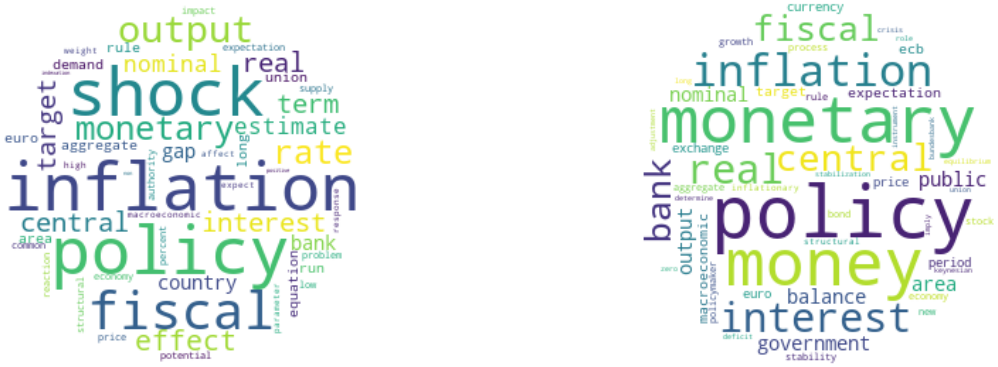
**Figure 6.5:** Word clouds for the topic "Foreign trade"

Key words for the topic labeled as `foreign trade` are shown in Figure 6.5. With average weights of 0.0276 and 0.0224, this topic was the 15th and the 11th most discussed topic in Polish and German text corpora respectively. An economic indicator used in the VAR associated with foreign trade was a share of net exports in GDP. This variable was selected as a primary indicator of developments in foreign trade.

The outcomes of Granger causality tests presented in Table 6.3 indicate two way causality and instantaneous causality found for Poland and one way causality for Germany. In the latter case, the changes in popularity of the subject in the scientific literature preceded real movements in the German share of net exports in GDP.

### Monetary Policy

The most important words associated with the topic on `monetary policy` are provided in Figure 6.6. This topic was quite popular in both text corpora – it had the 16th and 13th heighest mean weights in Polish and German collections, respectively. A natural economic indicator for this topic is the monetary policy rate. Granger causality test results based on VAR incorporating policy rate as well as topic trend (see Table 6.3) indicate causality from the indicator to topic weights for both countries. This implies that past changes in the




Poland Germany

**Figure 6.6:** Word clouds for the topic "Monetary policy"

monetary policy rate add predictive power to a dynamic model for explaining the development of topic weights over time.

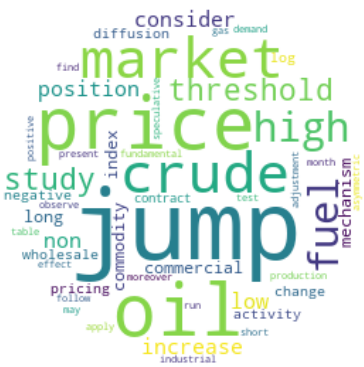
### Business Cycle


Poland Germany

**Figure 6.7:** Word clouds for the topic "Business cycle"

Contributions related to `business cycle` studies were identified in both text corpora according to the word clouds presented in Figure 6.7. This topic was the 19th and 15th most popular one in the Polish and German collection, respectively (mean weights were equal to 0.0232 and 0.0194). The bivariate VAR models were constructed for topic weights and the national output gap indicators. However, no Granger causality was found for neither corpus, nor was there an indication of instantaneous effects. Given the recurrence of business cycles, it might not come as a surprise that the scientific literature on the topic does not lead or follow actual business cycles. Thus, the result of no Granger causality appears sensible.

### Crude Oil Market




Poland Germany

**Figure 6.8:** Word clouds for the topic "Crude oil market"

Figure 6.8 shows the word clouds associated with the topic on the `crude oil market`, which exhibit a clear focus on oil price shocks. This theme was found to be relatively more important in the German corpus where it was the 9th most frequently discussed topic (with an average weight of 0.0259 – see Table 6.2). In the Polish text collection this topic was only the 33rd most important one out of 37 themes identified (with an average weight of 0.0121).

Given the word clouds, it is intuitive that the corresponding topic weights might be related to shocks on the global crude oil market rather than to an oil price series. Therefore, for this specific topic, the relevant economic indicator is not directly observable, but has to be derived in a first step. To construct a series of shocks we use the method described by Kilian (2009). It is based on a trivariate structural VAR, including log-differences of world crude oil production ($\Delta prod$), an index of global real economic activity ($rea$), and the real price of oil ($rpo$) (see Table D.2.3 for the description of the data used). As recommended by Kilian (2009), the model is estimated using monthly data. The model has the form

$$A_0 y_t = \nu + \sum_{i=1}^{24} A_i y_{t-i} + u_t\,, \qquad A_0^{-1} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

where $y_t = (\Delta prod_t, rea_t, rpo_t)'$. The three structural shocks labeled by Kilian (2009) as oil supply shocks, aggregate demand shocks, and oil-specific demand shocks are given by

$$e_t = A_0^{-1} \begin{bmatrix} u_t^{\textit{oil supply shock}} \\ u_t^{\textit{aggregate demand shock}} \\ u_t^{\textit{oil specific-demand shock}} \end{bmatrix},$$

and represent innovations to global oil production, global demand for industrial commodities, and precautionary demand for oil, respectively.

These estimated series of structural shocks are aggregated to annual frequency as used in the present study. Unlike in other VAR models, where topic weights from German and

Polish text collections were modelled jointly with economic indicators from Germany and from Poland, respectively, in the models used to study the topic of crude oil market, the same series of global shocks are used for both countries.

As shown in Table 6.3, for Poland, oil-specific demand shocks were found to be a Granger cause for the topic weights and instantaneous causality was found for oil supply shocks. This implies that future scientific discussion on the role of the crude oil market are stimulated by demand shocks, while supply shocks might be realized faster or – given the lag between scientific research and publications – even with some lead. For Germany, we expected a similar finding in particular due to the high interest in economic science devoted to the crude oil market after the supply shocks in the 1970s and early 1980s. However, since our sample starts in 1984, these events are not part of the sample period, which might explain that no Granger causality was found for German data for the models incorporating oil supply shocks and oil-specific demand shocks. In contrast, a surge of the interest to the topic in the period 2002-2005 correlates with large shocks to the index of global demand for industrial commodities (the Kilian index, see Kilian (2009)). Shocks to the global demand for industrial commodities can be transmitted (with a delay) to the crude oil prices and eventually to the discussion of oil prices in the JES, which might explain the finding that aggregate demand shocks were found to be a Granger cause for crude oil market topic weights.

### Banking and Credit


Poland Germany

**Figure 6.9:** Word clouds for the topic "Banking and credit"

The most important keywords for the topic labeled as `banking and credit` are shown in Figure 6.9. With average weight of 0.0192, it was the 23rd most covered theme in the Polish corpus. The mean weight for the German collection was 0.0177 which resulted in the 19th place of the most popular topics. In the Granger causality analysis, the topic weights were contrasted with an economic indicator for the importance of the banking sector given by the credit-to-GDP ratio. This variable turned out to be Granger causal for topic weights for the German corpus, i.e., a shock to the credit-to-GDP ratio preceded changes in the

importance of the topic in the scientific literature. However, no Granger causality was found for the Polish text collection, nor was there an indication of contemporaneous effects.

### Stock Market


Poland Germany

**Figure 6.10:** Word clouds for the topic "Stock market"

The topic related to the `stock market` (see Figure 6.10 for the word clouds) attracted more attention in the German corpus, where it occupied position 12 with an average weight of 0.0205. In the Polish text collection this topic was 29th and had a mean weight of 0.0156. This discrepancy in the relative interest was expected as the Warsaw Stock Exchange was not operating during the first part of our sample period (it became active in 1991, our sample comprises data from 1995 - 2020 for Poland).

In the case of this topic, two economic indicators were considered alternatively, as it was difficult to select the best suited one a priori. The two indicators are the stock market return computed as the growth rate of the annual average national stock market index and the stock price volatility calculated as the average of the 360-day volatility of the national stock market index. The two indices considered were the DAX and the WIG, for German and Polish corpora, respectively.

Results from Table 6.3 indicate two-way Granger causality between stock prices volatility and topic weights for the Polish corpus. For German text collection, topic weights were a Granger cause for stock market returns series.

### Labour Market

The most important words for the topic on the `labour market` are provided in Figure 6.11. It seems that the topic found from the German text corpus has a broader focus on the labour market and its functioning, while the topic from the Polish corpus focuses more on wage and (un-)employment. The weights of this topic were given by 0.0198 and 0.0092 for the Polish and German corpora, respectively. For both text collections the topic belonged to the second


Poland Germany

**Figure 6.11:** Word clouds for the topic "Labour market"

half of most popular themes (22nd place out of 37 for Poland and 39th place out of 60 for Germany). Topic weights were modelled jointly with unemployment rate. The outcomes of Granger causality tests were as follows: causality from indicator to topic weights was found for the German corpus and instantaneous causality was detected for Polish corpus.

### Energy Sector


Poland Germany

**Figure 6.12:** Word clouds for the topic "Energy sector"

Word clouds for the `energy sector` topic are exhibited in Figure 6.12. This topic was not very popular in neither corpus under investigation. With an average weight of 0.0159 it was 28th on the list of most frequently discussed topics for Poland. For the German corpus it was the third least important theme (with a mean weight of only 0.0053, it occupied the 57th place). The economic indicator selected for causality analysis was primary energy production.

Results from Table 6.3 show significant relations between topic weights and the economic indicator for both text corpora. The direction of causality is however different: for the

German corpus real developments precede and are simultaneous with the changes in topic weights, while there is causality from topic weights to the real indicator in the case of the Polish corpus.

## 6.5 Conclusions

We used text mining techniques to identify topics discussed in two scientific text corpora on economics from Germany and from Poland for the same period of 1984–2020. Thereby, our main focus was on the relative importance of these topics over time and its correspondence to central economic indicators related to the topics. For one theme corresponding to shocks to the crude oil market, the indicator had to be derived from an auxiliary model.

In order to determine whether economic research was leading or following real events, bivariate relationships between topic trends and economic variables were analyzed using vector autoregressive models and Granger and instantaneous causality tests. We considered 10 distinct topics which appeared in both text corpora. This allowed us to study whether Granger causality between topic trends and the corresponding indicators was frequent and also which direction of causality prevailed. Furthermore, the parallel analysis for German and Polish textual data and economic indicators made it also possible to compare results for these two countries.

The analysis indicated significant links between scientific literature and real developments for at least one country for all but one pair of topic and the corresponding economic indicator. This exception was given by the topic on `business cycle` and the output gap. In this case, scientific analysis might be focused more on the existence and stability of the recurring cycles than on each particular cycle, which might explain the lack of a significant link.

In general, the reason for not detecting causality between topics and indicators might also be due to a relatively short sample size, imperfect matching of topics from the German and Polish corpora and, consequently, problems with selecting an appropriate economic indicator fitting well with the topics for both countries. Therefore, the substantial number of significant links actually found indicates a strong focus of researchers publishing in the two corpora on empirical evidence. However, the results on the direction of Granger causality were mixed. Examples of instantaneous causality, both kinds of unidirectional causality and two-way causality were found. Overall, it could be concluded that economic indicators were leading topic trends more often than the other way round. Thus, economists tend to follow real developments in their analysis rather than predict them.

While the frequency of significant links between economic variables and topic popularity for the Polish and German corpora were comparable, causality patterns for specific topic-indicator pairs were often not the same. Some similarities included, for example, significant reactions by researchers of the topic `monetary policy` to changes in the policy rate and Granger causality from growth rate of gross fixed capital formation to the trend of the topic `capital and growth`. Interestingly, the discussion in economic science of `foreign trade` was preceding changes of the share of net exports in GDP in both countries, i.e.,

globalization was discussed by researchers prior to its realization. Some differences in the results across countries, e.g., for the topics on the `stock market` or the `energy sector` could be explained by alternative economic settings in Poland and in Germany during the years under investigation. For example, the stock exchange only emerged in Poland during the sample period and energy production mixes are not comparable, which might be reflected in the focus and content of economic discussion of these topics.

There are several limitations of our analysis, which might be considered as a first explorative study of the link between economic science and economic reality in a cross country comparison. Further research will have to address both methodological and content related issues. On the methodological side, alternative methods for identifying (common) topics and for quantifying their relevance over time might be considered. This also includes a more thorough analysis of the robustness of findings with regard to the specific method and parameter settings used. In addition, it would be instructive to conduct the analysis recursively, i.e., limiting the modelling of textual data to the information available up to a specific year. On the content side, it might be of interest to include other text corpora covering economic research for the selected countries, but also to enlarge the set of countries and, when feasible, the observation period. Further research should also focus on the link between topics and economic reality for specific topics, possibly using case studies for a better understanding of the driving forces behind the links found in our analysis.

# Appendix D

## D.1 Selection of Topics for Further Analysis

As can be seen from Table D.1.1, after the analysis of topic word clouds, 10 out of 24 topics were selected for further analysis. 9 topics labeled as: `forecasting`, `time series econometrics`, `econometric models`, `estimation methods`, `economic theories`, `input-output analysis`, `simulation methods`, `allocation optimization` and `welfare economics` were identified as methodological or theoretical and thus disregarded from this study. Further 4 topics corresponding to `firm growth`, `market competition`, `household income`, and `employment and wage` were rejected as well as they seemed to correspond to the micro economic level which might not be well reflected in aggregated observed time series employed in this paper. Eventually, a topic labeled as `East-West German studies` was disregarded as it deals with studying the consequences of a specific event.

| Topic label | Decision/Choice of indicator |
|---|---|
| Forecasting | Discard: Methodological |
| International economics | Trade share |
| Banking and credit | Credit-to-GDP ratio |
| Time series econometrics | Discard: Methodological |
| Business cycle | Output gap |
| Capital and growth | Growth rate of investment |
| Econometric models | Discard: Methodological |
| Labour market | Unemployment rate |
| Crude oil market | Oil market shocks |
| Monetary policy | Policy rate |
| Estimation methods | Discard: Methodological |
| Firm growth | Discard: Microeconomic data |
| Stock market | Stock market return, stock prices volatility |
| Market competition | Discard: Microeconomic data |
| Economic theories | Discard: Theory |
| East-West German studies | Discard: Specific event |
| Household income | Discard: Microeconomic data |
| Employment and wage | Discard: Microeconomic data |
| Input-Output analysis | Discard: Methodological |
| Simulation methods | Discard: Methodological |
| Foreign trade | Net export share |
| Energy sector | Total primary energy production |
| Allocation optimization | Discard: Theoretical |
| Welfare economics | Discard: Theoretical |

**Table D.1.1:** Summary of common topics and selected economic indicators

## D.2 Data Sources and Data Transformations

| Topic | Variable | Poland | | Germany | |
|---|---|---|---|---|---|
| | | Source | Span | Source | Span |
| Capital and growth | Investment Growth Rate | OECD.Stat | 1991–2020 | OECD.Stat | 1984–2020 |
| International economics | Trade-to-GDP Ratio | OECD.Stat | 1990–2020 | OECD.Stat | 1990–2020 |
| Foreign trade | Net Export-to-GDP Ratio | OECD.Stat | 1990–2020 | OECD.Stat | 1984–2020 |
| Monetary policy | Central Bank Discount Rate | National Bank of Poland | 1989–2020 | Bundesbank | 1984–2020 |
| Business cycle | Output gap (HP filter) | OECD.Stat | 1990–2020 | OECD.Stat | 1984–2020 |
| Banking and credit | Credit-to-GDP ratio (actual) | Bank of International Settlements | 1992–2020 | Bank of International Settlements | 1984–2020 |
| Stock market (returns) | Growth rate of annual average stock market index | World Bank | 1995–2020 | World Bank | 1989–2020 |
| Stock market (volatility) | Average of the 360-day volatility of the national stock market index | World Bank | 1995–2020 | World Bank | 1988–2020 |
| Labour market | Unemployment Rate | Statistics Poland (Central Statistical Office) | 1990–2020 | OECD.Stat | 1984–2020 |
| Energy sector | Primary energy production | Eurostat | 1990–2020 | Eurostat | 1990–2020 |

**Table D.2.2:** Data

| Variable | Source | Span |
|---|---|---|
| Index of global real economic activity | Federal Reserve Bank of Dallas | 1984–2020 |
| Global crude oil production | U.S. Energy Information Administration (EIA) | 1984–2020 |
| U.S. crude oil imported acquisition cost by refiners (dollars per barrel) | U.S. Energy Information Administration (EIA) | 1984–2020 |
| U.S. CPI (all items) seasonally adjusted | OECD | 1984–2020 |

**Table D.2.3:** Data for the model of oil shocks

**Figure D.2.1:** Economic indicators and topic weights

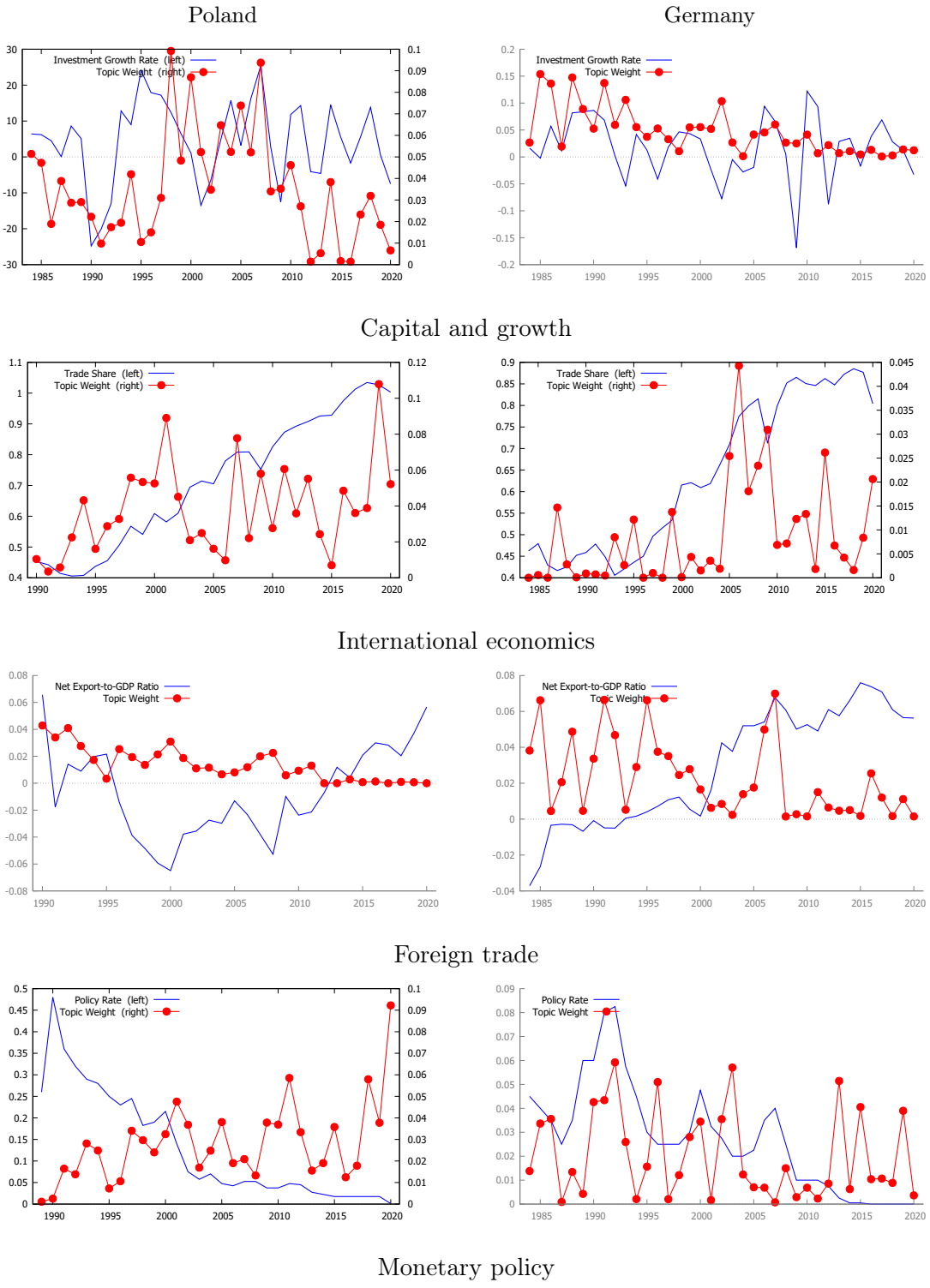


**Figure D.2.1:** (cont.) Economic indicators and topic weights

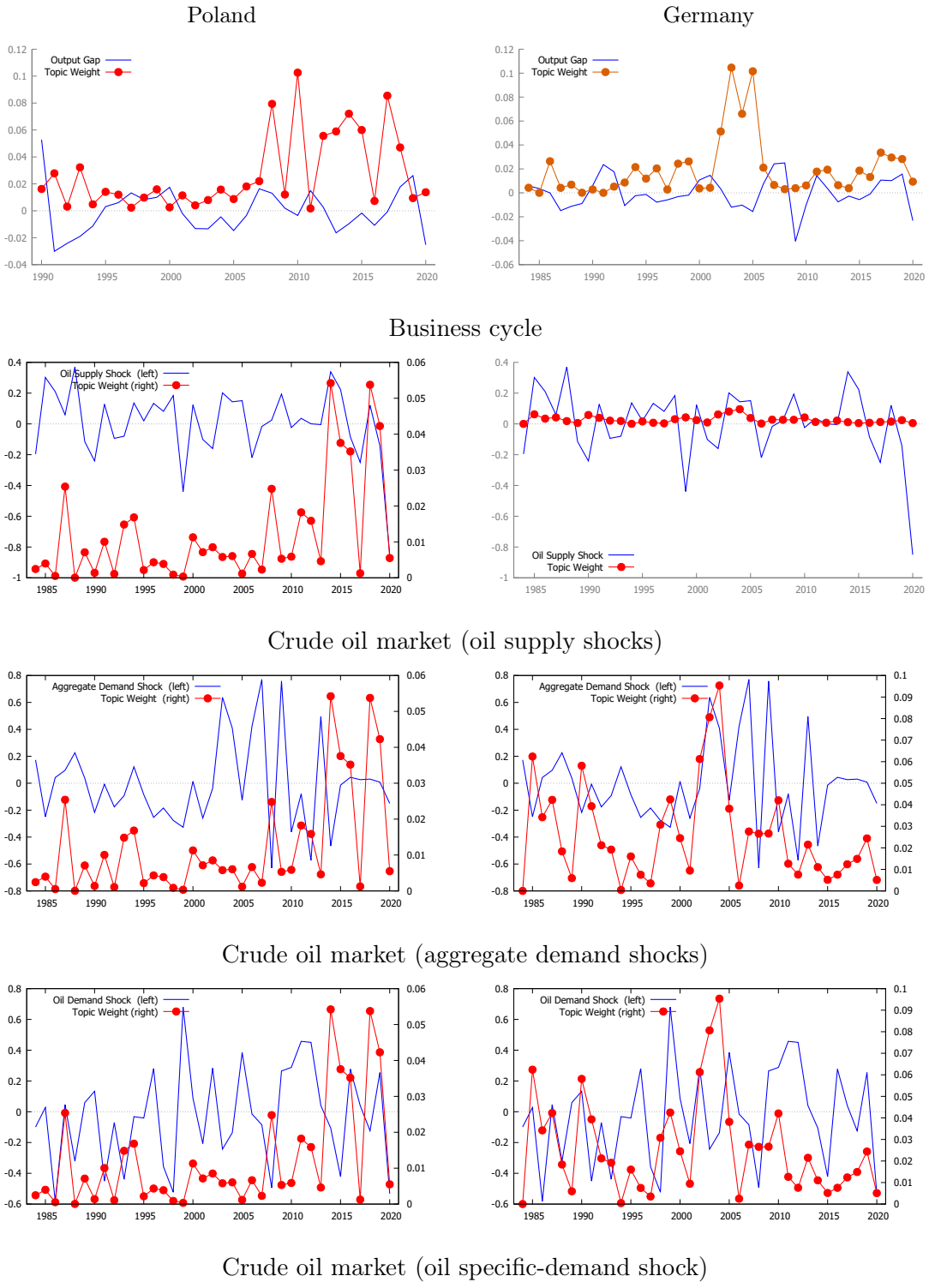
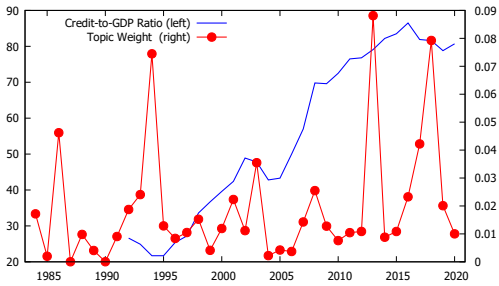
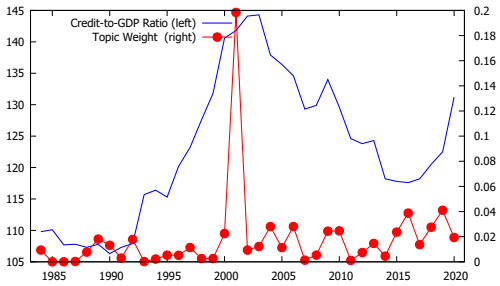


**Figure D.2.1:** (cont.) Economic indicators and topic weights

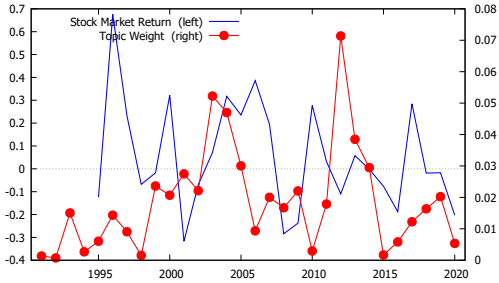
Poland

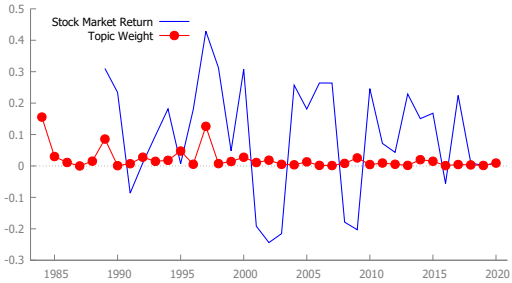



Germany

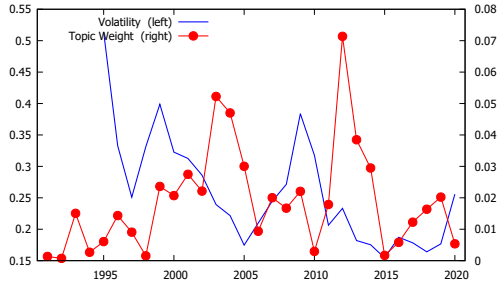


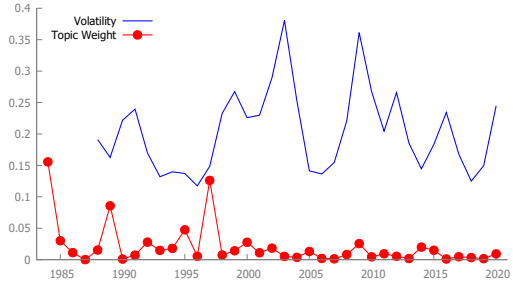

Banking and credit



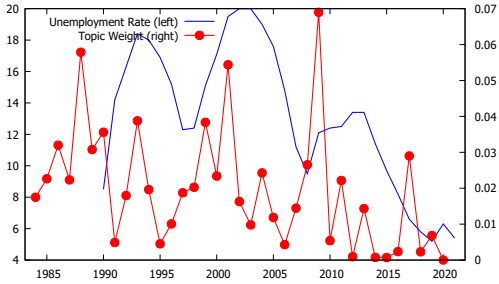



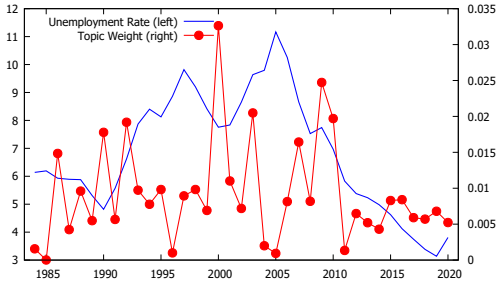

Stock market (stock market return)



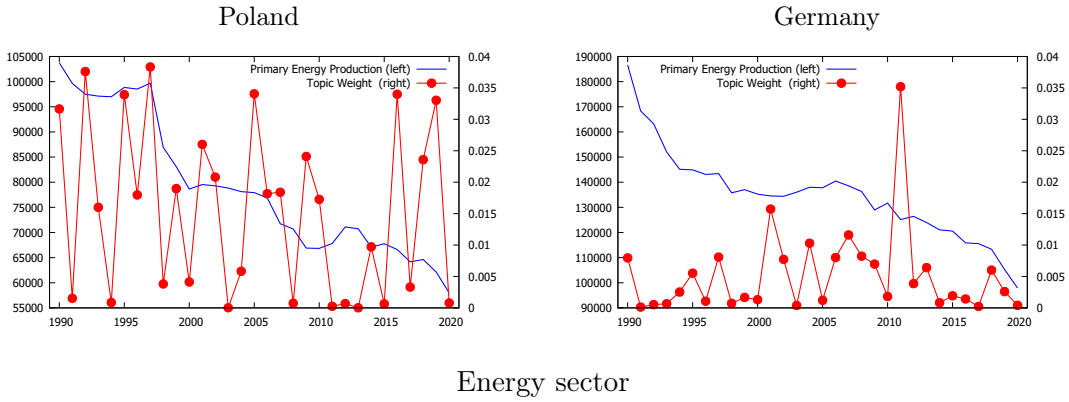

Stock market (stock prices volatility)




Labour market

**Figure D.2.1:** (cont.) Economic indicators and topic weights




| | Poland | | Germany | |
|---|---|---|---|---|
| Topic | topic weight transformation | economic indicator transformation | topic weight transformation | economic indicator transformation |
| Capital and growth | none | none | none | none |
| International economics | none | first difference | none | first difference |
| Foreign trade | none | first difference | none | first difference |
| Monetary policy | none | none | none | none |
| Business cycle | none | none | none | none |
| Crude oil market (oil supply shock) | none | none | none | none |
| Crude oil market (aggregate demand shock) | none | none | none | none |
| Crude oil market (oil specific-demand shock) | none | none | none | none |
| Banking and credit | none | first difference | none | first difference |
| Stock market (returns) | none | none | none | none |
| Stock market (volatility) | none | none | none | none |
| Labour market | none | first difference | none | first difference |
| Energy sector | none | first difference of logs | none | first difference of logs |

**Table D.2.4:** Data transformations

| | Poland | | | Germany | | |
|---|---|---|---|---|---|---|
| Topic | VAR lag order | deterministic terms | standard errors | VAR lag oder | deterministic terms | standard errors |
| Capital and growth | 4 | none | OLS | 2 | $t$ | HAC |
| International economics | 1 | $t$,$t^2$ | OLS | 1 | $t$,$t^2$ | OLS |
| Foreign trade | 3 | $t^2$ | HAC | 4 | $t$,$t^2$ | HAC |
| Monetary policy | 1 | $t$,$t^2$ | HAC | 1 | none | HAC |
| Business cycle | 1 | $t$,$t^2$ | HAC | 3 | $t$,$t^2$ | HAC |
| Crude oil market (oil supply shock) | 1 | $t$,$t^2$ | HAC | 1 | none | OLS |
| Crude oil market (aggregate demand shock) | 1 | $t$,$t^2$ | HAC | 1 | none | HAC |
| Crude oil market (oil specific-demand shock) | 1 | $t$,$t^2$ | OLS | 1 | none | HAC |
| Banking and credit | 1 | none | OLS | 1 | none | OLS |
| Stock market (returns) | 1 | none | HAC | 1 | none | HAC |
| Stock market (volatility) | 4 | $t$,$t^2$ | HAC | 1 | none | OLS |
| Labour market | 1 | none | OLS | 2 | $t$,$t^2$ | OLS |
| Energy sector | 1 | none | HAC | 4 | $t$,$t^2$ | HAC |

**Table D.2.5:** Model details

# Chapter 7

# Construction and Analysis of Uncertainty Indices based on Multilingual Text Representations

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Construction and analysis of uncertainty indices based on multilingual text representations |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 100%) |
| **Status:** | Published: *Economics Letters*, vol. 237, 2024, pp. 111653 |
| **Available from:** | `https://doi.org/10.1016/j.econlet.2024.111653` |

Earlier versions of this paper were presented at:

- Workshop "Digital Methods in History and Economics", October 14-15, Hamburg, Germany, 2021
- 4th Annual COMPTEXT Conference, 6-7 May, Dublin, Ireland, 2022

# Construction and Analysis of Uncertainty Indices based on Multilingual Text Representations

Viktoriia Naboka-Krell[‡,‖]

**Abstract.** The work by Baker et al. (2016), who propose a dictionary based method and estimate the level of *economic policy uncertainty* (EPU) based on the occurrence of specific terms in ten leading newspapers in the USA, is among the first ones to detect the potential of text data in economic research. Following this line of research, this paper proposes automated approaches to construction of EPU indices for different countries based on newspapers' texts. Multilingual fastText word embeddings, (S)BERT embeddings, and a novel multilingual topic modeling approach are used to construct EPU indices for Germany, Russia, and Ukraine. It is shown that constructed EPU indices based on multilingual word embeddings are Granger causal to the economic activity in all of the considered countries.

*Key Words: text-as-data, fastText emeddings, BERT, economic policy uncertainty, natural language processing*

‡ Faculty of Economics and Business Studies, Department of Statistics and Econometrics, Justus Liebig University Giessen, Licher Str. 64, 35394 Giessen, Germany

‖ Corresponding author: Viktoriia.Naboka-Krell@wi.jlug.de

## 7.1 Introduction

The work by Baker et al. (2016) is among the first ones to detect the potential of text data in economic research. Although dictionary based methods as in Baker et al. (2016) are widely used due to their simplicity and interpretability, recent advances in NLP offer many further possibilities to gain insights from text data. New approaches include the use of topic models such as Latent Dirichlet Allocation (LDA). Azqueta-Gavaldón (2017) proposes an LDA based procedure to build an uncertainty index that strongly resembles the index introduced by Baker et al. (2016) (BBD) index. The proposed EPU index is the aggregated time series based on the time series of the identified EPU related topics.

Some of them also make use of word embeddings, for example, to extend the EPU related term set as proposed by Ghirelli et al. (2019) for the case of Spain. The authors show that an unexpected shock in their modified EPU index leads to a significant decline in gross domestic product (GDP), private consumption, and investments. Algaba et al. (2020) follow this approach and construct an EPU index for Belgium using GloVe word embeddings. It has been shown that the constructed index negatively correlates (-0.62) with Consumer Confidence Indicator (CCI). Xie (2020) proposes a fully automated method to build an uncertainty index. The author applies the Wasserstein Index Generation model and uses word vectors to represent the analysed text units in a vector space.

These examples show that word embeddings might have a considerable impact on future applications and methods in economic literature. In contrast to other methods, word embeddings are able to capture the semantic and syntactic characteristics of words, which is very useful in numerous cases. The current work is dedicated to examination of word and text representations in context of EPU measurement and contributes to the growing area of text-as-data applications in economics, particularly uncertainty measurement, in several ways. First, it proposes several approaches to construction of EPU indices in the multilingual setting without any supervision. Second, it applies a novel zero-shot topic modeling approach that allows to train a topic model in one language and to predict topic distributions for documents in unseen languages. Third, the resulting uncertainty indices are evaluated with regard to their impact on economic activity in selected countries.

## 7.2 Text Representation Techniques

Multilingual word embeddings (MWE) are word vectors in multiple languages that are embedded in a shared vector space. These representations are characterized by the interpretability of the distances between them in different languages, meaning that similar words are closer to each other in the shared vector space. Several approaches have been proposed to train such multilingual word embeddings. One of the widely used approaches is the mapping based approach that relies on so-called off-the-shelf lexicons. Freely available

multilingual fastText[1] word representations were also learned following the mapping based approach proposed by Joulin et al. (2018) (bilingual mapping) and Grave et al. (2018) (multilingual mapping by defining a pivot language).

A great breakthrough in and a major contribution to the field of language model learning has been made with the publication of the work by Devlin et al. (2019). The authors present their novel approach to text representations BERT which differs substantially from existing models. BERT stands for Bidirectional Encoder Representations from Transformers and consists of a multi-layer Transformer encoder. BERT became the state-of-the-art in many NLP tasks. However, to overcome some capacity and time issues, Sentence BERT (SBERT) was introduced that was fine-tuned for semantic similarity search (Reimers & Gurevych, 2019).

Probabilistic topic modeling approaches are one further well-known and widely used tool for extracting and analysing latent themes behind the underlying unstructured text data in different areas. Bianchi et al. (2021) introduce a novel approach to topic modeling for the multilingual setting. Multilingual contextualized topic modeling (MCTM) allows to train a topic model in one model and to infer topic distributions for documents in unseen languages just relying on their SBERT representations.

Further details on methods used in this paper are provided in E.1.

## 7.3 Data

For the empirical analysis, three datasets of news articles in three different languages are used: DER SPIEGEL for Germany, Lenta.ru for Russia, and UNIAN for Ukraine. The following preprocessing steps were taken: remove punctuation, numbers, special characters, stopwords and lowercase. The final datasets contain 833,454 articles in the period from January 2000 to September 2020 for Germany, 864,481 articles in the period from September 1999 to September 2020 for Russia, and 785,750 articles in the period from January 2007 to September 2020 for Ukraine. Economic activity is measured by industrial production index, which is often used in academic literature as a high-frequency indicator of a country's economic activity.

## 7.4 Construction of Uncertainty Indices

Overall, four different approaches are proposed to identify articles related to uncertainty in economic policy. The first approach is a dictionary based one, which uses fastText multilingual word embeddings to either identify three term sets referring to the three components of the EPU concept (later referred to as *dic*1) or one combined term set that should describe the EPU concept as a whole (later referred to as *dic*2). While the former method searches for nearest neighbors to the three terms *economic*, *policy*, and *uncertainty*, the latter makes use of

[1] fastText is a free library for text classification and representation learning.

the additive feature of the word vectors and searches for nearest neighbors to the compound word vector *economic + policy + uncertainty*. The similarity of the word vectors is measured by cosine similarity.[2] The number of relevant words is controlled for automatically based on a threshold, namely the 99.99% percentile of all the cosine similarity values between a certain word in one language (e.g. English word "policy") and all the word vectors available in other language (e.g. German) as shown in Figure 7.1. The identified EPU related terms are presented in E.2.

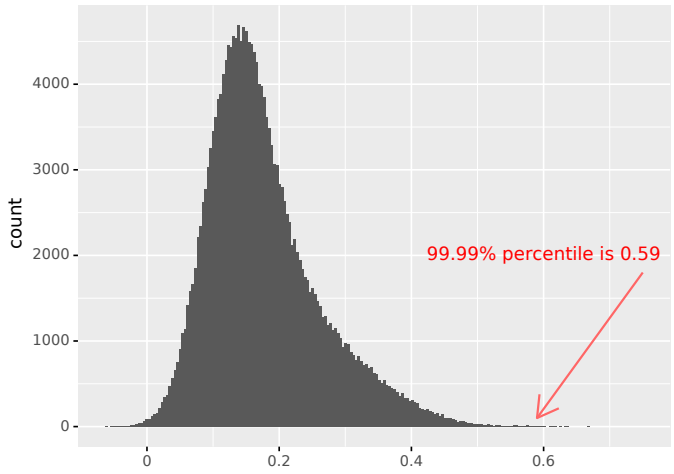



**Figure 7.1:** Cosine Similarity Values between *policy* and all German words

The second approach considers the articles as Bag-of-Words and represents them as the sum of the constituent words vectors, which results in aggregated document embeddings (later referred to as *art_emb* approach). The third approach applies Transformer based text embeddings to identify articles that relate to EPU (later referred to as *art_sbert*).[3] Finally, a novel language agnostic topic modeling technique called zero-shot cross-lingual topic modeling is applied. Thereby, a topic model was trained based on German SBERT embeddings of articles and the topic distributions for Russian and Ukrainian articles were then also inferred based on SBERT embeddings of articles. EPU related topics have been identified using the embeddings of the most frequent topic words. In the following, this approach is referred to as *MCTM_{k}_Topic*. *k* can stand for the topic number of a topic that is identified as an EPU related topic or have the designation *combined*, if the average topic frequency of all the EPU related topics is used. Overall, 10 different EPU time series are provided for each country.

[2] Cosine similarity is defined as the cosine of the angle between two vectors. The values range from -1 to 1. A cosine similarity value of 1 means that the vectors are pointing in the same direction.

[3] To train SBERT articles' embeddings, a pre-trained `distiluse-base-multilingual-cased-v2` model was used. Thereby, Python's implementation of SBERT, namely *Sentence-Transformers*, was used to load and apply the model.

## 7.5 Results

### 7.5.1 EPU Indices

This section presents the constructed indices. All the indices were normalized to have a mean of 100 and a variance of 1.

Figure 7.2 shows the indices resulting from the *dic*1 and *dic*2 approaches. In Germany, the peaks correspond with such events as the September 11 attacks, economic crisis in Germany, global financial crisis, and Corona virus outbreak. The spikes of the EPU in Russia between 2004 and 2005 as well as in the period from 2014 and 2018 could be explained by the Orange Revolution in neighbouring Ukraine and the Russia-Ukraine gas disputes, and by the Crimean crisis and the War in Donbass, respectively. Surprisingly, both of the constructed EPU indices for Ukraine show a downward trend. Some peaks can be identified at the beginning of 2007 (political crisis in Ukraine), in 2008-09 (global financial crisis), 2014 (beginning of the Crimean crisis and the War in Donbass), and at the beginning of 2019 (presidential and parliamentary elections). The Corona virus outbreak, instead, seems to have caused a relatively small increase in the uncertainty index. As this approach largely relates to that proposed by Baker et al. (2016), further analyses between the constructed indices and the available BBD indices have been carried out (see E.5).

Figures 7.3 and 7.4 show the *art_emd* and *art_sbert* indices for all three countries, respectively. There are some noticeable differences between the two, as for example, stronger interdependencies between Russia and Ukraine according to the *art_sbert* approach.

Finally, based on the results of the multilingual topic modeling five EPU related topics are identified in the current application. These are presented in Figure 7.5. Based on the qualitative assessment of the most common words of the topics, these were assigned the following labels: `government`, `stock market`, `political parties`, `elections`, `U.S. political leaders`. The values in brackets represent the cosine similarity values to the EPU embedding. The corresponding time series for each country are presented in E.3. For an additional robustness check for the U.S. data see E.4.

### 7.5.2 VAR Models

All the constructed EPU indices are tested within vector autoregressive (VAR) models with regard to their impact on the economic activity of the countries. The economic activity is measured by the industrial production index, which is often used in academic literature as a high-frequency indicator of a country's economic activity (Baker et al., 2016; Perić & Sorić, 2018; Čižmešija et al., 2017).[4]

For each country, 10 two-dimensional VAR models with seasonal dummies were estimated, each including one of the constructed EPU indices and the corresponding industrial production index.[5]

[4] The data for the analyses come from State Statistics Service of the considered countries.

[5] According to the performed stationarity tests, all the variables needed to be transformed to

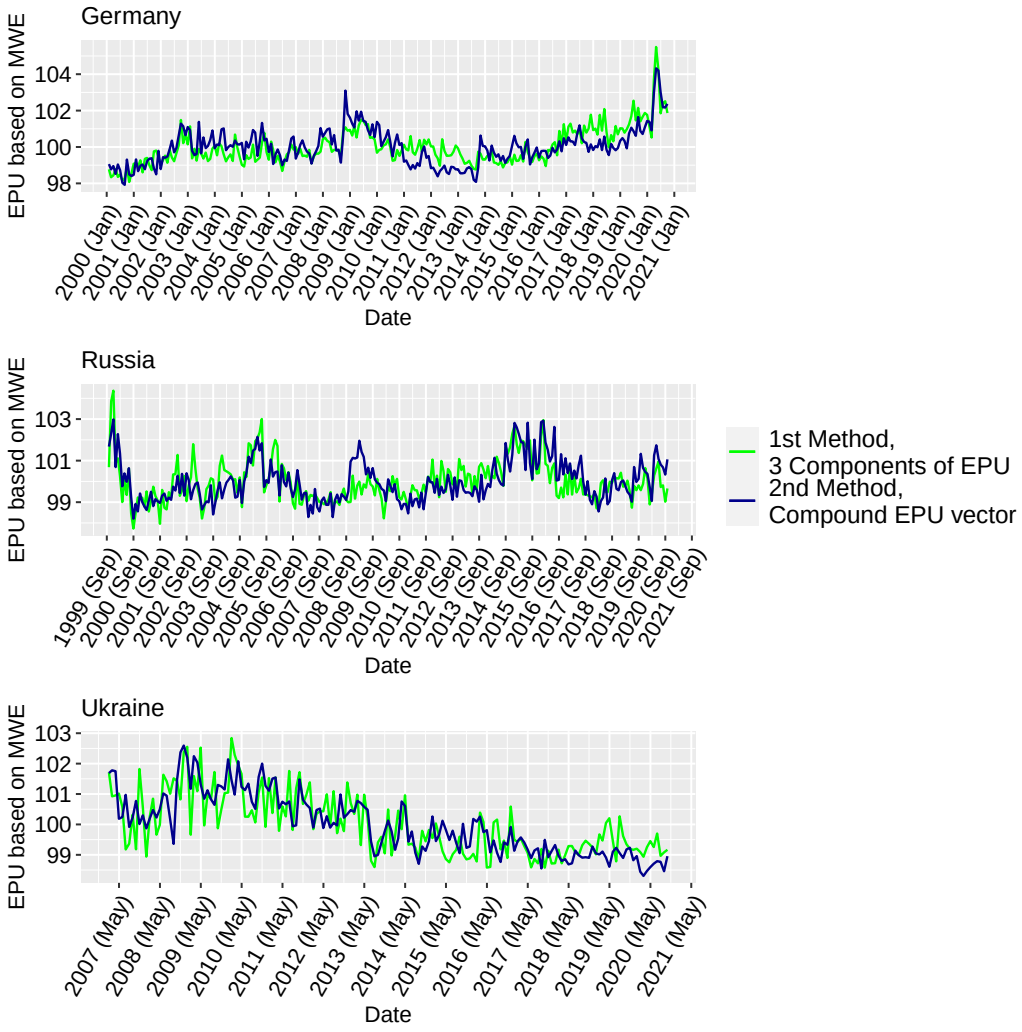



**Figure 7.2:** *dic*1 and *dic*2 EPU Indices

### Granger causality

The first set of analysis is dedicated to Granger causality tests, especially the null hypothesis "EPU does not Granger cause Industrial Production Index". According to Granger causality tests, the following EPU indices led to significant results at least in one country: *dic*1 (Ukraine), *dic*2 (all countries), *art_emb* (Germany, Ukraine), *art_sbert* (Germany), *MCTM_stock_market_Topic* (all countries).[6] The results of Granger causality test are summarized in E.6.

### Impulse Response Functions

A close look is taken on the *dic*2 EPU index, as this index has proved to be Granger causal to economic activity in all the considered countries. Figure 7.6 illustrates the responses of the industrial production indices to an *dic*2 EPU indices shock in all considered countries.

become stationary. For this reason, the first log differences of all the variables were calculated and used in all estimated VAR models.

[6] Results were considered significant if p-value is smaller than 10%.

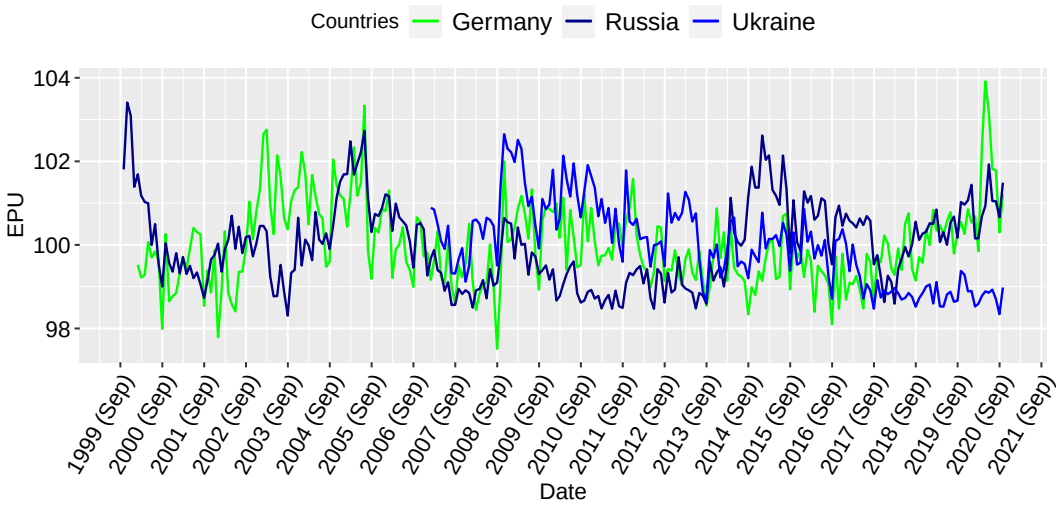



**Figure 7.3:** *art_emd* EPU Indices

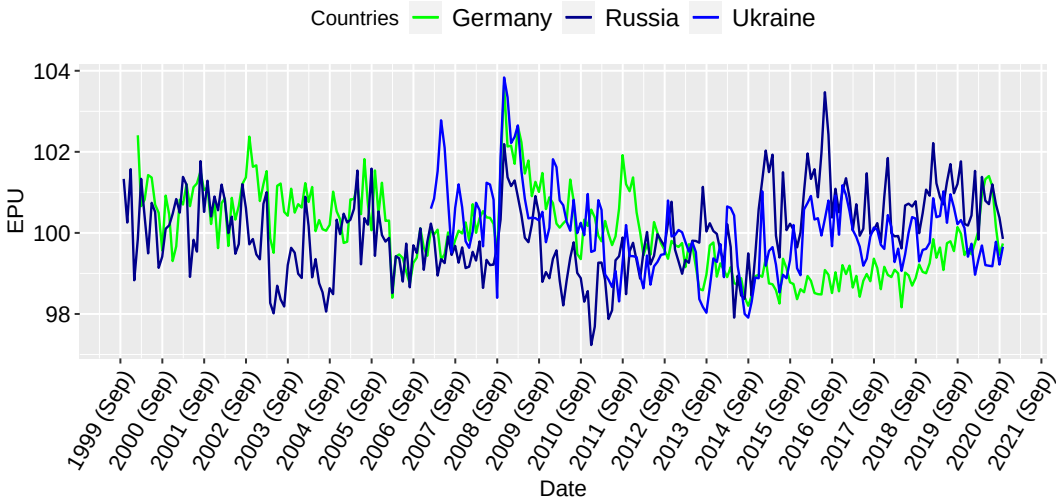



**Figure 7.4:** *art_sbert* EPU Indices

The shaded areas represent the 95% confidence bands. Thereby, the orthogonal impulse responses are considered meaning that contemporaneous effects are allowed. It can be inferred from the figure that one standard deviation shock in EPU leads to a significant drop of 0.1 and 0.44 percentage points in the industrial production index after one month in Germany and Russia, respectively. While in Germany there is only a short-term impact of the EPU shock on the industrial production, there is also a long-term significant negative impact of the EPU shock on the industrial production index (about 0.3 percentage points) in Russia. The pattern of the impulse response function (IRF) in Ukraine is similar but not significant over the entire period.

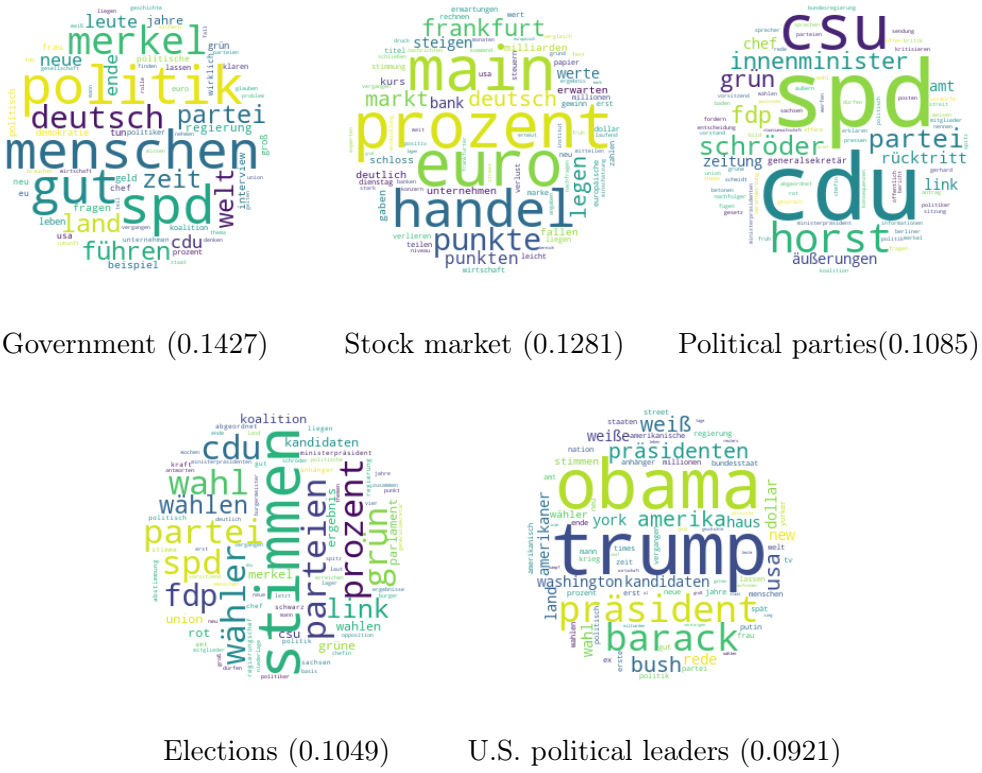



**Figure 7.5:** MCTM with 40 Topics: EPU Related Topics

## 7.6 Conclusions

One of the key finding of the current work is that the dictionary based approach in combination with multilingual word embeddings results in indices that are Granger causal to the economic activity in all of the considered countries. Further, to the best of my knowledge, the current paper is the first to apply a novel language agnostic topic modeling technique introduced by Bianchi et al. (2021) in economic context. One of the identified EPU related topics has proved to Granger cause the economic activity in all of the considered countries. This is one promising finding as the topic modeling approach by Bianchi et al. (2021) allows to predict topic distributions for texts in unseen languages just based on SBERT embeddings without renewed training of the model. Finally, it has been also shown that a sudden shock in the constructed EPU indices leads to significant short-term and/or long-term declines in industrial production. This finding indicates that the constructed EPU indices could be used as high frequency indicators of economic activity.

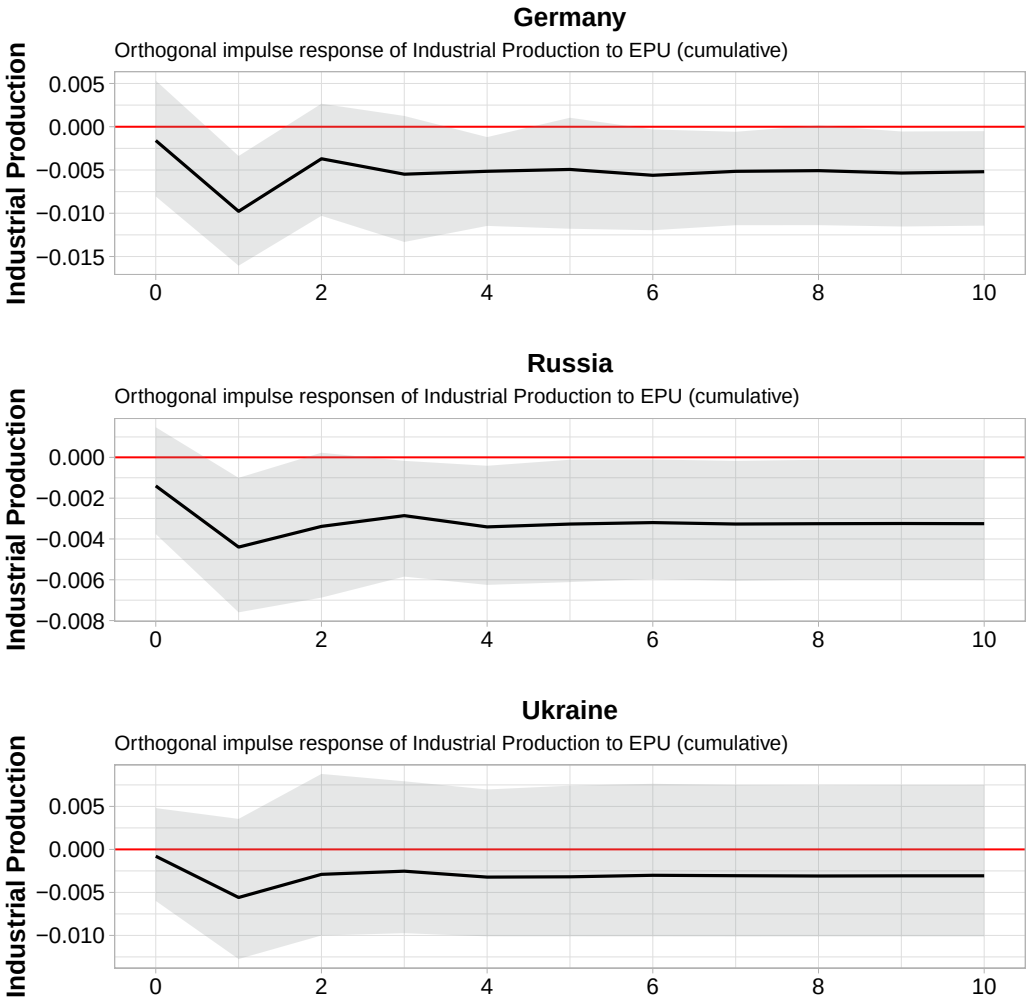
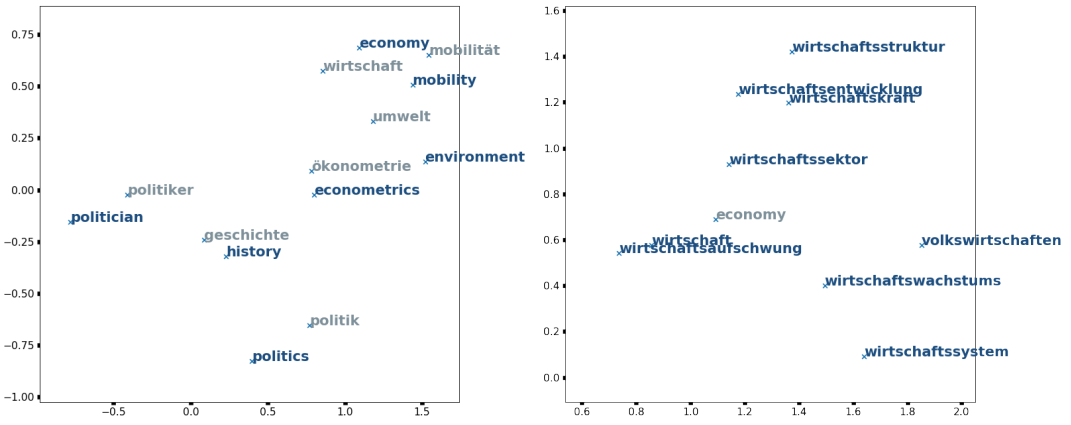



**Figure 7.6:** IRFs: *dic*2 EPU Index →Industrial Production Index

# Appendix E

## E.1 Methods

All existing text representation techniques aim for efficiency in the representation and understanding of natural language. All the techniques aim to capture different features of a word, a sentence or a document and represent these in a machine-readable form, i.e. as a vector, also called embedding. Figure E.1.1 provides an example of such multilingual word embeddings (MWE). The nearest English neighbours to selected German terms in a multilingual vector space were found based on cosine similarity between vectors. This led to the following pairs: *wirtschaft-economy*, *politik-policy*, *umwelt-environment*, *mobilität-mobility*, *ökonometrie-econometrics*, *geschichte-history*, *politiker-politician* (Figure E.1.1a). In Figure E.1.1b, the nearest German neighbours of the word *economy* are presented, e.g. *wirtschaft*, *wirtschaftsaufschwung*, *wirtschaftsektor*, *wirtschaftskraft*, *wirtschaftsentwicklung*. These examples illustrate the ability of MWE to capture the meaning of the words and find semantically similar words across languages.




**(a)** Nearest neighbours of selected terms **(b)** Nearest neighbours of *economy* in German

**Figure E.1.1:** Example of Multilingual Word Embeddings provided by fastText library, dimensions reduced using Principal Component Analysis

This paper is dedicated to a specific problem of measuring *economic policy uncertainty* (EPU). The use of MWE should automate, accelerate, and simplify the process of finding relevant terms describing the concept of EPU in different languages.

A great breakthrough in and a major contribution to the field of language model learning

has been made with the publication of the work by Devlin et al. (2019). The architecture builds on top of the Transformer model first introduced in 2017 by Vaswani et al.. An in-depth explanation of the Transformer model is out of scope of this work. However, a brief summary of the main characteristics is provided in the following.

As the original application was the machine translation, one can think of the Transformer as a black box (see Figure E.1.2) that receives the input text in one language and outputs the translation in another language (Alammar, 2018). There are two basic mechanisms behind the Transformer: encoder and decoder stacks that consist of several encoder and decoder blocks, respectively. Further, each encoder block consists of two layers: feed forward neural network (FFNN) and Self-Attention. The latter is especially important, as it allows the encoder to look at all the input words, for example, to predict the end of a sentence or the next word. The information from the Self-Attention layer flows then to the FFNN. Similar to the encoder blocks, each decoder block also consists of the two mentioned layers and one additional layer, Encoder-Decoder Attention.

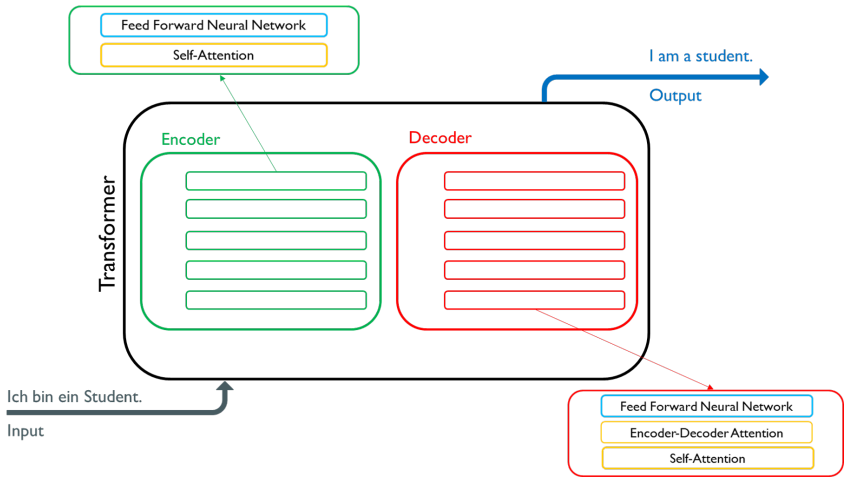



**Figure E.1.2:** The Transformer's Structure. Adapted from: Alammar (2018)

BERT is a multi-layer Transformer encoder containing 12 and 24 encoder blocks in the base and the large model, respectively (Devlin et al., 2019). BERT was trained to perform various downstream tasks such as sentence classification or sentiment analysis. For these examples, the input would be just a sentence (a movie review) and the output would be the sentence's class like "spam" or "not spam" (a sentiment label like "positive" or "negative"). To enable performing such downstream tasks, a general model is trained in the pre-training phase (predicting a masked token and next sentence prediction) and some of the parameters can be adjusted to a specific task (e.g. sentence classification or sentiment analysis) in the fine-tuning phase.

Although BERT delivers high performance on different NLP tasks, Reimers & Gurevych (2019) address one considerable drawback of the model. The authors give an example of simple semantic textual similarity task of finding the most similar sentence pair out of 10,000 sentences, which takes over 50 million inference computations and a lot of time (about 65

hours) using BERT. To accelerate the process of *semantic search*, Reimers & Gurevych (2019) developed SBERT. The idea behind the proposed approach is to map the input sentences to a shared vector space that captures the semantic relationships between the sentences. Such sentence embeddings allow then to measure the similarity based on cosine similarity or Euclidean distance. Although the SBERT embeddings were not trained to perform downstream tasks in the first place, they seem to capture not only the semantic relatedness of the sentences, but also the underlying sentiment of the sentences (Reimers & Gurevych, 2019). Reimers & Gurevych also extended their model to the multilingual setting in a follow-up project. The authors highlight that the proposed approach allows meaningful representations of sentences in several languages from various language families (for more details see Reimers & Gurevych (2020)). Python's library implementation of SBERT, namely the *Sentence-Transformers*[7] module provides several pre-trained multilingual models that were trained for different use cases. In the current work, `distiluse-base-multilingual-cased-v2` model, which was fine-tuned to find semantically similar sentences across languages, was used. This model was chosen because the focus in the current application is on semantic similarity between the search query ("economic policy uncertainty") and all the articles in different languages. According to the module's documentation, embeddings can be calculated not only for sentences but also for short phrases (here: "economic policy uncertainty") or for longer texts containing multiple sentences (here: news articles).

Overall, BERT based models have been used in several applications and the number of implementation possibilities is still growing. This concerns not only traditional natural language processing (NLP) tasks. The interest in sentence and text representations is also increasing in other fields. This work also uses multilingual SBERT embeddings in order to capture the semantics of the considered news articles.

These contextualized vectors have attracted attention in other text mining techniques, e.g. in topic modeling. In 2021, Bianchi et al. introduce multilingual contextualized topic modeling, a richer version of topic modeling that involves contextualised representations of the documents. Contextualised SBERT document representations are used in the multilingual setting to uncover latent topic structure behind the data (Bianchi et al., 2021).However, Bag-of-Words (BoW) representations of the documents in the source language are still used for the visualisations of the discovered topics. Let us consider the following example. There are two distinct datasets in German and English. SBERT representations of all the documents in both datasets can be obtained using one of the pre-trained multilingual models. First, one can train a MCTM on the German collection of texts passing their BoW and SBERT representations to the model. In the next step, the topic distributions for the unseen English documents can be inferred from the trained model, even though the model was trained on the German dataset. This process is referred to as zero-shot topic modeling. Thus, MCTM is fully language independent and the topic distribution can be predicted for each document

[7] The complete documentation and examples are available on `https://www.sbert.net/index.html`, last accessed on 28.02.2024.

given its contextualised SBERT representation. In quantitative and qualitative analyses, Bianchi et al. (2021) have shown that the proposed MCTM approach leads to coherent topics and allows meaningful assignment of the documents in unseen languages to trained topics in another language.

## E.2 EPU Terms

| | |
|---|---|
| Economic Terms | "economic" or "economy" |
| Uncertainty Terms | "uncertainty" or "uncertain" |
| Policy Terms | "Congress", "deficit", "Federal Reserve", "legislation", "regulation" or "White House" |

**Table E.2.1:** EPU Terms by Baker et al. (2016)

| | |
|---|---|
| Economic Terms | "wirtschaft" or "wirtschaftlich" |
| Uncertainty Terms | "unsicher" or "unsicherheit' |
| Policy Terms | "steuer" or "wirtschaftspolitik" or "regulierung" or "regulierungs" or "ausgaben" or "bundesband" or "ezb" or "zentralbank" or "haushat" or "defizit" or "haushaltsdefizit" |

**Table E.2.2:** EPU Terms for Germany by Baker et al. (2016)

| | |
|---|---|
| Economic Terms | "экономика" (economy) |
| Uncertainty Terms | "неопределённый" (uncertain) or "неопределённость" (uncertainty) |
| Policy Terms | "политика" (policy), "налог" (tax) |

**Table E.2.3:** EPU Terms for Russia by Baker et al. (2016)

| | |
|---|---|
| Economic Terms | "wirtschaftliche" (0.7934), "ökonomische" (0.7865), "wirtschaftspolitische" (0.7746), "wirtschaftsentwicklung" (0.7525), "volkswirtschaftliche" (0.7461), "gesamtwirtschaftliche" (0.7239), "wirtschaftswachstums" (0.7187), "marktwirtschaftliche" (0.717), "weltwirtschaft" (0.7098) |
| Policy Terms | "policy" (0.668), "informationspolitik" (0.6379), "richtlinienkompetenz" (0.6344), "ausländerpolitik" (0.6234), "grundsatzentscheidungen" (0.6161), "richtlinien" (0.6145), "neutralitätspolitik" (0.612), "ordnungspolitik" (0.6107), "wirtschaftspolitik" (0.6106), "rechtspolitik" (0.603), "beschäftigungspolitik" (0.5976), "industriepolitik" (0.5962), "migrationspolitik" (0.5939), "währungspolitik" (0.5934), "deutschlandpolitik" (0.5922), "regierungspolitik" (0.5921), "gesellschaftspolitik" (0.5914), "verteidigungspolitik" (0.5913), "politik" (0.5907) |
| Uncertainty Terms | "unsicherheit" (0.7323), "ungewissheit" (0.6751), "messunsicherheit" (0.6411), "eintrittswahrscheinlichkeit" (0.6298), "zufälligkeit" (0.6138), "wahrscheinlichkeiten" (0.6022), "zeitlichkeit" (0.5974), "messbarkeit" (0.5973), "voraussagen" (0.5956), "unbestimmtheit" (0.5943), "wahrscheinlichkeitsverteilung" (0.5877), "ausfallwahrscheinlichkeit" (0.5875), "erwartungswerte" (0.5846), "relativität" (0.5846), "berechenbarkeit" (0.5792), "bestimmtheit" (0.5787), "allgemeingültigkeit" (0.578) |

**Table E.2.4:** EPU Terms and their Cosine Similarity Values for Germany based on *dic*1 approach

| | |
|---|---|
| Economic Terms | “экономического” (economic, 0.7818), “экономик” (economies, 0.7466), “макроэкономических” (macroeconomics, 0.7326), “социально” (socially, 0.6664), “рыночных” (market, 0.6537) |
| Policy Terms | “небюрократия” (non-bureaucracy, 0.5588), “правила” (rules, 0.5532), “законодательства” (legislation, 0.547), “политике” (policy, 0.5465), “рекомендаций” (recommendations, 0.532), “ужесточении” (tightening, 0.5308), “неприемлемы” (unacceptable, 0.5215), “лоббизм” (lobbying, 0.5181), “ивп” (iwp, 0.5164), “недопустимости” (inadmissibility, 0.5147), “правилах” (rules, 0.5134) |
| Uncertainty Terms | “неопределённость” (uncertainty, 0.6886), “относительность” (relativity, 0.6237), “определённость” (certainty, 0.618), “противоречивость” (inconsistency, 0.5964), “неясность” (ambiguity, 0.5875), “неуверенность” (uncertainty, 0.5655), “неизбежность” (inevitability, 0.5632), “адекватность” (adequacy, 0.5543), “закономерность” (regularity, 0.5535), “субъективность” (subjectivity, 0.5494), “обоснованность” (validity, 0.5483), “согласованность” (consistency, 0.5482) |

**Table E.2.5:** EPU Terms and their Cosine Similarity Values for Russia based *dic*1 approach

| | |
|---|---|
| Economic Terms | “економіч” (economic, 0.7607), “економічна” (economic, 0.7348), “економік” (economies, 0.7208), “макроекономічної” (macroeconomics, 0.6612), “економіці” (economics, 0.645), “підприємництва” (entrepreneurship, 0.6332), “зовнішньополітична” (foreign policy, 0.6292), “зовнішньоекономічна” (foreign economic, 0.6281) |
| Policy Terms | “політики” (politics, 0.6205), “політикуму” (politicum, 0.5964), “зовнішньополітичний” (foreign policy, 0.5707), “законодавства” (legislation, 0.5596), “етнополітики” (ethnopolitics, 0.5475), “геополітики” (geopolitics, 0.5341), “політизації” (politicisation, 0.5284), “політиці” (politics, 0.5268), “мінагрополітики” (agricultural politics, 0.5228) |
| Uncertainty Terms | “невизначеність” (uncertainty, 0.6846), “невизначеності” (uncertainty, 0.6778), “визначеність” (certainty, 0.6074), “ймовірності” (probability, 0.5939), “імовірності” (probability, 0.5836), “визначеності” (certainty, 0.5833), “непередбачуваність” (unpredictability, 0.5774), “спостережуваного” (observable, 0.5761), “невизначеного” (uncertain, 0.574), “вірогідності” (probability, 0.5724), “неясність” (ambiguity, 0.5677), “взаємозалежність” (interdependence, 0.5647), “передбачуваного” (predictable, 0.5603), “нестабільність” (instability, 0.5573), “ймовірнісні” (probabilistic, 0.5545), “відносність” (relativity, 0.5545), “суперечливість” (inconsistency, 0.5542), “імовірність” (probability, 0.5523) |

**Table E.2.6:** EPU Terms and their Cosine Similarity Values for Ukraine based on *dic*1 approach

| | |
|---|---|
| Germany | "wirtschaftspolitische" (0.7891), "marktwirtschaftliche" (0.748), "industriepolitik" (0.7362), "beschäftigungspolitik" (0.7326), "währungspolitik" (0.7243), "konjunkturpolitik" (0.724), "fiskalpolitik" (0.7188), "gesellschaftspolitik" (0.7186), "entwicklungspolitik" (0.7123), "steuerpolitik" (0.7107), "ökonomische" (0.7105), "volkswirtschaftliche" (0.7105), "finanzpolitik" (0.7093), "regierungspolitik" (0.7054) |
| Russia | "экономического" (economic, 0.7026), "макроэкономических" (macroecnomic, 0.6971), "экономик" (economics, 0.6913), "либерализации" (liberalisation, 0.6514), "внешнеполитическая" (foreign policy, 0.6466), "внешнеэкономическая" (foreign economic, 0.6464), "предпринимательства" (entrepreneurship, 0.6431) |
| Ukraine | "зовнішньополітична" (foreign policy, 0.689), "економіч" (economic, 0.6869), "економічна" (economic, 0.6793), "економік" (economics, 0.6734), "макроекономічної" (macroeconomics, 0.6606), "політика" (policy, 0.6547), "зовнішньоекономічна" (foreign economic, 0.6389) |

**Table E.2.7:** EPU Terms and their Cosine Similarity Values for Germany, Russia, and Ukraine based on *dic*2 approach

## E.3 Topics Time Series

Figures E.3.3, E.3.4, and E.3.5 show the corresponding topic time series of the selected EPU related topics for Germany, Russia, and Ukraine, respectively. The horizontal red dashed lines represent the average topic probability (2.5%) if all of the 40 topics were equally distributed. In Germany, for example, the time series of `stock market` topic experience a considerable increase up to 9% in the period from the end of 2000 to 2002 that could be explained by the dot-com bubble. Time series of `elections` topic constantly fluctuates throughout the considered time period with some considerable spikes around major political events like parliamentary elections (2002, 2005, 2009. 2013). Finally, major spikes of the figure for the relative importance of `U.S. political leaders` are probably associated with the presidential elections in the USA.

At this point, it should be highlighted that BoW representations of the German texts are only used to visualize topics that were trained based only on SBERT embeddings. It means that topics were trained based on semantic similarity of the underlying texts. While these topic representations can be meaningfully interpreted for the German dataset, as it was used for training, one should be careful when describing inferred topic distributions for Russia and Ukraine. For example, Russian or Ukrainian articles assigned to `political parties` topic most probably do not only report on German political parties but also on domestic key political players. The analysis of the Ukrainian articles assigned to this topic revealed that they mainly report on the key events in Verkhovna Rada (unicameral parliament of Ukraine) and major political Ukrainian parties and coalitions. With regard to other EPU related topics, it has been observed that articles with the highest probability for `government` and `elections` topics discuss political themes in general, for `stock market` topic - movements of stock markets at home and worldwide, for `U.S. political leaders` topic - political leaders of the USA.

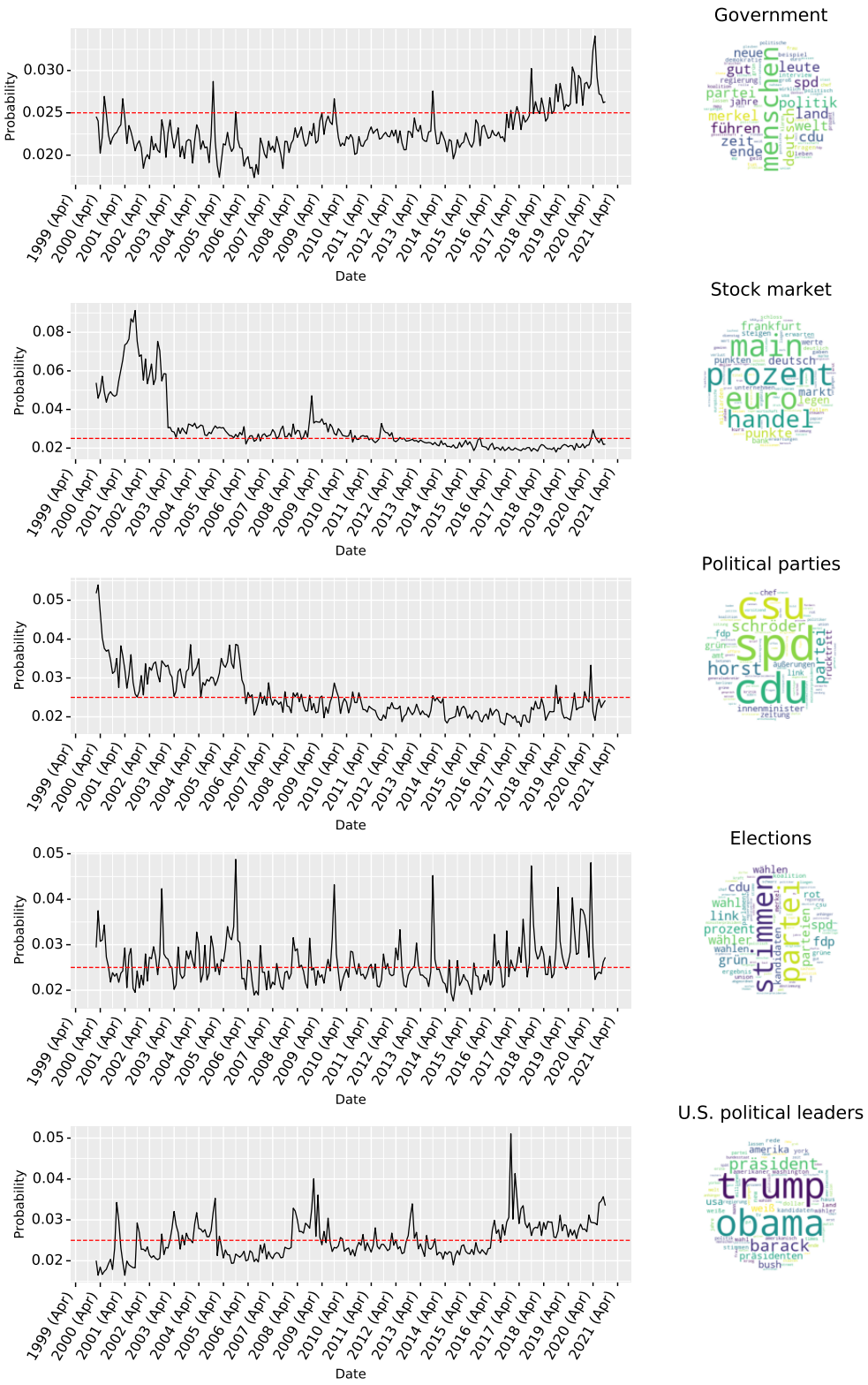



**Figure E.3.3:** Time Series of Selected EPU Topics in Germany

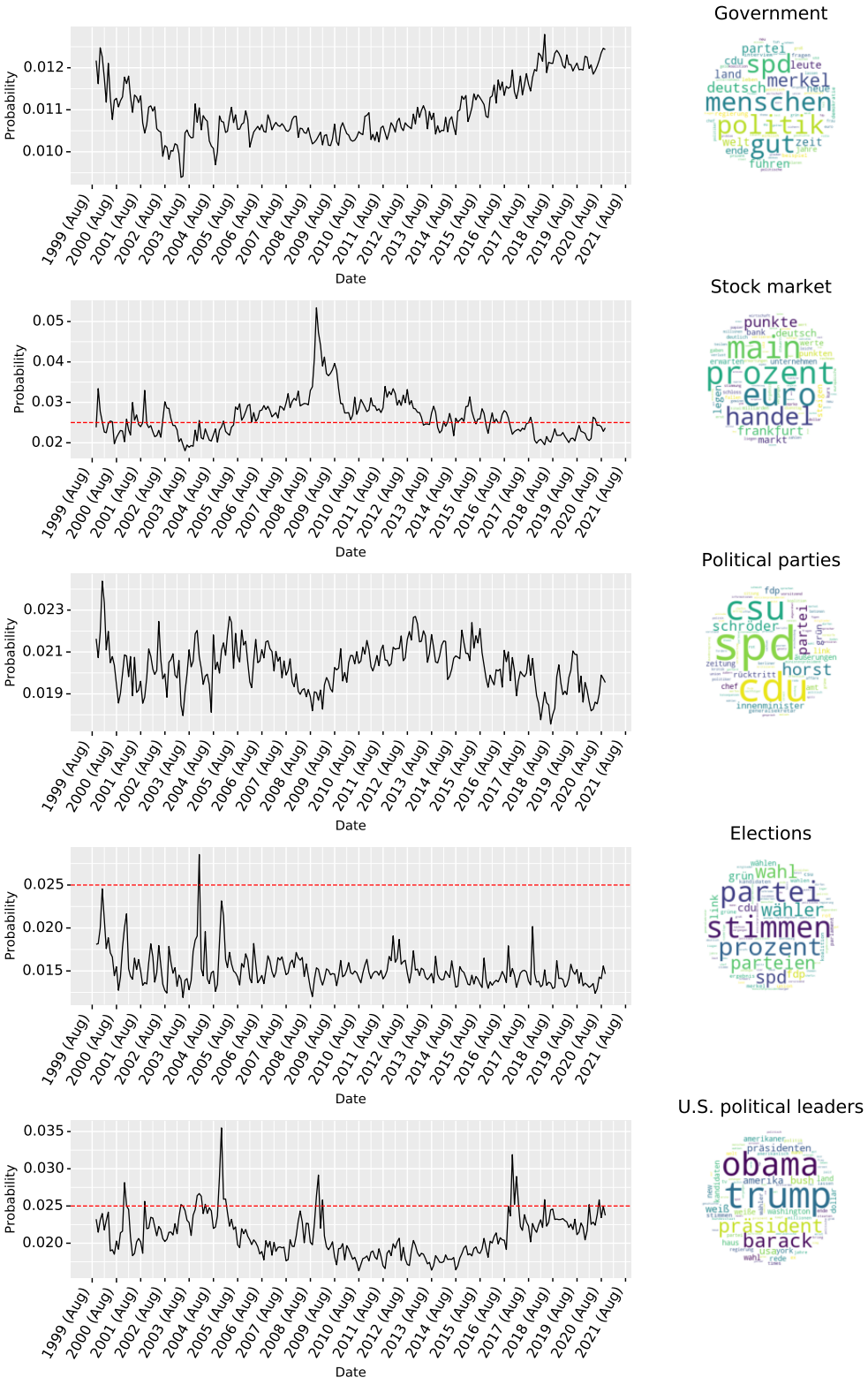



**Figure E.3.4:** Time Series of Selected EPU Topics in Russia

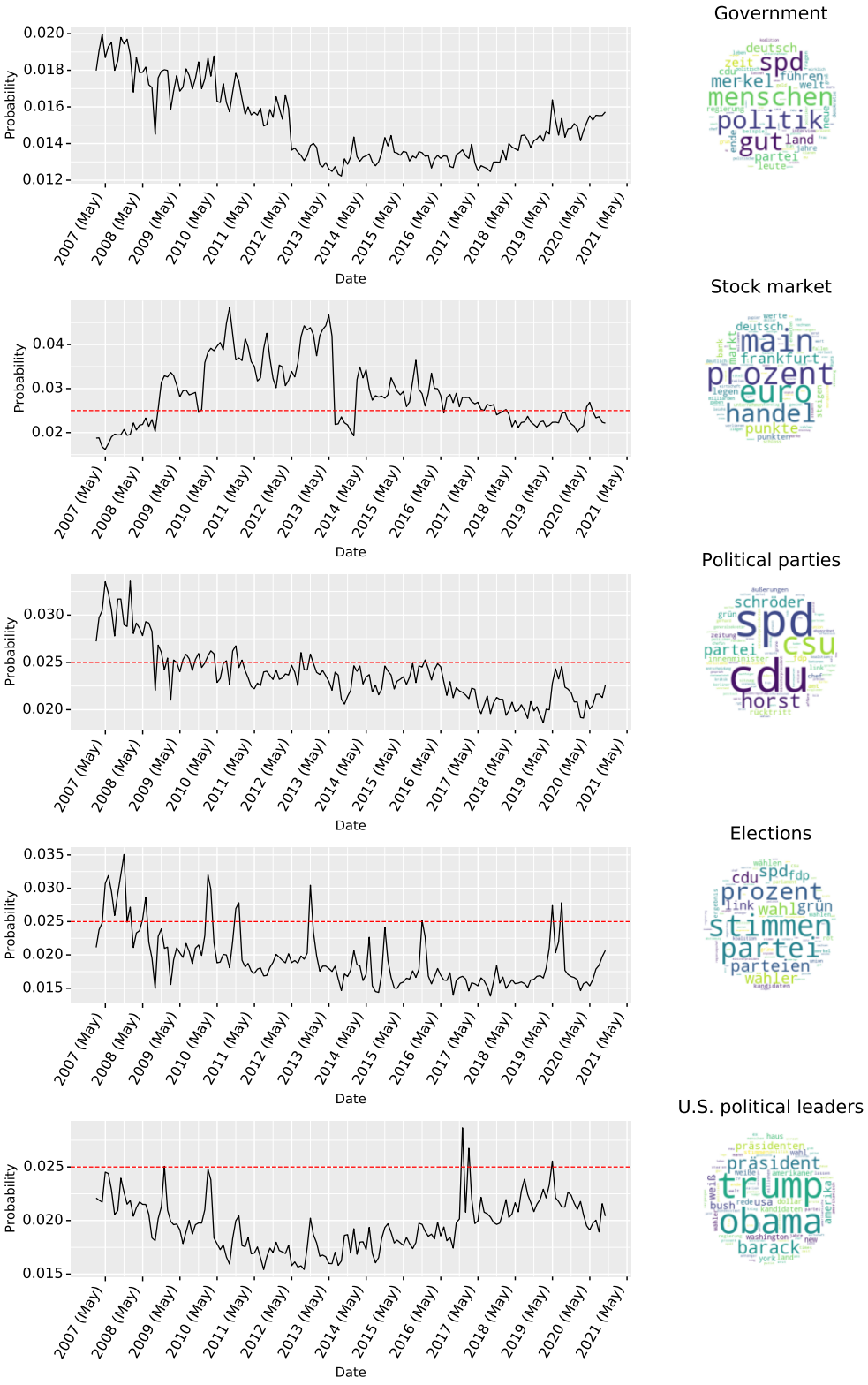



**Figure E.3.5:** Time Series of Selected EPU Topics in Ukraine

## E.4 Robustness check: Cross-lingual Assignments for U.S. Data

To further demonstrate the features of the applied zero-shot topic modeling, a robustness check was performed for the U.S. To do so, a small subset of articles from *The Washington Post* was collected. The data cover the period of March 2003 and contain a total of 3,156 articles. In a first step, SBERT articles' representations were obtained using a pre-trained model, namely `distiluse-base-multilingual-cased-v2`. Afterwards, the previously trained multilingual topic model was used to infer topic weights for these unseen articles in English. To analyse the quality of the assignments, for each article the most prevalent topic was determined (topic with the highest weight for the given article). As `stock market` topic proved to be Granger causal to economic activity in all the considered countries in the current study, this topic was also examined in more detail for the U.S. data. The top words describing this topic (translated from German) are, for example, *percentage*, *trade*, *market*, *points*, *bank*, *loss*, *rise*, *price*, *fall*, *expectations*, *sentiment*, *economy*, *million*, *profit* etc. The article from the U.S. dataset with the highest probability for this topic (72%) is the one with the title *"Stocks Rise On Positive News About Economy"* from March 1, 2003. The content of the article is as follows:

*Section: Business, E03 NEW YORK, Feb. 28 – Stocks rose again today as better-than-expected reports on economic growth and Midwest manufacturing boosted investor optimism that corporate profits will recover. "There's underlying strength in the economy, and that's supporting equities," said Charles White, president of Avatar Associates, which oversees $1.7 billion. Intel Corp. led the advance after a Lehman Brothers analyst said the semiconductor maker is benefiting from increased demand overseas and higher prices. The Standard & Poor's 500-stock index rose 3.87 points, or 0.5 percent, to 841.15. Computer-related stocks contributed half the gain. The Dow Jones industrial average climbed 6.09, or 0.1 percent, to 7891.08. The Nasdaq composite index advanced 13.58, or 1 percent, to 1337.52. Today's economic reports helped offset concern about a war with Iraq that drove the S&P to a third straight monthly decline. The benchmark fell 1.7 percent in February, extending its year-to-date drop to 4.4 percent. For the week, the S&P 500 was down 0.8 percent, the Dow 1.6 percent and the Nasdaq 0.9 percent. The gross domestic product grew at a 1.4 percent annual rate in the fourth quarter as consumer and business spending was higher than originally estimated, the Commerce Department said. Chicago purchasing managers' regional manufacturing index was 54.9 in February. Economists had predicted a reading of 52.5, according to a Bloomberg survey. A reading above 50 indicates growth. The University of Michigan's final index of consumer sentiment climbed to 79.9 in February from an initial 79.2. Intel rose 56 cents, to $17.26, after Lehman analyst Daniel Niles boosted his first-quarter profit by one penny to 13 cents a share and lifted his 2003 target to 63 cents from 60 cents. He rates the stock "overweight." Computer-related shares gained with Intel. Microsoft Corp. advanced 27 cents, to $23.87. Cisco Systems Inc. rose 23 cents, to $13.98.*

*Dell Computer Corp. climbed 45 cents, to $26.96. The Dow slipped 2 percent this month while the Nasdaq gained 1.3 percent amid fluctuating investor perceptions about the likelihood of war. "Most investors and portfolio managers are still taking a wait-and-see approach on the potential conflict in Iraq," said Gene Pisasale, senior investment officer at Wilmington Trust Co., which manages more than $25 billion. "I don't think we'll see a clear market trend until that's resolved." First Horizon Pharmaceutical Corp., which makes drugs to treat chronic diseases, tumbled $3.74 to $2.06. The company slashed its earnings forecast by almost 75 percent on slower sales of its Sular high-blood-pressure drug and Tanafed DP pediatric cold medicine. Novell Inc., whose programs manage computer networks, slid 48 cents to $2.60. The company reported a fourth-quarter loss on failed investments and falling sales. Symantec Corp., whose software deflects computer viruses, tumbled $6.89 to $40.47. From News Services The New York Stock Exchange composite index rose 22.54, to 4716.07; the American Stock Exchange index rose 2.43, to 830.63; and the Russell index of 2,000 small stocks fell 0.91, to 360.52. Advancing issues outnumbered declining ones by 4 to 3 on the NYSE, where trading volume rose to 1.3 billion shares, from 1.27 billion on Thursday. On the Nasdaq, advancers slightly outnumbered decliners and volume totaled 1.29 billion shares, up from 1.25 billion. The price of the Treasury's benchmark 10-year note rose $4.06 per $1,000 invested, and its yield fell to 3.69 percent, from 3.74 percent late Thursday. The dollar rose against the Japanese yen and fell against the euro. In late New York trading, a dollar bought 118.10 yen, up from 117.58 yen late Thursday, and a euro bought $1.0799, up from $1.0763. Light, sweet crude oil for April delivery settled at $36.60 a barrel, down 60 cents, on the New York Mercantile Exchange. Gold for current delivery rose on the Commodity Exchange division of the New York Mercantile Exchange to $350.20 a troy ounce from $346.10 on Thursday.*

Further articles' titles whose prevalent topic is `stock market` and the probability values are quite high (the values are given in the brackets) are presented in the following:

*"World Markets Continue To Rally" (60%), March 22, 2003*
*"Mortgage Foreclosures Up; Delinquencies Down" (38%), March 25, 2003*
*"ManTech's Revenue Rises, Driven by Security Demand" (34%), March, 2003*
*"Shaken by Uncertainty" (33%), March 30, 2003*
*"Stocks Soar on Verge of War" (33%), March 18, 2003*
*"VSE Shares Tumble 18% as Search Ends for Buyer" (32%), March 10, 2003*
*"Stock Indexes Retreat From War" (32%), March 11, 2003*

These example show that the unseen articles in English could be meaningfully assigned to the pre-trained topics. All the articles that show high probability for the `stock market` topic seem indeed deal with national and international stock market movements.

Some of the articles appeared to deal predominantly with the `elections` topic. The three titles of these articles are (probability values in the brackets):

*"Poll: Voters Divided Over Mayor" (22%), March 6, 2003*

*"Polls Open Today for Arlington Vote" (19%), March 11, 2003*
*"D.C. Democrats Put Off Deciding on Primary Date" (12%), March 26, 2003*

Also `U.S. political leaders` topic appeared in some articles with relatively high probabilities:

*"Clinton, Dole Ready for 120 Seconds" (16%), March 7, 2003*
*"Even With a War, Business as Usual" (16%), March 20, 2003*
*"Countering a Trend, Dodd Won't Join Democratic Field" (15%), March 4, 2003*
*"Political Reality: 'Clinton & Dole' No Match for a Sitting President" (14%), March 12, 2003*
*"Ridge's Rise from Adviser to 'Mr. Secretary"' (13%), March 2, 2003*

As for the other EPU related topics identified, these do not have a strong presence in the data set under consideration. Although topic `government` is the dominant topic in two articles, a look at the distribution across all the topics for these articles shows that the probability values are still relatively low. Topic `political parties` does not occur as a prevalent topic in any of the articles from March 2003. It is not surprising that not all the topics are present in the out-of-sample U.S. data, as the data only covers one month. The model, on the other hand, was trained based on the longer time horizon.

Besides the EPU related topics, other topics were examined for their prevalent appearance in the considered U.S. dataset. One of such topics is `science/history` topic (see E.4.6) that has a probability of 51% in an article with the title *"No Cataclysm Brought Down Maya"* from March 14, 2003. The most frequent words (translated from German) of this topic are, for example, *researcher*, *scientist*, *animals*, *earth*, *sun*, *discovered*, *degree*, *examined*, *human*, *energy*, *explain*, *experts* etc. All in all, the topic actually seems to describe the content of the article quite well.


**Figure E.4.6:** MCTM: "Science" topic

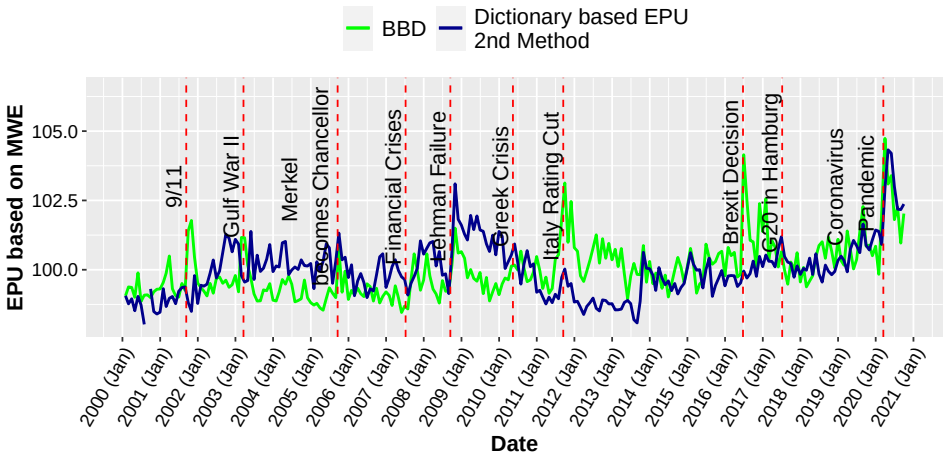
The qualitative robustness check presented above provides additional insights on the power of cross-lingual learning in this context. It also shows that pre-trained multilingual models can also be exploited in comparable text-as-data applications in the economic context.

## E.5 Comparison with BBD indices

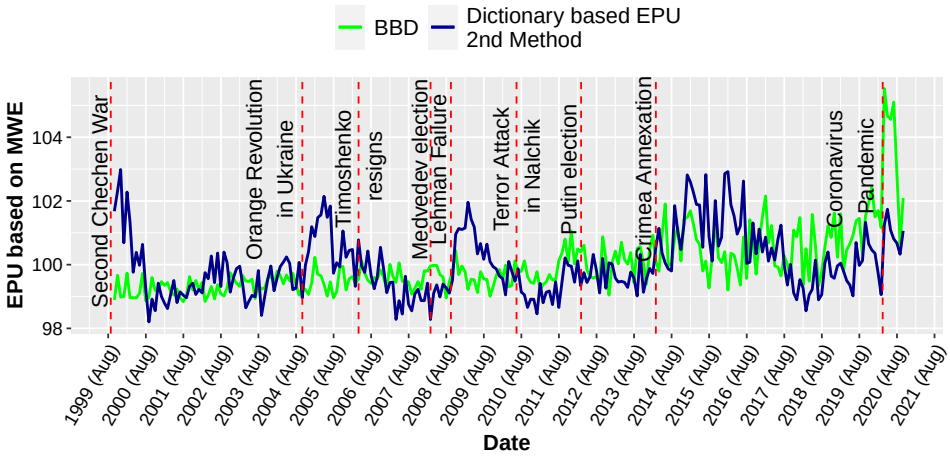
The index proposed by Baker et al. (2016) is often considered a baseline in the followed-up studies on measuring economic policy uncertainty. Thereby, the researchers usually report the correlation values between the indices as well as qualitatively compare their indices with available BBD indices (Azqueta-Gavaldón, 2017; Nyman & Ormerod, 2020; Xie, 2020). As the BBD indices for Germany and Russia are also provided online, such qualitative analyses were also conducted for these countries. Thereby, the constructed EPU indices based on the second method of the dictionary based approach (*dic*2) were used for these two countries. Figure E.5.7 illustrates the BBD indices (in green) and the EPU indices calculated according to the proposed automated approach using multilingual fastText word vectors *sic*2 (in dark-blue). The indices are normalised to a mean value of 100 and a variance of 1. For a better comparison, the events reported by Baker et al. (2016) in the online Appendix[8] are also illustrated in Figure E.5.7 by the vertical red dashed lines. In Germany, the two time series do not strongly correlate (0.3). However, both indices seem to show similar dynamics experiencing considerable spikes that might be associated with such events as September 11 attacks, Second Gulf War, German federal elections etc. The period from 2012 to 2015 are referred to as Eurozone stresses, at the beginning of which both indices show some discrepancies. Both indices also respond strongly to the outbreak of the Corona virus pandemic.

The BBD and *dic*2 indices for Russia both experience an increase, for instance, during Russian military conflicts (Second Chechen War), presidential elections, and global financial crisis. In general, increased uncertainty values in Russia are often associated with major political events in neighboring Ukraine. The last 6 years, beginning from Crimea annexation, are characterised by constant political tensions due to the Donbass War. The correlation coefficient of the two time series is also about 0.3.

[8] The online Appendix can be viewed under the following link `https://academic.oup.com/qje/article/131/4/1593/2468873#supplementary-data`, last accessed on 28.02.2024.




**(a)** Germany




**(b)** Russia

**Figure E.5.7:** BBD and *dic*2 EPU Indices

## E.6 Granger Causality Tests

| Country | | Dictionary Based | | Aggregated Document Embeddings | SBERT | Multilingual Contextualized Topic Modeling | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | separated terms | combined terms | | | Government | Stock Market | Political parties | Elections | U.S. political leaders | Combined |
| Germany | EPU →IND | 0.1153* | **0.02374*** | **0.07272*** | **0.009405** | 0.196* | **0.05637*** | **0.08755'** | 0.7107' | 0.2807 | 0.1853* |
| | IND →EPU | 0.9654* | **0.04103*** | 0.2034* | **0.01072** | 0.583* | 0.5979* | 0.3958' | 0.5694' | 0.4474 | **0.03061*** |
| Russia | EPU →IND | 0.2496" | **0.06071"** | 0.254' | 0.2504' | **0.01369*** | **0.008587'** | 0.164" | **0.003431'** | 0.7935' | 0.1468' |
| | IND →EPU | 0.7121" | 0.611" | 0.6064' | 0.2818' | 0.5302* | 0.7798' | 0.3228" | 0.2363' | 0.6583' | 0.4365' |
| Ukraine | EPU →IND | **0.08736*** | **0.08407*** | **0.00461*** | 0.1311* | 0.2746* | **0.06555*** | **0.02054*** | 0.5193* | 0.1874 | 0.2617* |
| | IND →EPU | 0.75381* | 0.2311* | 0.1903* | 0.9265* | 0.809*4 | 0.9312* | 0.07189* | 0.443* | 0.3294 | 0.4316* |

*: The residuals of the model are heteroscedastic.

": The residuals of the model are autocorrelated.

': The residuals of the model are heteroscedastic and autocorrelated.

The values in bold are significant at 10% significance level.

Dictionary based: *Separated terms* refers to the first method, where three terms sets are defined that describe three components of the EPU concept. *Combined terms* refers to the second method, where a single term set is defined to describe EPU.

Aggregated Document Embeddings: articles are represented as the sum of constituent words (fastText embeddings).

SBERT: SBERT embeddings are used to represent articles' texts.

MCTM: The uncovered `government`, `stock market`, `political parties`, `elections`, `U.S. political leaders` topics have been identified as EPU topics. Additionally, a combined topic has been constructed.

IND: the growth rates of the industrial production indices as they were used in VAR models.

EPU: the growth rates of the corresponding constructed EPU indices as they were used in VAR models.

**Table E.6.8:** Granger Causality Test Results (p-values)

# Chapter 8

# Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects

**The following chapter is based on the paper:**


| | |
|---|---|
| **Title:** | Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 30%), Albina Latifi (contribution: 30%), Peter Tillmann (contribution: 20%), Peter Winker (contribution: 20%) |
| **Status:** | Published: *European Economic Review*, vol. 168, 2024, pp. 104827 |
| **Available from:** | `https://doi.org/10.1016/j.euroecorev.2024.104827` |

Earlier versions of this paper were presented at:

- 24th International Conference on Computation Statistics (COMPSTAT 2022), Bologna, Italy, 2022
- 48th International Conference MACROMODELS, Wieliczka, Poland 2022
- 5th Annual COMPTEXT Conference, Glasgow, 2023 (presented by Co-Author)
- 1st NEAR Conference on Narrative Economics, Bochum, 2024 (presented by Co-Author)
- the annual meetings of the European Economic Association (EEA), Barcelona, 2023 (presented by Co-Author)
- the annual conference of the Verein für Socialpolitik, Regensburg, 2023 (presented by Co-Author)
- 1st NEAR Conference on Narrative Economics, Bochum, Germany, 2024 (presented by Co-Author)

# Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects*

Albina Latifi‡ Viktoriia Naboka-Krell‡
Peter Tillmann‡,‖ Peter Winker‡

**Abstract.** Fiscal policy is made in parliaments. We go to the roots of changes of fiscal policy in Germany and use a novel data set on all parliamentary speeches in the Bundestag from 1960 to 2021. We propose an embedding-based approach, which allows the representation of words and documents in a shared vector space, in order to measure fiscal policy-related sentiment in parliamentary debates at a scale from contractionary to expansionary. We also distinguish between sentiment related to exogenous and endogenous fiscal policy. We put fiscal sentiment into a series of recursively-identified vector autoregressive models to show that a change in fiscal sentiment causes a shift in government spending and has significant effects on the macroeconomy. The results support the notion that the debate in parliament contains information for the identification of government spending shocks.

*Key Words: text mining, word embeddings, VAR models, identification, government spending*

*JEL classification: C89, E60, E62*

* We are grateful to the editor and the associate editor of this journal as well as anonymous reviewers for excellent feedback on the first version of the paper. We also thank Leif Anders Thorsrud and Bernd Hayo who kindly shared his data series. Financial support from the German Research Foundation (DFG) (TI 594/4-1 and WI 2024/8-1) for the project MaFiText is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

‡ Department of Economics, Justus Liebig University Giessen

‖ Corresponding author: peter.tillmann@wirtschaft.uni-giessen.de

## 8.1 Introduction

Fiscal policy is made in parliaments. Legislating an increase or cut of federal public spending or taxes is a key prerogative of parliaments. This implies that changes in government spending are usually preceded by extensive debates in parliament and beyond, often stretching several quarters or even years. Hence, shifts in the tone of the parliamentary debate should indicate changes in government spending further down the road.

The empirical literature on the identification of government spending shocks and the estimation of their effects on the macroeconomy, see V. Ramey (2016) for a recent survey, has not yet made use of this information. In this paper, we use data on all speeches delivered in the Bundestag, the federal parliament of Germany, in order to measure fiscal policy. We argue that the measurement and the identification of fiscal policy impulses can be improved by examining the roots of fiscal policy-making, namely the parliamentary process itself. We believe that households and firms monitor the parliamentary process and adjust their expectations and decisions well before the law is eventually passed and comes into effect.

As a matter of fact, parliamentary speeches are multi-dimensional objects. Extracting quantitative information about fiscal policy is not straightforward. We exploit recent advances in natural language processing (NLP) to quantify fiscal sentiment based on a large data set of parliamentary speeches.

We proceed as follows: First, we propose an embedding-based approach, which allows the representation of words and documents in a shared vector space, in order to measure fiscal policy-related sentiment in parliamentary debates at a scale from contractionary to expansionary. For this purpose, we create a dictionary containing terms related to expansionary and contractionary fiscal policy measures. Specifically, we adopt Doc2Vec, an unsupervised method to represent natural language in a high-dimensional vector space (Mikolov, Chen, et al., 2013). The resulting text vectors (also known as embeddings) capture semantic characteristics of the texts. As the context in which fiscal policy measures are discussed may change over time, we adapt the approach proposed by Kapfhammer et al. (2020) and implement a rolling forecast architecture. This provides us with three series of fiscal sentiment: for the entire Bundestag, the governing parties and the opposition. We find large fluctuations in fiscal sentiment that fit the established historical narrative.

To the best of our knowledge, we are the first to use a large body of parliamentary texts to measure shifts in fiscal sentiment and, eventually, to estimate its effect on the macroeconomy.[1] Abercrombie & Batista-Navarro (2020) provide a comprehensive literature review of 61 studies, all of which deal with the automatic analysis of sentiments and opinions as well as the positions of speakers in parliamentary debates. In their research outlook, the authors regret that most of the studies only perform a rough positional analysis (e.g. left

[1] Allard et al. (2013) and Dybowski & Adämmer (2018) quantify fiscal sentiment in central bank documents and the communication of U.S. presidents. Instead, our paper is based on a much larger text corpus.

vs. right), instead of identifying policy preferences, which constrains possible applications. Furthermore, almost all studies in Abercrombie & Batista-Navarro (2020) are limited to the analysis of a single election period. This makes this paper one of the few existing studies that deal with the automated identification of political preferences over several election periods.

A recent strand of the macroeconomic literature proposes a "narrative approach" to the identification of exogenous changes to fiscal policy, i.e. Romer & Romer (2010), Mertens & Ravn (2012) and Cloyne (2013). These authors derive tax policy shocks from text documents such as presidential speeches or parliamentary reports.[2] Guajardo et al. (2014) use historical records such as budget speeches, central bank writings, IMF staff reports and OECD documents to identify changes in fiscal policy designed to consolidate public finances. Our paper goes beyond the selected numbers of text documents used in these studies. Instead, we use data on *all* parliamentary speeches to derive a measure of fiscal sentiment.

In a second step, we follow the notion of Romer & Romer (2010) and Cloyne (2013) and distinguish between parliamentary contributions on fiscal-policy related questions that are either exogenous or endogenous with respect to the current economic situation. We apply an Latent Dirichlet Allocation (LDA) model, a probabilistic topic model, which allows us to link speeches to latent topics (Blei et al., 2003). These topics can be grouped into topics related to exogenous fiscal policy (e.g. national defense, energy policy and social welfare), endogenous fiscal policy responses (e.g. the labor market, economic growth and public investment) or policies not immediately fiscally relevant. This provides us with series of exogenous and endogenous fiscal sentiment, respectively.

Third, we put the resulting sentiment series in a battery of standard Bayesian vector autoregressive (VAR) models. The models also include variables such as real government spending, real GDP, real private consumption or real investment. Our aim is to evaluate whether fiscal sentiment causes government expenditure and, hence, macroeconomic responses. Following the pioneering work of Blanchard & Perotti (2002), a large literature uses a recursive identification scheme in order to estimate the causal effects of government spending. This draws on the notion that within a quarter government spending is predetermined such that a feedback from GDP or other macro aggregates on the level of government spending should be excluded. Fortunately, our baseline series of fiscal sentiment lends itself to a straightforward extension of this identification scheme: we order sentiment last such that a change in sentiment cannot contemporaneously drive government spending. Sentiment, on the other hand, can immediately respond to economic developments. The decomposition of sentiment into exogenous and endogenous fiscal sentiment gives rise to an extension of this identification: exogenous fiscal sentiment is ordered first, macroeconomic reactions follow, while endogenous fiscal sentiment comes last.

We find that an unexpected increase in sentiment towards a more expansionary fiscal stance causes higher government spending, an expansion of real economic activity and an

[2] See Hayo & Uhl (2014), Hayo & Mierzwa (2022) and Christofzik et al. (2022) for similar approaches to tax policy shocks in Germany.

increase in private consumption. Several extensions of the model show that fiscal sentiment also increases investment and inflation and reduces unemployment and the budget balance, among other variables. These responses are in line with standard (New-)Keynesian business cycle models. Furthermore, fiscal sentiment has consequences for Germany as an open economy: more expansionary sentiment leads to a real appreciation and a deterioration of the trade balance. Hence, sentiment measured from parliamentary speeches has strong and robust macroeconomic effects. These results are mostly due to shifts in exogenous fiscal sentiment, while shocks to endogenous sentiment does not result in significant macroeconomic responses.

We also contribute to the literature on fiscal news shocks (V. Ramey, 2011; Ricco et al., 2016; Ben Zeev & Pappa, 2017), which argues that fiscal shocks are to some extent anticipated before they materialize. V. Ramey (2011) constructs two measures of news, one from a narrative account of reports about defense spending and one from professional forecasters, and shows that both are able to predict spending shocks from recursively identified VAR models. Ricco et al. (2016) also uses data from professional forecasters to measure the future path of government spending. Ben Zeev & Pappa (2017) identify a news shock as the shock that best explains future movements in defense spending, while at the same time being orthogonal to current defense spending. By detecting changes in fiscal sentiment in speeches of parliamentarians that precede actual legislation, we offer an alternative way to shed light on fiscal news.

Equipped with our sentiment series, we revisit the problem of fiscal foresight, which is often put forward as an argument to invalidate a recursive identification of government spending shocks, e.g. V. Ramey (2011), V. Ramey (2016) and Ellahie & Ricco (2017). From a standard recursively-identified VAR model in the spirit of Blanchard & Perotti (2002), Fatás & Mihov (2001), Galí et al. (2007), Born & Müller (2012), Auerbach & Gorodnichenko (2012), Ilzetzki et al. (2013) and others we obtain a series of structural government spending shocks. We then show that fiscal sentiment predicts these structural shocks six to eight quarters in advance. Hence, supposedly unanticipated government spending shocks are in fact anticipated once information from the parliamentary debate is taken into account.

The remainder of this paper is structured as follows. Section 8.2 describes the data from the German Bundestag as well as the preprocessing steps. Section 8.3 presents our approach for constructing the text-based fiscal sentiment indicator. In Section 8.4, we estimate a range of VAR models to understand the effects of fiscal policy on the German economy. Section 8.5 revisits the problem of fiscal foresight. Finally, Section 8.6 concludes. An online appendix contains additional material.

## 8.2 Text data

This section introduces the textual data from which we derive a measure of fiscal sentiment. Subsection 8.2.1 describes the underlying text data, i.e. the full set of parliamentary speeches. In subsection 8.2.2, we describe specific text data preprocessing steps which are needed for

the text mining methods applied later.

### 8.2.1 Bundestag speeches as a novel data source

The Bundestag is the German federal parliament. At the time of writing, it has more than 700 members with the exact number varying over time. Although we later focus on fiscal policy, we start from the full set of all parliamentary debates. The digitized debates of the plenary sessions, not the committees, are made available to the public by the German Bundestag.[3]

Unfortunately, this text data is not directly suitable for the application of NLP methods as the parliamentary speeches up to election period 18, i.e. until October 2017, are available in an unstructured form only. This means that the XML documents only have one "TEXT" tag in addition to some meta-information on the whole session such as the election period and the date. The content of the "TEXT" tag comprises the entire stenographic report – including the agenda items, the actual meeting and the annexes. Therefore, it is necessary to further structure the content of the "TEXT" tag so that each speech can be assigned to a speaker. Furthermore, each speaker should be assigned with his or her role and party affiliation. To structure the XML files, we use the workflow documented in more detail in Latifi (2024), which proceeds as follows:

First, we parse and clean the XML documents using a set of regular expressions. Second, a Named Entity Recognition (NER) model with a customised entity, which usually consists of a speaker's first and last name followed by his or her party affiliation in brackets, a colon and a newline, is developed to identify the begin of each speech. Thereby, a small hand-labelled data set is created to train the NER model. Eventually, the cleaned stenographic protocols can be split by each identified beginning of a speech. After that, one can extract roles and party affiliations of each speaker.

With the aim of promoting computer-assisted research on parliamentary data, the German Bundestag has been publishing the XML files in a structured form since election period 19, so that the precise extraction of relevant information involves comparatively little effort and manual reworking.[4] We convert these plenary protocols into a file format that can be used for further processing by using the python package `pybundestag` (Hruzik, 2019). In a final step, we unify the data set of election periods 1-18 with the data set of election period 19 in a consistent structure. The complete data set of the election periods 1 to 19 comprises a total of 877 140 speeches.

[3] All stenographic reports can be downloaded from this website as XML files packed in .zip files: `https://www.bundestag.de/services/opendata`.

[4] Further information on the structure of the XML documents as of election period 19 are described here: `https://www.bundestag.de/resource/blob/577234/f9159cee3e045cbc37dcd6de6322fcdd/dbtplenarprotokoll_kommentiert-data.pdf`.

### 8.2.2 Corpus and text preprocessing

Since most macroeconomic time series are available from 1970 onward and as we need a rolling 10-year training data set for our embedding approach, our data set begins in 1960. In a first step, speeches by the President of the Bundestag or an office holder of a similar function are excluded, as their main task is primarily to chair and moderate (including announcing voting results, calling up items on the agenda, calling up speakers) the plenary sessions. Furthermore, we exclude speeches from state ministers representing the federal states of Germany, guest speakers such as foreign dignitaries and other irregular speakers.

This leaves us with all speeches delivered by members of parliament, Chancellors, Federal Ministers and State Secretaries.[5] Furthermore, we remove very short and very long speeches from the data set. Short remarks are often made during swearing-in ceremonies or as interposed questions. We count the words of each speech and remove all speeches containing less than 100 tokens, i.e words, or more than 3 573 tokens. The latter threshold corresponds to the 99.5%-percentile of the word frequency distribution over the documents. After these exclusion steps, the data set contains 235 129 speeches covering the period from January 20, 1960 to September 07, 2021.

Figure 8.1 shows the number of speeches on a quarterly basis. Over the entire period, the average number of speeches per quarter is 959.71, though the number of speeches increases over time. Furthermore, there are seasonal fluctuations. The third quarter contains by far the fewest speeches in the data set, with a total of 26 141 speeches, which can be attributed to the summer break.[6] In addition, Figure 8.1 shows that members of the coalition parties forming the government deliver more speeches than members of the opposition parties. This is plausible given the distribution of seats. However, these differences have become smaller in the recent past.

We then prepare the corpus using common text preprocessing steps, such as those described in Grimmer & Stewart (2013) and Denny & Spirling (2018). We first lemmatize all speeches using the model *"de_core_news_lg"* from the python package `spaCy` (Honnibal et al., 2020) so that all words in the speeches are traced back to their root words. This also reduces the complexity of the corpus. In the following, we refer to a word as a token and a speech as a document. The corpus is the collection of all documents.

After lemmatizing, we convert all German umlauts and the eszett to exclude encoding errors. In addition, we remove line breaks, digits, blank sentences and special characters and convert all letters to lower case. We remove single-element tokens and tokens with more than 30 elements. The removal of further short tokens is not advisable for this domain, because meaningful tokens such as "is", "eg", "eu", "ki", "db" etc. are included in this corpus.

The next step is to create a list of stop words. Stop words are tokens that occur very

[5] Members of the government do not have to be members of parliament, though in most cases they are.

[6] For comparison, quarter I contains a total of 68 553 speeches, quarter II contains 73 412 speeches, and quarter IV contains 67 023 speeches.

frequently in the corpus, but contain little information and therefore do not contribute to the understanding of a text (e.g. personal pronouns, conjunctions, etc.). In the `nltk` package (Bird & Klein, 2009), there is a predefined list of stop words for the German language. Since a general stop words list does not include domain-specific terms such as “Bundestag", “Abgeordneter", “Deutschland", “Redezeit", "Drucksache“, etc., we create a list of stop words based on the inverse document frequency (*idf*) value of each unique word.[7] A low *idf* value implies that a word occurs in a very large number of documents and is therefore not very specific. All tokens that occur in 97% of all documents in each election period are considered potential stop words. The final list of stop words is manually expanded in an iterative process, so that it comprises a total of 1 405 terms. We report the final list of stop words in the appendix. These stop words were lemmatized analogously to the text and finally removed from the corpus.

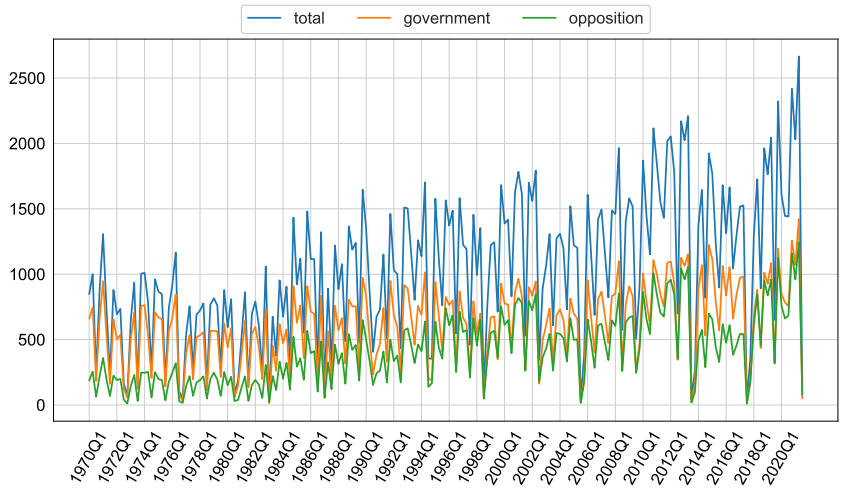



**Figure 8.1:** Number of speeches

*Notes:* The figure shows the number of speeches over our sample period aggregated to quarterly frequency.

Figure 8.2 shows the distribution of the length of a document before and after applying the described preprocessing steps. A document contains an average of 637.10 tokens before preprocessing, but only 164.83 tokens after preprocessing. The entire corpus contains 149 800 737 tokens before and 38 755 290 tokens after preprocessing. This means that the size of the corpus is reduced by almost 75% after preprocessing.

## 8.3 A text-based fiscal sentiment indicator

This section first documents the construction of a dictionary with fiscal policy-specific terms. Next, we introduce word and document embeddings. Finally, we propose an approach to construct a fiscal sentiment indicator based on these text representations.

[7] The *idf* value is calculated using the package `scikit-learn` (Pedregosa et al., 2011).

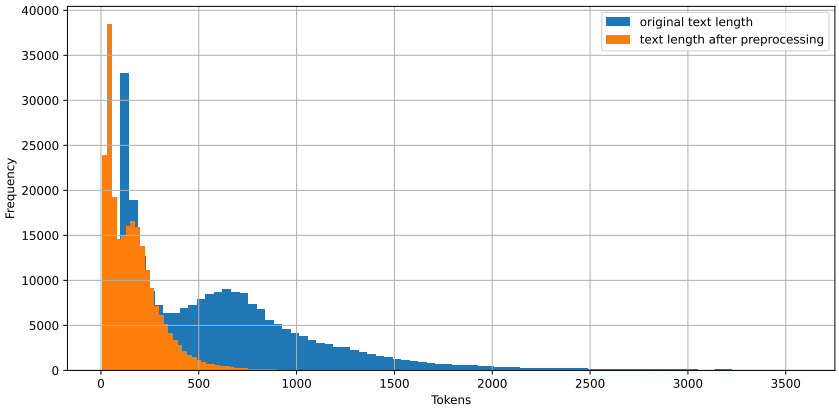



**Figure 8.2:** Length of speeches before and after preprocessing

*Notes:* The figure shows the distribution of the length of a document (speech) before and after preprocessing.

### 8.3.1 Compiling a dictionary on fiscal policy

Although dictionary-based approaches have been widely used in economics and finance lately, e.g. to quantify the sentiment of financial reports (Loughran & McDonald, 2011), central bank communication (Picault & Renault, 2017) or to construct newspaper-based indicators of economic policy uncertainty (Baker et al., 2016), there is no dictionary available for fiscal policy. Therefore, we first need to assemble a fiscal policy-related dictionary for Germany.

We construct a list of terms relevant to fiscal policy through extensive study of the speeches in the Bundestag. Since not all randomly selected speeches address the core of fiscal policy, we also take advantage of data from the Manifesto corpus (Burst et al., 2020), in order to identify particularly informative expressions. This data contains publicly available, thematically classified quasi-sentences from election manifestos of various parties from over 50 countries. The categories capture the most relevant political issues and goals and are assigned to the respective quasi-sentences according to a strict annotation scheme. The annotation scheme is described in Werner et al. (2011). The corpus can be downloaded via the R package `manifestoR` or via the website.[8] We select 19 categories related to fiscal policy.[9]

[8] The .csv files can be downloaded via this page: `https://visuals.manifesto-project.wzb.eu/mpdb-shiny/cmp_dashboard_dataset/`.

[9] This amounts to 39 .csv files of German election programs from 1998 to 2017. These files contain a total of 22 602 quasi-sentences classified with the selected categories. The following categories are selected: [303] Governmental and Administrative Efficiency, [401] Free Market Economy, [402] Incentives, [403] Market Regulation, [404] Economic Planning, [408] Economic Goals, [409] Keynesian Demand Management, [410] Economic Growth: Positive, [412] Controlled Economy, [414] Economic Orthodoxy, [416] Anti-Growth Economy: Positive, [503] Equality: Positive, [504] Welfare State Expansion, [505] Welfare State Limitation, [501] Environmental Protection: Positive, [701] Labour Groups: Positive, [702] Labour Groups: Negative, [416.1] Anti-Growth

In this way, we construct a preliminary list of 163 keywords. After that, we expand and refine the keywords, considering also the characteristics specific to the German language, such as composite words and synonyms. After this extension, the fiscal dictionary comprises a total of 322 words. We then label terms as expansionary, neutral or contractionary. It is important that we identify terms as uniquely as possible in terms of the expansionary or contractionary sentiment conveyed. In case an expression is ambiguous, we decide based on the majority vote among the authors. Eventually, our list consists of 218 terms or compound terms, of which 122 are classified as “expansionary" and 96 as “contractionary".[10]

### 8.3.2 Doc2Vec approach

Doc2Vec is an extension of Word2Vec, which is an unsupervised, neural network-based method to represent natural language in a high-dimensional vector space (Mikolov, Chen, et al., 2013). It could be considered a black box that “translates” semantic features of natural language into dimensions of a vector space. Almost no supervision is needed to train such text representations as the algorithm learns semantics from original texts. One of the most characteristic features of resulting embeddings is the interpretability of mathematical operations between them. Let us consider one specific example in the context of fiscal policy. Assume that we obtained 100-dimensional vectors for the words “staatliche” (government) and “Investitionen” (investments) each. Then, we sum up these two vectors and look for the nearest neighbors to this sum (based on cosine similarity). We obtain the following words: *oeffentlich (public)*, *investieren (invest)*, *privat (private)*, *Aufbau (development)*, *gesetzlich (legal)*, *Innovation (innovation)*, *nachhaltig (sustainable)*, *langfristig (long-term)*. A further example considers the word “Steuerentlastung” (tax relief). Its nearest neighbors are the following: *Steuererhoehung (tax increase)*, *Entlastung (relief)*, *Progression (progression)*, *Muetterrente (mothers' pension)*, *Soli (solidarity tax)*, *Kinderfreibetrag (allowance for children)*, *Elterngeld (parental allowance)*, *Neuverschuldung (new debt)*, *Steuervereinfachung (tax simplification)* etc. These examples demonstrate the ability of embeddings to capture semantic similarity of natural language. The exemplary words and their neighbors are indeed likely to appear in the same context.

Doc2Vec allows representing sentences, paragraphs and whole documents as vectors (Le & Mikolov, 2014; Mikolov, Sutskever, et al., 2013). The distances between these representations could be interpreted meaningfully. In the appendix, we graphically illustrate the approach based on a a selected speech of Olaf Scholz, the Federal Minister of Finance, on March 25, 2020.[11]

We propose a Doc2Vec approach to construct a fiscal policy index for the German Bundestag. Thereby, we require that values of the index at time $t$ are based solely on

Economy: Positive, [4012] Control of Economy: Negative.

[10] The list of expansionary and contractionary terms used can be found in the appendix.

[11] Due to described features of word and text vectors, these are used in different text-as-data applications, e.g. Rheault & Cochrane (2020), Gennaro & Ash (2021) and Rodriguez & Spirling (2022).

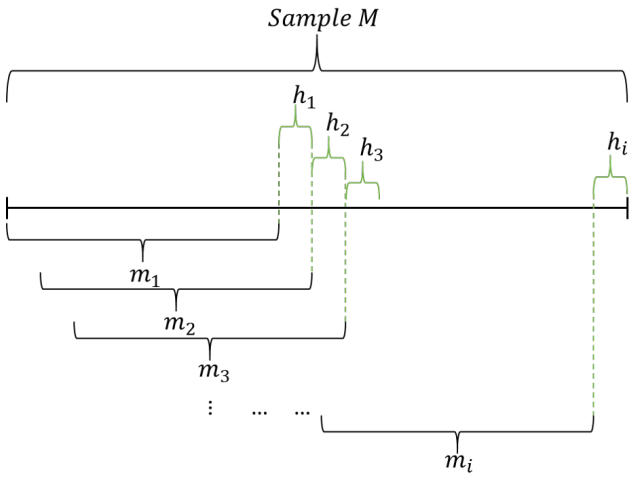



**Figure 8.3:** Rolling-window Doc2Vec

*Notes*: This figure represents the training procedure of the dynamic Doc2Vec approach. Each training period $m$ covers the same period of time (e.g. quarter, year, 10 years). Each forecast period $h$ is also of fixed size. The shift length equals the defined forecast length.

information up to time $t$. Furthermore, the model should allow for possible changes in language usage over time. New concepts with regard to fiscal policy may occur and some of them may disappear over time. Also, the general context in which fiscal policy measures are discussed may change over time.

To address the first feature described above, we propose a rolling forecast architecture. Figure 8.3 represents the training procedure for the rolling window setting. Given the sample $M$ that corresponds to the corpus presented in Subsection 8.2.2, we divide the period into training periods $m_i$. For each period $m_i$, we train a Doc2Vec model using lemmatized and preprocessed texts. For the subsequent period $h_i$, document vectors are inferred based on the trained Doc2Vec model. This means that documents in the forecast period $h$ are actually new to the model. The vectors for these documents are predicted based on the trained word dependencies and relationships. Each training and forecast period is of fixed size. However, the number of observations in each period $m_i$ and $h_i$ might differ depending on the number of documents.

To address the second feature, we adapt the approach proposed by Kapfhammer et al. (2020). The authors use word embeddings to measure climate change transition risk and investigate how the media speaks about risk and how the context changes over time. To address the changing context, the authors divide the time period into sub-periods and estimate separate Word2Vec models for each sub-period. Kapfhammer et al. (2020) argue that making the word embedding methodology dynamic can capture changes of the relationships between specific words over time.

Overall, the dynamic approach proceeds in four steps:

1. For a defined period of time, train the Doc2Vec model using preprocessed texts.

2. For the given training sub-period, construct an expansionary and a contractionary

vector as the average vector of the identified fiscal policy-related terms. Thereby, their representation is different in each training period depending on which words occur in the learned vocabulary. For two-words terms such as "Steuern senken", "Arbeitsplätze schaffen" and others, the average of the corresponding word vectors is used to represent these terms.

3. Based on the pre-trained model, infer speeches vectors for the subsequent forecast period. Calculate the cosine similarities between the inferred document vectors and the fiscal policy vectors. Each speech receives two scores: similarity to an expansionary stance of fiscal policy and similarity to a contractionary stance of fiscal policy.

4. Construct a continuous indicator by taking the difference between the similarity to the expansionary vector and the contractionary vector. This results in a value that falls into the range from -1 (very contractionary) to 1 (very expansionary).

This procedure is fully unsupervised and language agnostic. It can be applied to other corpora and other languages in case a fiscal policy dictionary is available. Additional technical details are described in the appendix.

### 8.3.3 Baseline sentiment

We use a training length of 40 quarters and a forecast length of one quarter. As mentioned before, this limits the sample for which we will obtain the sentiment series to 1970 to 2021. Each training sub-period contains 37 889 speeches on average, while each forecast period contains 1 037 speeches on average. As described in the previous subsection, each speech receives two scores that correspond to the cosine similarities between the single speeches and the constructed expansionary and contractionary vectors. We build a continuous fiscal policy index by subtracting the similarity to the contractionary vector from the similarity to the expansionary vector. For example, a fiscal sentiment of zero could imply that a speech discusses both expansionary and contractionary fiscal policy measures and the cosine scores to both fiscal vectors cancel out.

We interpret the resulting series of fiscal sentiment as a measure of the propensity of parliamentarians to engage in expansionary fiscal policy. A higher sentiment score means that policy is more inclined to implement expansionary policies.

Figure 8.4 shows the final sentiment series at quarterly frequency. These series will enter the estimated VAR models in the next section. We show one series for the entire Bundestag, i.e. for speeches of all members, one for speeches delivered by members of the governing parties and one based on speeches from opposition members. The figure also highlights recessions in Germany as identified by the German Council of Economic Experts.[12] We find a close co-movement of government and opposition sentiment. The evolution of sentiment is consistent with the established historical narrative: First, the shift from a center-left to

[12] See `https://www.sachverstaendigenrat-wirtschaft.de/en/topics/business-cycles-and-growth/konjunkturzyklus-datierung.html` for the recession dates.

center-right coalition in 1982, which was also motivated by concerns about fiscal sustainability, is clearly visible as a drop in fiscal sentiment of the government towards a more restrictive policy stance. Second, when in 2003 the European Commission triggered the excessive deficit procedure against Germany, which is specified in the Stability and Growth Pact, fiscal sentiment deteriorated. Third, we spot a fall in fiscal sentiment starting in 2014 when the coalition government pushed its policy of budget surpluses ("schwarze Null"). Towards the end of our sample period, i.e. after the 2017 election, there is a remarkable upward trend in sentiment.

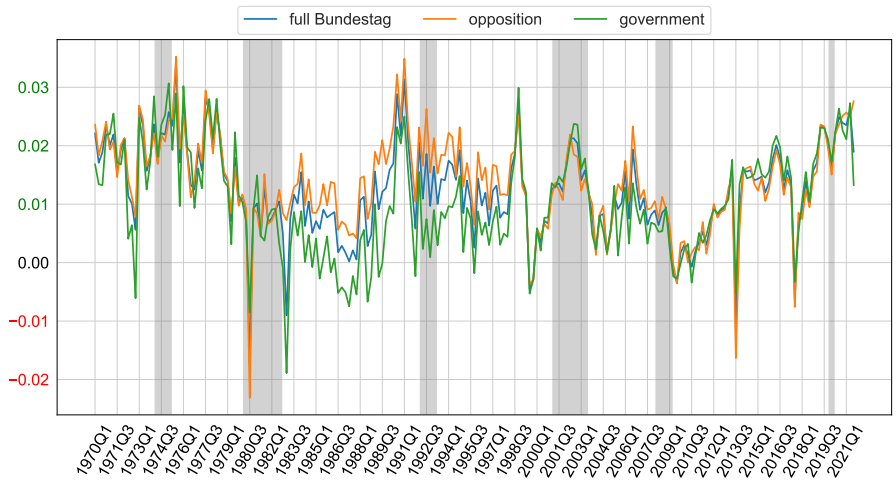



**Figure 8.4:** Fiscal policy sentiment

*Notes:* The figure shows the fiscal sentiment of the entire Bundestag as well as the government and the opposition. The shaded areas highlight recessions identified by the German Council of Economic Experts.

We further analyse fiscal sentiment with regard to the changing coalition governments. Figure 8.5 presents the average sentiment in the German Bundestag for each election period. In most periods, the government exhibits a more expansionary sentiment than the opposition. The only exception to this is the 2017-2021 coalition with Finance Minister Wolfgang Schäuble pushing the "schwarze Null". The distance between government and opposition varies over time and reaches a maximum in the early 1980s.

In the appendix, we compare our sentiment index with indices based on fiscal policy-specific dictionaries.

### 8.3.4 Exogenous vs endogenous fiscal sentiment

The approach so far allows us to construct sentiment time series covering all aspects of fiscal policy. In the empirical estimation below, we need to impose restrictions on the model in order to identify shocks to fiscal sentiment. The text analysis can help us in the identification of sentiment shocks and make the restrictions imposed on the VAR model more credible.

For that purpose, we exploit the range of issues members of the Bundestag discuss in

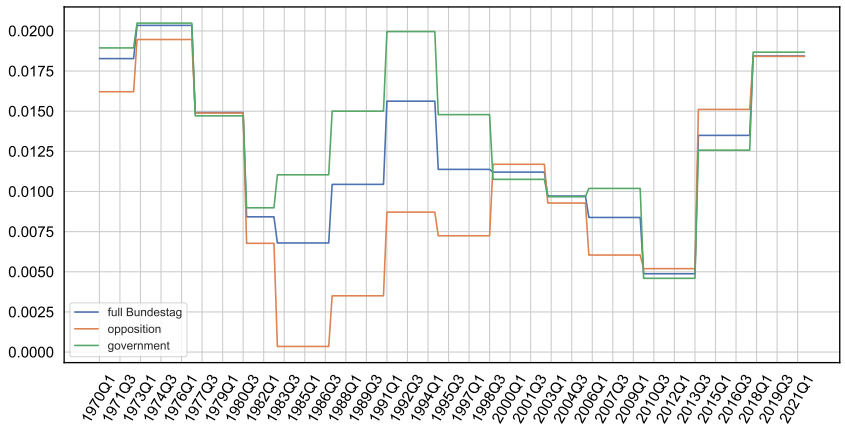



**Figure 8.5:** Average fiscal sentiment per election period

*Notes:* The figure shows the average fiscal sentiment of the entire Bundestag the government and the opposition in each election period from 1970Q1 to 2021Q3.

their speeches. Some issues are clearly reflecting the current economic situation, while others reflect structural decisions that have fiscal consequences but are unrelated to the current state of the economy. Consider one parliamentary speech on the state of the German military and another speech on the government's budget. Both speeches impact government spending and contribute to fiscal sentiment. However, the latter is a response to the state of the business cycle, while the former is unrelated to the cycle.

We follow the notion of Cloyne (2013), who picks up an approach of Romer & Romer (2010), that an "exogenous policy decision is one that was not designed to offset other macroeconomic shocks" (p. 1511). Consequently, we refer to the sentiment expressed in speeches on the German military and the budget as examples for exogenous and endogenous sentiment, respectively. Fluctuations in exogenous fiscal sentiment should reflect changes in the preferences of politicians or long-term structural issues such as infrastructure, the military, the development of the new federal states, social justice, etc., which are not an immediate response to the business cycle. Fluctuations in endogenous fiscal sentiment are a consequence of the economic cycle.

Using topic modelling, it is possible to uncover topics within the large corpus of Bundestag speeches. These topics can be categorized as either exogenous or endogenous branches of policies. We can then categorize speeches as reflecting exogenous or endogenous policy and calculate the sentiment for each speech.[13]

One of the most popular topic modelling technique is the Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). The LDA model is considered the prototype of probabilistic generative topic models. The model assumes that all documents in a corpus (which altogether possess exactly $K$ predetermined topics) are generated from a random

[13] Our distinction between exogenous and endogenous fiscal policy rests on a classification of topics. In contrast, the pioneering work of Romer & Romer (2010) is based on a careful reading of actual tax laws. In a useful validation exercise, one could check whether a topic model is able to replicate the Romer & Romer (2010) shock series. We leave this for future research.

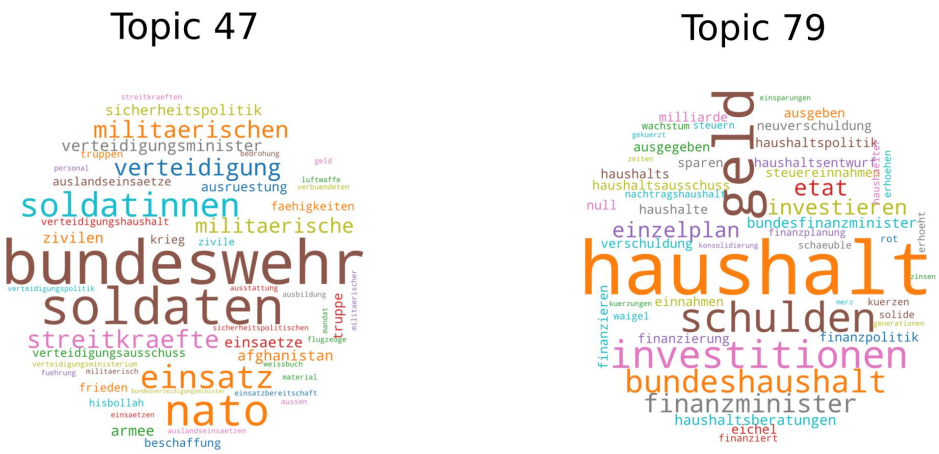



**Figure 8.6:** Word clouds of topics “Bundeswehr" (topic 47) and “Budget, Debt, Investment" (topic 79)

*Notes:* The figure shows the wordmix of a topic visualized as a word cloud of the 50 most important words. The larger the font size, the more important this word is for this topic. Topic 47 (“Bundeswehr") is classified as exogenous fiscal policy. Topic 79 (“Budget, Debt, Investment") is classified as endogenous fiscal policy.

mix of different latent topics. A topic, in turn, represents a specific mix of words from the vocabulary. Each document $d$ can contain words linked to any of these $K$ topics. However, the documents (i.e. parliamentary speeches) differ with regard to the weights of the topics. Thus, a document can be described as a probability distribution $\theta_d$ over topics, and a topic can be seen as a probability distribution $\beta_k$ over the given vocabulary. Since the topics are not known in advance, the goal of the LDA model is to learn the topic mix $\theta_d$ in each document and the word mix $\beta_k$ in each topic, from the data. For more technical details see Blei & Lafferty (2009).

We estimate an LDA model with the assumption of $K = 100$ topics.[14] We manually classify the 100 topics as either exogenous or endogenous topics, based on the notion of Cloyne (2013), or as topics unrelated to fiscal policy. This classification is guided by our thorough reading of the learned topic-word distributions $\beta_k$ of all topics. These topic-word distributions are usually visualized using word clouds. Figure 8.6 shows, for example, two word clouds that represent the topics “Bundeswehr" on the left-hand and “Budget, Debt, Investment" on the right-hand side. We can classify the “Bundeswehr" topic as exogenous fiscal policy and the “Budget, Debt, Investment" topic as endogenous fiscal policy.

A total of nine topics are classified as endogenous fiscal policy, while 25 topics are classified as exogenous fiscal policy.[15] Based on this classification, we can aggregate the topic-document-probability $\theta_d$ for each speech and obtain three scores summing up to one that provide us with the affiliation of each speech with exogenous or endogenous policy or

[14] We apply the `gensim`-package of Řehůřek & Sojka (2010).

[15] The thematic interpretation of the topics and the classification into exogenous and endogenous policy are listed in the appendix.

issues unrelated to fiscal policy. The first two scores are used to derive endogenous and exogenous sentiment series. To this end, we multiply the fiscal sentiment time series with the corresponding score and obtain exogenous and endogenous fiscal sentiment series by aggregation over all speeches.

Figure 8.7 shows the evolution of these fiscal sentiment time series for the entire Bundestag. For most of the sample period, both series closely move together. The peak in exogenous sentiment coincides with the German re-unification in 1990. Endogenous sentiment is particularly high during the oil crises of the 1970s. In 2009, both series start to diverge. One reason for this could be the increased number of exogenous shocks with large fiscal consequences, i.e. the European sovereign debt crisis, the increase in migration due to the war in Syria and the Covid-19 pandemic.

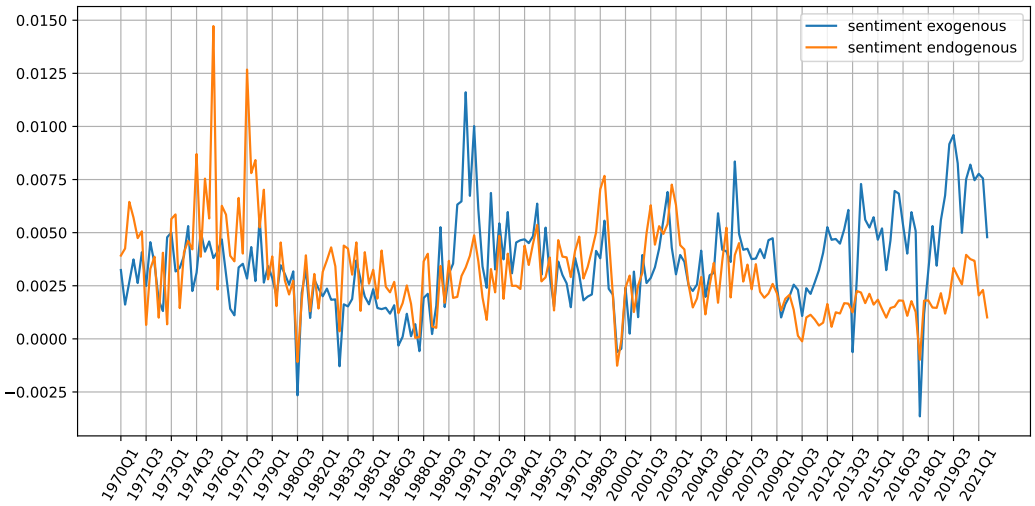



**Figure 8.7:** Exogenous and endogenous fiscal sentiment

*Notes:* The figure shows exogenous and endogenous fiscal sentiment reflected in speeches of all members of the Bundestag.

Figure 8.8 shows the correlation of our main sentiment indices with business cycle variables as well as the main series from the influential survey conducted by the ifo Institute and the World Uncertainty Index for Germany at leads and lags.[16] The aggregate sentiment series is positively correlated with growth and inflation and negatively correlated with unemployment. As expected, the endogenous sentiment is more closely correlated with growth, inflation and unemployment than the exogenous sentiment. This difference is particularly striking for the correlation with the ifo series: exogenous sentiment is procyclical, while endogenous sentiment is countercyclical. An increase in uncertainty coincides with higher exogenous sentiment and lower endogenous sentiment.

To further illustrate the evolution of exogenous and endogenous fiscal sentiment, Figure

[16] We merged the ifo series on business climate and business expectations, respectively, from the pre-1991 period with the series from the post-1991 period. The raw data is available at `https://www.ifo.de/en/ifo-time-series`. The uncertainty index is available at `https://worlduncertaintyindex.com/`.

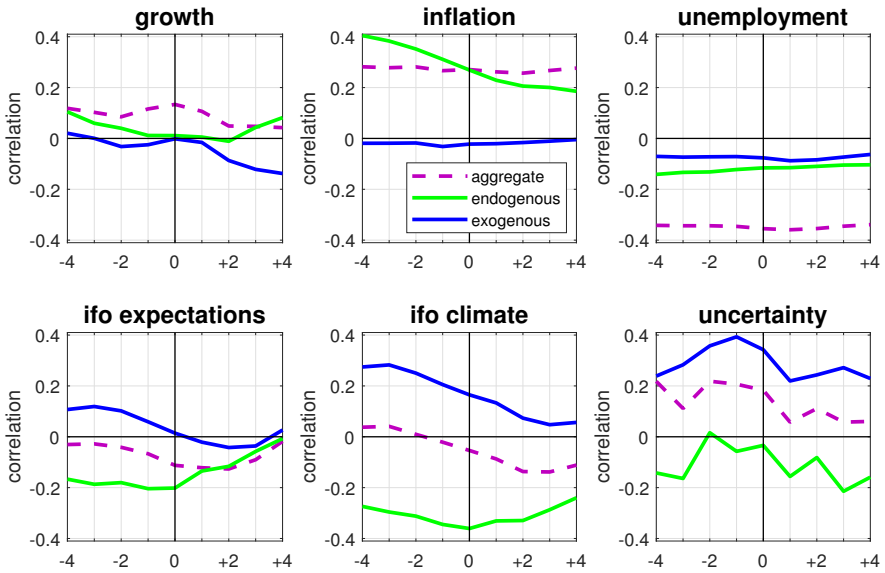



**Figure 8.8:** Cross-correlation of sentiment with macroeconomic variables

*Notes:* The figure shows the correlation of the sentiment indices in quarter $t$ with key macroeconomic variables as well as the indicators from the survey conducted by the ifo Institute and the World Uncertainty Index in $t \pm k$ with $k = 0, ..., 4$.

8.9 highlights selected episodes. We show sentiment following four large adverse economic shocks that could be considered exogenous: the first oil crisis in 1973, the second in 1979, the collapse of Lehman Brothers at the peak of the global financial crisis and the Covid-19 pandemic. Since we normalize sentiment to one at the beginning of each episode, this figure is not informative about the level of sentiment. Rather, it showcases the responses to these events. In each of the four episodes, we see a stronger increase in sentiment about endogenous fiscal policy topics compared to exogenous fiscal policy topics. This is intuitive as the large adverse shocks should elicit an endogenous fiscal stabilization as a response.

## 8.4 Estimating the macroeconomic effects

We now study the macroeconomic effects of exogenous changes in fiscal sentiment as reflected in the textual data. For that purpose, we augment a relatively standard VAR model by our new sentiment series.

### 8.4.1 VAR model

We estimate a reduced-form VAR model with $p$ lags

$$y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + C x_t + \varepsilon_t, \tag{8.4.1}$$

where $y_t$ is the $n \times 1$ vector of endogenous variables, $A_1, ..., A_p$ are $n \times n$ coefficient matrices and $x_t$ is a vector of exogenous regressors such as constant terms, dummies and a time trend. The vector of error terms, $\varepsilon_t$, follows a multivariate normal distribution, $\varepsilon_t \sim N(0, \Sigma)$, where $\Sigma$ is the variance-covariance matrix with $E(\varepsilon_t \varepsilon_t') = \Sigma$. The residuals are mutually uncorrelated at all leads and lags.

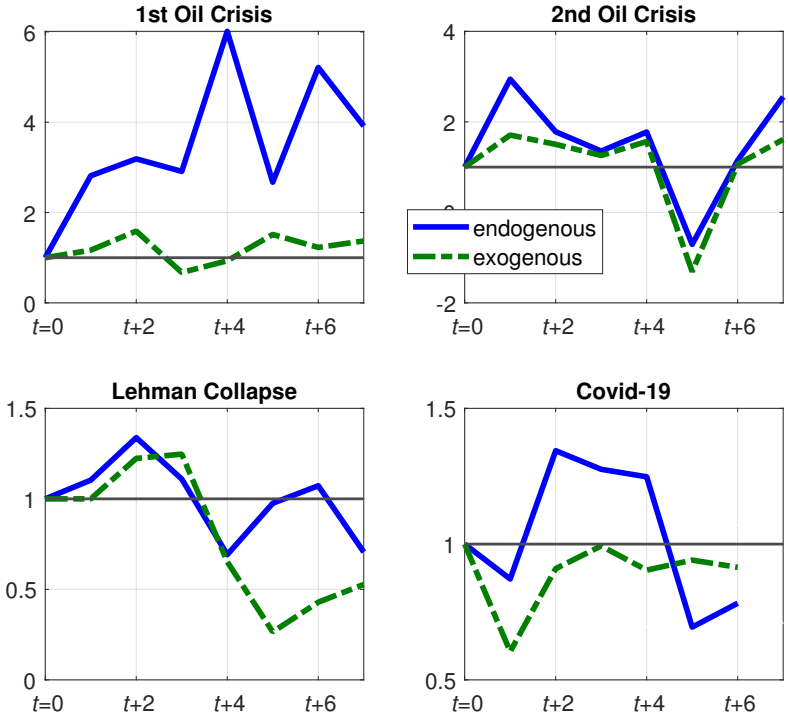



**Figure 8.9:** Fiscal sentiment in selected episodes

*Notes:* The figure shows the standardized fiscal sentiment for the government and the opposition during selected episodes. We normalize both sentiment series to one in 1973Q3 (1st oil crisis), 1979Q2 (2nd oil crisis), 2008Q2 (collapse of Lehman Brothers) and 2019Q4 (Covid-19).

The VAR model is estimated using Bayesian methods, thus treating parameters as random variables drawn from an underlying probability distribution. We assume a Normal-Wishart prior, though our results remain unchanged for alternative priors specifications.[17]

As discussed before, we adopt a recursive identification scheme. Let us write the model in its structural form

$$D_0 y_t = D_1 y_{t-1} + ... + D_p y_{t-p} + F x_t + \eta_t, \tag{8.4.2}$$

where $\eta \sim N(0, \Gamma)$ is the vector of structural shocks and the $D$ matrices are defined appropriately. With $D = D_0^{-1}$, the reduced-form error terms and the structural shocks are linked by $\varepsilon_t = D\eta_t$. We assume that $D$ is lower triangular, thus imposing restrictions on the contemporaneous interdependencies between the endogenous variables.

### 8.4.2 Data

In our baseline model, we include four endogenous variables: the log of real government expenditure, $Gov_t$, the log of real GDP, $GDP_t$, the log of real private consumption, $Cons_t$, and one of the three alternative indicators of fiscal sentiment derived in the previous sections, $Senti_t^j$, with $j \in (Bundestag, Government, Opposition)$. Hence, the vector of endogenous

[17] In order to estimate the model, we rely on the BEAR toolbox for MATLAB, see `https://www.ecb.europa.eu/pub/research/working-papers/html/bear-toolbox.en.html`.

variables is

$$y_t' = \begin{bmatrix} Gov_t & GDP_t & Cons_t & Senti_t^j \end{bmatrix}. \tag{8.4.3}$$

In four alternative specifications, we augment the baseline VAR by additional variables. First, we include the log of real private investment and the employment rate as two additional variables reflecting the domestic business cycle. Second, we add government revenues and the real interest rate. Third, we include the federal budget balance and business expectations from the ifo survey. Fourth, we add two variables to the model that reflect the open-economy transmission of fiscal policy, i.e. the log of the real effective exchange rate and the trade balance relative to GDP. All log series are multiplied by 100. The estimation frequency is quarterly and the data spans 1970Q1 - 2021Q3.

We also include a time trend and an impulse dummy that is one in 2020Q2 and zero otherwise. This dummy capture the extreme drop in real economic activity due to the Covid-19 pandemic and the ensuing lockdown. As Germany went into lockdown in the second half of March 2020, we choose to set the dummy to one in the second quarter of 2020. We estimate the VAR model for $p = 8$ lags.

The core time series are taken from the OECD data file: real GDP, real private consumption and real government consumption. In an extended model, we also use real gross fixed capital formation, i.e. investment, and real government revenues (interpolated to quarterly frequency). Both are also taken from the OECD. All series are seasonally adjusted. The data for the federal budget balance is taken from the Bundesbank and interpolated to quarterly frequency.[18]

### 8.4.3 Identification

We draw on the extensive literature on the identification of exogenous fiscal policy shocks pioneered by Blanchard & Perotti (2002) and applied, among others, by Fatás & Mihov (2001), Galí et al. (2007), Born & Müller (2012), Auerbach & Gorodnichenko (2012) and Ilzetzki et al. (2013) and impose a recursive ordering onto the variables.[19] The ordering of the variables as in (8.4.3) implies that in a given quarter government expenditure is predetermined. Changes in GDP or consumption, respectively, do not contemporaneously affect government spending. Our specific application lends itself to a straightforward extension of this line of literature. The starting point of our analysis is that fiscal policy is made in parliaments and that parliamentary decisions take time. This is exactly why spending is predetermined in a given quarter. Our text data reflects this parliamentary debate. In fact, as argued by

[18] The additional data series are taken from the FRED database of the Federal Reserve Bank of St. Louis. The series ID are: unemployment rate (LMUNRRTTDEQ156S), consumer price index (DEUCPIALLMINMEI), long-term bond yield ( IRLTCT01DEQ156N), short-term interbank rate (IR3TIB01DEQ156N), real effective exchange (CCRETT01DEQ661N), nominal exports (DEUGDPNQDSMEI), nominal imports (DEUIMPORTQDSMEI) and nominal GDP (DEUEXPORTQDSMEI).

[19] Tenhofen et al. (2010) apply this identification for VAR model with German data.

Mertens & Ravn (2010), V. Ramey (2011) and V. Ramey (2016), among others, changes in government spending could be anticipated several quarters in advance. Ordering government spending first thus implies that VARs do not identify unanticipated government spending shocks. We will revisit this issue in the next section.

Including information on the fiscal sentiment expressed in parliament alleviates this concern. We order sentiment last. Hence, a change in fiscal sentiment as expressed in Bundestag speeches should not contemporaneously drive either government expenditure, nor real GDP or real consumption. At the same time, fiscal sentiment is contemporaneously responding to the business cycle.[20] If sentiment is informative about fiscal policy, we should expect that an exogenous increase in sentiment, i.e. a shift towards a more expansionary policy stance, raises government expenditure and economic activity. This effect should be more pronounced for the sentiment of speakers of the parties forming the government compared to opposition speakers.

### 8.4.4 Results

Figure 8.10 shows the responses of our endogenous variables to an increase in the fiscal sentiment of the entire Bundestag one standard deviation in size. All figures also include probability bands that cover 68% and 90% of all draws. As a consequence of the shock, government expenditure increase strongly by about 0.5%. This response is highly significant and very persistent.[21] Hence, a shift towards a more expansionary policy stance as reflected in the speeches held in the Bundestag does indeed cause a subsequent increase in government spending. We also see that the increase in spending needs time to unfold: the increase becomes significant six quarters after the impulse to sentiment. The additional spending has real economic effects: real GDP as well as real private consumption increase by about 0.4%.

In Figure 8.11, we depict the responses to fiscal sentiment as reflected in the speeches of the members of parliament who belong to the governing parties. While the increase in government spending is slightly smaller than in Figure 8.10, the overall macroeconomic effects are somewhat larger. Again, expansionary fiscal policy has a strong impact on income and consumption. If we include only sentiment in those speeches that are delivered by the opposition parties, see Figure 8.12, the increase in government spending is similar. Nevertheless, the implied fiscal multiplier, i.e. the response of GDP relative to the response of spending, is higher for a shock to government sentiment compared to opposition sentiment as we will show below.

[20] The results remain unchanged if we order fiscal sentiment second, i.e. after government spending but before GDP and consumption.

[21] Fisher & Peters (2010) also find quite persistent effects of (defense) spending.

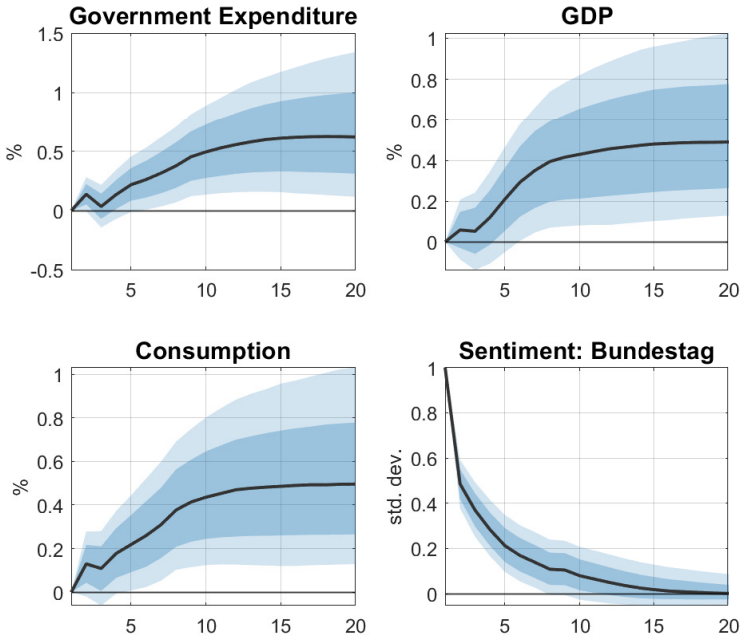



**Figure 8.10:** Response to fiscal sentiment (entire Bundestag)

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

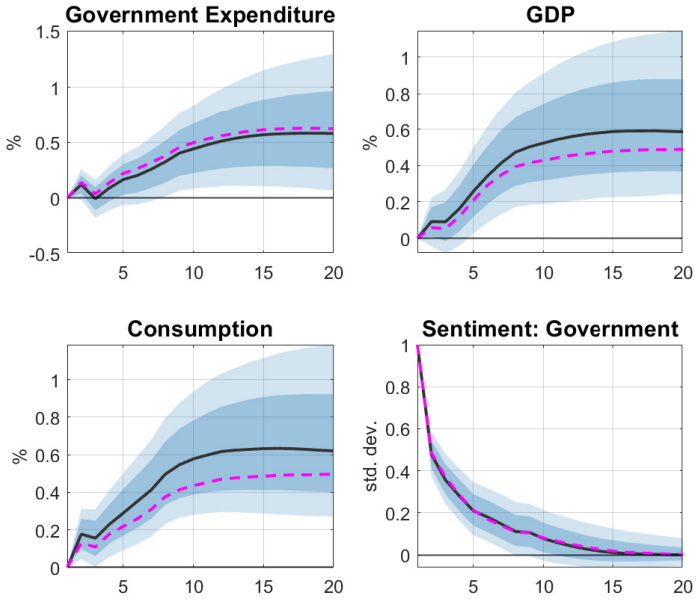



**Figure 8.11:** Response to fiscal sentiment (government)

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of members of the parties forming the government. The dashed line is the response from the model with speeches from the entire Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

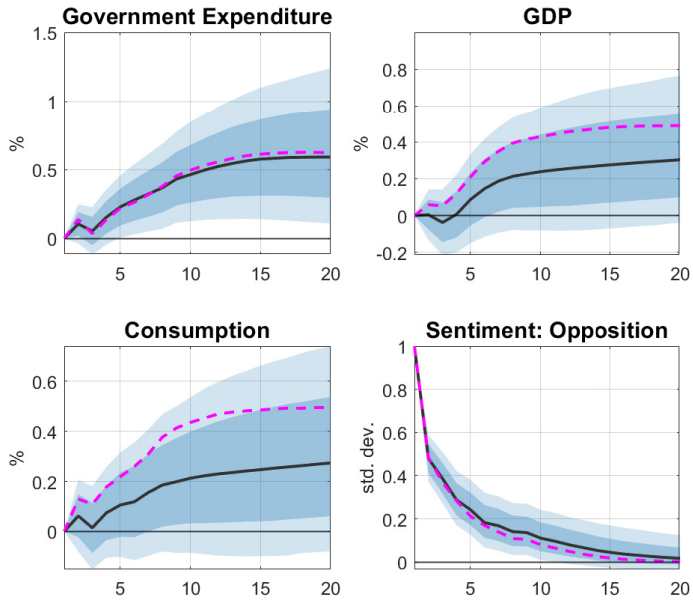



**Figure 8.12:** Response to fiscal sentiment (opposition)

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members in the opposition. The dashed line is the response from the model with speeches from the entire Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

We now extend the baseline model by additional variables. In the first alternative model, we include real investment and unemployment. Both variables are also ordered behind government expenditure but before our sentiment indicator. Figure 8.13 reports the corresponding impulse response functions. The shift in sentiment causes a strong increase in government expenditure and a significant increase in private consumption and investment. As expected, the response of investment is larger than the response of consumption. The unemployment rate falls after the shock, which is in line with the economic expansion.

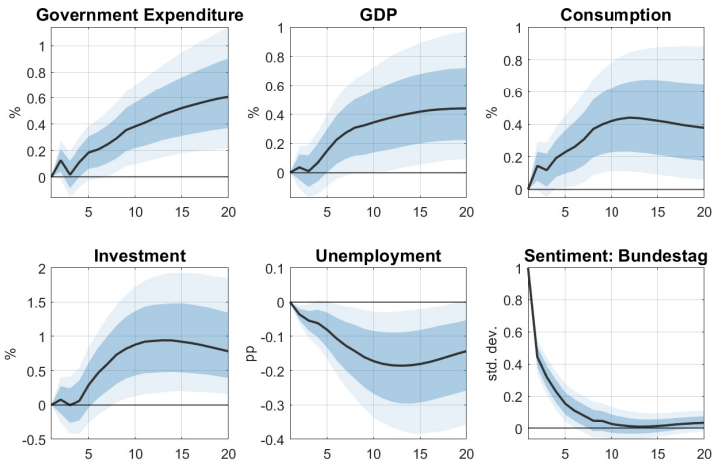



**Figure 8.13:** Response to fiscal sentiment (Bundestag): extended VAR

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

In the second alternative model, we augment the baseline variables by government revenues and the real interest rate measured by the difference between the long-term German bond yield and the year-on-year inflation rate. Figure 8.14 shows that more expansionary sentiment causes an increase in government revenues. As our sentiment variables reflects both sentiment on spending and taxation, this result suggests that the expansion of economic activity, which should raise revenues, is stronger than the drop in revenues, which could results from lower taxes. As in V. Ramey (2016), a fiscal expansion leads to a drop in the real interest rate. In the appendix, we show that this is because the increase in inflation overcompensates the response of the nominal interest rate. Hence, the expansionary fiscal impulse is leading to inflationary pressure, which is in line with standard New-Keynesian models.

In a third extension, we include the central government's budget balance and the ifo index of business expectations. Due to a lack of quarterly data, we had to interpolate the annual data series of the budget balance in percent of GDP. When sentiment in speeches turns more expansionary, the budget balance falls significantly and business expectations improve, see Figure 8.15.

We study a fourth alternative model specification that reflects the open-economy transmission of fiscal policy. In addition to the four variables of the baseline model, we include

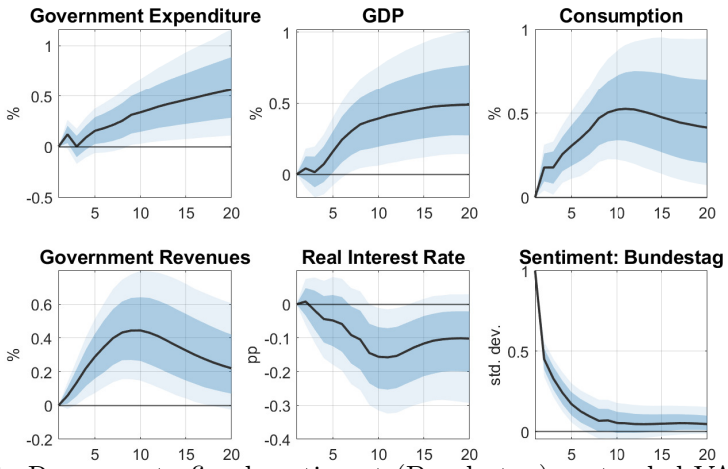



**Figure 8.14:** Response to fiscal sentiment (Bundestag): extended VAR

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.


**Figure 8.15:** Response to fiscal sentiment (Bundestag): extended VAR

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

the real effective exchange rate and the trade balance relative to GDP. Again, both variables are ordered after government expenditure but before fiscal sentiment. Textbook models of an open economy suggest that a fiscal expansion, here reflected by a shifts towards a more expansionary sentiment, causes a real appreciation of the domestic currency and a deterioration of the trade balance. Figure 8.16 shows that the impulse responses are perfectly in line with standard models. Germany experiences a real appreciation of about 0.7% after 10 quarters as well as a drop in the trade balance by 0.2 percentage points.[22]

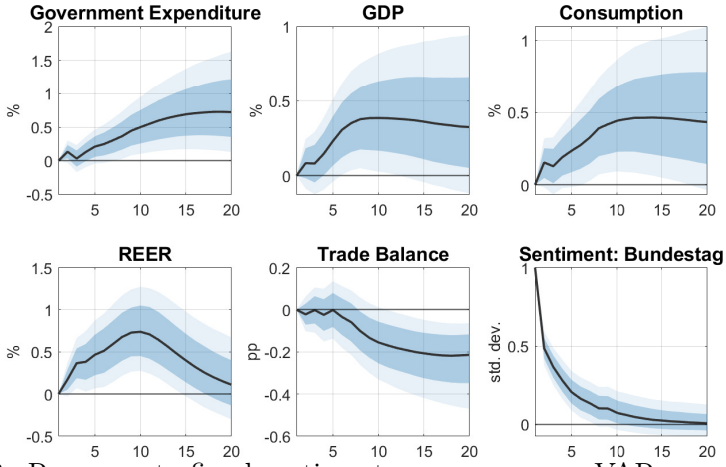



**Figure 8.16:** Response to fiscal sentiment: open-economy VAR

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

Taken together, these findings suggest that a change in the fiscal sentiment expressed in the Bundestag does indeed have real economic effects. Our results lend support to the Keynesian paradigm, i.e. suggesting that expansionary fiscal policy does indeed increase income and consumption.[23]

### 8.4.5 Robustness

In our baseline model, we include the log-levels (times 100) of macroeconomic aggregates such as GDP, consumption and government expenditure. An alternative would be to detrend the three macroeconomic variables. We follow Gordon & Krenn (2010), V. Ramey (2016), V. A. Ramey & Zubairy (2018) and Ilori et al. (2012) and detrend each variable using the trend in real GDP, i.e. we include variable $x_t$, which is either real GDP, real consumption or real government expenditure, as $100 \times (\ln x_t - \ln y_t^{trend})$, where $y_t^{trend}$ is the estimated trend in real GDP. We derive $y_t^{trend}$ either from fitting a quadratic trend to log GDP as in Gordon & Krenn (2010), V. Ramey (2016) and V. A. Ramey & Zubairy (2018) or from applying the

[22] Thus, our findings are in line with theory and do not exhibit a puzzling depreciation after an expansionary policy, see Forni & Gambetti (2016) and Ferrara et al. (2021) for this debate.

[23] In the appendix, we show that the presence of fiscal sentiment in the VAR model also tends to weaken the transmission of shocks to government spending to GDP.

Hamilton (2018) filter to log real GDP as in Ilori et al. (2012).

The results based on the quadratic trend in GDP are shown in Figure 8.17. An unexpected increase in fiscal sentiment raises government expenditure, GDP as well as private consumption. All responses are distinct from zero and look similar to the results based on the log-level variables presented in the previous section. In Figure 8.18, we show the impulse responses based on the GDP trend derived from the Hamilton (2018) detrending procedure. The change in fiscal sentiment still pushes up private consumption.

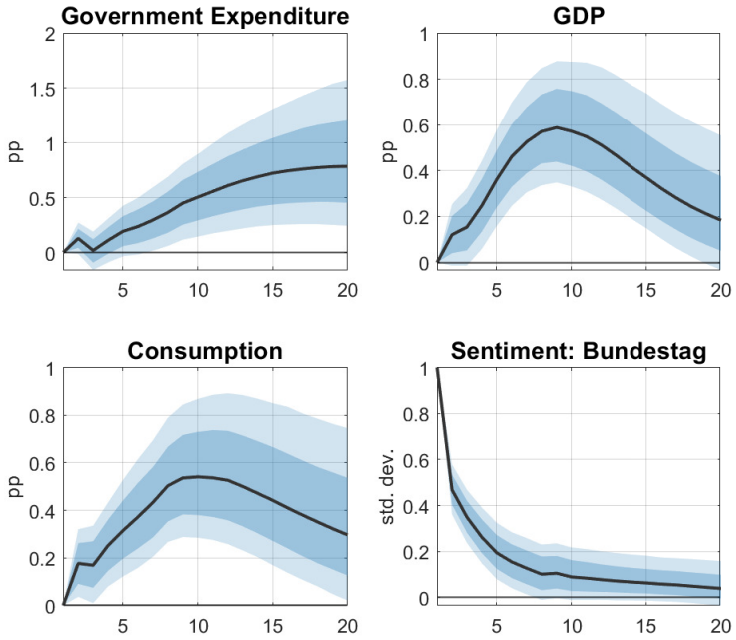



**Figure 8.17:** Response to fiscal sentiment: detrending

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment. The three macroeconomic variables are detrended by the quadratic trend in GDP. All responses are derived from a recursively identified Bayesian VAR model with alternative lags orders $p$ and Normal-Wishart priors. The purple responses are derived from the model estimated by OLS. The shaded areas cover 68% and 90% of all draws.

How strong is the response of GDP relative to the increase in government spending? We address this question by calculating the cumulative response of GDP divided by the cumulative response of government expenditure. This calculation is often used to quantify the fiscal multiplier, e.g. V. A. Ramey & Zubairy (2018). However, it should be emphasized that this is an *implied* multiplier only as we do not look at the consequences of an increase in government spending but rather an increase in fiscal sentiment. Figure 8.19 reports these multipliers for each horizon of the impulse response. For the baseline model, the multiplier remains slightly below one.[24] It is larger for the Gordon & Krenn (2010) detrending and lower if we apply the Hamilton (2018) filter. Interestingly, the implied multiplier is much

[24] The size of the multiplier is in line with the literature: V. A. Ramey (2011) argues that the spending multiplier is between 0.8 and 1.5 for the U.S. economy. For Germany, Tenhofen et al. (2010) estimates multipliers larger than ours, while Berg (2016) finds smaller multipliers. Our results are consistent with the literature on fiscal news shocks. V. Ramey (2011) and Ben Zeev & Pappa (2017) also find that spending multipliers for anticipated fiscal spending are rather

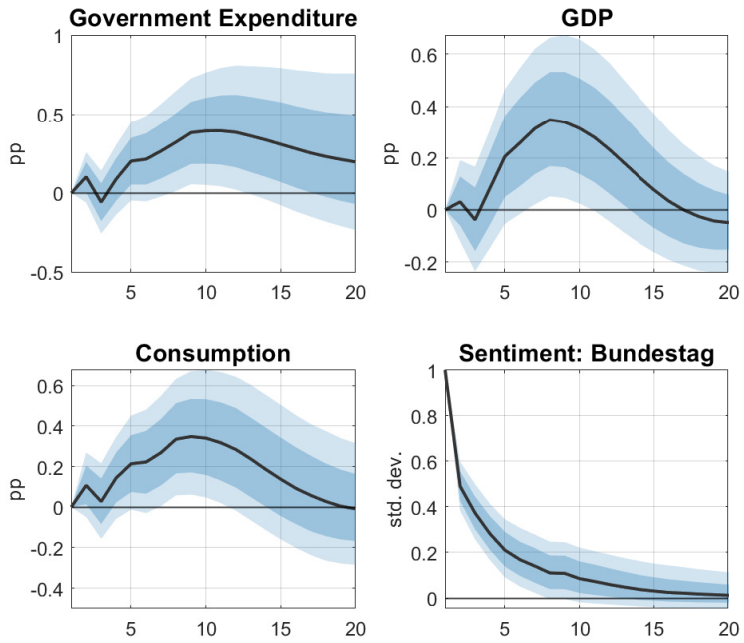



**Figure 8.18:** Response to fiscal sentiment: detrending

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment. The three macroeconomic variables are detrended by the Hamilton (2018) trend in GDP. The black responses are derived from a recursively identified Bayesian VAR model with alternative lags orders $p$ and Normal-Wishart priors. The purple responses are derived from the model estimated by OLS. The shaded areas cover 68% and 90% of all draws.

larger following an increase in government sentiment compared to opposition sentiment. One potential reason for that could be that the composition of spending is different when the increase in spending is triggered by opposition sentiment compared to government sentiment.

### 8.4.6 Exogenous and endogenous fiscal sentiment

In the estimated VAR model of the previous section, we ordered sentiment last, thus treating each aspect of fiscal policy alike by assuming that it drives real economic activity with a lag of at least one quarter. We now estimate the model based on the distinction between sentiment about exogenous and endogenous fiscal policy, respectively.

Exogenous sentiment, $Senti_t^{j,exo}$, is ordered first as it is not contemporaneously driven by the remaining variables.[25] Endogenous sentiment, $Senti_t^{j,endo}$, is ordered last as it is contemporaneously responsive to the other variables in the model. The vector of variables becomes

$$y_t' = \left[ Senti_t^{j,exo} \quad Gov_t \quad GDP_t \quad Cons_t \quad Senti_t^{j,endo} \right]. \tag{8.4.4}$$

Figure 8.20 reports the estimated responses to an increase in exogenous sentiment. The shock to exogenous sentiment causes a significant increase in each of the four other variables. The responses are quantitatively similar to the responses in the baseline model with a

large.

[25] It should be stressed that despite the labeling the sentiment series is still endogenous in the sense that it potentially responds to the other variables with a delay of one quarter.

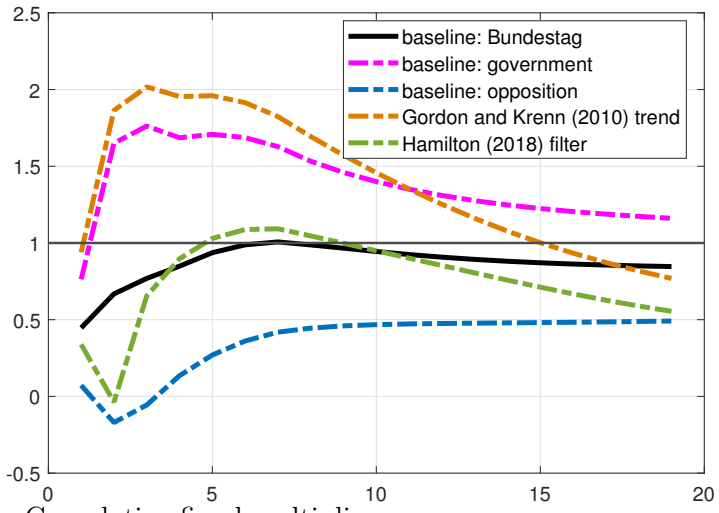



**Figure 8.19:** Cumulative fiscal multipliers

*Notes:* The figure shows the cumulative responses of GDP to an increase in fiscal sentiment relative to the cumulative response of government expenditure for alternative model specifications.

cumulative multiplier below one. An increase in the endogenous fiscal sentiment, Figure 8.21, has no significant effect on the other variables. Think about shocks to endogenous sentiment as deviations of fiscal stabilization from the implicit policy rule. These deviations are not reflected in higher expenditure and have no effect on GDP and private consumption. The results suggest that most of our baseline findings are in fact driven by shifts in exogenous fiscal sentiment.

## 8.5 Fiscal foresight revisited

The literature on government spending shocks argues that fiscal foresight invalidates the recursive Blanchard-Perotti identification, see Mertens & Ravn (2010), V. Ramey (2011) and V. Ramey (2016) and Ellahie & Ricco (2017). Our data set allows us to examine the degree to which fiscal sentiment expressed in parliamentary speeches allows the public to forecast government spending shocks. In other words, we check whether the estimated government spending shock in a Blanchard & Perotti (2002) style model with government spending ordered first in a recursive VAR is predicted by our fiscal sentiment index. We estimate a VAR model with the three core variables used before: government spending, GDP and consumption. We use all three alternative treatments of the variables, i.e. log-levels, quadratic detrending and Hamilton-detrending and do not include sentiment at this stage. Importantly, we adopt the recursive Blanchard-Perotti identification scheme that orders government spending first. Hence, government spending is predetermined with respect to output and consumption. This provides us with three alternative series of structural government spending shocks - one for each treatment of the endogenous variables.

In the next step, we assess whether sentiment contains information that allows us to predict future government spending shocks. The following model in the spirit of Jordà (2005)

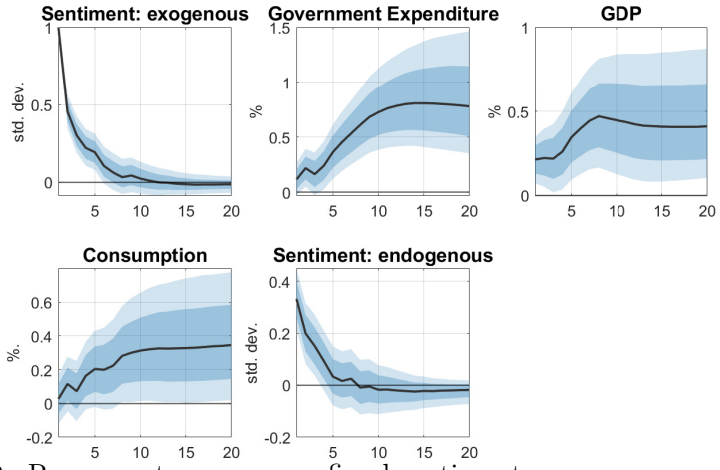



**Figure 8.20:** Response to exogenous fiscal sentiment

*Notes:* The figure shows the responses of the endogenous variables to an increase in exogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

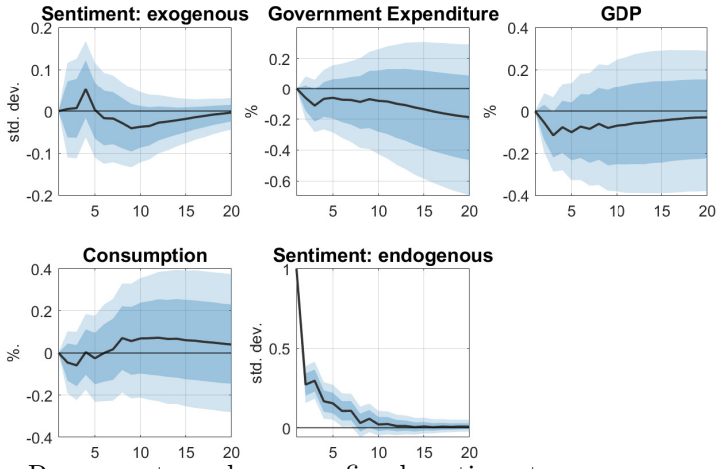



**Figure 8.21:** Response to endogenous fiscal sentiment

*Notes:* The figure shows the responses of the endogenous variables to an increase in endogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

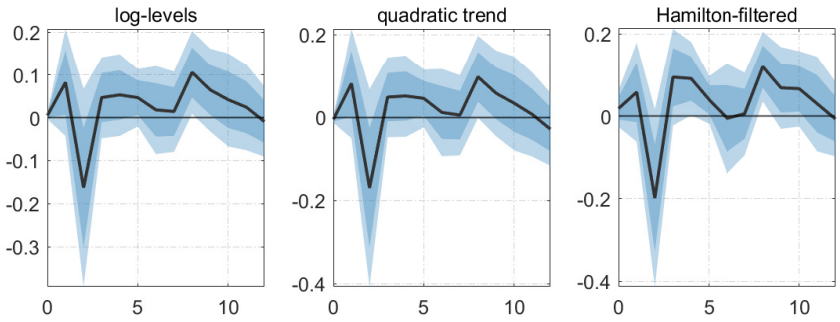
regresses the government spending shock for model $k$ at time $t+h$ on the parliamentary sentiment at time $t$

$$shock^k_{t+h} = \alpha_h + \beta_h Senti^j_t + \gamma_h X_{t-1} + \varepsilon_{t+h} \tag{8.5.1}$$

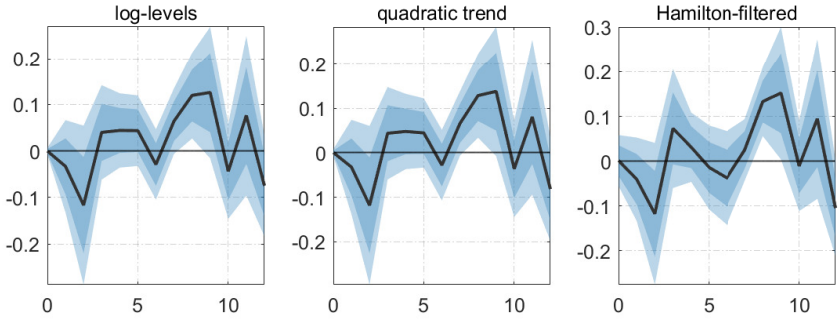
with $j \in (Bundestag, Government, Opposition)$. A significant $\beta_h$ would indicate that current sentiment predicts future government spending shocks. We estimate this model for each shock $k$ as well as for the standardized sentiment of the government, the opposition and the entire Bundestag. The vector $X_t$ contains contemporaneous and lagged realizations of GDP, consumption and government expenditure as control variables. As the dependent variable is the result of a structural identification, it should be orthogonal to these control variables. Nevertheless, we include these variables as controls.

Panel (a) of Figure 8.22 plots the estimated $\beta_h$ as a function of $h$ for the sentiment of the entire Bundestag. The results are consistent across the alternative treatments of the variables: A shift towards a more expansionary fiscal sentiment in $t$ predicts an increase in government spending six to eight quarters later. Hence, the government spending shocks are predictable from parliamentary speeches. The results are weakly significant at the 90% level. When we narrow the set of speeches to the members of the governing parties, see panel (b) of the figure, we obtain similar results. Information from speeches of politicians from the opposition parties, see panel (c), does not predict future government spending shocks.

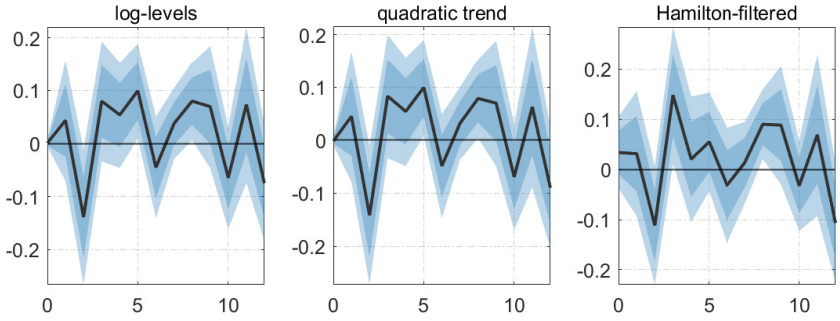
Overall, this section supports the notion that government spending shocks identified from a recursive ordering can indeed be anticipated. This also underlines the relevance of the parliamentary process as a source of information for upcoming changes to fiscal policy.




**(a)** Entire Bundestag




**(b)** Government




**(c)** Opposition

**Figure 8.22:** Response of government expenditure shock to sentiment

*Notes:* The figure shows the responses of the recursively identified government expenditure shock to an increase in fiscal sentiment. The shaded areas reflect 68% and 90% confidence bands constructed from Newey-West standard errors.

## 8.6 Conclusions

This paper went to the roots of fiscal policy-making - the debate in parliament. We use the full set of parliamentary speeches delivered in the German Bundestag as a source of information about fiscal preferences. An embedding-based approach using the latest advances in text mining provides us with a sentiment index on a scale from expansionary to restrictive that summarizes the debate about fiscal policy. This sentiment series has real economic effects: recursively identified VAR models suggest that an increase in fiscal sentiment towards a more expansionary policy stance increases government spending, output and consumption. Hence, a change in fiscal sentiment has macroeconomic effects consistent with standard New-Keynesian business cycle models.

We draw two main conclusions: First, we believe textual data to be very informative about economic policy-making. The rich information incorporated in parliamentary speeches is particularly promising for researchers interested in fiscal policy. In this paper, we focused on the consequences of fiscal sentiment for government expenditure and the macroeconomy. In follow-up work, we will study the consequences of disagreement about fiscal policy between the government and the opposition. Using this data set to assess the consequences of sentiment on the revenue side of public finances could also be an interesting way forward.

Second, the identification of government spending shocks often rests on the assumption that the part of government spending not forecastable from lags of the endogenous variable is a suitable exogenous shock. Information from parliamentary debates about fiscal policy might help enhance this identification. As parliamentary speeches partly forecast future expenditure, only the part of government expenditure that is orthogonal to lags of business cycle variables as well as lags of textual information from the parliamentary debates should qualify as a government expenditure shock.

# Appendix F

## F.1 Further Details on Doc2Vec

### F.1.1 Technical Details on Doc2Vec

The main idea behind the Word2Vec algorithm (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) is to learn distributed representations, whereby semantically similar words are grouped together (linguistic regularities and patterns are encoded). The main model behind the algorithm is the so-called skip-gram model whose aim is to predict surrounding words based on a given center word. For more details on the skip-gram model, see the original work by Mikolov, Chen, et al. (2013).

We build both word and document representations. Doc2Vec models are trained using Python's `gensim` package (Řehůřek & Sojka, 2010). We fit each model with a learning rate of 0.025 (default) and 100 epochs to achieve convergence. We set the dimension of the resulting vectors space to 100. We use a window of five words before and after the center word during the learning phase.

### F.1.2 Illustrating the method

Let us consider the speech given by Olaf Scholz, the Federal Minister of Finance, on March 25, 2020 as an example.[26] This speech is classified as rather expansionary with a sentiment score of 0.06. The speech is about the Covid-19 pandemic and the economic stimulus packages needed. Figure F.1.1 shows a two-dimensional representation of the fiscal policy-related vectors from the first quarter of 2020. Here, we drastically reduce the dimensions in order to illustrate the approach. We chose a projection such that the resulting expansionary and contractionary vectors are orthogonal. They are shown as green and red arrows, respectively. Correspondingly, green and red points represent single expansionary and contractionary terms, which are used to construct the aggregated vectors. The angle bisector (grey dashed line) allows for a rough division of the points into expansionary (area below grey dashed line) and contractionary (area above grey dashed line).

The speech given by Scholz is represented by the diamond-shaped marker. This representation suggests that the speech is rather related with an expansionary path of fiscal policy,

[26] The speech is available at `https://www.bundesregierung.de/breg-de/suche/rede-des-bundesministers-der-finanzen-olaf-scholz--1735392`.

which is also in line with human interpretation and our own understanding based on the content of the speech. However, one should keep in mind that the visualization based on the reduction of 100 to just two dimensions cannot fully capture the trained characteristics of the presented word and documents.

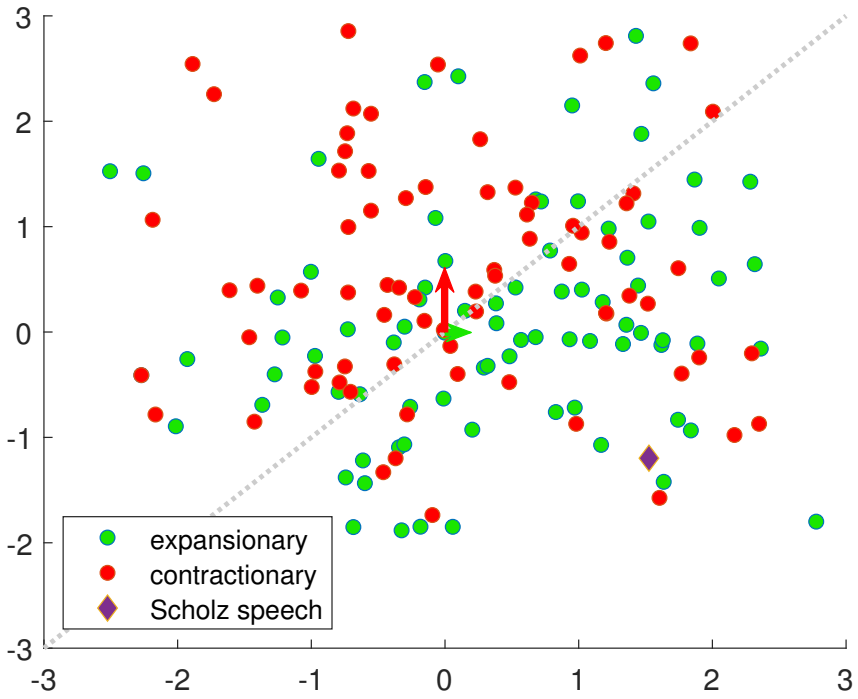



**Figure F.1.1:** Visualization of trained text vectors for 2020Q1

*Notes:* The figure shows a two-dimensional projection of the constructed expansionary (green arrow) and contractionary (red arrow) vectors as well as selected fiscal policy-related terms in 2020Q1. Further, a speech given by Olaf Scholz is represented by the diamond-shaped marker.

### F.1.3 Semantic Features of Doc2Vec

The most important advantage of text embeddings is their ability to capture associative meaning between words and documents. As described in the main paper, we apply a rolling window architecture and divide the whole period into training and forecast periods. This way we obtain 207 distinct training periods, which is equivalent to 207 different models. This allows us to additionally study the dynamics in the usage of language related to certain concepts over time. To demonstrate this feature, we take a closer look at the term "Steuerenlastung" (tax relief). Table F.1.1 summarizes its nearest neighbors based on cosine similarity in different training periods. While in the training period from 1960Q1 to 1969Q4 such words as "Haushaltsmehrbelastung", "Praemienhoechstbetrag", "Verteilerplan" show up in the top 5, "Steuerreform", "Einkommenssteuerreform", "Entlastung", "Mittelstand" show up in the top 5 in the training period from 2000Q1 to 2019Q4. This is also in line with the argumentation by Kapfhammer et al. (2020) who highlight that not only the frequency of a certain concept matters, but also the context.

| | |
|---|---|
| 1960 – 1969 | steuerentlastung, haushaltsmehrbelastung, praemienhoechstbetrag, verteilerplan, aufgedeckt, ausgleichzahlungen, vorversicherungszeiten, lohnsteuerpflichtigen, tagessaetzen, pauschaliert, abschoepfungsvorteile, uebernachtung, kostenvolumen, bagatellfaellen, englaenderin, gesamtvertragssumme, favorisierten, kaufkraftausgleiche, bewerbes, wiedergutmachungen |
| 1970 – 1979 | steuerentlastung, steuerentlastungen, entlastungsmassnahmen, steuersenkungen, mehrwertsteuererhoehung, steuererleichterungen, steuerreform, wuchtig, ortsgemeinderat, entlastung, belatung, uebergangsjahr, beziehern, konjunkturpolitisch, steuersenkung, nettokreditaufnahme, ausbildungsfreibetrages, kinderadditiven, steueraenderungen, haefele |
| 1980 – 1989 | steuerentlastung, steuersenkung, steuerreform, steuersenkungen, steuerentlastungen, tarifreform, grundfreibetrag, steuerpolitik, verbrauchsteuererhoehungen, steuerlast, spitzenverdiener, entlastungsvolumen, steuererhoehungen, familiensplitting, grundfreibetrages, neuverschuldung, entlastung, einkommensteuer, familienlastenausgleich, steuergeschenke |
| 1990 – 1999 | steuerentlastung, steuerentlastungen, nettoentlastung, steuererhoehungen, entlasten, steuersenkung, entlastung, steuersenkungen, konsolidierung, steuerlasten, steuererhoehung, unternehmensteuerreform, steuereinsparungen, inflationstendenzen, mehrwertsteuererhoehung, besserverdienenden, kindergelderhoehung, sozialtransfers, beschaeftigungspaket, abschreibungsmoeglichkeiten |
| 2000 – 2009 | steuerentlastung, steuerreform, einkommensteuerreform, entlastung, mittelstand, kapitalgesellschaften, steuerentlastungen, steuersenkungen, grundfreibetrag, personenunternehmen, realeinkommen, personengesellschaften, mittelstandes, elterngeld, freibetrag, gewerbesteuer, mehreinnahme, familienfoerderung, entlastungs, einkommensteuer |
| 2010 – 2019 | steuerentlastung, steuererhoehung, entlastung, umlage, progression, muetterrente, soli, kinderfreibetrag, autofahrerinnen, solidaritaetszuschlags, elterngeld, vervollstaendigung, neuverschuldung, fleissig, kommensteuer, steuervereinfachung, mittelstandsbauch, strukturverbesserungen, kindergeldes, energieeinsparmassnahmen |

**Table F.1.1:** Term usage over time

*Notes:* These examples show the ability of Doc2Vec to capture the change of semantics over time on the example of the term "Steuerentlastung". The first word is the word itself. The other words are ordered based on their cosine similarity to the defined word.

## F.2 Dictionary-based sentiment

We want to compare our index to alternative measures derived from textual information. However, as our paper is the first to construct a long time series reflecting the fiscal sentiment of the German Bundestag, no further indices are available for comparison. In standard sentiment analysis, sentiment scores are often obtained based on counting the occurrence of specific words, e.g. words carrying a positive vs negative meaning. Here, we use a

dictionary-based approach for comparison. Instead of applying Doc2Vec and working with text representations, we count the relevant expansionary (*#expansionary*) and contractionary (*#contractionary*) terms and divide this count by the total number of words (*#words*) in the document as in equation (F.2.1) or by the sum of expansionary and contractionary words as in equation (F.2.2), i.e.

$$\text{Dictionary}_1 = \frac{\#expansionary - \#contractionary}{\#words} \tag{F.2.1}$$

or

$$\text{Dictionary}_2 = \frac{\#expansionary - \#contractionary}{\#expansionary + \#contractionary}. \tag{F.2.2}$$

We then aggregate the series to quarterly frequency. Figure F.2.2 presents the standardized fiscal policy sentiment indicator for the entire Bundestag and two dictionary-based alternatives. The overall evolution of the two dictionary-based series shares some similarities with our Doc2Vec benchmark. However, both series are only weakly correlated with our Doc2Vec-based sentiment. The contemporaneous correlation is 0.11 ($p$=0.10) with $Dictionary_1$ and 0.26 ($p$=0.00) with $Dictonary_2$. In addition, there are notable and persistent discrepancies between them, e.g. in the mid-1980s or the 2000s.

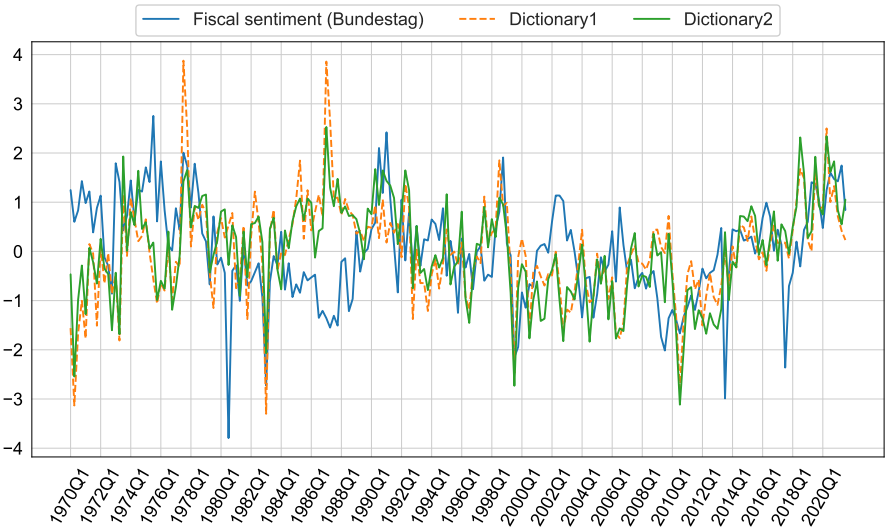



**Figure F.2.2:** Comparing fiscal policy sentiment in the Bundestag across methods

*Notes:* The figure shows the standardized Doc2Vec-based fiscal policy sentiment indicator for the entire Bundestag as well as the two dictionary-based reference series.

We now replace our baseline sentiment series in the estimated VAR models with the series derived from the two alternative dictionary-based approaches. Figure F.2.3 shows the response to an increase in fiscal sentiment (based on method I) for the entire Bundestag. As in the main part of the paper, the responses of the remaining three endogenous variables are positive and significant. The same holds for the alternative dictionary-based sentiment series (method II) shown in Figure F.2.4.

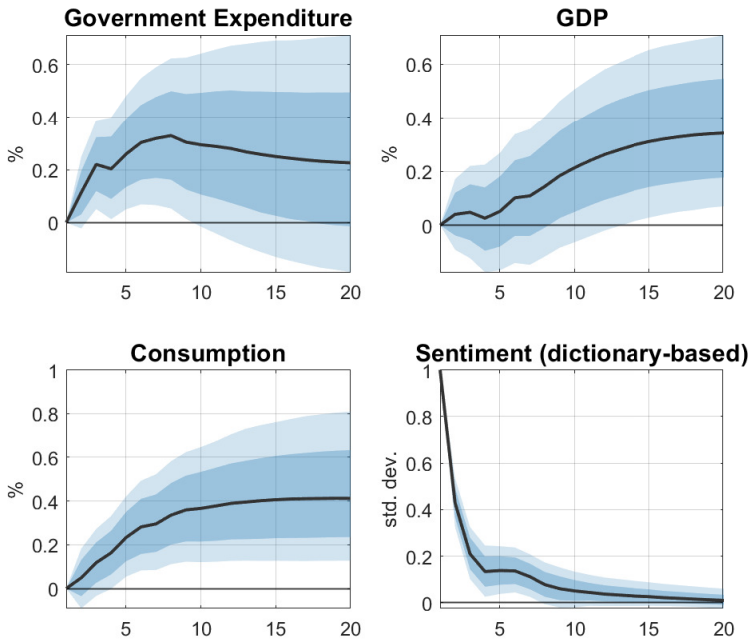



**Figure F.2.3:** Response to fiscal sentiment (dictionary I)

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

Based on the topics delivered by the LDA approach as described in more detail in the main part, we derive dictionary-based indicators for endogenous and exogenous fiscal sentiment, respectively. Using these indicators in the VAR models, see Figure F.2.5, Figure F.2.6, Figure F.2.7 and Figure F.2.8, the qualitative findings remain broadly unchanged when compared to the results from the main part of the paper. This supports the robustness of our results.

Nevertheless, we prefer the Doc2Vec approach presented in the main part of the paper for two reasons. The first advantage is the property of Doc2Vec of not only putting (high) weights to words from the dictionary. Rather, it also takes into account further semantically related words even if these terms are not included in the dictionary. This is achieved by exploiting the implicit relationships and structures within the documents. The second advantage is that the Doc2Vec approach is able to capture dynamics in language usage. In particular, changes in concepts of fiscal policy will be taken into account, which is not possible with the dictionary-based approach.

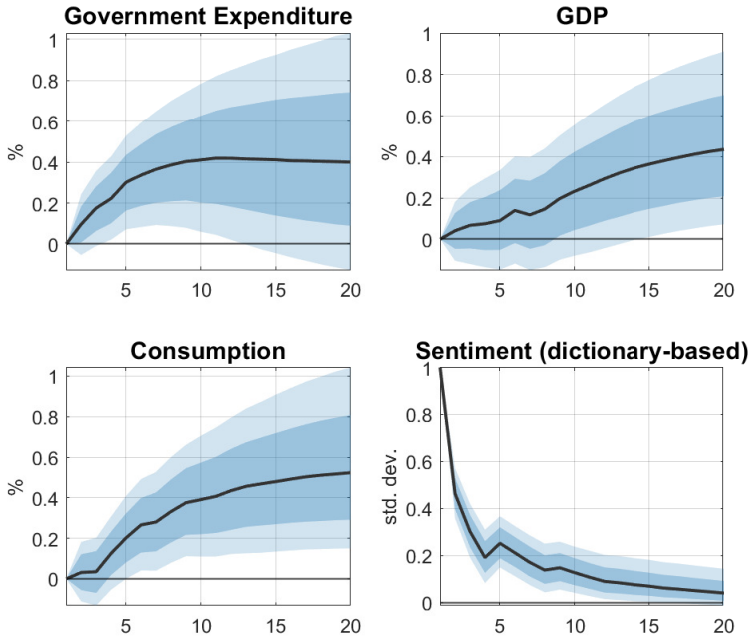



**Figure F.2.4:** Response to fiscal sentiment (dictionary II)

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

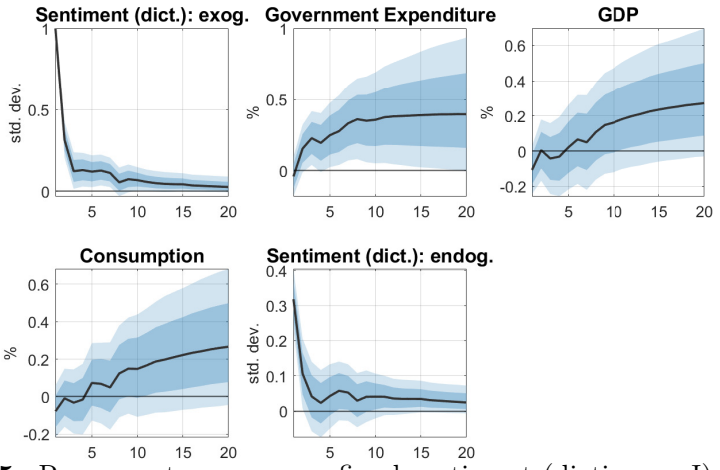



**Figure F.2.5:** Response to exogenous fiscal sentiment (dictionary I)

*Notes:* The figure shows the responses of the endogenous variables to an increase in exogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

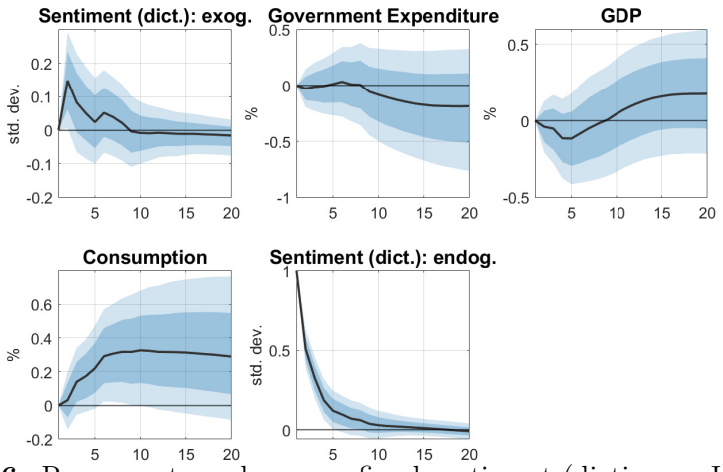



**Figure F.2.6:** Response to endogenous fiscal sentiment (dictionary I)

*Notes:* The figure shows the responses of the endogenous variables to an increase in exogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

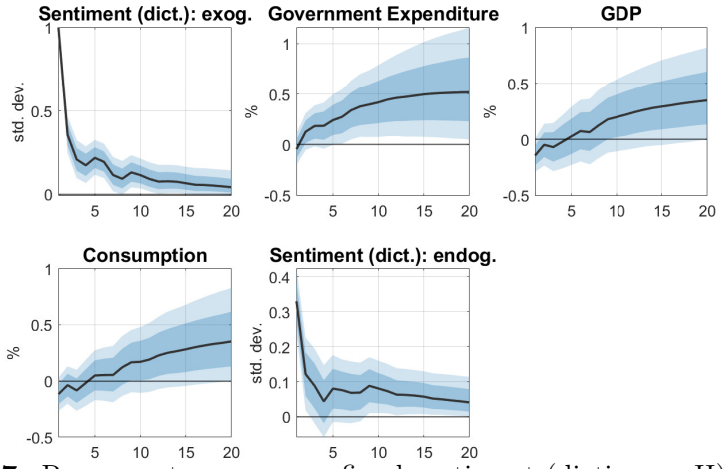



**Figure F.2.7:** Response to exogenous fiscal sentiment (dictionary II)

*Notes:* The figure shows the responses of the endogenous variables to an increase in exogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

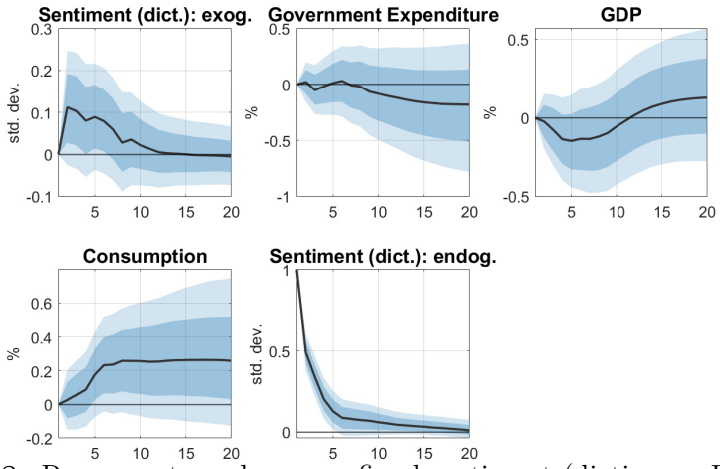



**Figure F.2.8:** Response to endogenous fiscal sentiment (dictionary II)

*Notes:* The figure shows the responses of the endogenous variables to an increase in exogenous fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

## F.3 Stop words

The following list contains stop words:

ab, aber, abg, abgegeben, abgelehnt, abgeordnete, abgeordneten, abgeordneter, abs, absicht, abstimmen, abstimmung, abzulehnen, aendern, aenderung, aenderungsantrag, aktuellen, all, alle, allein, allem, allen, aller, allerdings, alles, allgemein, allgemeine, allgemeinen, als, also, alten, am, an, ander, andere, anderem, anderen, anderer, andererseits, anderes, anderm, andern, anderr, anders, anfang, anfrage, angeht, angelegenheit, angenommen, angesichts, angesprochen, anlage, anlass, anliegen, annahme, annehmen, ansatz, ansicht, anspruch, anteil, antraege, antrag, antrage, antrags, antwort, antworten, april, arbeit, arbeiten, arbeitnehmer, arbeitsplaetze, art, artikel, auch, auf, auffassung, aufgabe, aufgaben, aufgrund, aufmerksam, aufmerksamkeit, auftrag, augen, augenblick, august, aus, ausdruck, ausdrucklich, ausdruecklich, ausfuehrungen, ausgaben, ausgefuehrt, ausgehen, ausgesprochen, ausschuessen, ausschuss, ausschusse, ausschusses, ausschuss, aussprache, auswaertigen, auswirkungen, ausserdem, ausserordentlich, bald, beamten, beantragt, beantworten, bedarf, bedenken, bedeutet, bedeutung, bedingungen, befasst, beginn, begruenden, begruendet, begruendung, begruessen, behandeln, behandelt, behandlung, bei, beide, beiden, beim, beispiel, beispiele, beispielsweise, beitrag, bekannt, bekommen, bemerkung, bemerkungen, bemuehen, bemueht, bemuehungen, beraten, beratung, beratungen, bereich, bereichen, bereit, bereits, bericht, berichterstatter, berlin, beruecksichtigt, beschaeftigen, beschaeftigt, beschlossen, beschluss, beschlussempfehlung, besondere, besonderen, besonders, besser, bessere, bestehen, besteht, bestimmt, bestimmte, bestimmten, bestimmung, bestimmungen, beteiligt, beteiligten, betrag, betreiben, betriebe, betrieben, betrifft, betroffenen, bevoelkerung, bevolkerung, bevor, bewusst, bewusst, beziehungen, bezueglich, bezug, bin, bis, bisher, bisherigen, bisschen, bist, bitte, bitten, bisschen, bleiben, bleibt, blick, bonn, brauchen, braucht, bringen, bringt, buendnis, buerger, buergerinnen, bund, bundes, bundeskanzler, bundeskanzlerin, bundeslaender, bundeslaendern, bundesminister, bundesministerium, bundesrat, bundesregierung, bundesrepublik, bundestag, bundestages, bundnis, burger, burgerinnen, but, bzw, cdu, chance, cronenberg, csu, da, dabei, dadurch, dafuer, dafur, dagegen, daher, dahin, damals, damen, damit, danach, dank, dankbar, danke, dann, daran, darauf, daraus, darf, darin, daruber, darueber, darum, das, dass, dasselbe, davon, dazu, dass, debatte, dehler, dein, deine, deinem, deinen, deiner, deines, dem, dementsprechend, demokratie, demokratischen, demselben, den, denen, denke, denken, denn, dennoch, denselben, der, deren, derer, derselbe, derselben, derzeit, des, deshalb, desselben, dessen, deswegen, deutlich, deutsche, deutschen, deutscher, deutschland,

deutschlands, dezember, dich, die, diejenigen, dienen, dienst, dienstag, dies, diese, dieselbe, dieselben, diesem, diesen, dieser, dieses, dieter, dinge, dingen, dir, diskussion, diskutieren, diskutiert, dm, doch, donnerstag, dort, dr, draussen, drei, dringend, dritte, dritten, drittens, drucksache, drucksachen, du, duerfen, durch, durchaus, durchfuehrung, durchgefuehrt, durfen, eben, ebene, ebenfalls, ebenso, egal, ehemaligen, eher, ehlers, ehrlich, eigene, eigenen, eigentlich, ein, eindeutig, eindruck, eine, einem, einen, einer, eines, einfach, einfuehrung, eingebracht, eingehen, eingehend, eingeraumt, eingesetzt, einheit, einig, einige, einigem, einigen, einiger, einiges, einkommen, einmal, einstimmig, eintreten, einverstanden, einzelne, einzelnen, einzige, ende, endlich, enthalten, enthaltungen, entscheiden, entscheidende, entscheidung, entscheidungen, entschieden, entsprechend, entsprechende, entsprechenden, entspricht, entstehen, entwicklung, entwurf, entwurfs, er, erfahrungen, erfolg, erfolgen, erfolgt, erforderlich, erfuellen, erfuellt, ergeben, ergebnis, ergibt, erhalten, erheben, erhebliche, erhoehung, erinnern, erkennen, erklaeren, erklaert, erklaerung, erklaren, erleben, erledigt, ermoglichen, erneut, ernst, erreichen, erreicht, erscheint, erst, erste, ersten, erstens, erster, erwaehnt, erwahnt, erwarten, es, etwa, etwas, euch, euer, eure, eurem, euren, eurer, eures, euro, europa, europaeische, europaeischen, europaische, europaischen, faelle, faellen, fall, falle, fallen, falsch, familie, fassung, fast, fdp, februar, ferner, fest, festgestellt, feststellen, feststellung, festzustellen, finde, finden, findet, foerderung, folge, folgen, folgende, folgendes, fordern, forderung, forderungen, form, formuliert, forsten, frage, fragen, fraktion, fraktionen, frankreich, frau, frauen, freien, freiheit, freitag, freue, freunde, froh, frueher, frueheren, fuehren, fuehrt, fuenf, fuer, fuhren, fuhrt, funcke, funf, funktioniert, fur, fuer, gab, ganz, ganze, ganzen, gar, geaendert, gebe, geben, gebiet, gebiete, gebieten, gebracht, gedanken, geehrte, geehrten, geehrter, gefahr, gefordert, gefragt, gefuehrt, gefuhrt, gefunden, gegeben, gegen, gegenprobe, gegenteil, gegenuber, gegenueber, gehabt, gehalten, gehen, gehoeren, gehoert, gehoren, gehort, geht, gekommen, gelegenheit, geleistet, gelingt, gemacht, gemeinden, gemeinsam, gemeinsame, gemeinsamen, gemeinschaft, genannt, genannten, genau, genauso, genommen, genug, gerade, gerecht, gern, gerne, gerstenmaier, gesagt, gesamte, gesamten, geschaeftsordnung, geschaffen, geschehen, geschichte, geschieht, gesehen, gesellschaft, gesetz, gesetze, gesetzentwurf, gesetzes, gesetzt, gesichtspunkt, gesprochen, gestatten, gestellt, gestern, getan, getragen, getroffen, gewesen, gewisse, gewissen, geworden, gewuenscht, gezeigt, gibt, gilt, ging, glaube, glauben, gleich, gleiche, gleichen, gleichermassen, gleichzeitig, grosse, grossen, grosser, gruende, gruenden, gruenen, grund, grunde, grunden, grundgesetz, grundgesetzes, grundlage, grundsaetzlich, grundsatz, grune, grunen, gut, gute, guten, guter, gutes, hab, habe, haben, haelt, haette, haetten, halte, halten, haltung, hand, handeln, handelt, handzeichen, hans, hat, hatte, hatten, haus, hause, hauses, heisst, helfen, her, heraus, herr, herren, herrn, herzlichen, heute, heutigen, hier, hierher, hierzu, hilfe, hin, hinaus,

hinblick, hingewiesen, hinsichtlich, hinter, hintergrund, hinweisen, hinzu, hoch, hoehe, hoere, hoeren, hoffe, hoffen, hohe, hohen, horen, ich, ihm, ihn, ihnen, ihr, ihre, ihrem, ihren, ihrer, ihres, im, immer, immerhin, in, indem, inhalt, innerhalb, innern, ins, insbesondere, insgesamt, insofern, interesse, interessen, internationale, internationalen, inzwischen, ist, ja, jaeger, jahr, jahre, jahren, jahres, januar, je, jede, jedem, jeden, jedenfalls, jeder, jedes, jedoch, jemand, jene, jenem, jenen, jener, jenes, jetzt, juli, junge, juni, kam, kann, kaum, kein, keine, keinem, keinen, keiner, keines, keineswegs, kennen, kenntnis, klar, klare, klaren, klaus, kleine, kleinen, koalition, koenne, koennen, koennte, koennten, kohl, kollege, kollegen, kollegin, kolleginnen, komme, kommen, kommenden, kommission, kommt, kommunen, konkret, konnen, konnte, konnten, konzept, kosten, kraft, kritik, kuenftig, kunftig, kurz, koennen, koennte, laender, laendern, laengst, laesst, laesst, lage, land, lande, lander, landern, landes, lange, langer, langst, lassen, lasst, last, laufe, laut, least, leben, lediglich, legen, legislaturperiode, leicht, leider, leisten, leistungen, lesen, lesung, letzte, letzten, leute, liebe, lieber, liegen, liegt, linie, linke, linken, loesen, loesung, losen, losung, machen, macht, maerz, mag, mai, mal, man, manche, manchem, manchen, mancher, manches, manchmal, mann, mark, masse, massnahmen, mehr, mehrfach, mehrheit, mein, meine, meinem, meinen, meiner, meines, meinung, meisten, menschen, mich, milliarden, millionen, mindestens, minister, ministerin, ministerium, mir, mit, miteinander, mitglieder, mittel, mitteln, mittlerweile, mittwoch, mochte, moechte, moechten, moeglich, moeglicherweise, moeglichkeit, moeglichkeiten, moeglichst, moglich, monaten, montag, morgen, mueller, muesse, muessen, muesste, muessten, muss, mussen, musste, mussten, muss, musste, nach, nachdem, nachgefragt, nachher, nachste, nachsten, nachster, naechsten, naemlich, namen, namens, namlich, natuerlich, naturlich, neben, nehme, nehmen, nein, nennen, neu, neue, neuen, nicht, nichts, nie, niemand, nimmt, noch, noetig, not, notwendig, notwendige, notwendigen, notwendigkeit, november, nr, nun, nunmehr, nur, nutzen, ob, obwohl, oder, oeffentliche, oeffentlichen, oeffentlichkeit, of, offen, offenbar, offensichtlich, offentlichen, oft, ohne, ohnehin, ok, oktober, opposition, ordnung, ort, ost, paar, parl, parlament, parlamentarischer, parlaments, partei, parteien, passiert, pds, peter, pflicht, plenum, politik, politiker, politisch, politische, politischen, position, praesident, praesidenten, praesidentin, praktisch, prasident, prasidentin, praxis, presse, pro, problem, probleme, professor, programm, prozent, pruefen, pruefung, punkt, punkte, rahmen, rahmenbedingungen, raum, rechnen, rechnung, recht, rede, reden, redezeit, redner, reform, regelung, regelungen, regierung, reicht, reihe, renger, richtig, richtige, richtung, rolle, ruecksicht, rufe, rund, sache, sage, sagen, sagt, sagte, sagten, samstag, satz, schaffen, schauen, scheint, schliesslich, schluss, schluss, schmid, schmidt, schmitt, schnell, schoen, schoettle, schon, schritt, schutz, schwer, schwierigkeiten, sehe, sehen, sehr, sei, seien, sein, seine, seinem, seinen, seiner, seines, seit, seitdem,

seite, seiten, selber, selbst, selbstverstaendlich, september, setzen, setzt, sich, sicher, sicherheit, sicherlich, sicht, sie, sieht, sind, sinn, sinne, sinnvoll, situation, sitzung, so, soeben, sofort, sogar, sogenannte, sogenannten, solche, solchem, solchen, solcher, solches, soll, sollen, sollte, sollten, sondern, sonntag, sonst, sorge, sorgen, soweit, sowie, sowohl, sozialdemokraten, soziale, sozialen, spaeter, spater, spd, sprechen, spricht, staat, staaten, staatsminister, staatssekretaer, staatssekretar, staerker, stand, standpunkt, stark, starken, starker, statt, stehen, steht, stelle, stellen, stellt, stellung, stellungnahme, stimme, stimmen, stimmkarte, stimmt, strauss, stuck, stuecklen, stunde, suessmuth, system, taetigkeit, tag, tage, tagen, tagesordnung, tat, tatsache, tatsachlich, tatsaechlich, teil, teilen, teilweise, the, thema, themen, tragen, treffen, treten, trotz, trotzdem, tun, tut, uber, uberhaupt, ubrigen, ubrigens, ueber, ueberhaupt, ueberlegen, ueberlegungen, ueberzeugt, ueberzeugung, uebrigen, uebrigens, uhr, um, umdruck, umfang, umgesetzt, umsetzung, umstaenden, umwelt, unbedingt, und, union, unmoeglich, uns, unser, unsere, unserem, unseren, unserer, unseres, unter, unterhalten, unternehmen, unterschied, unterstuetzen, unterstuetzung, usw, verabschiedet, verabschiedung, verantwortung, verbessern, verbesserung, verehrte, verehrten, vereinbart, verfahren, verfuegung, verfugung, vergangenen, vergangenheit, vergessen, vergleich, verhaeltnis, verhaeltnisse, verhalten, verhandlungen, verhindern, verlangen, verlangt, verordnung, verpflichtet, verpflichtung, verschiedenen, verstaendnis, verstanden, verstehen, versuch, versuchen, versucht, vertrag, vertreten, vertreter, verwaltung, verwiesen, viel, viele, vielen, vieles, vielleicht, vielmehr, vier, vizepraesident, vizepraesidentin, vizeprasidentin, vockenhausen, voellig, volk, volkes, voll, vollig, vom, von, vor, voraussetzung, voraussetzungen, vorgelegt, vorgenommen, vorgeschlagen, vorgesehen, vorgetragen, vorhanden, vorher, vorhin, vorlage, vorliegen, vorliegenden, vorliegt, vorschlaege, vorschlag, vorschlage, vorschriften, vorstellen, vorstellungen, vorwurf, waehrend, waere, waeren, wahl, wahlperiode, wahrend, wahrheit, wahrscheinlich, wann, war, ware, waren, warst, warum, was, weder, weg, wege, wegen, wehner, weil, weise, weit, weiter, weitere, weiteren, weiterer, weiteres, weiterhin, weiternun, weiterreden, weitgehend, weiss, welche, welchem, welchen, welcher, welches, welt, weltweit, wenig, wenige, wenigen, weniger, wenigstens, wenn, wer, werde, werden, wert, wesentlich, wesentliche, wesentlichen, wichtig, wichtige, wichtigen, wichtiger, widerspruch, wie, wieder, wiederholen, wiederholt, will, willen, wir, wird, wirklich, wirklichkeit, wirst, wirtschaft, wirtschaftliche, wirtschaftlichen, wissen, wo, wobei, woche, wochen, wohl, wolfgang, wollen, wollte, wollten, womoglich, worden, wort, worte, worten, wortmeldungen, woruber, wuenschen, wuenscht, wuerde, wuerden, wunsch, wurde, wurden, waehrend, wuerde, wuerden, zahl, zahlen, zehn, zeigen, zeigt, zeit, zeitpunkt, ziehen, ziel, ziffer, zitiere, zitieren, zitiert, zu, zudem, zugestimmt, zukunft, zuletzt, zum, zumindest, zunachst, zunaechst, zur, zuruck, zurueck, zusaetzlich, zusaetzliche, zusaetzlichen, zusammen, zu-

sammenarbeit, zusammenhang, zusatzfrage, zusatzlich, zustaendigen, zustand, zustimmen, zustimmung, zuzustimmen, zwar, zweck, zwei, zweifel, zweifellos, zweite, zweiten, zweitens, zwischen, zwischenfrage, ueber

## F.4 Terms related to fiscal policy

### Terms reflecting expansionary policy

The following list contains terms reflecting expansionary fiscal policy:

Abgaben senken, zu hohe Abgaben, Abgabenlast beschränken, Abgabenlast vermindern, zu hohe Abgabenlast, Arbeitsbeschaffungsmaßnahme, Arbeitsplätze schaffen, öffentliche Aufträge, Ausgaben erhöhen, mehr ausgeben, Beschäftigungsprogramm, niedrige Besteuerung, deficit spending, Defizitfinanzierung, Einnahmen senken, öffentliche Einnahmen senken, Entlastung, Entlastungsvolumen, entlasten, Beitragsentlastung, Kostenentlastung, Steuerentlastung, fiskalische Belastung vermindern, Fördermaßnahmen, Förderpaket, Förderprogramm, staatliche Fördermaßnahmen, Nachtragshaushalt, Hilfspaket, Hilfsprogramm, Investitionen, Investitionslücke, staatliche Investitionen, Neuinvestitionen, Investitionen erhöhen, Investitionsstau, mehr investieren, Kapitalertragsteuer senken, Kapitalertragsteuer abschaffen, Kaufanreiz, Keynes, keynesianisch, Konjunkturmaßnahme, Konjunkturpaket, Konjunkturprogramm, Konsum erhöhen, Konsumanreiz, Konsumbereitschaft, Konsumbereitschaft erhöhen, Konsumneigung, Konsumneigung erhöhen, Kosten senken, Arbeitskosten senken, Lohnnebenkosten senken, Kosten abschaffen, Kreditaufnahme, Kurzarbeit, Mehrwertsteuer senken, Mehrwertsteuersenkung, Nachfrage erhöhen, nachfragesteigernd, Neuverschuldung, höhere Neuverschuldung, schuldenfinanziert, Schulden aufnehmen, Schulden erhöhen, höhere Staatsschulden, Schuldenbremse abschaffen, Schwarze Null abschaffen, Soli abschaffen, Soli abbauen, Soli reduzieren, Soli streichen, Solidaritätszuschlag abschaffen, Solidaritätszuschlag abbauen, Sondervermögen, Sozialabgaben senken, Staatshilfen, Abschaffung der Steuer, Steuerbelastung reduzieren, Steuerentlastung, auf Steuererhöhung verzichten, Steuererleichterung, Steuerlast senken, Steuern senken, Steuersenkung, Quellensteuer senken, Quellensteuer abschaffen, Reichensteuer senken, Reichensteuer abschaffen, Reichensteuersatz senken, Unternehmenssteuer senken, Umsatzsteuer senken, Verbrauchsteuer senken, Vermögenssteuer senken, Vermögensteuer senken, Einfuhrumsatzsteuer verringern, Einfuhrumsatzsteuer erstatten, heimliche Steuererhöhung abschaffen, kalte Progression abschaffen, Spitzensteuersatz senken, steuerliche Anreizmodelle schaffen, Abzug der Steuerschuld, steuerlich berücksichtigen, negative Einkommenssteuer einführen, Freibetrag erhöhen,

Stimulation, Stimulus, stimulieren, stimulierend, Wirtschaft stimulieren, Subventionen, mehr Subventionen, Thesaurierungsbegünstigung, Transferleistungen erhöhen, Transferzahlungen erhöhen, Wirtschaftskrise abwenden, Wirtschaftskrise abbremsen, Zuschuss, Zuschüsse, crowding in, expansiv

### Terms reflecting contractionary policy

The following list contains terms reflecting contractionary fiscal policy:

Abbau des Defizits, Abgaben erhöhen, Vermögensabgabe, Abgaben einführen, Abgaben fordern, Ausgaben kürzen, Ausgaben senken, Ausgabendisziplin, zu viel ausgeben, zu viele Ausgaben, hohe Besteuerung, Budgetdisziplin, Defizitabbau, Defizitgrenze, Einnahmen erhöhen, öffentliche Einnahmen erhöhen, Einsparungen, Fiskalpakt, Fiskalregel, fiskalische Belastung, Haushaltsdefizit begrenzen, Haushaltsdisziplin, Haushaltsgesetz, Haushaltskonsolidierung, Konsolidierung, Haushaltsregel, Investitionsstopp, weniger investieren, Kapitalertragsteuer erhöhen, Kapitalertragsteuer einführen, Konjunktur, Kosten erhöhen, Kürzungen, Maastricht, Mehrwertsteuer erhöhen, Mehrwertsteuererhöhung, nachfragesenkend, Primärüberschuss, Rücklagen, Rückstellungen, Schuldenabbau, Schuldenbremse, Schuldengrenze, Schuldenregel, Schuldentilgung, Schuldlast senken, Schulden senken, Schuldenaufbau vermeiden, Schwarze Null, Soli einführen, Solidaritätszuschlag einführen, Sozialabgaben erhöhen, sparen, Sparanstrengung, Sparmaßnahme, Sparpolitik, Sparprogramm, Sparaufforderung, Auffordern zum Sparen, Staatsdefizit begrenzen, Stabilitätspakt, Einführung der Steuer, Steuereinführung, Steuererhöhung, Steuerlast erhöhen, Steuern erhöhen, Steuerschlupflöcher schließen, Quellensteuer erhöhen, Quellensteuer einführen, Reichensteuer erhöhen, Reichensteuer einführen, Reichensteuersatz erhöhen, Unternehmenssteuer erhöhen, Umsatzsteuer erhöhen, Börsenumsatzsteuer einführen, Gewerbesteuer, Ökosteuer, Luftverkehrssteuer, Verbrauchsteuer erhöhen, Vermögenssteuer erhöhen, Vermögensteuer erhöhen, Tobinsteuer einführen, Spitzensteuersatz erhöhen, Steuervergünstigungen abbauen, Steuervergünstigungen streichen, Steuersubventionen streichen, Steuersubventionen abbauen, Zusatzsteuer, Subventionsabbau, keine Subventionen, Transferleistungen senken, Transferzahlungen senken, Verschuldungsregel, crowding out, restriktiv, kontraktiv

## F.5 Endogenous and exogenous fiscal policy-related topics

Tables F.5.2 and F.5.3 report the classification of estimated topics into endogenous and exogenous topics. The word clouds visualizing the most important words (in German) in each topic are available upon request.

| Topic No. | Interpretation |
|---|---|
| 3 | Investment, Tax Reforms |
| 6 | Banks |
| 22 | Labor Market |
| 27 | Municipalities |
| 45 | Growth, Business Cycle |
| 63 | Economic Policy, Bundesbank |
| 67 | Labor Market Policy, Minimum Wage |
| 79 | Budget, Debt, Investment |
| 91 | Greece |

**Table F.5.2:** Endogenous fiscal policy-related topics in Bundestag speeches

| Topic No. | Interpretation |
|---|---|
| 7 | Coal, Energy Policy |
| 14 | Developing Countries |
| 15 | Development Policy |
| 16 | Petitions |
| 17 | Barbara Hendricks |
| 24 | Heating Systems |
| 28 | Care |
| 31 | Health |
| 34 | Environmental Policy |
| 37 | Digitalization |
| 38 | Social Welfare |
| 40 | Poverty |
| 44 | Transportation |
| 47 | Bundeswehr |
| 50 | Compulsory Service |
| 53 | Climate Change |
| 54 | Housing Construction |
| 58 | Bundeswehr Missions |
| 59 | East Germany |
| 64 | Disarmament, Nuclear Weapons |
| 70 | German Democratic Republic |
| 72 | Competition |
| 90 | Higher Education |
| 96 | Nuclear Energy |
| 97 | Research |

**Table F.5.3:** Exogenous fiscal policy-related topics in Bundestag speeches

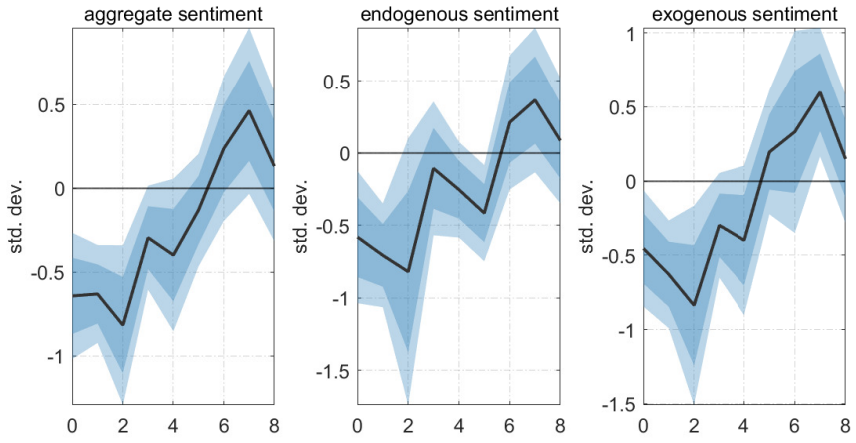



**Figure F.6.9:** Response of fiscal sentiment to tax shocks

*Notes:* The figure shows the responses of fiscal sentiment to the tax shocks from Hayo & Uhl (2014). The shaded areas reflect 68% and 90% confidence bands constructed from Newey-West standard errors.

## F.6 The response of sentiment to tax shocks

In order to investigate how our sentiment indices respond to the tax shocks of Hayo & Uhl (2014), we estimate the following local projection (Jordà, 2005) model

$$senti^{j}_{t+h} = \alpha_h + \beta_h shock_t + \gamma_h X_{t-1} + \varepsilon_{t+h} \tag{F.6.1}$$

with $j \in (aggregate, exogenous, endogenous)$, where $shock_t$ is the tax shock of Hayo & Uhl (2014). We expect the coefficient $\beta_h$ to be negative, such that a restrictive tax shocks causes a decline in fiscal sentiment. The vector $X_t$ contains the contemporaneous and four lagged realizations of GDP, consumption and government expenditure as well as four lags of the dependent variable. The availability of the tax shock series limits the sample to the period 1970Q1 – 2011Q4. Figure F.6.9 shows that sentiment deteriorates after a restrictive tax shock.

## F.7 The response of inflation and monetary policy

In the main part of the paper, we show that the real interest rate, measured as the difference between the long-term German bond yield and the inflation rate, declines after an increase in fiscal sentiment. We now take a look at the response of the short-term interest rate, i.e. on the monetary policy response by the Bundesbank (before 1999) and the European Central Bank (after 1999), as well as the response of the inflation rate. Figure F.7.10 reveals that the short-term nominal rate falls in the first six quarters after the shock, while inflation significantly increases two years after the shock.

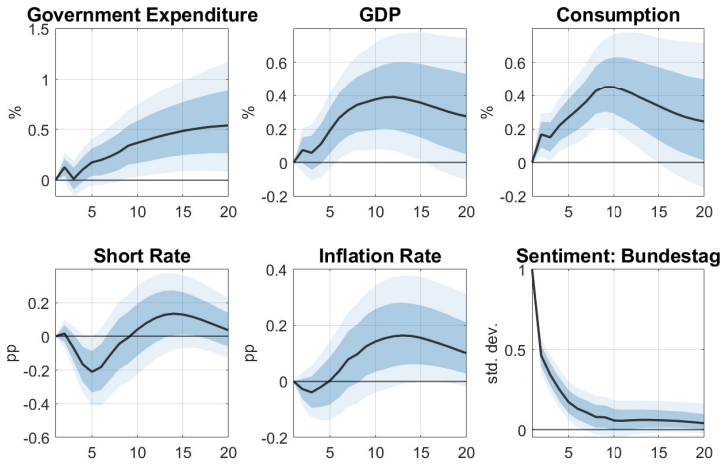



**Figure F.7.10:** Response to fiscal sentiment (Bundestag): inflation and monetary policy

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment as reflected in speeches of all members of the Bundestag. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

## F.8 Robustness of baseline VAR

We assess the robustness of the findings with respect to the lag order $p$ of the VAR model. Figure F.8.11 reports the impulse responses for the baseline model with $p = 8$ as well as three alternative models with $p \in 4, 6, 10$. While the responses of the three macroeconomic variables tend to be smaller for $p = 4$, they remain qualitatively unaffected when changing the lag order from $p = 8$ to $p = 6$ or $p = 10$. We conclude that our results are robust with respect to plausible alternative lag orders.

## F.9 The role of sentiment for the transmission of spending shocks

Our main results show that shocks to fiscal sentiment have strong effects on government spending and the macroeconomy. This could imply that including sentiment in the VAR model changes the transmission of shocks to government spending. We investigate this conjecture by comparing a VAR specification with and without sentiment. Specifically, we estimate a model with government spending ordered first, followed by GDP, consumption, inflation and (government) fiscal sentiment and derive impulse responses to an increase in spending based on the identifying assumption that spending does not respond contemporaneously to the remaining variables. We then estimate the same model but exclude fiscal sentiment.

Figure F.9.12 compares the resulting impulse responses. In the absence of fiscal sentiment, an increase in spending has a positive and significant effect on GDP, private consumption and inflation. Interestingly, the effect is much smaller compared to the implicit fiscal multipliers

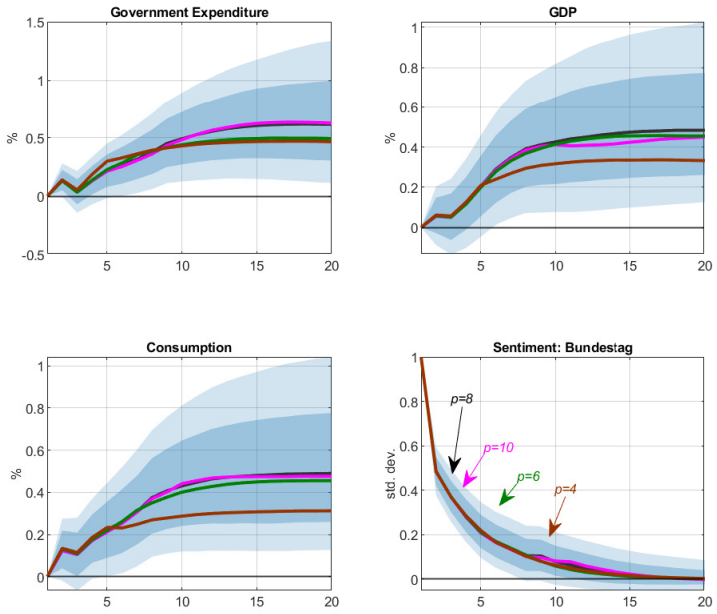



**Figure F.8.11:** Response to fiscal sentiment (entire Bundestag): lag order

*Notes:* The figure shows the responses of the endogenous variables to an increase in fiscal sentiment. All responses are derived from a recursively identified Bayesian VAR model with alternative lags orders $p$ and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

derived from the model with sentiment shown in the main part of the paper. Once we include sentiment in the model, the responses become markedly smaller.[27] This finding suggests that including sentiment does indeed change the transmission of conventionally identified government spending shocks.

[27] We do not show the probability bands around the model with sentiment in order to keep the figure readable. We also do not show the response of sentiment to the spending shock.

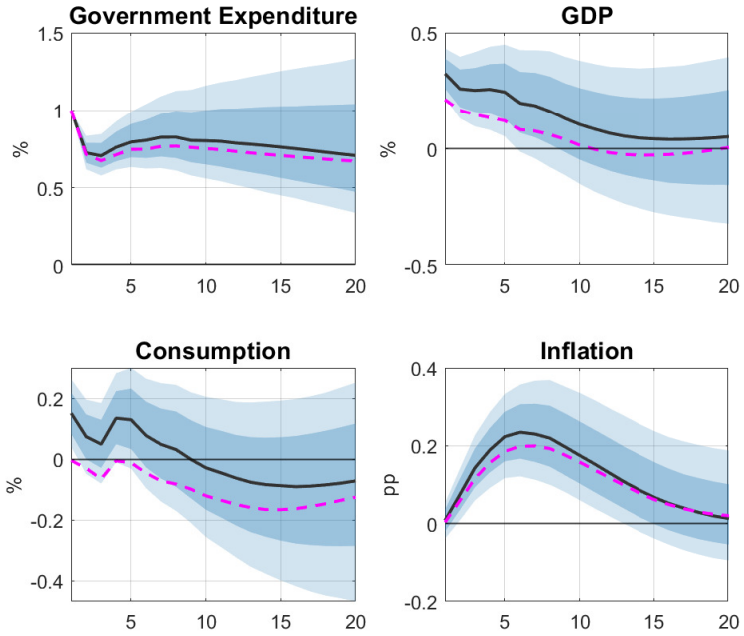



**Figure F.9.12:** Response to government spending

*Notes:* The figure shows the responses of the endogenous variables to an increase in government spending in the absence of fiscal sentiment from the VAR model. All responses are derived from a recursively identified Bayesian VAR model with 8 lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws. The magenta line is the response from a model that includes sentiment.

# Chapter 9

# Disagreement about Fiscal Policy

| **The following chapter is based on the paper:** | |
|---|---|
| **Title:** | Disagreement about Fiscal Policy |
| **Authors:** | Viktoriia Naboka-Krell (contribution: 30%),<br>Albina Latifi (contribution: 30%),<br>Peter Tillmann (contribution: 30%),<br>Peter Winker (contribution: 10%) |
| **Status:** | *Working Paper*; submitted to *Oxford Bulletin of Economics and Statistics* |

This paper will be presented at:

- Franco-German Fiscal Policy Seminar, Berlin, Germany, 2024

# Disagreement about Fiscal Policy*

Albina Latifi‡ Viktoriia Naboka-Krell‡
Peter Tillmann‡,‖ Peter Winker‡

**Abstract.** Politicians disagree about fiscal policy. This disagreement should have economic effects beyond the effects of government spending and taxation. We use the full set of speeches in the German Bundestag since 1960 and apply state-of-the art natural language processing techniques to construct two series of fiscal disagreement starting in 1970: disagreement about fiscal policy between the government and the opposition and disagreement within the coalition government. Both series fluctuate strongly and peak in 1982/83, but are also notably different from each other. In a VAR analysis, we show that within-coalition disagreement has adverse effects on real activity. Higher disagreement causes a drop in industrial production, manufacturing orders and stock prices. In contrast, government-opposition disagreement does not affect the business cycle.

*Key Words: fiscal policy, disagreement, uncertainty, business cycle, natural language processing*

*JEL classification: C89, E60, E62*

* Financial support from the German Research Foundation (DFG) (TI 594/4-1 and WI 2024/8-1) for the project MaFiText is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

‡ Department of Economics, Justus Liebig University Giessen

‖ Corresponding author: peter.tillmann@wirtschaft.uni-giessen.de

## 9.1 Introduction

Fiscal policy is made in parliaments, often after long and fierce debates. The extent of disagreement about fiscal policy among policymakers is likely to have economic effects besides government spending and taxation themselves. Disagreement about the path of fiscal policy should hamper real economic activity, and private investment in particular, for two reasons (Azzimonti, 2018). To the extent new projects are irreversible and subject to fixed costs, firms postpone investment if they perceive political disagreement. This factor will depress economic activity. In addition, disagreement prevents an efficient fiscal stabilization of exogenous shocks.

In Germany, the dispute about the budget intensified after the Federal Constitutional Court declared in November 2023 that parts of the 2021 federal budget violated the constitution. The ongoing fight in the coalition government is indeed seen as a burden for the economy. *The Economist* chooses the headline: "German business is fed up with a government in disarray" (December 14, 2023).[1] *Reuters* writes: "Budget crisis shakes industry's confidence in Germany" (November 25, 2023).[2] Hence, providing empirical evidence on the depressing effect of fiscal disagreement on the economy is important.[3]

In this paper, we propose a new way to measure fiscal disagreement in the Bundestag, the German parliament, and estimate its macroeconomic effects. We construct series of disagreement based on the textual data of all speeches delivered in the Bundestag, the German parliament. Rather than measuring fiscal disagreement from newspaper reporting (Azzimonti, 2018) or private forecasts (Ricco et al., 2016), we measure disagreement directly at its source: the debate in parliament. Using high-dimensional representations, so-called embeddings, of the speeches delivered in the German Bundestag, we quantify relevant information from these texts and construct time series of fiscal sentiment. In our companion paper (Latifi et al., 2024), we explain in detail how we measure fiscal sentiment. Here, we go beyond the measurement of fiscal sentiment and make two contributions.

First, from the data about fiscal sentiment we compute two series of disagreement. The first is the disagreement between the government and the opposition. This measure is the absolute difference in average fiscal sentiment expressed in parliamentary speeches of members of the governing parties and members of the opposition party. Owing to the nature of the German political system, disagreement could arise not just between the government and the opposition, but also among the parties forming the coalition government. Therefore, the second measure focuses on the disagreement within the governing coalition. This measure is

[1] See `https://www.economist.com/business/2023/12/14/german-business-is-fed-up-with-a-government-in-disarray`.

[2] See `https://www.reuters.com/markets/europe/budget-crisis-shakes-industrys-confidence-germany-2023-11-24/`.

[3] On August 25, 2024, the *Frankfurter Allgemeine Zeitung* provides an overview over the history of budget fights in German coalition governments: `https://www.faz.net/aktuell/wirtschaft/wo-der-kanzler-irrt-frueher-war-das-regieren-nicht-einfacher-19942934.html`.

given by the absolute difference in the fiscal sentiment between the speakers from the two parties forming the government. Treating the CDU/CSU as one party, all governments in our sample were formed by at most two parties. By distinguishing between the two types of disagreement, we take account of the specifics of the parliamentary system in Germany.[4]

Throughout the paper, we refer to these two series as measures of disagreement, because we believe this is indeed what we measure in the data. Alternative labels used in the literature such as partisanship, partisan conflict, polarization or government division are related, but do not fit the German context. A key advantage of our data is that we can go back long in history starting in 1970, thus covering the 1982/83 break-up of the coalition due to a fight over economic and fiscal policies as well as the reunification episode. We find that both series of disagreement fluctuate over the sample period 1970 - 2021 and peak in 1982/83, when the Liberal Democrats left the coalition with the Social Democrats. The series of within-coalition disagreement exhibits another peak in 1999, when Oskar Lafontaine, the Finance minister of the new SPD-led coalition resigned over a fight about fiscal policy with Chancellor Gerhard Schröder. We also find a cycle in the four-year election periods: after an initial honeymoon at the beginning of the election period, disagreement strongly increases towards the end of the election period when the next election is looming.

Second, we estimate the macroeconomic effects of fluctuations of disagreement. We put each series of disagreement into a standard vector autoregression (VAR) model together with business cycle variables such as GDP, private investment, government expenditures, manufacturing orders, stock prices and others. In order to identify a shock to fiscal disagreement, we assume that disagreement affects the business cycle with a lag of one quarter, i.e. invoke a recursive identification scheme. Our results suggest that an increase in disagreement within the coalition government has a contractionary effect on the economy. It causes a drop in industrial production, a fall in manufacturing orders and a decline in the sentiment reflected in surveys. Furthermore, stock prices fall significantly after an increase in disagreement. The response of investment is negative, but is not statistically different from zero. Interestingly, the responses are different after an unexpectedly higher level of disagreement between the government and the opposition. In this case, an unexpected increase in disagreement causes an increase in GDP, investment and sentiment in the manufacturing survey, though these responses are statistically significant at the 68% level only. We believe disagreement within the governing coalition should be more relevant for firm investment and, ultimately, for the business cycle than disagreement between the government and the opposition. The latter is a feature of any functioning democracy. In contrast, the former can cause a delay of fiscal decisions, an inefficient allocation of resources and an ineffective fiscal stabilization.

The paper closest to ours is Azzimonti (2018). She assembles a short dictionary of keywords associated with partisanship in U.S. politics and filters a large corpus of newspapers for articles in which these keywords appear. Based on the selected articles, she constructs the Partisan Conflict Index, which tracks political disagreement at the federal level of U.S.

[4] We disregard disagreement among the opposition parties.

politics. In a second step, she provides evidence, both from a VAR model estimated on aggregate data as well as firm-level data, that an increase in partisanship has contractionary effects on real activity and investment.[5] Hong et al. (2024) also resort to newspaper reporting in order to construct country-specific indices of fiscal policy uncertainty for a very large set of countries. When ordered first in an otherwise standard VAR model, shifts in fiscal policy uncertainty can have large negative effects on economic activity. Silgado-Gómez (2024) extracts an uncertainty indicator about public finances in Spain from very large time series data set. He finds that higher fiscal uncertainty dampens economic activity.[6]

Fernández-Villaverde et al. (2015) offer another perspective on fiscal uncertainty.[7] They estimate statistical processes for the time series of taxes and government spending in the U.S. and allow for time-varying volatility. In a second step, the authors include the time series of time-varying fiscal volatility, which they interpret as a proxy for fiscal uncertainty, into a recursively identified VAR model. Volatility is ordered first such that it can contemporaneously drive the remaining variables. An unexpected increase in volatility has strong adverse effects on the economy.

In contrast, Ricco et al. (2016) use the disagreement of forecasts about fiscal variables to construct a measure of disagreement about future U.S. fiscal policy. Furthermore, they show that fiscal policy impulses are more effective on the real economy if fiscal disagreement is low. Beckmann & Czudaj (2021) employ forecasts for the German budget balance in order to construct a measure of forecast disagreement. The authors interpret this as an index of fiscal policy uncertainty. A VAR model with fiscal uncertainty ordered first generates a strong effect of uncertainty on the growth rate of industrial production.[8] The study of Hantzsche (2022) is also based on the uncertainty reflected in forecasts of fiscal variables, though the author uses forecasts from international institutions such as the IMF, the OECD or the European Commission rather than private sector forecasts. He finds that uncertainty about future fiscal policy drives sovereign credit risk.[9]

In contrast to both Azzimonti (2018) and Ricco et al. (2016), we measure the extent of disagreement directly from the speeches of policymakers in parliament. We do not rely

[5] Hacioğlu-Hoke (2024) uses the Partisan Conflict Index as an instrument in a VAR model to show that a fall in disagreement is expansionary.

[6] Our paper also rests on the voluminous political economy literature on the determinants and consequences of political business cycles and divided governments (Alesina et al., 1999; Alesina & Rosenthal, 1995). Specifically, Roubini & Sachs (1989) show that a divided government, i.e. two parliamentary chambers controlled by different political parties, leads to inefficient stabilization. Poterba (1994) finds that the deficit is reduced less under divided than unified government.

[7] Born & Pfeifer (2014) derive a measure of time-varying fiscal policy risk, i.e. the volatility of government spending and taxation, from an estimated New Keynesian model.

[8] Anzuini & Rossi (2021) provide similar evidence for the U.S. economy.

[9] Other influential papers on political uncertainty and stock prices are Pástor & Veronesi (2012) and Papamichalis et al. (2024). For the nexus between political uncertainty and firm investment, see Julio & Yook (2012), Jens (2017) and Gulen & Ion (2016).

on newspaper reporting, nor on forecast data. Hence, our notion of disagreement is more directly related to the protagonists of fiscal policy, not their reception in the media, and is fundamentally different from the notion of forecast uncertainty employed by Ricco et al. (2016) and Beckmann & Czudaj (2021). It is also important to emphasize that we directly measure fiscal disagreement, which we interpret as a driver of fiscal uncertainty, though we remain agnostic and do not aim at empirically distinguishing between the consequences of disagreement and the effects of uncertainty.[10] No research has been done so far to study the effects of disagreements on the German government on economic variables such as income, household consumption, firm investment and the business climate.

The remainder of this paper is organized as follows. Section 9.2 briefly introduces the corpus containing the speeches delivered in the German Bundestag in the period from 1960 to 2021. It also describes the process of constructing a fiscal sentiment index for Germany as well as the disagreement indices. Section 9.3 presents the constructed disagreement indices, government-opposition and within coalition. In section 9.4 we estimate the macroeconomic effects of the disagreement indices using VAR models. Section 9.5 expands this analysis and presents alternative model specifications. Section 9.6 summarizes the main findings. An online appendix contains additional material.

## 9.2 Constructing indicators of disagreement from textual data

In this section, we briefly introduce the text corpus used in this analysis as well as the process of constructing fiscal policy (disagreement) indices. The text data for this analysis is based on the corpus presented in Latifi (2024), in which the stenographic protocols of the German Bundestag were made accessible for natural language processing applications using an innovative algorithm based on a custom Named Entity Recognition model for reliable identification of speech segments. The corpus includes all speeches delivered in the German Bundestag from 1949 to 2021, along with relevant metadata such as the date of the speech, the speaker's name, party affiliation and role, encompassing a total of 877,140 speeches. We restrict the corpus to speeches from 1960 onwards, as most macroeconomic time series data for Germany are only available starting from 1970.[11] We focus on all speeches delivered by members of the Bundestag that can be clearly attributed to a political party, thus excluding speeches by non-affiliated ("parteilos") or independent ("fraktionslos") members from this analysis.

The pre-processing follows Latifi et al. (2024). Hence, we remove very short and very long speeches. The dataset now includes 165,148 speeches, of which 79,611 belong to speakers

[10] The literature on forecasting carefully distinguishes between the disagreement among forecasts and the uncertainty of forecasts, see Lahiri & Sheng (2010) and Zohar (2024).

[11] The 10-year difference between the start of the corpus and the macroeconomic time series is due to the dynamic Doc2Vec approach, which uses a rolling window of 10 years.

from governing parties and 85,537 to speakers from opposition parties. We lemmatize all speeches. We then convert all German umlauts and the eszett to avoid encoding errors. Further, we remove line breaks, digits, empty sentences, and special characters, and convert all letters to lowercase. We also remove single-element tokens and tokens with more than 30 elements. Moreover, we use the domain-specific stopwords list presented by Latifi et al. (2024), which includes 1,405 terms particularly relevant to the terminology used in Bundestag speeches. These stopwords are also removed from the corpus.

For the next step, we use the dictionary developed by Latifi et al. (2024). It contains fiscal policy-relevant terms, divided into two groups, namely *expansionary* (in total 122 compound terms) and *contractionary* (in total 96 compound terms).

Further, to construct a fiscal sentiment index for Germany, we propose to use a Doc2Vec-based approach (Le & Mikolov, 2014). In a nutshell, Doc2Vec describes a neural network-based technique to represent whole text paragraphs and documents in a high-dimensional vectors space. It means that textual characteristics of a text are translated into vectors with specific features. One of the most important features of such text vectors, also called embeddings, is that the distances between them directly correspond to the semantic characteristics of the underlying texts. In other words, it means that synonyms or semantically related sentences can be identified by looking at the distanced between these items in the shared vector space.

The proposed Doc2Vec approach can be summarized in three steps. First, we divide the whole observation period into training and forecast periods. The patterns are learned based on all the speeches in each training period and then embeddings for the speeches in the subsequent forecast periods are inferred. The purpose is to avoid future information to be available to build embeddings for the current forecast period (for more details and an illustration of the approach see Latifi et al. (2024)). Second, we construct expansionary and contractionary vectors for each period based on the fiscal policy dictionary. Third, based on the distances between a single speech and the expansionary and contractionary vectors, the final score for each document is calculated as the difference between the similarity to the expansionary vector and to the contractionary vector. These scores are then aggregated at the quarterly basis which represents the final fiscal policy sentiment index for Germany.

In the current project, we expand our analysis and construct indices of disagreement about fiscal policy in Germany. Since the information on fiscal sentiment as well as the information on party affiliation of the single speakers are available on the document level, we are able to construct two different disagreement indices. First, we consider disagreement between government and opposition. On the quarterly basis, we calculate a time series of fiscal sentiment scores for each party in the Bundestag as the mean of fiscal sentiment of all speeches from all members of the respective party. To obtain the fiscal disagreement score in period $t$, we calculate the absolute difference between the average sentiment score of the parties forming the government at time $t$, $Sentiment_t^{Gov}$, and the average sentiment score of the parties in the opposition at time $t$, $Sentiment_t^{Opp}$,

$$Dis_t = |Sentiment_t^{Gov} - Sentiment_t^{Opp}|. \tag{9.2.1}$$

This implies that the indicator does not address the direction of differences, i.e. whether government or opposition use a more expansionary vocabulary, but solely the distance between both.

Second, we also construct an index measuring the disagreement about fiscal policy within the government coalition. Throughout our sample period, Germany was governed by coalition governments of two parties (when treating CDU/CSU as one party). To this end, on quarterly basis, we average over fiscal sentiment scores of all speeches from members of the two parties forming the government. The final within coalition-disagreement is calculated as the absolute difference between the two parties supporting the government at time $t$, $Sentiment_t^{GovI}$ and $Sentiment_t^{GovII}$,

$$Dis_t = |Sentiment_t^{GovI} - Sentiment_t^{GovII}|. \tag{9.2.2}$$

The resulting disagreement indices for Germany are presented in the subsequent section. In Appendix G.2, we present a third measure of disagreement: the absolute gap between the most expansionary and the most restrictive fiscal policy stance among the parties in the Bundestag in each quarter. An increase of type of disagreement does not have economic effects. We interpret this exercise similar to a placebo test: it is not the difference between parties in the Bundestag as such that has economic consequences, but the disagreement between government and opposition and, even more so, within the government.

## 9.3 Some stylized facts on fiscal policy disagreement

Figure 9.1 shows the two standardized series of disagreement.[12] Both series exhibit some c-movement in specific episodes, although with a correlation coefficient of 0.22, their overall co-movement is rather low.

Both series peak around 1982/83, when the Liberal Democrats left the coalition with the Social Democrats, ousted chancellor Schmidt and entered a new coalition with the Christian Democrats led by chancellor Kohl. The discontent of the Liberals with the fiscal and economic policies of the coalition was the main motivation of this step, which until today is the only case of one party leaving a coalition government in order to form a new government. In the jargon of political observers, this “Bonner Wende" is still a metaphor for extreme forms of fiscal disagreement.

In the 1980s and early 1990s, disagreement between the government and the opposition is more pronounced than disagreement within the coalition, which is also in line with the established historical narrative. Within-coalition disagreement spikes in 1998/1999, when the dissent about fiscal priorities led to the resignation of finance minister Lafontaine of the newly formed coalition between the Social Democrats and the green party. Overall, the level of disagreement seems to decline over time. After the 1980s and 1990s, in which disagreement was relatively high, the period since the early 2000 witnesses a lower level of

[12] The non-standardized series, which allow for a comparison of the levels of disagreement, are shown in the online appendix.

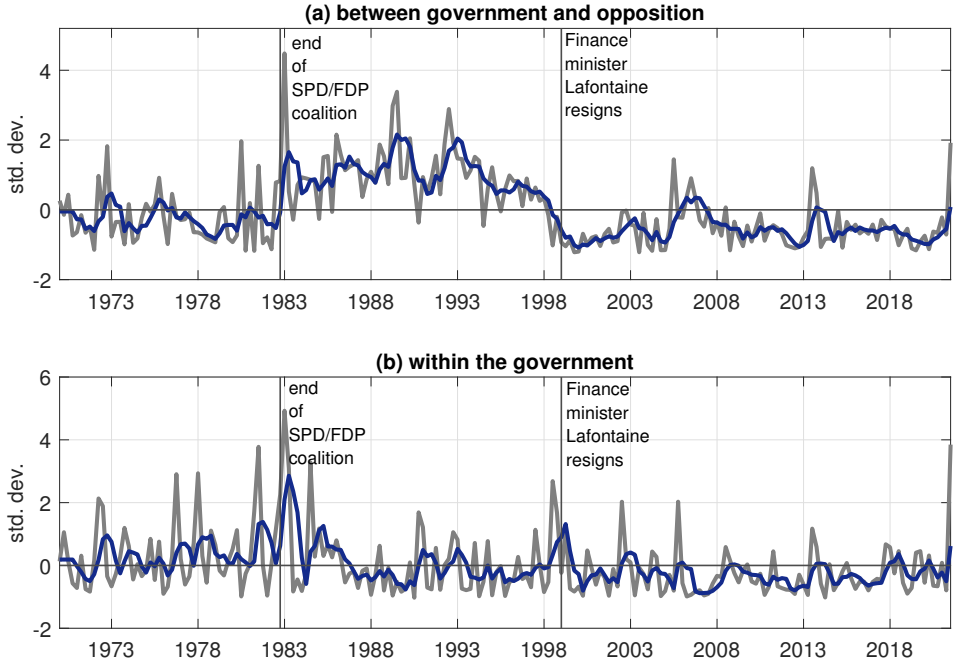



**Figure 9.1:** Disagreement in the Bundestag

*Notes*: The gray lines show the standardized series of fiscal policy disagreement. The blue lines are four-quarter (backward) moving averages.

disagreement. In future work, it could be interesting to investigate whether the adoption of the debt brake in 2009, which constrains the federal borrowing and is codified in the German constitution, had an effect on the debate about fiscal policy in the Bundestag and the dynamics of disagreement. The increase in disagreement in Q3 2021, the last observation in our sample period, results from the fact that we have only two parliamentary sessions in Q3 2021 in our data set. The Bundestag debated about the fiscal consequences of the flooding in the Ahr Valley. In Appendix G.4, we show that the results remain unchanged when excluding this outlier (see Appendix G.4).

It is likely that disagreement does not only fluctuate over the full sample period, but also exhibits a systematic pattern over the election periods. Each election period, the period from the constitution of the newly elected Bundestag to its last session before the next election, lasts four years. In 1993 and 2003, the election period was only three years due to snap elections.[13] In Figure 9.2, we express the level of disagreement in quarter $\tau$ of the election period minus the level of disagreement in the quarter $\tau = 0$ in which the election period starts. Note that due to the quarterly frequency of the data, we cannot distinguish sharply between the old and the new election period. Often, the election campaign, the actual election and the beginning of the election period, i.e. the constitution of the Bundestag, falls

[13] The data on the election periods is taken from `https://de.wikipedia.org/wiki/Liste_wichtiger_Wahltermine_und_Wahlperioden_in_Deutschland#Bundesrepublik_Deutschland`.

in the same quarter. This has to be taken into account when interpreting the figure. In some elections periods (EP), most notably in EP 7, EP 15, EP 17 and EP 19, both series of fiscal disagreement evolve similarly. In others, e.g. EP 10, EP 12 or EP 13, the behavior of both series is markedly different. A common feature is that disagreement tends to increase towards the end of the election period. Hence, the debate about fiscal policy intensifies as the next general election is approaching.

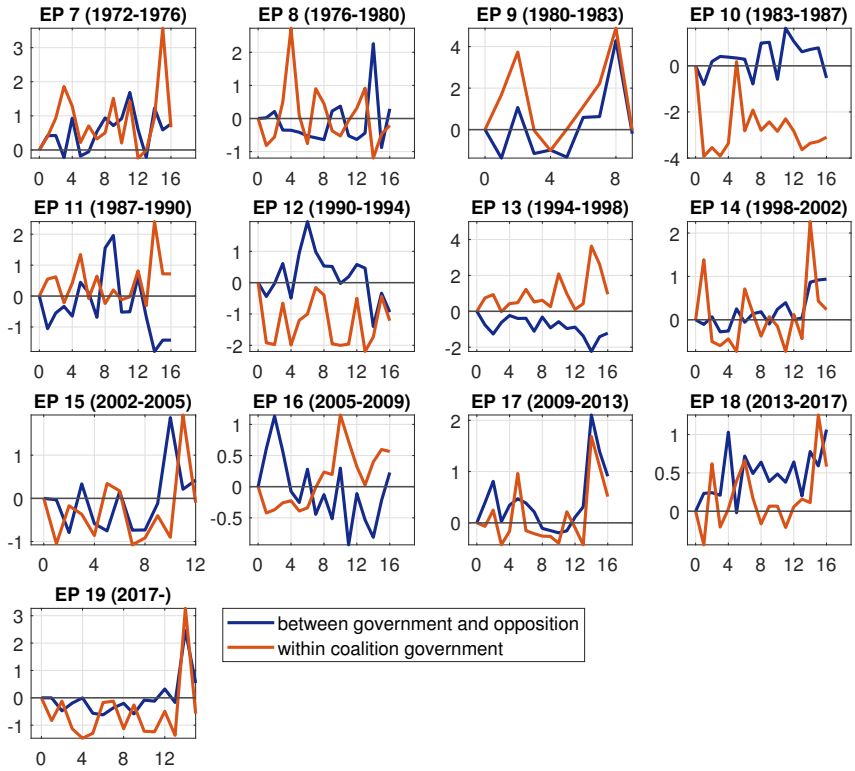



**Figure 9.2:** Disagreement in each election period

*Notes*: This figure shows the standardized series of fiscal policy disagreement in the election periods (EP) in our sample period. We express the level of disagreement in quarter $\tau$ of the election period minus the level of disagreement in the quarter $\tau = 0$ when the election period starts.

To spot the general pattern of the election periods, Figure 9.3 shows the two series of disagreement averaged over the course of each election period in our sample. The figure reveals an interesting cycle of disagreement over the span of an election period: in the first six quarters, both measures of disagreement are lower than at the beginning of the election period. Afterwards government-opposition disagreement strongly increases. Towards the end of the election period when the upcoming election is looming, within-coalition disagreement jumps upwards. In order to account for this trajectory of disagreement over the election period, we will also estimate a VAR model that allows for election period-specific cubic time trends.

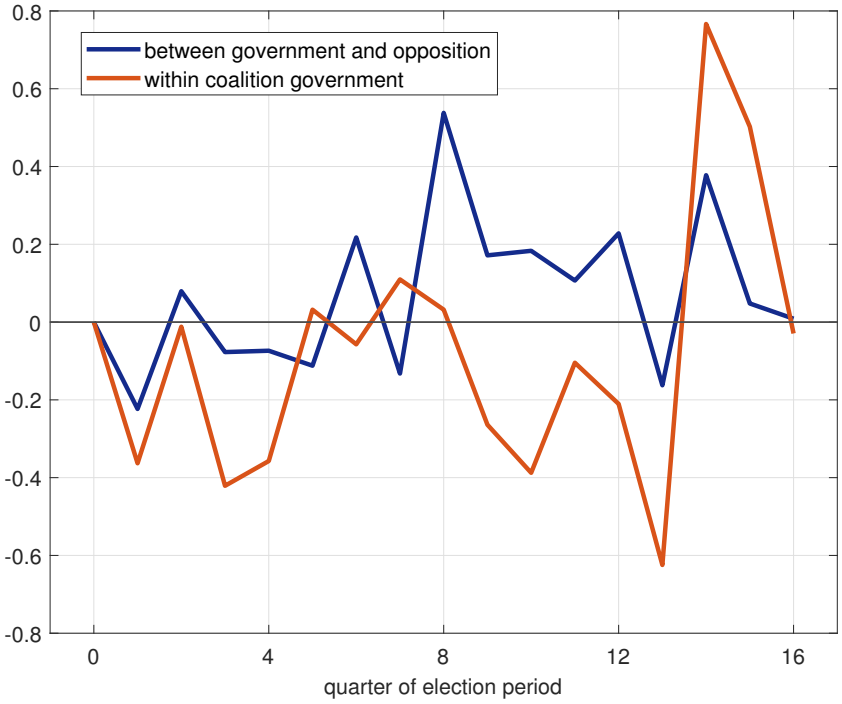



**Figure 9.3:** Average disagreement across all election periods

*Notes*: This figure shows the average of the standardized series of fiscal policy disagreement across the election periods in our sample period. We express the level of disagreement in quarter $\tau$ of the election period minus the level of disagreement in the quarter $\tau = 0$ when the election period starts.

We also analyze the evolution of disagreement in the six recession periods since 1970. In Figure 9.4, we plot disagreement, normalized in $\tau = 0$, i.e. the last full quarter before the recession begins.[14] In all cases, government-opposition disagreement increases after the recession hit. This increase is strongest during the two oil crises and became weaker thereafter. There is no clear pattern of within-coalition disagreement across recession episodes.

In EP 16 (2005-2009), 18 (2009-2013) and 19 (2017-2021), the ruling coalition was a grand coalition of CDU and SPD. In the online appendix, we estimate a simple regression of either indicator on a dummy variable that is one for a Grand Coalition and zero otherwise. We also control for the business cycle and an election period cubic time trend. The results are clear-cut: both types of disagreement are significantly lower under a Grand Coalition. In all election periods other than EP 16 and EP 19, two of the three grand coalition governments, the Finance minster and the Chancellor belonged to the same political party. It seems that the level of disagreement is lower in the grand coalitions even though the Chancellor and the Finance minister are not from the same party.

[14] The recession dates are provided by the German Council of Economic Experts at `https://www.sachverstaendigenrat-wirtschaft.de/themen/konjunktur-und-wachstum/konjunkturzyklus-datierung.html`.

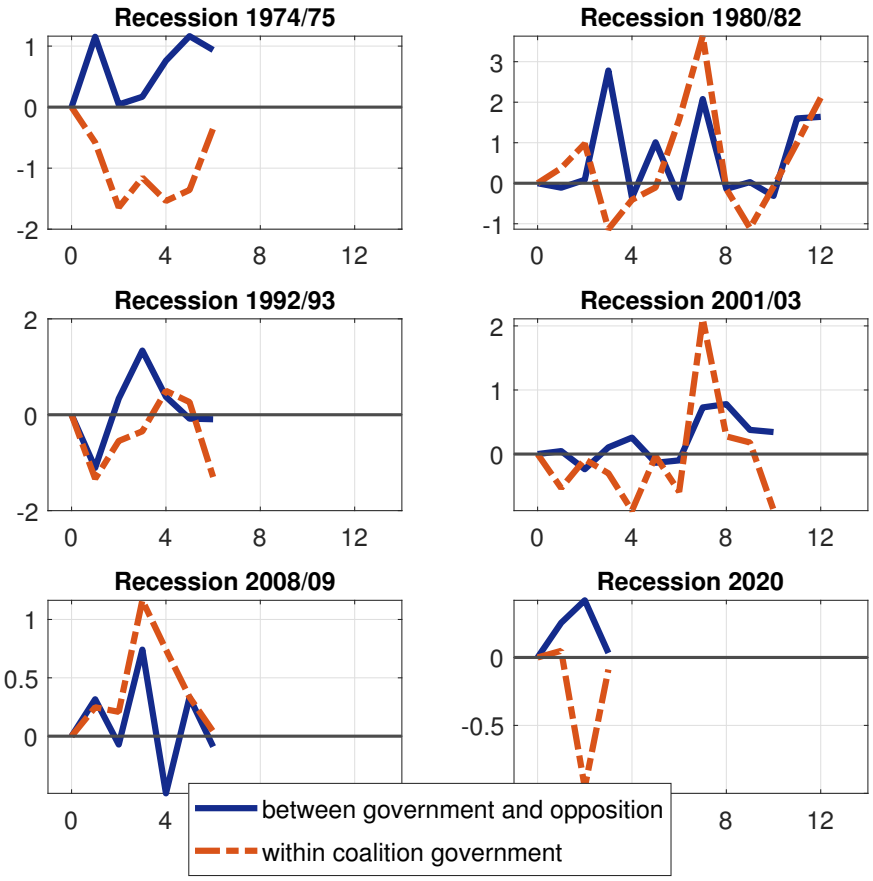



**Figure 9.4:** Disagreement during recessions

*Notes*: This figure shows the average of the standardized series of fiscal policy disagreement across the recessions in our sample period. We express the level of disagreement in quarter $\tau$ of the recession period minus the level of disagreement in the last quarter $\tau = 0$ before the recession.

## 9.4 The effects of fiscal disagreement

In this section, we estimate the macroeconomic effects of fiscal policy disagreement on the German economy. We concentrate on the transmission of fiscal disagreement through firms' investment decisions.

### 9.4.1 A VAR model

We estimate a straightforward VAR model on quarterly data for Germany. The vector of endogenous variables is

$$y_t' = (GDP_t \quad Exp_t \quad IP_t \quad Inv_t \quad Ord_t \quad Surv_t \quad Dis_t \quad Stocks_t), \tag{9.4.1}$$

where $Dis_t$ is our series of disagreement, either between the government and the opposition or within the ruling coalition. Here we use the raw series, not the moving average. We include real GDP ($GDP_t$), real government spending ($Exp_t$), industrial production ($IP_t$),

real Gross Fixed Capital Formation ($Inv_t$), the stock of manufacturing orders ($orders_t$), the survey sentiment of manufacturing firms ($survey_t$) and the DAX stock price index ($Stocks_t$). All series other than stock prices are seasonally adjusted. The series other than the survey sentiment are included in natural logs multiplied by 100. The survey sentiment is measured in percentage balances, i.e. the share of firms reporting an improvement of the situation minus the share of firms reporting a deterioration.

The VAR model is identified by imposing a recursive ordering of the contemporaneous interactions between the variables. We order the variables as in equation (9.4.1). This implies that fiscal disagreement has a contemporaneous effect only on stock prices, while the remaining variables respond to fiscal disagreement with a lag of one quarter. Disagreement responds contemporaneously to GDP, Investment and the other macroeconomic variables ordered before disagreement.

Blanchard & Perotti (2002) propose a recursive identification scheme of government spending shocks based on the notion that spending is predetermined. As a consequence, they order government spending first. This approach inspired a huge literature (Fatás & Mihov, 2001; Galí et al., 2007; Born & Müller, 2012; Ilori et al., 2012; V. A. Ramey & Zubairy, 2018; Tenhofen et al., 2010). In Latifi et al. (2024), we propose an alternative ordering based on the idea that the fiscal sentiment expressed in Bundestag speeches should be responsive to economic circumstances within the same quarter. Hence, we ordered fiscal sentiment last. As our disagreement measure is a derivative of the data used in Latifi et al. (2024), we adopt the same ordering here. Fiscal disagreement contemporaneously responds to the business cycle, government spending, manufacturing orders and the manufacturing survey. Within a given quarter, only stock prices are allowed to respond to fiscal disagreement.[15]

We estimate the VAR model using Bayesian methods based on a Normal-Wishart prior for the time period 1970Q1 to 2021Q3 including four lags of the endogenous variables.[16]

### 9.4.2 Results

We now show the dynamic responses of the six endogenous variables to a shock to fiscal policy disagreement. In each figure, we show the estimated impulse response as well as the 68% and 90% probability bands. Figure 9.5 shows the consequences of an increase in disagreement between the government and the opposition. We find that real GDP, real investment and the survey sentiment tend to increase. However, it is important to stress that none of these responses is statistically different from zero when judged by the 90% probability bands. Likewise, the responses of industrial production, manufacturing orders and stock prices are also indistinguishable from zero. Finally, government spending does not respond to fiscal disagreement between the government and the opposition. Thus, it seems that disagreement between the government and the opposition has no effect on the business

[15] Below, we will show that our results are in no way dependent on this specific ordering.

[16] In order to estimate the model, we rely on the BEAR toolbox for MATLAB, see `https://www.ecb.europa.eu/pub/research/working-papers/html/bear-toolbox.en.html`.

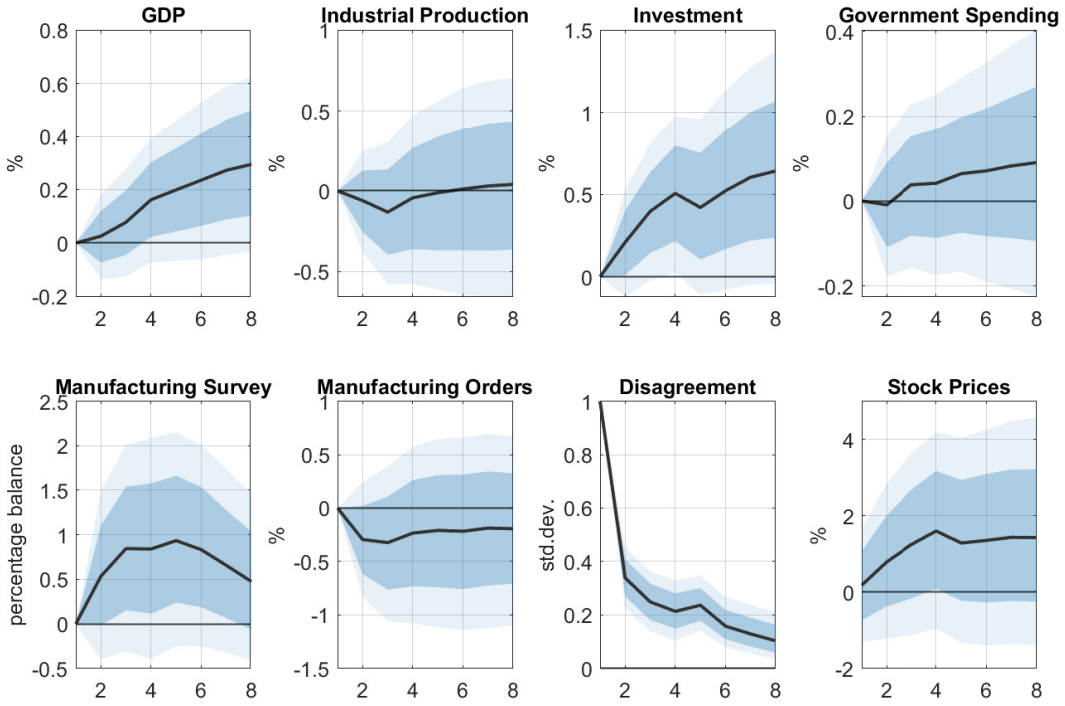



**Figure 9.5:** Response to disagreement between the government and the opposition

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

cycle.

The responses to an increase in disagreement within the coalition, see Figure 9.6, provides a different picture. Higher disagreement causes a drop in industrial production, the manufacturing sentiment and the stock of manufacturing order. In each case, this drop is statistically different from zero. Hence, higher disagreement among the parties forming the ruling coalition causes a decline in real economic activity. This is the main result of this paper. The effect is also quantitatively relevant: an increase in disagreement of one standard deviation causes a drop in industrial production after six quarters of about 0.5%.

Importantly, government spending does not respond to disagreement. Thus, we can rule out that real activity variables fall just because the government cuts-back spending.

### 9.4.3 The role of the majority in the Bundesrat majority

To become effective, most federal laws in Germany need to pass both the Bundestag and the Bundesrat, the second chamber of the parliamentary system. In the Bundesrat, each federal state has a specific number of votes depending on the size of its population and exercised by its state government. It is frequently the case that the parties forming the federal government have the majority in the Bundestag, but face a majority of the opposition parties in the Bundesrat. In order to avoid a gridlock in such a situation, the government needs to bargain with the federal states – and effectively with the opposition

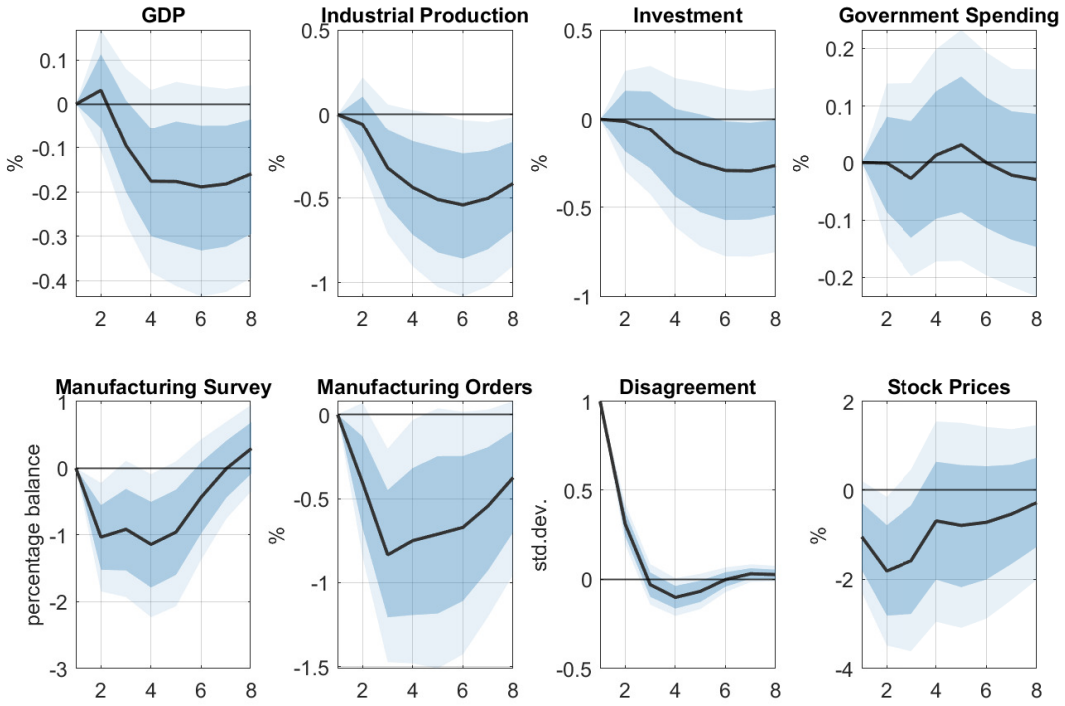



**Figure 9.6:** Response to disagreement within the government

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

parties. The government eventually gets the support from the federal states only if transfers to the states increase, either direct transfers or implicit transfers through a re-calibration of the fiscal burden-sharing in other policy areas. Hence, a high level of government-opposition disagreement could reflect a situation in which the government faces an unfavorable allocation of votes in the Bundesrat. The additional resources spent in order to acquire the votes in the Bundesrat could be expansionary.

Therefore, we now control for the allocation of votes in the Bundesrat.[17] We remain agnostic about the specific mechanism through which the votes in the Bundesrat affect disagreement. Figure 9.7 shows the series of disagreement between the government and the opposition against the share of the votes in the Bundesrat controlled by the government. We find that both series tend to co-move in the long-run with a correlation coefficient of 0.22. Periods with high disagreement, such as the 1980s or the mid 2000s, also exhibit a larger share of votes controlled by the government.

We now re-estimate the VAR model from the previous subsection, but include the share of votes in the Bundesrat from Figure 9.7 as an exogenous variable in the model. For the purpose of this exercise we believe we can treat the allocation of votes as an exogenous variable. Of course, the votes are the outcome of elections and the formation of coalitions in the federal states. Figures 9.8 and 9.9 report the estimated impulse response functions. Our

[17] The data is taken from `https://www.wahlen-in-deutschland.de/bBundesrat.htm`.

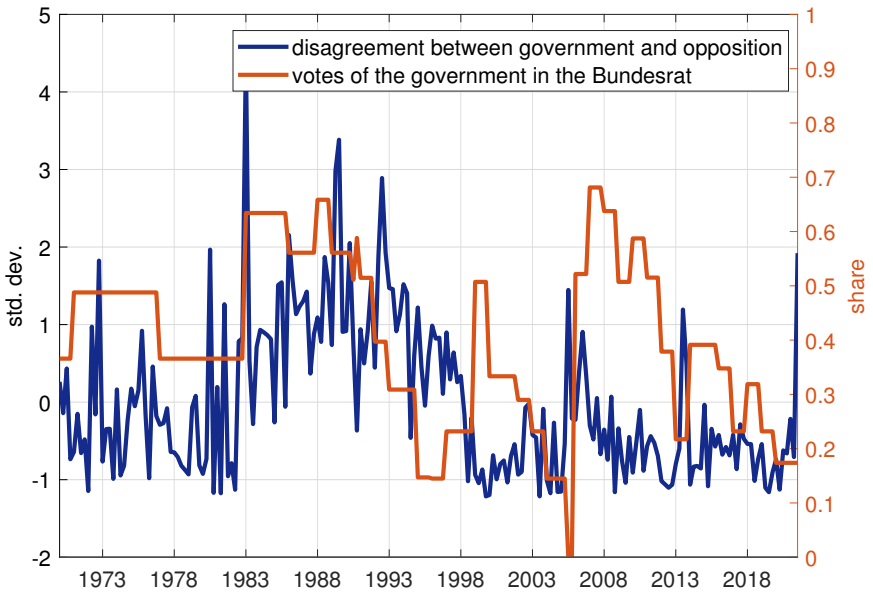



**Figure 9.7:** Disagreement between the government and the opposition and the votes in the Bundesrat

*Notes:* The figure shows the series of government-opposition disagreement (left axis) and the share of the votes in the Bundesrat controlled by the governing coalition parties (right axis).

main findings remain unaffected. Shocks to disagreement between the government and the opposition do not seem to drive economic activity. In contrast, an increase in disagreement among the coalition parties reduces industrial production, manufacturing sentiment and manufacturing orders. The results remain robust with respect to including the vote share in the Bundesrat.

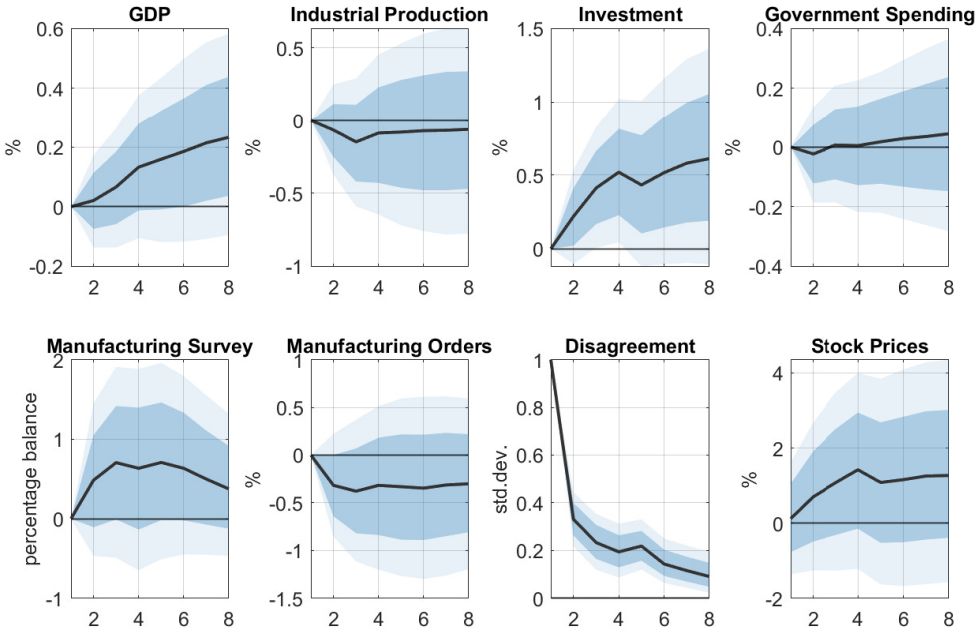



**Figure 9.8:** Response to disagreement between the government and the opposition: controlling for Bundesrat votes

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

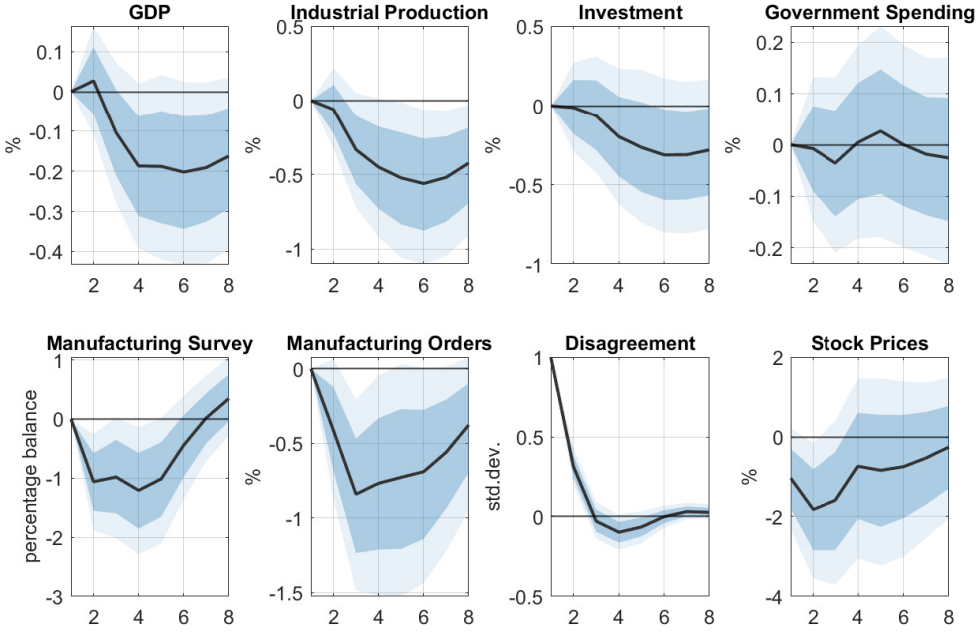



**Figure 9.9:** Response to disagreement within the government (controlling for Bundesrat votes)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

## 9.5 Alternative Specifications of VAR model

In this section, we study the results of alternative VAR models in which we (i) change the ordering of the variables, (ii) include election-period time trends, (iii) include the federal budget balance as a share of GDP and (iv) include both indicators of disagreement jointly as well as the German 10-year bond yield.

In our baseline VAR model, we ordered disagreement after the macroeconomic variables. Such a scheme assumes that GDP and the other time series respond to disagreement with a time lag of one quarter. We now estimate a VAR model with an alternative ordering. Specifically, we follow the literature on the identification of government spending shocks (Blanchard & Perotti, 2002) and order fiscal sentiment first. All other details of the VAR model are left unchanged.

Figures 9.10 and 9.11 show the estimated impulse response functions. We do not find any material difference compared to our baseline model. An increase in government-opposition disagreement tends to have an expansionary effect, while higher disagreement within the coalition is restrictive.

One of the stylized facts discussed before is that disagreement evolves in a U-shaped way over the four-year election period. To take account of this disagreement cycle, we include a cubic election period-specific time trend as a deterministic variable. Specifically, we include $\tau$, $\tau^2$ and $\tau^3$ for $\tau = 1, ..., 16$, where $\tau$ is the time index in each election period. This time trend should capture the initial honeymoon period, the decline in disagreement, and the subsequent increase when the election period comes to its end. Figures 9.12 and 9.13 show the estimated impulse response functions for this alternative model specification for government-opposition and within-coalition disagreement. An increase in disagreement between the government parties and the opposition parties has no discernible impact on the endogenous variables. As before, GDP and investment increase, but these responses lack statistical significance. Higher disagreement among the coalition parties causes a fall in industrial production, the manufacturing sentiment and in the volume of manufacturing orders.

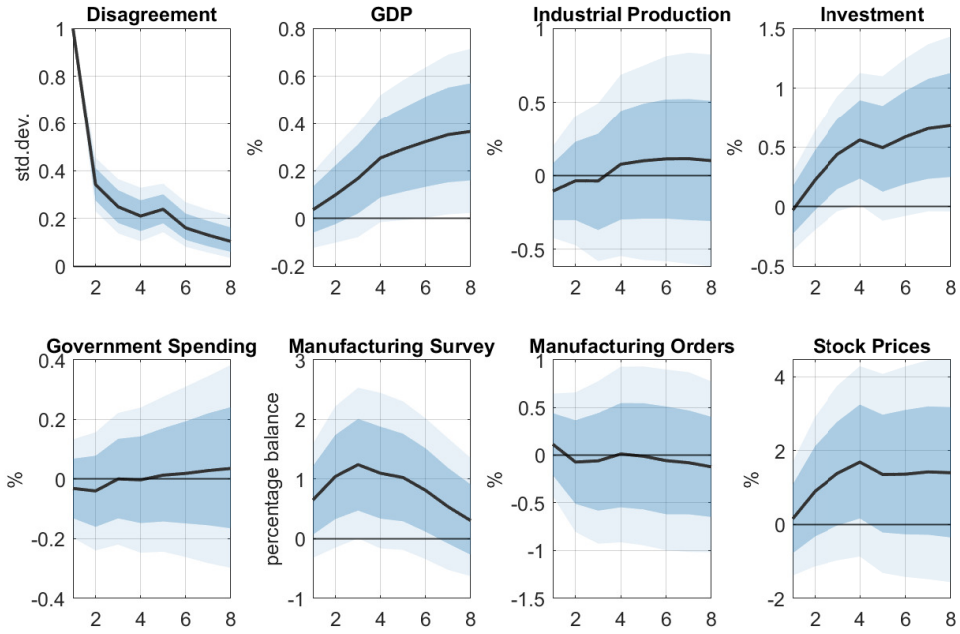



**Figure 9.10:** Response to disagreement between the government and the opposition (alternative ordering)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

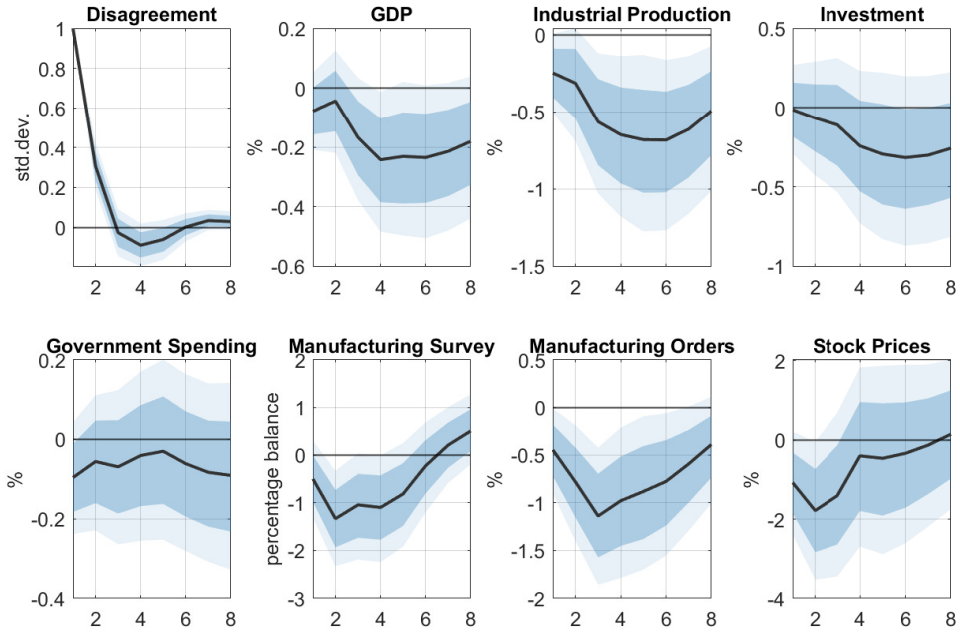



**Figure 9.11:** Response to disagreement within the government (alternative ordering)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

As our third robustness check, we replace the time series of government spending with the government's budget balance in percent of GDP. This series is interpolated from annual to quarterly frequency. Figure 9.14 shows the response to an increase in disagreement between the government and the opposition. Again, this shock tends to increase GDP and investment, though both responses are significant only at the 68% level. The response of the budget balance offers a perspective on why this type of disagreement is expansionary: we see that the budget balance increases significantly. A one standard deviation increase in disagreement causes an increase of the budget balance of 0.2 percentage points of GDP. Apparently, the higher disagreement is driving a shift to a more expansionary fiscal stance. This effect is absent when we look at the responses to within-coalition disagreement as shown in Figure 9.15. The higher disagreement does not change the budget balance. Industrial production, sentiment in the manufacturing industry as well as the manufacturing order volume fall as a result of higher disagreement. In addition, stock prices fall is consistent with a drop in the present value of future economic profits.

Lastly, we estimate a VAR model that includes both indicators of fiscal disagreement jointly. In a recursively identified VAR model, this requires a decision about the relative position of each indicator in the vector of endogenous variables. We decide to order government-opposition disagreement before within-government disagreement, though reversing the relative ordering of these series does not affect our results. This implies that within one quarter, government-opposition disagreement can drive withing-coalition disagreement, but not vice versa. As a second modification, we replace the DAX stock market index by the interest rate on 10-year German government bonds. To the extend higher disagreement causes an increase in risk premia, bond yields should increase.

Figure 9.16 shows the estimated responses. A shock to government-opposition disagreement causes an increase in disagreement withing the governing coalition, while the opposite is not true. We also find that bond yields increase following a higher disagreement between the government and the opposition, but not after a higher level of disagreement in the coalition. One reason for this discrepancy could be that 10-year yields, and the risk-premia they include in particular, are sensitive to uncertainty in the log-run, not the disagreement in a given election period. Diverging views between the government and the opposition could be a better predictor for long-run uncertainty than disagreement in the current coalition. The other responses remain qualitatively unchanged compared to the baseline model.

To conclude, all of our major results are robust with respect the alternative model specifications considered here. The online appendix includes the results from additional specifications.

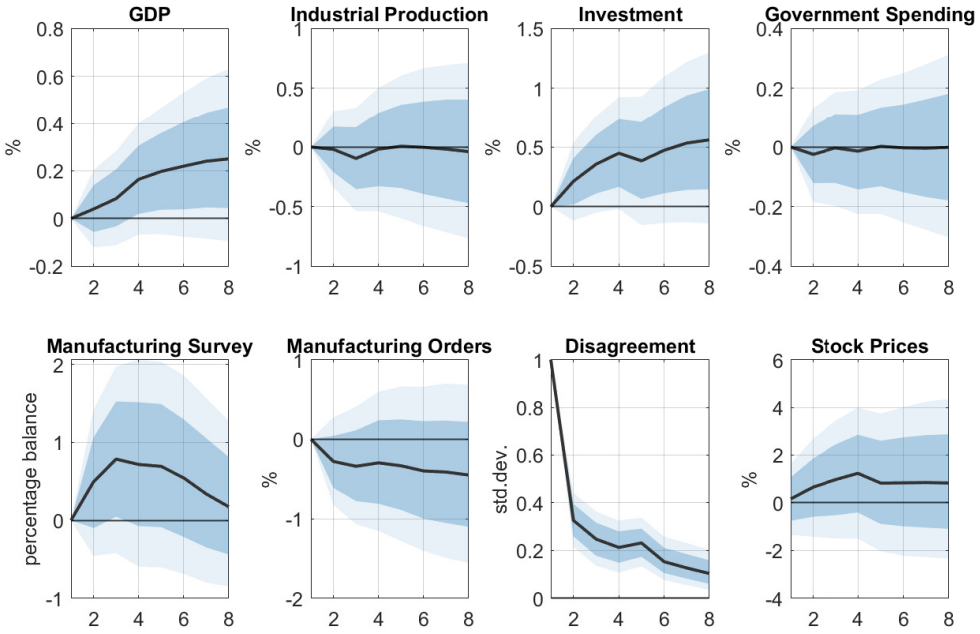



**Figure 9.12:** Response to disagreement between the government and the opposition (including election period-specific time trends)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

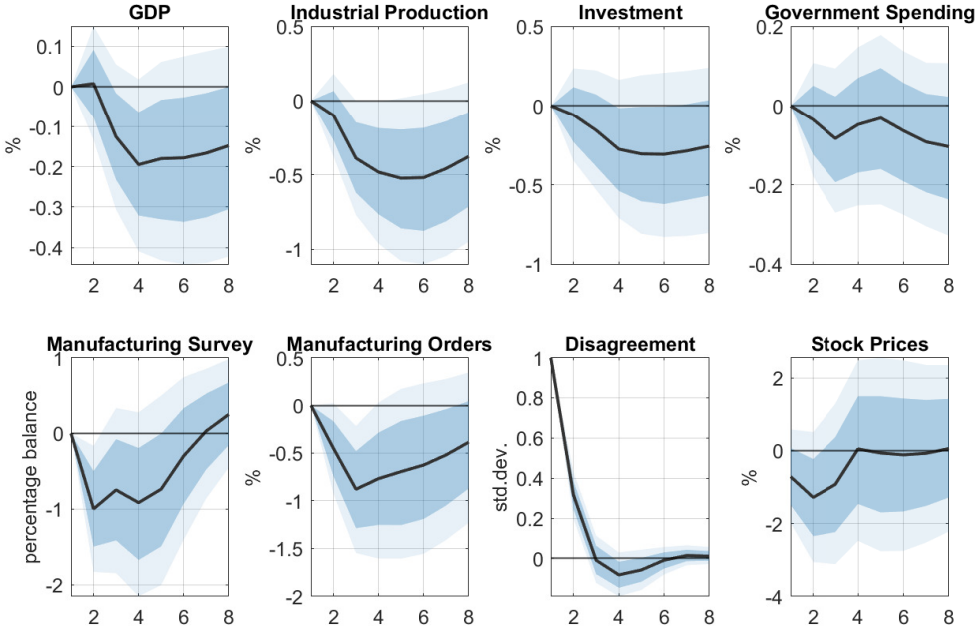



**Figure 9.13:** Response to disagreement within the government (including election period-specific time trends)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

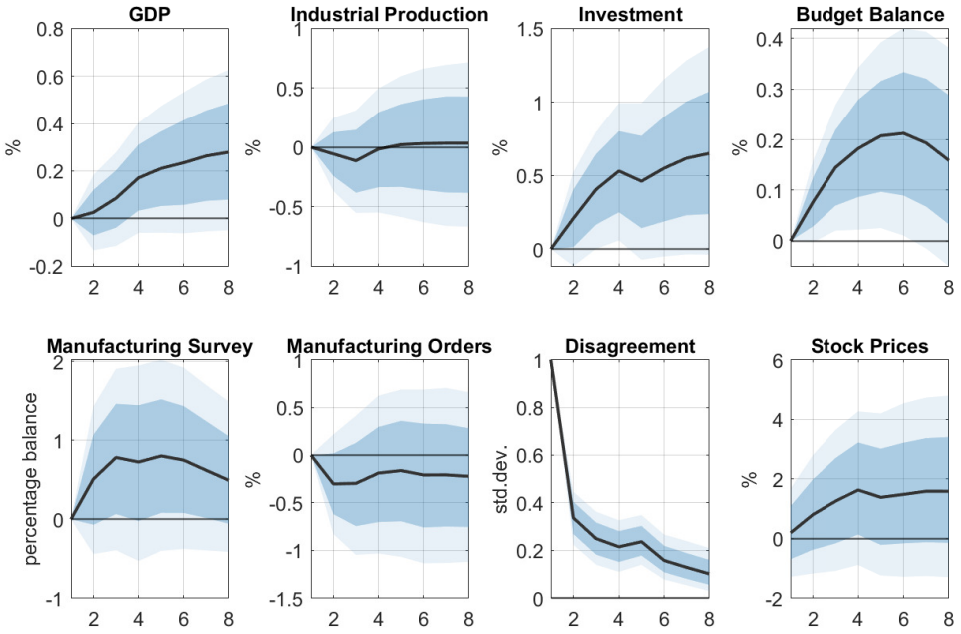



**Figure 9.14:** Response to disagreement between the government and the opposition (including the budged balance)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

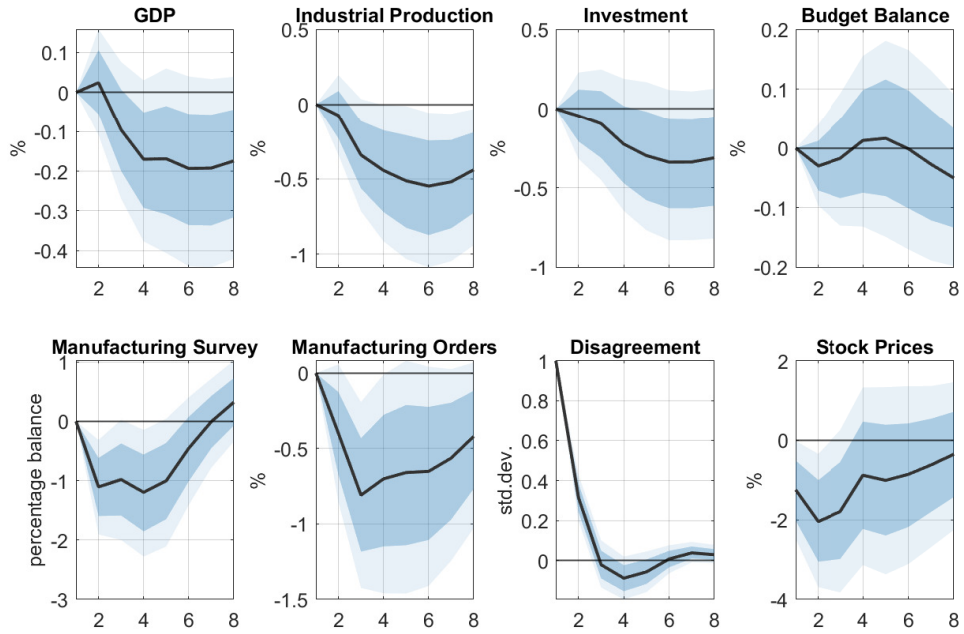



**Figure 9.15:** Response to disagreement within the government (including the budget balance)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

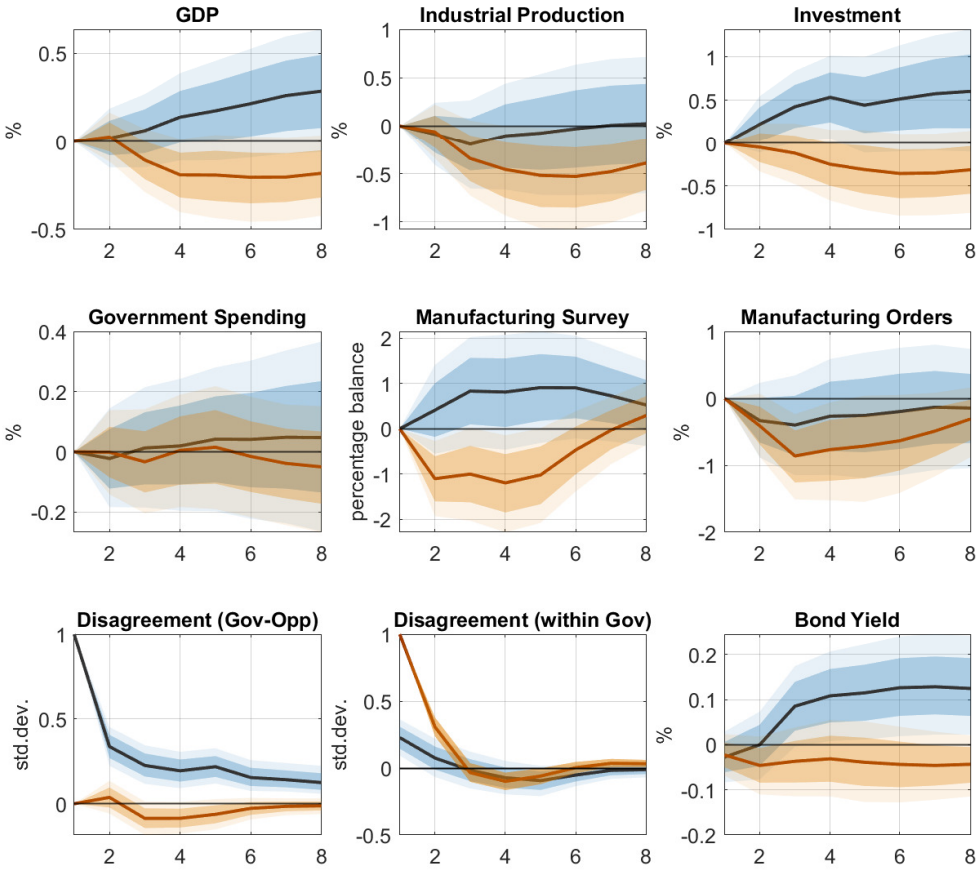



**Figure 9.16:** Response to disagreement

*Notes:* The figure shows the responses to an increase in fiscal disagreement between the government and the opposition (blue) and within the government (orange). All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

## 9.6 Conclusions

Disagreement about the path of fiscal policy is a pervasive feature of democracies. While the effect of fiscal policy as such, i.e. taxing and spending, is widely studied in the empirical literature, we know little about the effects of disagreement about these policies. In this paper, we construct two series of disagreement among German parliamentarians: the disagreement between the members of the Bundestag supporting the government and members forming the opposition and the disagreement between members of the parties forming the coalition government. We build these indicators from a textual analysis of all speeches delivered in the Bundestag since 1960. In a second contribution, we study whether fluctuations in fiscal policy disagreement have macroeconomic consequences. While higher government-opposition disagreement does not appear to affect the business cycle, higher within-government disagreement has a contractionary effect on real economic activity. We

find that industrial production, manufacturing orders, manufacturing sentiment and stock prices fall if this type of disagreement increases.

These results suggest that disagreement is itself a determinant of economic activity besides government spending and taxation. A cacophony of fiscal views among parliamentarians of the ruling parties is detrimental to economic activity. Speaking with one voice, i.e. reducing within-coalition disagreement, could have expansionary effects.

Future work could study whether the effectiveness of fiscal policy measures such as increases or decreases of government spending depends on the prevailing level of disagreement in the Bundestag. That is, studying the effects of fiscal policy *conditional* on the level of disagreement as in Ricco et al. (2016) could be fruitful. In addition, future work could differentiate between parliamentary debates about domestic fiscal policy and the design of fiscal policy and fiscal institutions at the level of the European Union. Given the central role of Germany as Europe's largest economy, shifts in the disagreement about domestic and supra-national fiscal policy could also have effects on other member states of the European Union.

# Appendix G

## G.1 Properties of disagreement

To shed light on potential determinants of fiscal disagreement, we regress each disagreement index on the binary recession indicator from the German Council of Economic Experts, the manufacturing survey used in the main part of the paper, a dummy that is one if the coalition is a grand coalition", e.g. a coalition of CDU and SPD, and zero otherwise, and the election period-specific cubic time trend.

| | dep. variable: disagreement | | | |
|---|---|---|---|---|
| | government-opposition | | within government | |
| | I | II | III | IV |
| recession | −0.193 | | 0.234 | |
| | (0.252) | | (0.145) | |
| tendency survey | | 0.010 | | −0.009 |
| | | (0.008) | | (0.005*) |
| grand coalition | −0.652 | −0.690 | −0.355 | −0.321 |
| | (0.181***) | (0.176***) | (0.142**) | (0.141**) |
| EP trend (cubic) | yes | yes | yes | yes |
| # obs. | 207 | 207 | 207 | 207 |
| $R^2$ | 0.088 | 0.095 | 0.073 | 0.075 |

**Table G.1.1:** Determinants of disagreement

*Notes:* The dependent variable is the the indicator of disagreement. Newey-West standard errors in parenthesis. A significance level of 10%, 5% and 1% is indicated by *, **, ***, respectively.

The results are shown in Table G.1.1. Disagreement does not appear sensitive to the

state of the business cycle. We find a clear pattern across all four regressions: either measure of disagreement is significantly lower under a grand coalition government. As a matter of fact, these results are interesting correlations only. We do not claim to estimate a causal effect.

## G.2 Additional indicators of disagreement

The two series of disagreement used in the main part of the paper have been standardized. A consequence of standardization is that we cannot compare the levels of disagreement between the two indicators. To shed light on the relative magnitude of disagreement, Figure G.2.1 presents the non-standardized series.

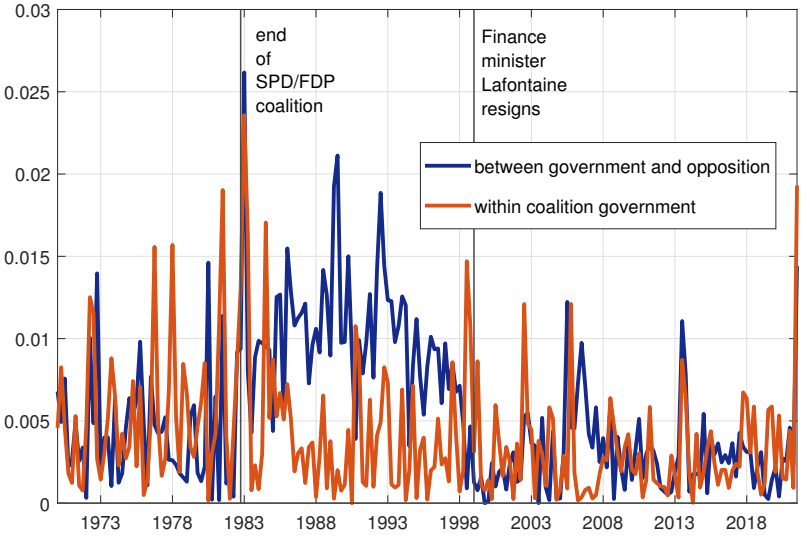



**Figure G.2.1:** Disagreement in the Bundestag: non-standardized data

*Notes*: The figure shows the two non-standardized series of fiscal policy disagreement.

During the chancellorship of Helmut Kohl and his CDU/CSU/FDP coalition, 1983 - 1998, the disagreement between the government and the opposition was much higher than the disagreement withing the governing coalition. Throughout the remaining sample, both measures of disagreement have similar magnitudes.

In the main part of the paper we studied the disagreement between the average fiscal sentiment of speakers belonging to one specific party and compare the averages between parties in the opposition and parties in the government as well as between the parties forming the government. We now introduce an alternative notion of disagreement: the absolute sentiment gap between the most expansionary party and the most restrictive party at each point in time, again both based on average fiscal sentiment in the speeches of their parliamentarians. Thus, we measure the widest distance in the fiscal sentiment across parties at a given point in time, whether the party is in the government or not. We refer to this

disagreement as maximum/minimum disagreement,

$$Dis_t = |Sentiment_t^{max} - Sentiment_t^{min}|. \tag{G.2.1}$$

Panel a of Figure G.2.2 shows the maximum and minimum fiscal sentiment among the parties in the Bundestag for each quarter. Panel b reports the standardized series of maximum/minimum disagreement. This supports our notion that the disagreement within the coalition matters for the economy, but not the dispersion of views across all parties.

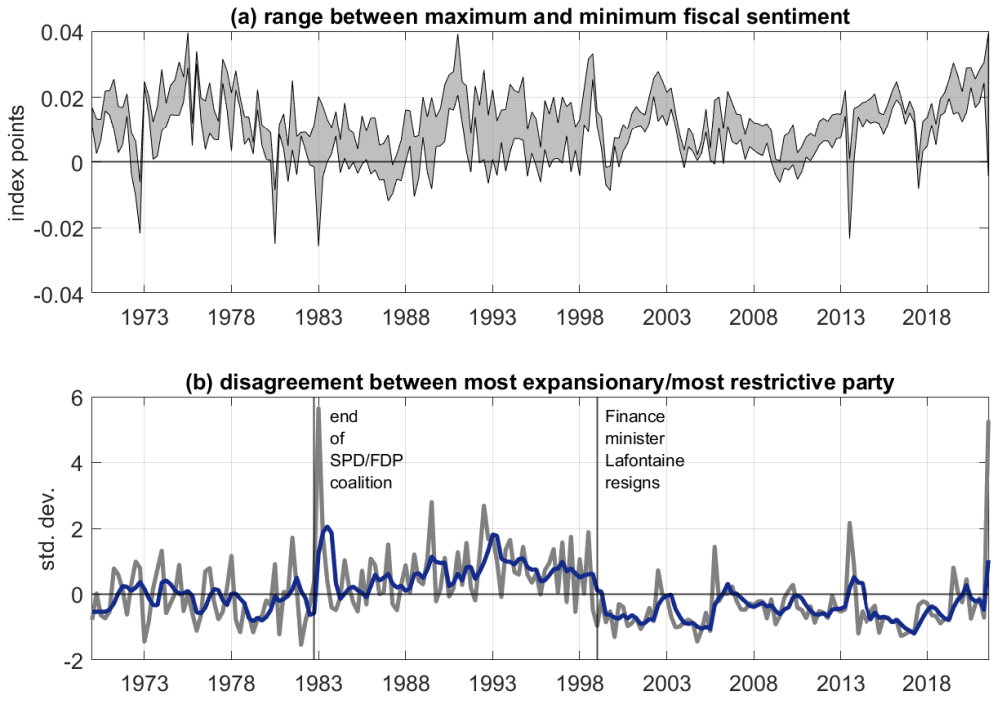



**Figure G.2.2:** Disagreement in the Bundestag: maximum vs minimum sentiment

*Notes*: The shaded area in panel (a) is range between the maximum and the minimum, i.e. the most expansionary and the most restrictive, fiscal sentiment across parties. The grey line in panel (b) shows the standardized series of fiscal policy disagreement. The blue line is the four-quarter (backward) moving average.

We now estimate our baseline model with the maximum/minimum disagreement index. As shown in Figure G.2.3, maximum/minimum disagreement has no significant effect on macroeconomic and financial variables.

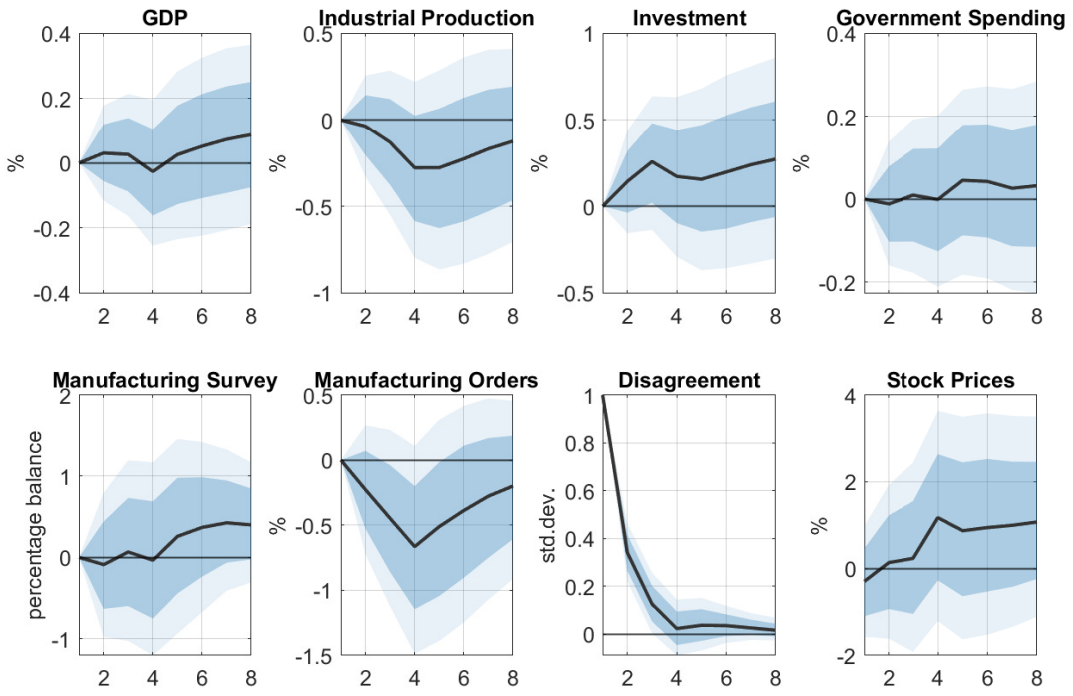



**Figure G.2.3:** Response to maxmimum/minimum disagreement

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

## G.3 Data Sources

In this section, we provide information on the sources and transformations of the macroeconomic time series used in the estimated VAR model.

1. *GDP*: GDP at constant prices, seasonally adjusted. The series is taken from `https://fred.stlouisfed.org/` ( NAEXKP01DEQ661S). Transformation: in natural logs times 100.

2. *IP*: Industrial production, seasonally adjusted. The series is taken from `https://fred.stlouisfed.org/` (DEUPROMANQISMEI). Transformation: in natural logs times 100.

3. *Investment*: Gross Fixed Capital Formation at constant prices, seasonally adjusted. The series is taken from `https://fred.stlouisfed.org/` (NAEXKP04DEQ661S). Transformation: in natural logs times 100.

4. *Expenditure*: Government Final Consumption Expenditure at current prices, seasonally adjusted. The series is taken from `https://fred.stlouisfed.org/` (DEUGFCEQDSMEI). Deflated by the seasonally adjusted GDP Deflator (DEUGDPDEFQISMEI). Transformation: in natural logs times 100.

5. *Orders*: Total manufacturing orders, seasonally adjusted. The series is taken from `https://fred.stlouisfed.org/` (ODMNTO01DEQ661S). Transformation: in natural logs times 100.

6. *Survey*: Business tendency survey for the manufacturing sector. The units are percentage balances. The series is taken from `https://fred.stlouisfed.org/` (BSPRTE02DEM460S). Transformation: none.

7. *Stocks*: German stock market index DAX. The series is taken from DATASTREAM. Transformation: in natural logs times 100.

8. *Budget Balance*: Federal budget balance as a share of GDP. The units are percentage points. The series is taken from the historical time series data base of the Bundesbank available at `https://www.bundesbank.de/en/statistics/sets-of-indicators/long-time-series`. Transformation: interpolated to quarterly frequency.

## G.4 Additional VAR results

As mentioned in the main text, the level of fiscal sentiment, and hence the extent of disagreement, in the last observation of the sample period (Q3 2021) is imprecisely estimated. This is because we only have two parliamentary sessions in this quarter in our data set. To show that this outlier does not drive our findings, we re-estimate the model and set disagreement in Q3 2021 to zero. The results are shown in Figures G.4.4 and G.4.5. All results remain unchanged.

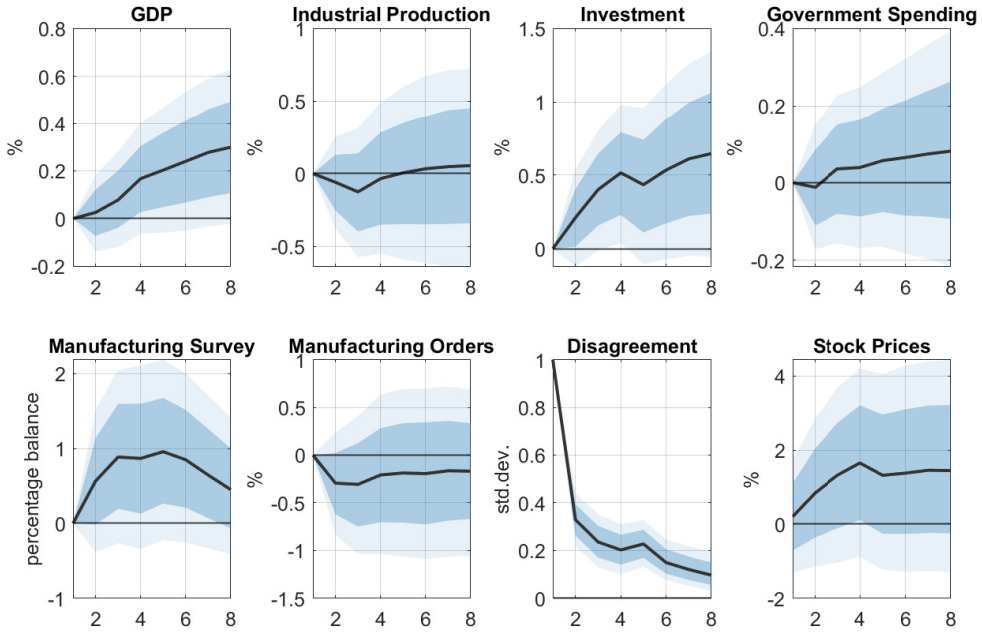



**Figure G.4.4:** Response to disagreement between the government and the opposition (correction outlier)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.


**Figure G.4.5:** Response to disagreement within the government (correcting outlier)

*Notes:* The figure shows the responses to an increase in fiscal disagreement. All responses are derived from a recursively identified Bayesian VAR model with four lags and Normal-Wishart priors. The shaded areas cover 68% and 90% of all draws.

# Part IV

# Conclusion

# Chapter 10

# Conclusion

As outlined in the introduction, the work presented in this thesis covers both the methodological and the application side of the text-as-data perspective. In the following, the final part of this thesis, most import findings are summarized and general conclusions are drawn.

The second part of this thesis presents papers on methodological advances in relation to a widely used topic modelling technique, the Latent Dirichlet Allocation (LDA). All of these papers aim to support more informed and empirically grounded decisions in the process of LDA model estimation. For example, the first paper, based on an extensive Monte Carlo simulation study, shows that around 30% of terms could be removed without any loss of quality in the resulting topics, which has practical benefits such as reduced computational costs for large datasets. The findings suggest that future research should systematically evaluate combinations of preprocessing steps, such as vocabulary pruning and stemming, to optimise text preprocessing strategies. The second paper in this section introduces methods to measure and visualize sampling uncertainty in topic models. Using non-parametric bootstrapping, it proposes to use graphical tools like word clouds with confidence bands and confidence bands for the topic weights over time to better communicate uncertainty in word importance and topic trends. The proposed methods and measures enhance the interpretability of results, highlighting that point estimates alone can be misleading and emphasizing the need to consider variability in the analysis. Future research could explore joint confidence regions and alternative bootstrap sampling techniques to improve uncertainty estimation. The third paper analyses scientific publications from Germany and Poland. LDA topic modelling is applied to identify main topics and compare them across the two corpora using proposed topic matching methods. Key contributions include using the sBIC criterion to select the optimal number of topics, developing a cosine similarity-based matching procedure to identify meaningful topic pairs, and introducing a language-agnostic approach leveraging multilingual word embeddings. Building on the results of this paper, the fourth paper investigates the performance of different model selection criteria for estimating the number of topics in Latent Dirichlet Allocation (LDA) models, with a focus on the singular Bayesian information criterion (sBIC). The results show that the performance of sBIC is robust across different

data generating processes (DGPs) considered, while the performance of other metrics seems to depend strongly on specific characteristics of the underlying data. All in all, by increasing awareness of factors that contribute more or less to the final outcomes, these advancements enhance the interpretability and reliability of LDA results, providing practical tools and insights for improved applications in diverse text-as-data applications.

The third part of this thesis is dedicated to text-as-data applications in economics. The papers presented in this part use different types of text data to construct text-based indicators and analyse their impact on, or general relationships with, real economic variables. The first paper analyses economic topics in German and Polish scientific publications from 1984–2020, examining their relevance over time and links to economic indicators through Granger and instantaneous causality tests. Results reveal significant connections between topic trends and indicators for most pairs, with economic developments generally leading scientific analysis, though some exceptions and country-specific differences were observed. Future work should be done in order to refine the proposed approach, e.g. extend the scope of the study and modify the methods used. The second paper uses news data and proposes different language-agnostic approaches to construct economic policy uncertainty (EPU) indices. These approaches build on different text representation techniques, also called embeddings. The resulted EPU indices are found to capture relevant signals and to be Granger causal to economic activity in three countries considered, Germany, Ukraine and Russia. Papers three and four in this section are related. Both are based on parliamentary speeches from the German Bundestag. Paper three presents an embedding-based approach to construct a fiscal policy sentiment index that takes into account the evolution of language over time and relies only on past information. The results of estimated vector autoregressive (VAR) models suggest a significant impact of the constructed index on the German economy. In the work that follows, the fourth paper, the impact of disagreements between the government and the opposition, as well as within the coalition government, is examined. While government-opposition disagreement has no apparent business cycle effects, increased within-government disagreement is contractionary, reducing industrial production, orders, sentiment, and stock prices. The results highlight the economic importance of the degree of agreement on fiscal policy measures within the governing coalition, suggesting that reducing intra-coalition disagreement could support economic growth. Future research could explore how disagreement impacts the effectiveness of fiscal policy and the debate at national versus EU-level. Overall, the findings in this part underscore the utility of text-based methods for understanding economic dynamics. These findings pave the way for future research to refine methodologies, expand cross-country studies, and explore the interplay between textual data and economic outcomes in other contexts.

# Bibliography

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270. doi: https://doi.org/10.1007/s42001-019-00060-w

Adämmer, P., & Schüssler, R. A. (2020). Forecasting the Equity Premium: Mind the News! *Review of Finance*, *24*(6), 1313–1355. doi: https://doi.org/10.1093/rof/rfaa007

Adämmer, P., Prüser, J., & Schüssler, R. A. (2025). Forecasting macroeconomic tail risk in real time: Do textual data add value? *International Journal of Forecasting*, *41*(1), 307-320. (Forthcoming) doi: https://doi.org/10.1016/j.ijforecast.2024.05.007

Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*, 319–335. doi: https://doi.org/10.1007/s10588-018-9266-8

Alammar, J. (2018). *The Illustrated Transformer.* Retrieved 2024-12-01, from `https://jalammar.github.io/illustrated-transformer/`

Alesina, A., & Rosenthal, H. (1995). *Partisan Politics, Divided Government, and the Economy.* Cambridge: Cambridge University Press.

Alesina, A., Roubini, N., & Cohen, G. D. (1999). *Political Cycles and the Macroeconomy.* London: MIT Press.

Algaba, A., Borms, S., Boudt, K., & van Pelt, J. (2020). The Economic Policy Uncertainty Index for Flanders, Wallonia and Belgium. *SSRN Electronic Journal*, BFW digitaal/RBF numérique 2020/6. doi: https://dx.doi.org/10.2139/ssrn.3580000

Allard, J., Catenaro, M., Vidal, J.-P., & Wolswijk, G. (2013). Central bank communication on fiscal policy. *European Journal of Political Economy*, *30*, 1–14. doi: https://doi.org/10.1016/j.ejpoleco.2012.12.001

Al Sharou, K., Li, Z., & Specia, L. (2021). Towards a better understanding of noise in natural language processing. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 53–62). Held Online: INCOMA Ltd. Retrieved 2024-12-04, from `https://aclanthology.org/2021.ranlp-1.7`

Anzuini, A., & Rossi, L. (2021). Fiscal policy in the US: a new measure of uncertainty and its effects on the American economy. *Empirical Economics*, *61*, 2613–2634. doi: https://doi.org/10.1007/s00181-020-01984-3

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., ... Zhu, M. (2013). A Practical Algorithm for Topic Modeling with Provable Guarantees. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (Vol. 28, pp. 280–288). Atlanta, Georgia, USA: PMLR. Retrieved 2024-12-04, from `https://proceedings.mlr.press/v28/arora13.html`

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 391–402). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: https://doi.org/10.1007/978-3-642-13657-3_43

Auerbach, A., & Gorodnichenko, Y. (2012). Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, *4*, 1-27. doi: https://doi.org/10.1257/pol.4.2.1

Azqueta-Gavaldón, A. (2017). Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. *Economics Letters*, *158*, 47–50. doi: https://doi.org/10.1016/j.econlet.2017.06.032

Azzimonti, M. (2018). Partisan conflict and private investment. *Journal of Monetary Economics*, *93*, 114-131. doi: https://doi.org/10.1016/j.jmoneco.2017.10.007

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. doi: https://doi.org/10.1093/qje/qjw024

Barushka, A., & Hajek, P. (2020). The Effect of Text Preprocessing Strategies on Detecting Fake Consumer Reviews. In *Proceedings of the 2019 3rd International Conference on E-Business and Internet* (pp. 13–17). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/3383902.3383908

Beckmann, J., & Czudaj, R. (2021). Fiscal policy uncertainty and its effects on the real economy: German evidence. *Oxford Economic Papers*, *73*, 1516—1535. doi: https://doi.org/10.1093/oep/gpab009

Ben Zeev, N., & Pappa, E. (2017). Chronicle of a war foretold: the macroeconomic effects of anticipated defence spending shocks. *The Economic Journal*, *127*, 1568-1597. doi: https://doi.org/10.1111/ecoj.12349

Berg, T. O. (2016). Time varying fiscal multipliers in Germany. *Review of Economics*, *66*, 13–46. doi: https://doi.org/10.1515/roe-2015-0103

Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. *PLoS ONE*, *12*(4), 1–22. doi: https://doi.org/10.1371/journal.pone.0176310

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1676–1683). Online: Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2021.eacl-main.143

Bird, E. L., Steven, & Klein, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media Inc.

Blanchard, O., & Perotti, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics*, *117*, 1329–1368. Retrieved 2024-12-04, from `http://www.jstor.org/stable/4132480`

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering and applications* (pp. 71–94). Boca Raton, Florida, USA: CRC Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Born, B., & Müller, G. (2012). Government spending shocks in quarterly and annual time series. *Journal of Money, Credit and Banking*, *44*, 507-517. Retrieved 2024-12-04, from `https://www.jstor.org/stable/41487806`

Born, B., & Pfeifer, J. (2014). Policy risk and the business cycle. *Journal of Monetary Economics*, *68*, 68–85. doi: https://doi.org/10.1016/j.jmoneco.2014.07.012

Burst, T., Krause, W., Lehmann, P., Lewandowski, J., Matthieß, T., Merz, N., ... Zehnter, L. (2020). *Manifesto corpus. Version: 2020-1.* Berlin.

Bystrov, V., Naboka, V., Staszewska-Bystrova, A., & Winker, P. (2022). Cross-Corpora Comparisons of Topics and Topic Trends. *Journal of Economics and Statistics*, *242*(4), 433–469. doi: https://doi.org/10.1515/jbnst-2022-0024

Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2023). Analysing the Impact of Removing Infrequent Words on Topic Quality in LDA Models. *arXiv*, *abs/2311.14505*. Retrieved 2024-12-04, from `https://arxiv.org/abs/2311.14505`

Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024a). Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria. *Journal of Machine Learning Research*, *25*(79), 1–30. Retrieved 2024-12-04, from `http://jmlr.org/papers/v25/23-0188.html`

Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A., & Winker, P. (2024b). Comparing Links between Topic Trends and Economic Indicators in the German and Polish Academic Literature. *Comparative Economic Research. Central and Eastern Europe*, *2*, 7–28. doi: https://doi.org/10.18778/1508-2008.27.10

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7), 1775–1781. doi: https://doi.org/10.1016/j.neucom.2008.06.011

Chen, Y.-C., Wang, Y. S., & Erosheva, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. *The Annals of Applied Statistics*, *12*(2), 846 – 876. doi: https://doi.org/10.1214/18-AOAS1169

Christofzik, D. I., Fuest, A., & Jessen, R. (2022). Macroeconomic effects of the anticipation and implementation of tax changes Germany: Evidence from a narrative account. *Economica*, *89*, 62-81. doi: https://doi.org/10.1111/ecca.12389

Čižmešija, M., Lolić, I., & Sorić, P. (2017). Economic policy uncertainty index and economic activity: what causes what? *Croatian Operational Research Review*, *8*(2), 563–575. doi: https://doi.org/10.17535/crorr.2017.0036

Cloyne, J. (2013). Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom. *American Economic Review*, *103*, 1507-28. doi: https://doi.org/10.1257/aer.103.4.1507

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word Translation Without Parallel Data. *arXiv*, *abs/1710.04087*. Retrieved 2024-12-04, from `https://arxiv.org/abs/1710.04087`

Dahlke, J., Bogner, K., Becker, M., Schlaile, M. P., Pyka, A., & Ebersberger, B. (2021). Crisis-driven innovation and fundamental human needs: A typological framework of rapid-response COVID-19 innovations. *Technological Forecasting & Social Change*, *169*, 120799. doi: https://doi.org/10.1016/j.techfore.2021.120799

Debnath, R., & Bardhan, R. (2020). India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling. *PLoS ONE*, *15*(9), e0238972. doi: https://doi.org/10.1371/journal.pone.0238972

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. doi: https://doi.org/10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N19-1423

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. doi: https://doi.org/10.1162/tacl_a_00325

Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *PLoS ONE*, *17*(2), e0263898. doi: https://doi.org/10.1371/journal.pone.0263898

Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(2), 323–380. doi: https://doi.org/10.1111/rssb.12187

Dybowski, T. P., & Adämmer, P. (2018). The economic effects of U.S. presidential tax communication: Evidence from a correlated topic model. *European Journal of Political Economy*, *55*, 511–525. doi: https://doi.org/10.1016/j.ejpoleco.2018.05.001

Edison, H., & Carcel, H. (2021). Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters*, *28*(1), 38–42. doi: https://doi.org/10.1080/13504851.2020.1730748

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1 – 26. doi: https://doi.org/10.1214/aos/1176344552

Egbert, J., & Plonsky, L. (2020). Bootstrapping techniques. In *A Practical Handbook of Corpus Linguistics* (pp. 593–610). Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-46216-1_24

Ellahie, A., & Ricco, G. (2017). Government purchases reloaded: Informational insufficiency and heterogeneity in fiscal VARs. *Journal of Monetary Economics*, *90*, 13-27. doi: https://doi.org/10.1016/j.jmoneco.2017.06.002

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. doi: https://doi.org/10.1002/jae.2859

Fatás, A., & Mihov, I. (2001). *The effects of fiscal policy on consumption and employment: Theory and Evidence* (Working Paper). Fontainebleau, France: INSEAD.

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., & Rubio-Ramírez, J. (2015). Fiscal Volatility Shocks and Economic Activity. *American Economic Review*, *105*, 3352-3384. doi: https://doi.org/10.1257/aer.20121236

Ferrara, F. M., Masciandaro, D., Moschella, M., & Romelli, D. (2022). Political voice on monetary policy: Evidence from the parliamentary hearings of the European Central Bank. *European Journal of Political Economy*, *74*, 102143. doi: https://doi.org/10.1016/j.ejpoleco.2021.102143

Ferrara, L., Metelli, L., Natoli, F., & Siena, D. (2021). Questioning the puzzle: fiscal policy, real exchange rate and inflation. *Journal of International Economics*, *133*, 103524. doi: https://doi.org/10.1016/j.jinteco.2021.103524

Fisher, J. D. M., & Peters, R. (2010). Using stock returns to identify government spending shocks. *The Economic Journal*, *120*, 414–436. doi: https://doi.org/10.1111/j.1468-0297.2010.02355.x

Foltas, A. (2022). Testing Investment Forecast Efficiency with Forecasting Narratives. *Journal of Economics and Statistics*, *242*(2), 191–222. doi: https://doi.org/doi:10.1515/jbnst-2020-0027

Forni, M., & Gambetti, L. (2016). Government spending shocks in open economy VARs. *Journal of International Economics*, *99*, 68–84. doi: https://doi.org/10.1016/j.jinteco.2015.11.010

Galí, J., López-Salido, D., & Vallés, J. (2007). Understanding the effects of government spending on consumption. *Journal of the European Economic Association*, *5*, 227-270. doi: https://doi.org/10.1162/JEEA.2007.5.1.227

Gennaro, G., & Ash, E. (2021). Emotion and reason in political language. *The Economic Journal*, *132*(643), 1037-1059. doi: https://doi.org/10.1093/ej/ueab104

Ghirelli, C., Pérez, J. J., & Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters*, *182*, 64–67. doi: https://doi.org/10.1016/j.econlet.2019.05.021

Gordon, R. J., & Krenn, R. (2010). *The end of the Great Depression: VAR insight on the roles of monetary and fiscal policy* (Working Paper No. No 16380). Cambridge, Massachusetts, USA: National Bureau of Economic Research.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved 2024-12-04, from `https://www.aclweb.org/anthology/L18-1550`

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. doi: https://doi.org/10.1073/pnas.030775210

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. doi: https://doi.org/10.1093/pan/mps028

Guajardo, J., Leigh, D., & Pescatori, A. (2014). Expansionary austerity? International evidence. *Journal of the European Economic Association*, *12*(4), 949–968. doi: https://doi.org/10.1111/jeea.12083

Gulen, H., & Ion, M. (2016). Policy uncertainty and corporate investment. *The Review of Financial Studies*, *29*, 523-564. doi: https://doi.org/10.1093/rfs/hhv050

Hacioğlu-Hoke, S. (2024). Macroeconomic effects of political risk shocks. *Economics Letters*, *242*, 111877. doi: https://doi.org/10.1016/j.econlet.2024.111877

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, *15*(5), e0232525. doi: 10.1371/journal.pone.0232525

Hamilton, J. (2018). Why you should never use the Hodrick-Prescott filter. *The Review of Economics and Statistics*, *100*, 831-843. doi: https://doi.org/10.1162/rest_a_00706

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, 114–133. doi: https://doi.org/10.1016/j.jinteco.2015.12.008

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870. doi: https://doi.org/10.1093/qje/qjx045

Hantzsche, A. (2022). Fiscal uncertainty and sovereign credit risk. *European Economic Review*, *148*, 104245. doi: https://doi.org/10.1016/j.euroecorev.2022.104245

Hartmann, P., & Smets, F. (2018). The European Central Bank's Monetary Policy during Its First 20 Years. *Brookings Papers on Economic Activity*, *Fall 2018*, 1–146. Retrieved 2024-12-04, from `https://www.brookings.edu/articles/the-first-20-years-of-the-european-central-bank-monetary-policy/`

Hayashi, N. (2021). The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation. *Neural Networks*, *137*, 127–137. doi: https://doi.org/10.1016/j.neunet.2021.01.024

Hayo, B., & Mierzwa, S. (2022). Legislative tax announcements and GDP: evidence from the United States, Germany, and the United Kingdom. *Economics Letters*, *216*, 110548. doi: https://doi.org/10.1016/j.econlet.2022.110548

Hayo, B., & Uhl, M. (2014). The macroeconomic effects of legislated tax changes in Germany. *Oxford Economic Papers*, *66*, 397–418. Retrieved 2024-12-04, from `https://www.jstor.org/stable/43772871`

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1833–1842). doi: https://doi.org/10.1109/HICSS.2014.231

Herwartz, H., & Lange, A. (2020). Bootstrapping in Macroeconometrics. In B. Anindya (Ed.), *Oxford Research Encyclopedia of Economics and Finance.* Oxford University Press. doi: https://doi.org/10.1093/acrefore/9780190625979.013.165

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 856–864). Red Hook, NY, USA: Curran Associates, Inc.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(1), 1303–1347. Retrieved 2024-12-04, from `https://jmlr.org/papers/v14/hoffman13a.html`

Hong, G., Ke, S., & Nguyen, A. D. (2024). *The Economic Impact of Fiscal Policy Uncertainty: Evidence from a New Cross-Country Database* (Working Paper). Washington, D.C., USA: International Monetary Fund.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Retrieved 2024-12-04, from `https://spacy.io/`

Hruzik, J. (2019). *pybundestag.* GitHub. Retrieved 2023-01-17, from `https://github.com/Jhruzik/pybundestag`

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*, *64*(6), 2833–2855. doi: https://doi.org/10.1287/mnsc.2017.2751

Ilori, A. E., Paez-Farrell, J., & Thoenissen, C. (2012). Fiscal policy shocks and international spillovers. *European Economic Review*, *141*, 103969. doi: https://doi.org/10.1016/j.euroecorev.2021.103969

Ilzetzki, E., Mendoza, E. G., & Vegh, C. A. (2013). How big (small?) are fiscal multipliers? *Journal of Monetary Economics*, *60*, 239–254. doi: https://doi.org/10.1016/j.jmoneco.2012.10.011

Jens, C. E. (2017). Policy uncertainty and investment: causal evidence from U.S. gubernatorial elections. *Journal of Financial Economics*, *124*, 563-579. doi: https://doi.org/10.1016/j.jfineco.2016.01.034

Jentsch, C., Lee, E. R., & Mammen, E. (2020). Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, *142*, 106813. doi: https://doi.org/10.1016/j.csda.2019.106813

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, *95*, 161-182. doi: https://doi.org/10.1257/0002828053828518

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2979–2984). Brussels, Belgium: Association for Computational Linguistics. doi: https:/doi.org/10.18653/v1/D18-1330

Julio, B., & Yook, Y. (2012). Political uncertainty and corporate investment cycles. *The Journal of Finance*, *67*, 45-83. doi: https://doi.org/10.1111/j.1540-6261.2011.01707.x

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). *Making text count: economic forecasting using newspaper text* (Bank of England working papers

No. 865). Bank of England. Retrieved 2024-12-04, from https://ideas.repec.org/p/boe/boeewp/0865.html

Kapfhammer, F., Larsen, V. H., & Thorsrud, L. A. (2020). *Climate risk and commodity currencies* (Working Papers No. No 10/2020). Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School. Retrieved 2024-12-04, from https://ideas.repec.org/p/bny/wpaper/0093.html

Kilian, L. (2009). Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. *American Economic Review*, *99*(3), 1053–1069. doi: https:/doi.org/10.1257/aer.99.3.1053

Kilian, L., & Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press. doi: 10.1017/9781108164818

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring Emotions in the COVID-19 Real World Worry Dataset. *arXiv*, *abs/2004.04225*. Retrieved 2024-12-04, from https://arxiv.org/abs/2004.04225

Kontoghiorghes, L., & Colubi, A. (2023). New metrics and tests for subject prevalence in documents based on topic modeling. *International Journal of Approximate Reasoning*, *157*, 49-69. doi: https://doi.org/10.1016/j.ijar.2023.02.009

Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2020). *The Digital Layer: How Innovative Firms Relate on the Web* (Tech. Rep. No. 20-003). ZEW - Centre for European Economic Research. Retrieved 2024-12-04, from https://ssrn.com/abstract=3530807

Kuhn, H. W., & Yaw, B. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 83–97. doi: https://doi.org/10.1002/nav.3800020109

Lahiri, K., & Sheng, X. (2010). Measuring forecast uncertainty by disagreement: the missing link. *Journal of Applied Econometrics*, *25*, 514-538. doi: https://doi.org/10.1002/jae.1167

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*, 203–218. doi: https://doi.org/10.1016/j.jeconom.2018.11.013

Larsen, V. H., & Thorsrud, L. A. (2022). Asset returns, news topics, and media effects. *The Scandinavian Journal of Economics*, *124*(3), 838–868. doi: https://doi.org/10.1111/sjoe.12469

Latifi, A. (2024). MaFiText-Bundestag speeches: Processing stenographic protocols of the German Bundestag. *Unpublished. Work in Progress*.

Latifi, A., Naboka-Krell, V., Tillmann, P., & Winker, P. (2024). Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects. *European Economic Review*, *168*, 104827. doi: https://doi.org/10.1016/j.euroecorev.2024.104827

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 1188–1196). Bejing, China: PMLR. Retrieved 2024-12-04, from `https://proceedings.mlr.press/v32/le14.html`

Lenz, D., & Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models. *PLoS ONE*, *15*(1), 1–18. doi: https://doi.org/10.1371/journal.pone.0226685

Lewis, C., & Grossetti, F. (2022). A Statistical Approach for Optimal Topic Model Identification. *Journal of Machine Learning Research*, *23*, 1–20. Retrieved 2024-12-04, from `https://jmlr.org/papers/v23/19-297.html`

Lin, A. Y.-T., & Katada, S. N. (2022). Striving for greatness: status aspirations, rhetorical entrapment, and domestic reforms. *Review of International Political Economy*, *29*(1), 175–201. doi: https://doi.org/10.1080/09692290.2020.1801486

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

Loureiro, S. M. C., Guerreiro, J., & Tussyadiah, I. (2021). Artificial intelligence in business: State of the art and future research agenda. *Journal of Business Research*, *129*, 911–926. doi: https://doi.org/10.1016/j.jbusres.2020.11.001

Lu, K., Cai, X., Ajiferuke, I., & Wolfram, D. (2017). Vocabulary size and its effect on topic representation. *Information Processing & Management*, *53*(3), 653-665. doi: https://doi.org/10.1016/j.ipm.2017.01.003

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254–277. doi: https://doi.org/10.1093/pan/mpu019

Lüdering, J., & Tillmann, P. (2020). Monetary policy on Twitter and asset prices: Evidence from computational text analysis. *The North American Journal of Economics and Finance*, *51*, 100875. doi: https://doi.org/10.1016/j.najef.2018.11.004

Lüdering, J., & Winker, P. (2016). Forward or Backward Looking? The Economic Discourse and the Observed Reality. *Journal of Economics and Statistics*, *236*(4), 483–515. doi: doi:10.1515/jbnst-2015-1026

Lütkepohl, H., Staszewska-Bystrova, A., & Winker, P. (2020). Constructing joint confidence bands for impulse response functions of VAR models – a review. *Econometrics and Statistics*, *13*, 69–83. doi: https://doi.org/10.1016/j.ecosta.2018.10.002

Lütkepohl, H., Staszewska-Bystrova, A., & Winker, P. (2015). Comparison of methods for constructing joint confidence bands for impulse response functions. *International Journal of Forecasting*, *31*(3), 782-798. doi: https://doi.org/10.1016/j.ijforecast.2013.08.003

MacKinnon, J. G. (2009). Bootstrap Hypothesis Testing. In *Handbook of Computational Econometrics* (pp. 183–213). John Wiley & Sons, Ltd. doi: https://doi.org/10.1002/9780470748916.ch6

Maier, D., Baden, C., Stoltenberg, D., Vries-Kedem, M. D., & Waldherr, A. (2022). Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections. *Communication Methods and Measures*, *16*(1), 19–38. doi: https://doi.org/10.1080/19312458.2021.1955845

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*, 93–118. doi: https://doi.org/10.1080/19312458.2018.1430754

Mamaysky, H. (2023). News and Markets in the Time of COVID-19. *Journal of Financial and Quantitative Analysis*, 1–37. doi: https://doi.org/10.1017/S002210902300131X

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. Jackson (Ed.), *Communication Theory* (pp. 486–502). Princeton: Academic Press.

Mertens, K., & Ravn, M. (2010). Measuring the impact of fiscal policy in the face of anticipation: a structural VAR approach. *The Economic Journal*, *120*, 393-413. doi: https://doi.org/10.1111/j.1468-0297.2010.02361.x

Mertens, K., & Ravn, M. (2012). Empirical evidence on the aggregate effects of anticipated and unanticipated US tax policy shocks. *American Economic Journal: Economic Policy*, *4*, 145-181. doi: https://doi.org/10.1257/pol.4.2.145

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR.* Scottsdale, Arizona, USA. Retrieved 2024-12-04, from `http://arxiv.org/abs/1301.3781`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26, pp. 3111–3119). Curran Associates, Inc. Retrieved 2024-12-04, from `https://arxiv.org/abs/1310.4546`

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved 2024-12-04, from `https://aclanthology.org/D11-1024`

Morstatter, F., & Liu, H. (2018). In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. *Journal of Machine Learning Research*, *18*(169), 1–32. Retrieved 2024-12-04, from `http://jmlr.org/papers/v18/17-069.html`

Niekler, A., & Jähnichen, P. (2012). Matching Results of Latent Dirichlet Allocation for Text. In N. Rußwinkel, U. Drewitz, & H. van Rijn (Eds.), *Proceedings of the 11th International Conference on Cognitive Modeling* (pp. 317–322). Berlin: Universitaetsverlag der TU Berlin.

Nyman, R., & Ormerod, P. (2020). Text as data: a machine learning-based approach to measuring uncertainty. *arXiv*, *abs/2006.06457*. Retrieved 2024-12-04, from `https://arxiv.org/abs/2006.06457`

Papamichalis, T., Ryu, D., & Wilson, M. (2024). *Divided government and the stock market* (Working Paper). Cambridge, United Kingdom: University of Cambridge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved 2024-12-04, from `https://www.jmlr.org/papers/v12/pedregosa11a.html`

Perić, B. Š., & Sorić, P. (2018). A Note on the "Economic Policy Uncertainty Index". *Social Indicators Research*, *137*(2), 505–526. doi: https://doi.org/10.1007/s11205-017-1609-1

Picault, M., & Renault, T. (2017). Words are not all created equal: a new measure of ECB communication. *Journal of International Money and Finance*, *79*, 136-156. doi: https://doi.org/10.1016/j.jimonfin.2017.09.005

Polyzos, E., & Wang, F. (2022). Twitter and market efficiency in energy markets: Evidence using LDA clustered topic extraction. *Energy Economics*, *114*, 106264. doi: https://doi.org/10.1016/j.eneco.2022.106264

Poterba, J. (1994). State Responses to Fiscal Crises: The Effects of Budgetary Institutions and Politics. *Journal of Political Economy*, *102*, 799-821. Retrieved 2024-12-04, from `https://www.jstor.org/stable/2138765`

Pástor, L., & Veronesi, P. (2012). Uncertainty about government policy and stock ürices. *The Journal of Finance*, *67*, 1219-1264. doi: http://dx.doi.org/10.2139/ssrn.1625845

Ramey, V. (2011). Identifying Government Spending Shocks: It's all in the Timing. *The Quarterly Journal of Economics*, *126*, 1-50. doi: https://doi.org/10.1093/qje/qjq008

Ramey, V. (2016). Chapter 2 - macroeconomic shocks and their propagation. In J. B. Taylor & H. Uhlig (Eds.), (Vol. 2, p. 71-162). Elsevier. doi: https://doi.org/10.1016/bs.hesmac.2016.03.003

Ramey, V. A. (2011). Can government purchases stimulate the economy? *Journal of Economic Literature*, *49*, 673–685. doi: https://doi.org/10.1016/j.jmoneco.2017.06.002

Ramey, V. A., & Zubairy, S. (2018). Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data. *Journal of Political Economy*, *126*, 850-901. doi: https://doi.org/10.1086/696277

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. doi: https://doi.org/10.13140/2.1.2393.1847

Reimann, S., & Dakota, D. (2021). Examining the effects of preprocessing on the detection of offensive language in German tweets. In K. Evang, L. Kallmeyer, R. Osswald, J. Waszczuk, & T. Zesch (Eds.), *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)* (pp. 159–169). Düsseldorf, Germany: KONVENS 2021 Organizers. Retrieved 2024-12-04, from `https://aclanthology.org/2021.konvens-1.14`

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. doi: https:/doi.org/10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv*, *abs/2004.09813*. Retrieved 2024-12-04, from `https://arxiv.org/abs/2004.09813`

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133. doi: https://doi.org/10.1017/pan.2019.26

Ricco, G., Callegari, G., & Cimadomo, J. (2016). Signals from the government: policy disagreement and the transmission of fiscal shocks. *Journal of Monetary Economics*, *82*, 107–118. doi: https://doi.org/10.1016/j.jmoneco.2016.07.004

Rieger, J., Jentsch, C., & Rahnenführer, J. (2020). Assessing the Uncertainty of the Text Generating Process Using Topic Models. In e. a. Koprinska Irena (Ed.), *ECML PKDD 2020 Workshops* (pp. 385–396). Cham: Springer International Publishing. Retrieved 2024-12-04, from `https://link.springer.com/chapter/10.1007/978-3-030-65965-3_26`

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003. doi: https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1–40. Retrieved from `https://www.jstatsoft.org/index.php/jss/article/view/v091i02` doi: 10.18637/jss.v091.i02

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What works, What doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101–115.

Romer, C. D., & Romer, D. H. (2010). The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review*, *100*, 763-801. doi: https://doi.org/10.1257/aer.100.3.763

Roubini, N., & Sachs, J. (1989). Political and economic determinants of budget deficits in the industrial democracies. *European Economic Review*, *33*, 903-938. doi: https://doi.org/10.1016/0014-2921(89)90002-0

Savin, I., Ott, I., & Konop, C. (2022). Tracing the evolution of service robotics: Insights from a topic modeling approach. *Technological Forecasting & Social Change*, *174*, 121280. doi: https://doi.org/10.1016/j.techfore.2021.121280

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 432–436). Valencia, Spain: Association for Computational Linguistics. Retrieved 2024-12-04, from `https://aclanthology.org/E17-2069`

Schwartz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved 2024-12-04, from `https://www.jstor.org/stable/2958889`

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*. doi: https://doi.org/10.1016/j.jeconom.2020.07.053

Silgado-Gómez, E. (2024). Sovereign uncertainty. *International Economic Review*, *65*(4). doi: https://doi.org/10.1111/iere.12718

Stoltenberg, D., Maier, D., Niekler, A., & Wiedemann, G. (2020). How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models. *Computational Communication Research*, *2*, 139–152. doi: https://doi.org/10.5117/CCR2020.2.001.MAIE

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 190–198). Bejing, China: PMLR. Retrieved 2024-12-04, from `https://proceedings.mlr.press/v32/tang14.html`

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. doi: https://doi.org/10.1198/016214506000000302

Tenhofen, J., Wolff, G. B., & Heppke-Falck, K. H. (2010). The Macroeconomic Effects of Exogenous Fiscal Policy Shocks in Germany: A Disaggregated SVAR Analysis. *Jahrbücher für Nationalökonomie und Statistik*, *230*, 328–355. doi: https://doi.org/10.1515/jbnst-2010-0305

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. doi: https://doi.org/10.1080/07350015.2018.1506344

Tiba, S., Rijnsoever, F. J. v., & Hekkert, M. P. (2018). Firms with benefits: A systematic review of responsible entrepreneurship and corporate social responsibility literature. *Corporate Social Responsibility and Environmental Management*, *26*(2), 265–284. doi: https://doi.org/10.1002/csr.1682

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. Retrieved 2024-12-04, from `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, *94*, 236–250. doi: https://doi.org/10.1016/j.techfore.2014.10.006

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why Priors Matter. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (pp. 1973–1981). Red Hook, NY, USA: Curran Associates, Inc.

Wang, F., Zhang, J. L., Li, Y., Deng, K., & Liu, J. S. (2021). Bayesian Text Classification and Summarization via A Class-Specified Topic Model. *Journal of Machine Learning Research*, *22*(89), 1–48. Retrieved 2024-12-04, from `http://jmlr.org/papers/v22/18-332.html`

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory.* Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511800474

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, *28*(4). doi: https://doi.org/10.1145/1852102.1852106

Wehrheim, L. (2019). Economic history goes digital: topic modeling the Journal of Economic History. *Cliometrica*, *13*(1), 83–125. doi: https://doi.org/10.1007/s11698-018-0171-7

Werner, A., Lacewell, O., & Volkens, A. (2011). Manifesto coding instructions (4th fully revised edition), may 2011 [Computer software manual]. Berlin.

Winker, P. (2023). Visualizing Topic Uncertainty in Topic Modelling. *arXiv*, *abs/2302.06482*. Retrieved 2024-12-04, from `https://arxiv.org/abs/2302.06482`

Xie, F. (2020). Wasserstein Index Generation Model: Automatic generation of time-series index with application to Economic Policy Uncertainty. *Economics Letters*, *186*, 108874. doi: https://doi.org/10.1016/j.econlet.2019.108874

Zohar, O. (2024). Cyclicality of uncertainty and disagreement. *Journal of Monetary Economics*, *143*, 103544. doi: https://doi.org/10.1016/j.jmoneco.2023.12.002