Nowcasting GDP in Real-Time with a Tone-Adjusted, Time-Varying Layered Topic Model

Jasper de Winter* & Dorinth van Dijk‡

De Nederlandsche Bank (DNB) & VU University Amsterdam*
De Nederlandsche Bank‡

2021 Joint Statistical Meetings, Seattle (USA), 12 August, 2021

Introduction

Research question

- Can we use newspaper text to track the business cycle and nowcast GDP growth?

Motivation

- Understand what drives business cycle fluctuations
- Important to have a point of departure: nowcast current pace of growth economy

Idea

- Extract topics from newspaper articles using Latent Dirichlet Allocation (LDA), see Blei et al. (2003)
- Extract newspaper sentiment from newspaper articles using lexicon-based method
- Combine LDA with sentiment: **tone-adjusted topic model**, see e.g. Larsen en Thorsrud (2018) and Thorsrud (2020)

DeNederlandscheBank EUROSYSTEEM

General idea

Traditional approach

- Use survey data, financial market indices, hard indicators
- Use a nowcasting model to forecast GDP growth.

New datasources

- A number is a fact, but the media in which it is presented/discussed/opinionated adds to the information
- Putting text into models like LDA is somewhat new (to economists) - until recently, see e.g. Hansen et al. (2018) and Thorsrud (2020)

Main contributions

- Analyze unique new source of Dutch newspaper texts
- Extend base-LDA by including **time-variation** and **layering** in topics
- Analyze forecasting quality time series of **tone-adjusted time-varying layered topics** in nowcasting model. See e.g. Jansen en de Winter (2018) and Ellingsen et al. (2021)

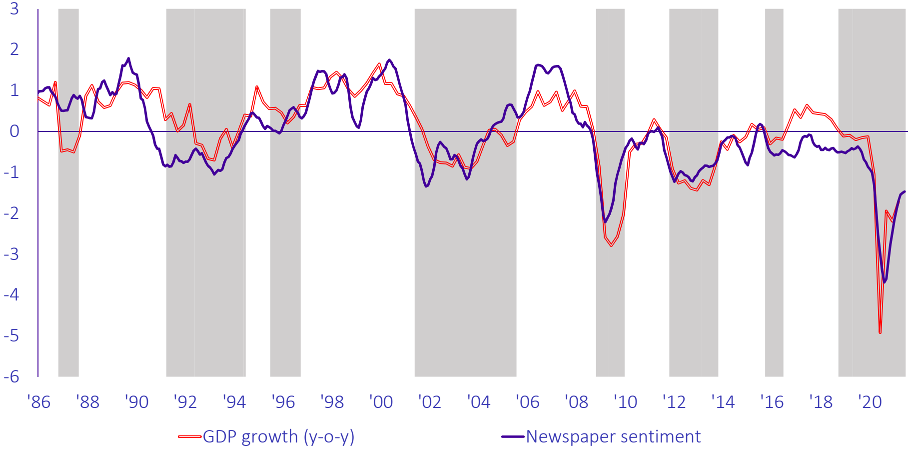
DeNederlandscheBank
EUROSYSTEEM

Outline presentation

Outline

- **Data**: Newspaper and preparation for topic model
- **Model**: Intuition topic model and extensions
- **Sentiment**: Measurement and indicator of newspaper sentiment
- **Nowcasting**: Usefulness newspaper sentiment for nowcasting GDP growth
- **Wrap up**

Sneak preview



DeNederlandscheBank
EUROSYSTEEM

Data: Financieele Dagblad

Source

- Complete full-text archive of Dutch “Financial Times”
- Language newspaper: Dutch
- Strong focus on financial-economic news and socio-economic (politics)
- Period | 36 years | January 1^{st} 1985 - December 31^{st} 2020
- $\pm$ 1.1 million full-text articles

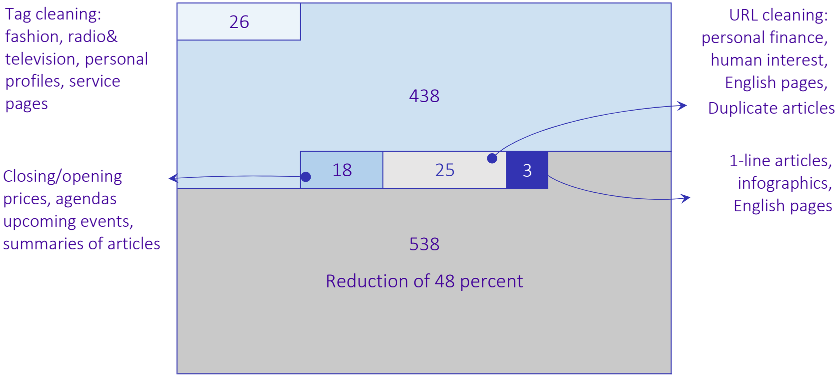
Coverpage FD, August 9 2021

Coverpage FD, August 6 2021

DeNederlandscheBank
EUROSYSTEEM

Data: Financieele Dagblad - cleanup articles

Cleaning database, X 1,000 articles



DeNederlandscheBank
EUROSYSTEEM

Data: Financieele Dagblad - cleanup text

Remove

- HTML-tags, numbers, punctuation
- Stopwords:

  R-package `snowballC` (Porter, 2001) and custom list

Adjust

- Collocations: Private equity | current account | Royal Dutch Shell
- Synonyms: Insufficient & inadequate | increase & enlarge | nice & fine
- Stemming:

  Dutch-stemmers have high error (e.g. Porter, 1980)
  Python-module `Pattern` (De Smedt en Daelemans, 2012) and custom list
  Verbs: 20,058 | **Nouns/adjectives**: frequency $> 2,000$

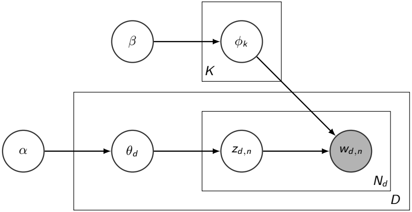
Create vocabulary for topic model

- 2,153 unique stemmed words | minimum frequency: $1,500$
- **No** verbs, sentiment words, very specific words

DeNederlandscheBank
EUROSYSTEEM

Model: Latent Dirichlet Allocation - intuition

- Pick the overall theme of articles by randomly giving them a distribution over topics, i.e.: choose $\theta_d \sim \text{Dir}(\alpha)$, where $d \in \{1, \ldots, D\}$.
- Pick the word distribution for each topic by giving them a distribution over words, i.e.: choose $\phi_k \sim \text{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$.
- For each of the words, $w_{d,n}$ where $n \in \{1, \ldots, N_d\}$, and $d \in \{1, \ldots, D\}$
  - From the topic distribution chosen in 1, randomly pick one topic, i.e.: choose a topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$
  - Given that topic, randomly choose a word from this topic, i.e.: choose a word $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$,



Latent Dirichlet Allocation using plate notation

Model: Latent Dirichlet Allocation - Bayesian inference

MCMC: Collapsed Gibbs Sampler

Bayesian inference to infer **document-topic** distribution (θ_d) and **topic-word** distribution (ϕ_k) via collapsed Gibbs sampler, approximation by Steyvers and Griffiths (2007)

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

Layered topic model:

10,000 iterations (burnin': 1,000) for plain-vanilla and layered topic model

Time-varying topic model:

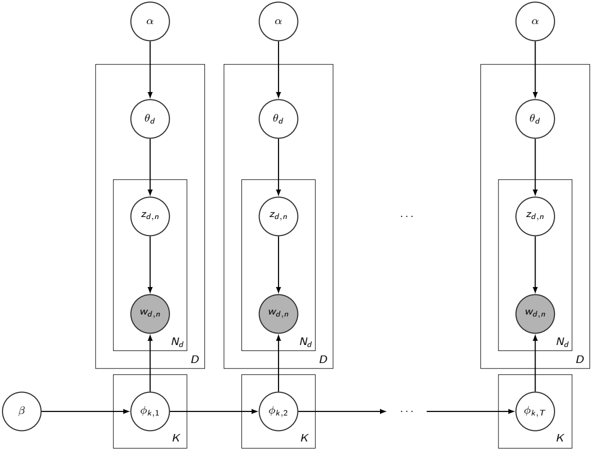
5,000 iterations (burnin': 1,000) for 1^{st} time slice

1,000 iterations for 2^{nd} until last time slice

$$\hat{\phi}_{ij} = \frac{\mathbf{C}^{WT}_{ij} + \beta}{\sum_{k=1}^{W} \mathbf{C}^{WT}_{kj} + W\beta}, \qquad \hat{\theta}_{dj} = \frac{\mathbf{C}^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} \mathbf{C}^{DT}_{dk} + T\alpha}$$

Model: Time-varying Latent Dirichlet Allocation - intuition

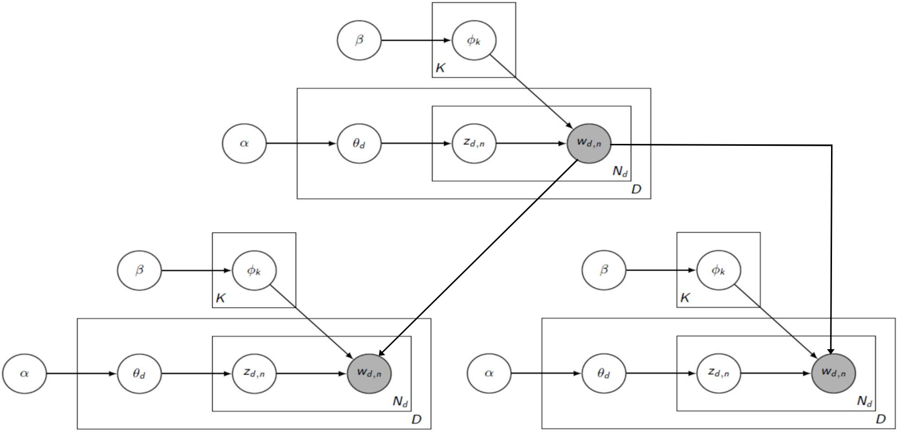
Example Time-Varying LDA using plate notation | **1,...T time slices; 15 year window**



DeNederlandscheBank
EUROSYSTEEM

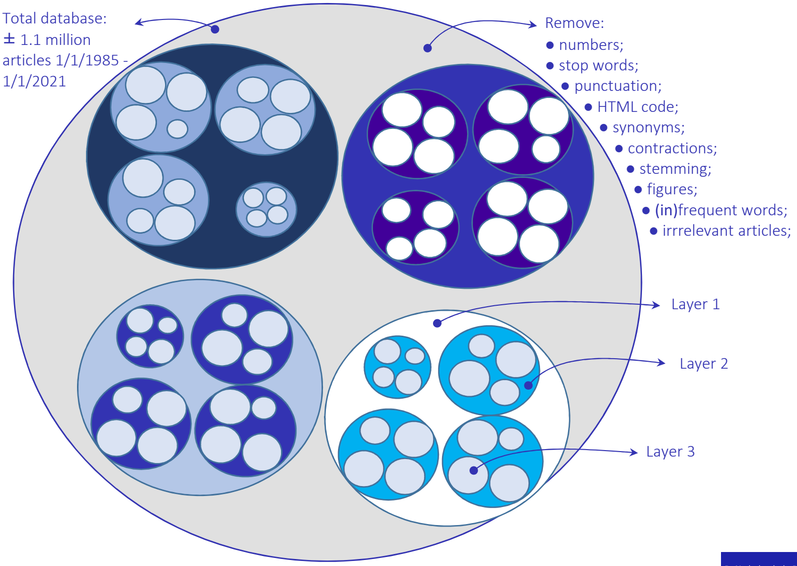
Model: Layered Latent Dirichlet Allocation - intuition

Example Layered LDA using plate notation | **Layer 1: 1 topic** | **Layer 2: 2 topics**

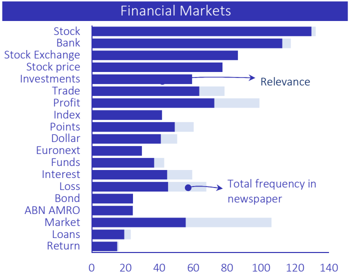


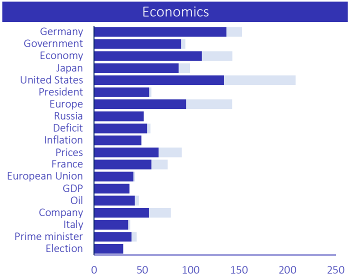
DeNederlandscheBank
EUROSYSTEEM

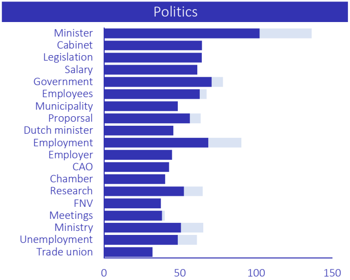
Model: Time-varying layered Latent Dirichlet Allocation



Model: Topics in Layer 1







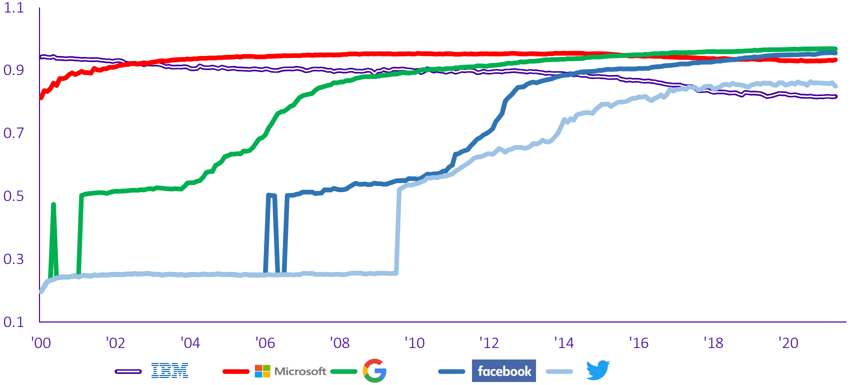
Model: Topics in Layer 1 through 3

Layer 1	Financial Markets	Firms	Economics	Politics
Layer 2	Markets	Infrastructure	Elections	Parliament
Layer3	1. Raw materials	17. Chemical & pharma	33. Elections	49. Politics
	2. Exchanges	18. Indices	34. Easten Europe	50. Budgettary policy
	3. International	19. Mobility	35. Africa & Asia	51. Cabinets
	4. Monetary policy	20. Company results	36. United States	52. Ministries
Layer 2	Financials	**Multinationals**	Indicators	National
Layer3	5. Corporate finance	21. Telecom	37. International	53. Justice
	6. Financials (international)	22. Customers	38. Europe	54. Pensions & health care
	7. Banks (national)	**23. Big-tech**	39. Trading partners	55. Supervision
	8. Insurance companies	24. Media	40. Fiscal policy	56. Education & research
Layer 2	News	Construction & Energy	Raw Materials	Lower Government
Layer3	9. Emissions	25. Construction	41. Asia	57. Housing
	10. Take-overs	26. Logistics	42. Oil & gas	58. Public-private
	11. Trade	27. Energy	43. Conflicts	59. Agriculture & fishery
	12. Insurers	28. Industry	44. Emerging economies	60. Transport
Layer 2	Fin. Indices	Demography	European Union	Social Partners
Layer3	13. Stock market	29. Retail	45. Germany	61. Wage negotiations
	14. Euronext	30. Bankruptcies	46. European Union	62. Labor market
	15. Analists	31. Listed	47. Italy & Spain	63. Entrepeneurs
	16. Results	32. International	48. France	64. Social security & pensions

DeNederlandscheBank
EUROSYSTEEM

Model: Time variation

Layer 1	Firms
Layer 2	Multinationals
Layer3	23. Big-tech



DeNederlandscheBank EUROSYSTEEM

Sentiment: Extraction from newspaper articles

Sentiment Lexicon

- No good financial-economic sentiment lexicon in Dutch language
- Base: Loughran and McDonald (2011) lexicon & Google Translate, DeepL
- Extend with: customized list based on word frequencies newspaper
- Check for double negations i.e.: deficit decreased, unemployment decreased

  Total: 1,532 words | Positive:468 | Negative: 1,063

Sentiment Calculation

- Calculate sentiment per article (see e.g. Tetlock, 2007 and Shapiro et al., 2020)
- Weighted sentiment-score (WSS):

$$\frac{(\#\text{positive words} - \#\text{negative words})/(\#\text{sentiment words})}{\#\text{words in article}}$$

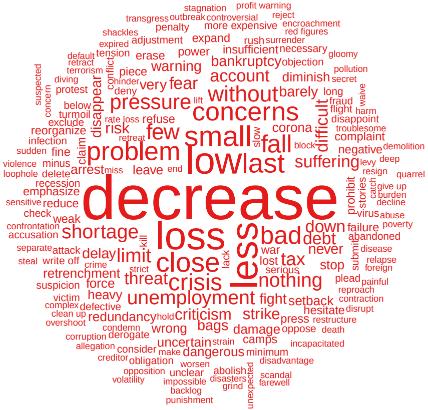
- Base: 6-month moving average of WSS. Many alternatives, see e.g. Algaba et al., 2020

DeNederlandscheBank EUROSYSTEEM

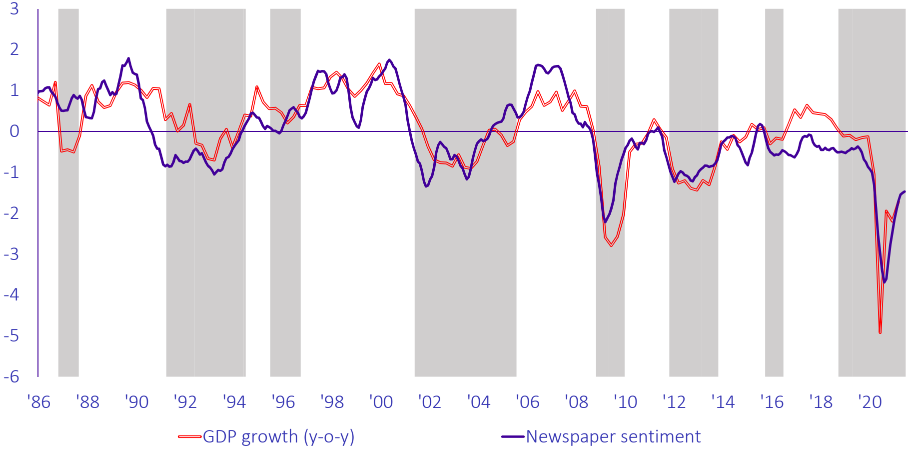
Sentiment: frequency sentiment words

Positive words

Negative words

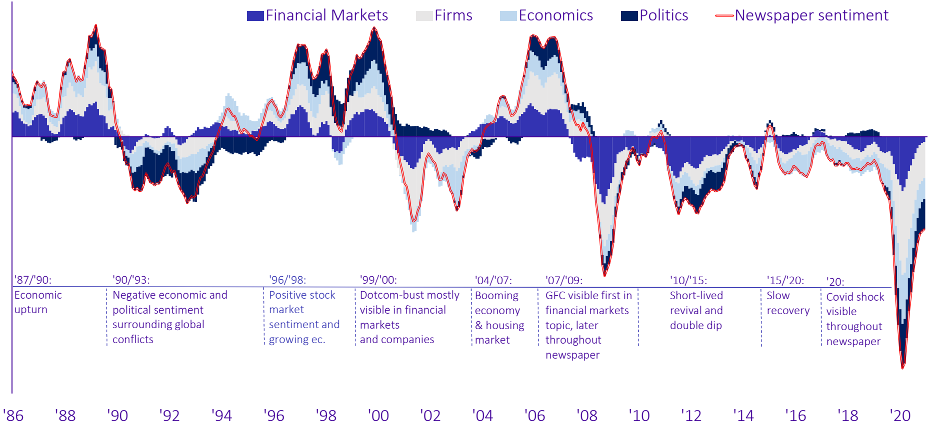


Sentiment: newspaper sentiment and GDP growth



DeNederlandscheBank
EUROSYSTEEM

Sentiment: newspaper sentiment per topic



Nowcasting horse-race forecasting quarterly GDP growth (q-o-q)

- Pseudo real-time exercise: using state-of the art nowcasting model **with** and **without** textual data | Estimation starts 1996M1 | Evaluation: 2003Q1–2020Q3
- Workhorse model: **dynamic factor model** (for technique see e.g. Jansen en de Winter, 2018; Hindrayanto, Koopman and de Winter, 2016; Jin, Jansen en de Winter, 2016)

Timing of forecast exercise for third quarter

Nr.	Forecast type	Month	Forecast made in middle of
1	Two-quarter ahead	1	January
2		2	February
3		3	March
4	One-quarter ahead	1	April
5		2	May
6		3	June
7	Nowcast	1	July
8		2	August
9		3	September
10	Backcast	1	October
11		2	November

DeNederlandscheBank
EUROSYSTEEM

Main outcome

1. TaTVL topic model increases the forecasting accuracy of the dynamic factor model
2. Particularly when it pertains to current or previous quarter
3. An much more so in crisis-times | Fin. crisis: 2009Q1 | COVID-crisis: 2020Q2-2020Q3

		backcast		nowcast			forecast		
	structure	m2	m1	m3	m2	m1	m3	m2	m1
Total sample									
Time varying layered topic model	4 X 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Baseline DFM		1.22	1.11	1.10	1.01	1.00	1.01	1.00	1.00
Excluding financial crisis and COVID-crisis									
Time varying layered topic model	4 X 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Baseline DFM		1.05	1.04	1.04	1.03	1.03	1.02	0.99	0.99

RMSFE Time varying layered topic model < alternative model

RMSFE Time varying layered topic model > alternative model

DeNederlandscheBank EUROSYSTEEM

Nowcasting: formal testing of forecasting accuracy

Layers and time-variation in topics

1. Layering and time-variation generally add little to forecasting accuracy in normal times
2. But sometimes add somewhat in times of crisis
3. Advantage: better "story-line" which can be important (especially during crisis)

		backcast		nowcast			forecast		
	structure	m2	m1	m3	m2	m1	m3	m2	m1
Total sample									
Topic model	16	0.98	1.02	1.09	1.07	1.06	1.00	1.00	1.00
Time varying topic model	16	0.98	1.01	1.05	1.05	1.03	1.00	1.00	1.00
Layered topic model	4 X 4	1.01	1.03	1.04	1.03	1.02	1.00	1.00	1.00
Time varying layered topic model	4 X 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Baseline DFM		1.22	1.11	1.10	1.01	1.00	1.01	1.00	1.00
Excluding financial crisis and COVID-crisis									
Topic model	16	1.01	1.00	1.00	1.00	1.01	1.02	1.02	1.01
Time varying topic model	16	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00
Layered topic model	4 X 4	1.01	1.01	1.00	0.99	1.00	1.00	0.99	0.99
Time varying layered topic model	4 X 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Baseline DFM		1.05	1.04	1.04	1.03	1.03	1.02	0.99	0.99

RMSFE Time varying layered topic model < alternative model

RMSFE Time varying layered topic model > alternative model

DeNederlandscheBank EUROSYSTEEM

Wrap up

Main takeaways

1. We introduce an extended tone-adjusted topic model with **layering** and **time-variation**
2. **Newspaper sentiment increases forecasting accuracy** nowcasting model
3. Layering and time-variation of topic model is mainly helpful for **interpretation**
4. Develop lexicon and cleaning strategies for Dutch financial-economic texts

DeNederlandscheBank
EUROSYSTEEM

Thank you for your attention!

J.M.de.Winter@dnb.nl