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

Disclaimer: This work should not be reported as representing the views of De Nederlandsche Bank. The views expressed are those of the authors and do not necessarily reflect those of de Nederlandsche Bank.
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)
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)
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

GDP growth (y-o-y) vs. Newspaper sentiment

Jasper de Winter
Nowcasting GDP TaTVL Topic Model
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 1st 1985 - December 31st 2020
- ± 1.1 million full-text articles
Data: Financieele Dagblad - cleanup articles

Cleaning database, X 1,000 articles

Tag cleaning:
- fashion, radio& television, personal profiles, service pages

Closing/opening prices, agendas upcoming events, summaries of articles

URL cleaning:
- personal finance, human interest, English pages, Duplicate articles

Reduction of 48 percent

1-line articles, infographics, English pages

26

438

18

25

3

538

Reduction of 48 percent
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 Daeleman, 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
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 ($\theta_d$) and **topic-word** distribution ($\phi_k$) via collapsed Gibbs sampler, approximation by Steyvers and Griffiths (2007)

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

**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 1st time slice
1,000 iterations for 2nd until last time slice

$$\hat{\phi}_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta}, \quad \hat{\theta}_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$
Example Time-Varying LDA using plate notation | $1, \ldots, T$ time slices; 15 year window
Model: Layered Latent Dirichlet Allocation - intuition

Example Layered LDA using plate notation | **Layer 1: 1 topic** | **Layer 2: 2 topics**
Model: Time-varying layered Latent Dirichlet Allocation

Total database: ± 1.1 million articles 1/1/1985 – 1/1/2021

Remove:
- numbers;
- stop words;
- punctuation;
- HTML code;
- synonyms;
- contractions;
- stemming;
- figures;
- (in)frequent words;
- irrelevant articles;

Layer 1
Layer 2
Layer 3
### Model: Topics in Layer 1 through 3

<table>
<thead>
<tr>
<th>Layer 1</th>
<th>Financial Markets</th>
<th>Firms</th>
<th>Economics</th>
<th>Politics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 2</td>
<td>Markets</td>
<td>Infrastructure</td>
<td>Elections</td>
<td>Parliament</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>Financials</th>
<th>Multinationals</th>
<th>Indicators</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8. Insurance companies</td>
<td>24. Media</td>
<td>40. Fiscal policy</td>
<td>56. Education &amp; research</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>News</th>
<th>Construction &amp; Energy</th>
<th>Raw Materials</th>
<th>Lower Government</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12. Insurers</td>
<td>28. Industry</td>
<td>44. Emerging economies</td>
<td>60. Transport</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>Fin. Indices</th>
<th>Demography</th>
<th>European Union</th>
<th>Social Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15. Analists</td>
<td>31. Listed</td>
<td>47. Italy &amp; Spain</td>
<td>63. Entrepreneurs</td>
</tr>
<tr>
<td></td>
<td>16. Results</td>
<td>32. International</td>
<td>48. France</td>
<td>64. Social security &amp; pensions</td>
</tr>
</tbody>
</table>
Layer 1 Firms
Layer 2 Multinationals
Layer 3 23. Big-tech
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):

  \[
  \frac{(#\text{positive words} - #\text{negative words})}{(#\text{sentiment words})}
  \]

  \[
  \frac{\#\text{words in article}}{}
  \]

- Base: 6-month moving average of WSS. Many alternatives, see e.g. Algaba et al., 2020
Sentiment: newspaper sentiment and GDP growth

GDP growth (y-o-y) and newspaper sentiment over the years.
Sentiment: newspaper sentiment per topic

- **'87/'90:** Economic upturn
- **'86 '88 '90 '92 '94 '96 '98 '00 '02 '04 '06 '08 '10 '12 '14 '16 '18 '20:** Growing economic
- **'90/'93:** Negative economic and political sentiment surrounding global conflicts
- **'96/'98:** Positive stock market sentiment and growing economy
- **'99/'00:** Dotcom-bust mostly visible in financial markets and companies
- **'04/'07:** Booming economy & housing market
- **'07/'09:** GFC visible first in financial markets topic, later throughout newspaper
- **'10/'15:** Short-lived revival and double dip
- **'15/'20:** Slow recovery
- **'20:** Covid shock visible throughout newspaper
Nowcasting: testing of forecasting accuracy

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

- 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)

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

Timing of forecast exercise for third quarter
**Main outcome**

- **TaTVL topic model increases the forecasting accuracy of the dynamic factor model**
- Particularly when it pertains to current or previous quarter

<table>
<thead>
<tr>
<th></th>
<th>backcast</th>
<th>nowcast</th>
<th>forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time varying layered topic model</td>
<td>4 X 4</td>
<td>1.00 1.00</td>
<td>1.00 1.00 1.00</td>
</tr>
<tr>
<td>Baseline DFM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.22 1.11</td>
<td>1.10 1.01 1.00</td>
</tr>
<tr>
<td><strong>Excluding financial crisis and COVID-crisis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**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)**

<table>
<thead>
<tr>
<th>Structure</th>
<th>Backcast</th>
<th>Nowcast</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic model</td>
<td>16</td>
<td>m2 0.98</td>
<td>m3 1.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>m1 1.02</td>
<td>m2 1.07</td>
</tr>
<tr>
<td>Time varying topic</td>
<td>16</td>
<td>m3 1.05</td>
<td>m3 1.05</td>
</tr>
<tr>
<td>Layered topic</td>
<td>4 X 4</td>
<td>m3 1.04</td>
<td>m3 1.04</td>
</tr>
<tr>
<td>Time varying</td>
<td>4 X 4</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Baseline DFM</td>
<td>16</td>
<td>1.22</td>
<td>1.11</td>
</tr>
</tbody>
</table>

### Excluding financial crisis and COVID-crisis

<table>
<thead>
<tr>
<th>Structure</th>
<th>Backcast</th>
<th>Nowcast</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic model</td>
<td>16</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>Time varying topic</td>
<td>16</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Layered topic</td>
<td>4 X 4</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>Time varying layered topic</td>
<td>4 X 4</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Baseline DFM</td>
<td>16</td>
<td>1.05</td>
<td>1.04</td>
</tr>
</tbody>
</table>

- **Green**: RMSFE Time varying layered topic model < alternative model
- **Red**: RMSFE Time varying layered topic model > alternative model
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
Thank you for your attention!

J.M.de.Winter@dnb.nl