THE SENTOMETRICS APPROACH FOR TEXTUAL SENTIMENT BASED PREDICTION

Samuel Borms
UniNE, VUB & *Sentometrics*

European Commission JRC Ispra, May 2019

CONTEXT

Optimal economic decision making involves making **accurate predictions**

Predictions are a function of the **information** available

But which information set?

CONTEXT

But which information set?

Numeric data:

- Prices, earnings, macro-data, …, at various frequencies
- EUROSTAT, COMPUSTAT, DATASTREAM, FACTSET, …

Qualitative data:

- Texts:
  - Reports, news articles, blogs, communications, …
  - LexisNexis, Factiva, Twitter, …
- Audio
- Video

CONTEXT

The information gold of the 21st Century?

TEXTS!

NLP tools to extract information from textual data.

Often we are talking about **topic** extraction and/or **sentiment** analysis.

Around **80%** of usable business information from unstructured (mostly textual) form

CONTRIBUTION TO THE WAVE

"Modern" academic view

Text as Data

Matthew Gentzkow

Bryan T. Kelly

Matt Taddy

JOURNAL OF ECONOMIC LITERATURE (FORTHCOMING)

Words are the New Numbers: A Newsy Coincident Index of the Business Cycle

Leif Anders Thorsrud

To cite this article: Leif Anders Thorsrud (2018): Words are the New Numbers: A Newsy Coincident Index of the Business Cycle, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2018.1506344

CONTRIBUTION TO THE WAVE

Exploiting longitudinal dependence

Inference about latent variables

Dealing with high dimensions

Transforming text to data

sentometrics

Transforming text to actionable insights

CONTRIBUTION TO THE WAVE

sentometrics

ANDRES

DAVE

JULIETTE

KEVEN

KRIS

LAURA

MUZAFER

SAM

SELECTED PROJECTS

Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values

David Ardia [a,b], Keven Bluteau [a,c,*], Kris Boudt [c,d,e]

When Econometrics Meets Sentiment...

Andres Algaba[a], David Ardia[b,c], Keven Bluteau[c,a], Samuel Borms[c,a,*], Kris Boudt[d,a,e]

The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment

David Ardia
University of Neuchâtel
HEC Montréal

Keven Bluteau
University of Neuchâtel
Vrije Universiteit Brussel

Samuel Borms
University of Neuchâtel
Vrije Universiteit Brussel

Kris Boudt
Ghent University
Vrije Universiteit Brussel
Vrije Universiteit Amsterdam

OUTLINE

1. What makes Sentometrics interesting from a methodological viewpoint?
2. Forecasting US economic growth
3. The R package sentometrics

OUTLINE

1. **What makes Sentometrics interesting from a methodological viewpoint?**
2. Forecasting US economic growth
3. The R package sentometrics

TEXT2DATA

Make use of data providers…

Or DIY!

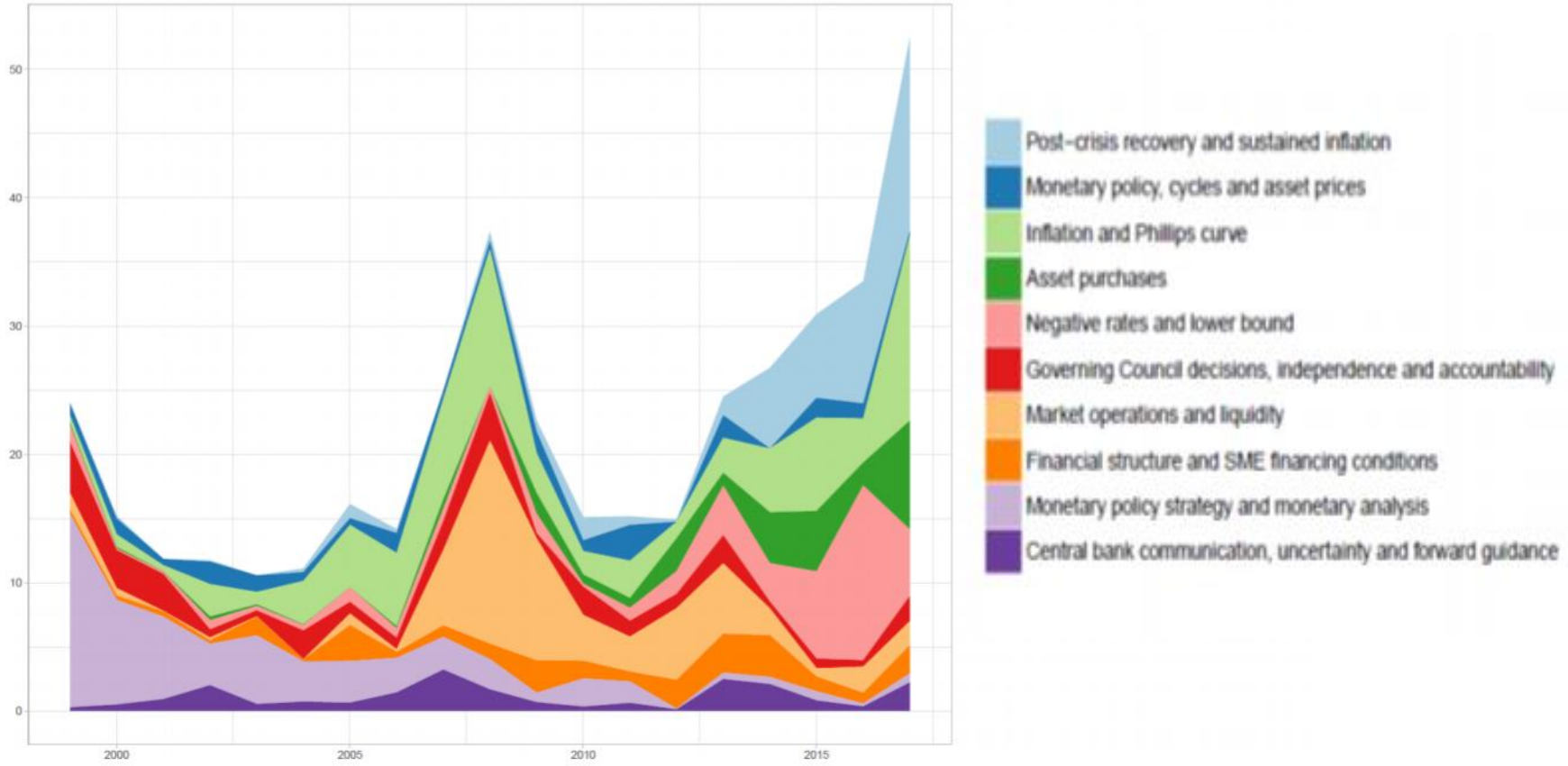
CHALLENGES

Number and type of documents

CHALLENGES

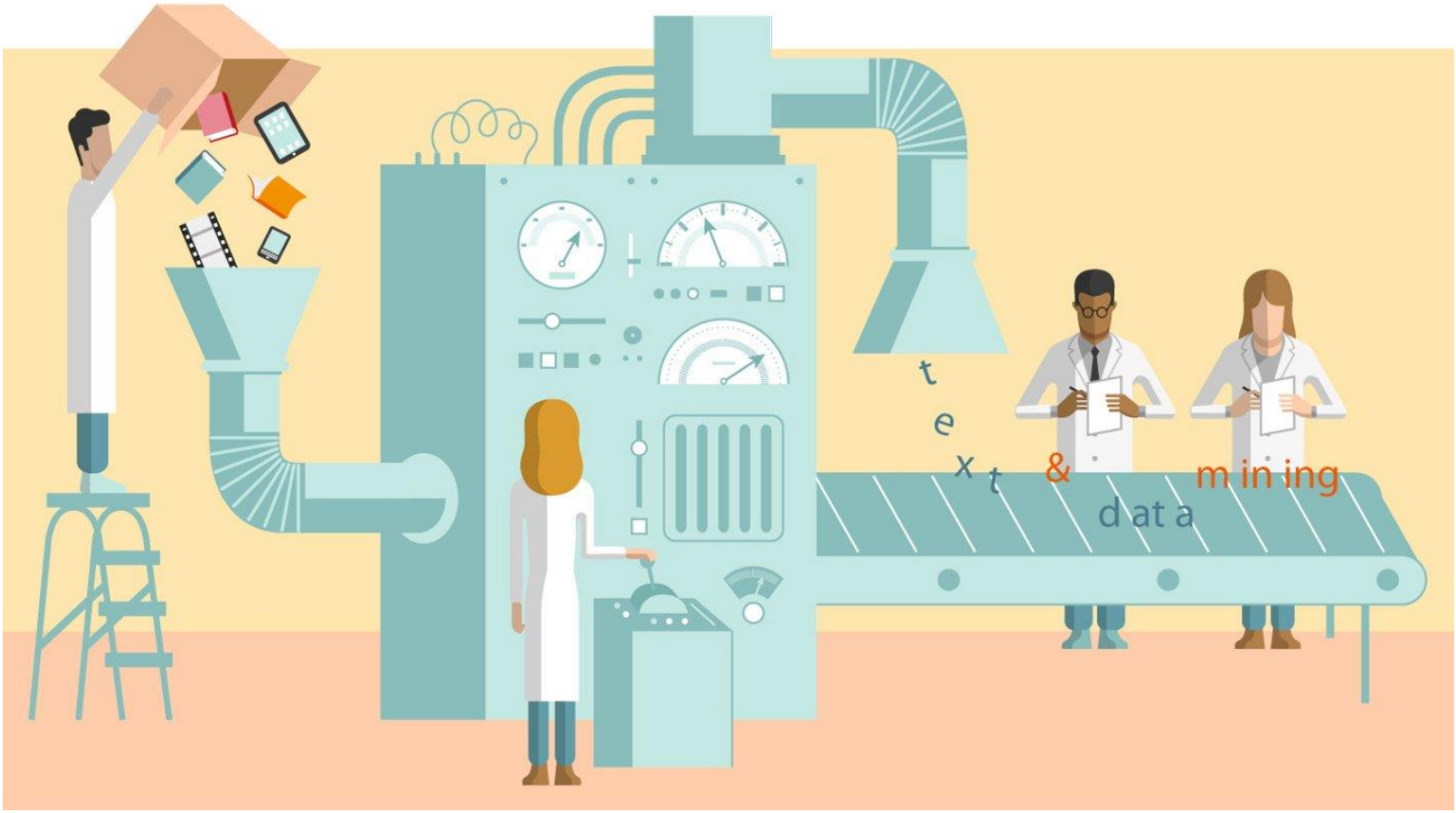
Number of variables that can be associated to a text

Figure 2: Speeches by ECB Executive Board members on monetary policy and inflation and their decomposition in topics (number of speeches per annum)



CHALLENGES

Number of sentiment implementations

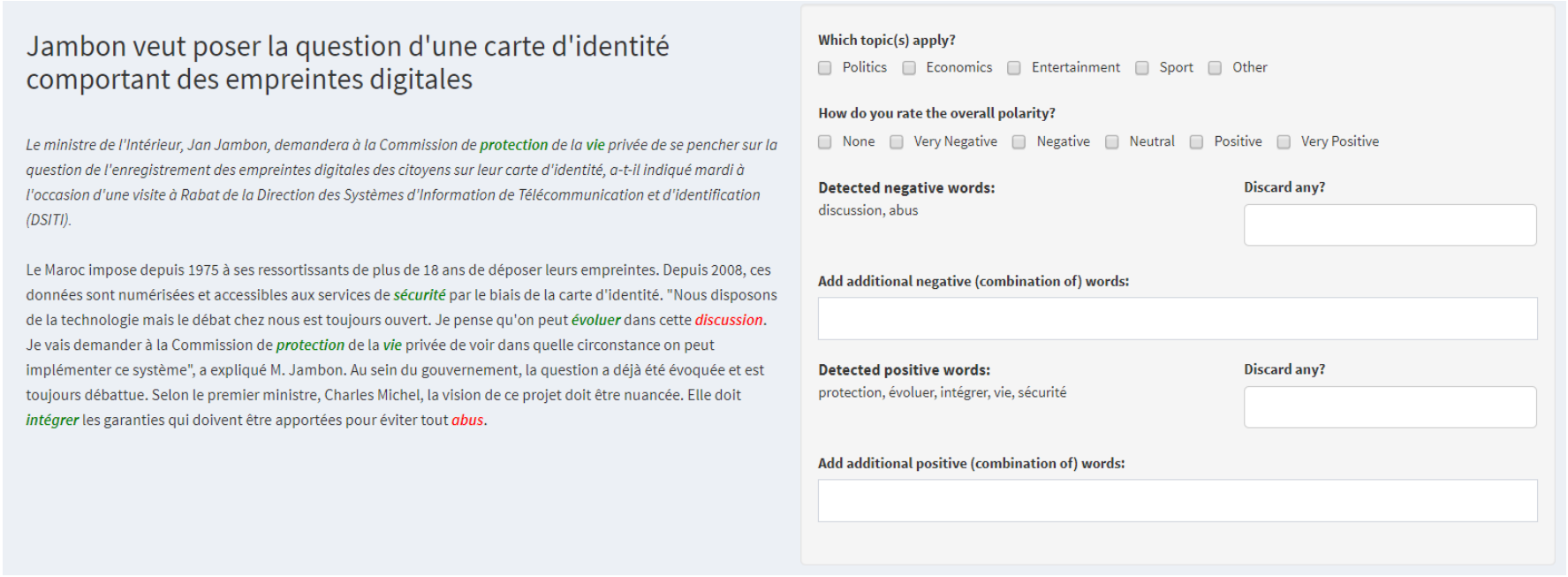


CHALLENGES

Number of sentiment implementations

Why *not* machine learning? Lexicon–based (bag–of–words) calculation is simple, effective, and widely used

Annotation needed to create **domain–specific** lexicons



CHALLENGES

DIMENSIONALITY

Number of documents

x

Number of variables that can be associated to a text

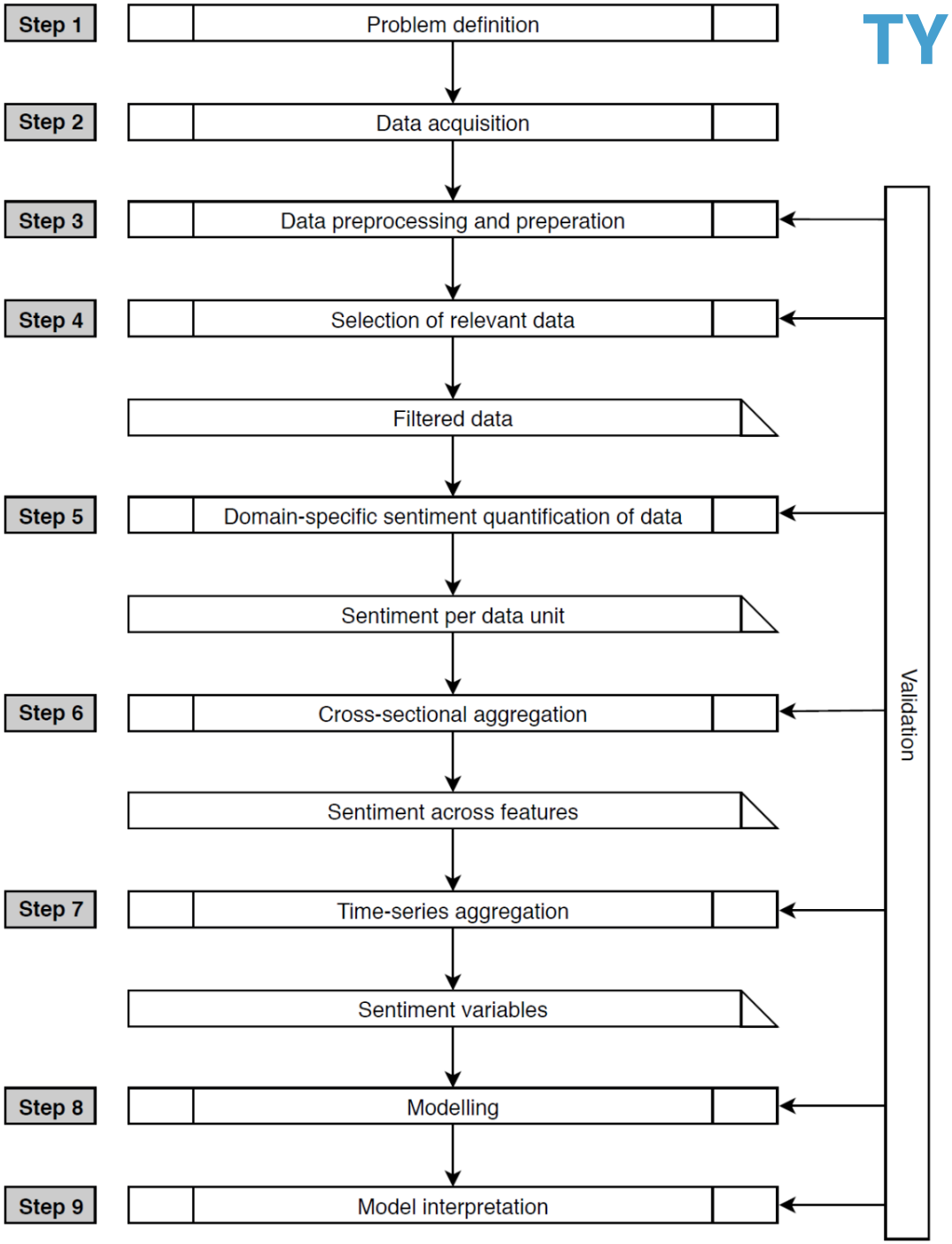
x

Number of implementations

HUGE!

Risk of data snooping
Amplified by difficulty in replication and sometimes lack of theory

TYPICAL WORKFLOW



DATA

SENTIMENT

MODEL

INTERPRETATION

OUTLINE

1. What makes Sentometrics interesting from a methodological viewpoint?
2. **Forecasting US economic growth**
3. The R package sentometrics

MOTIVATION

Traditional macro, financial, and survey–based indices are often used to forecast economic growth

Drawbacks to these traditional variables:

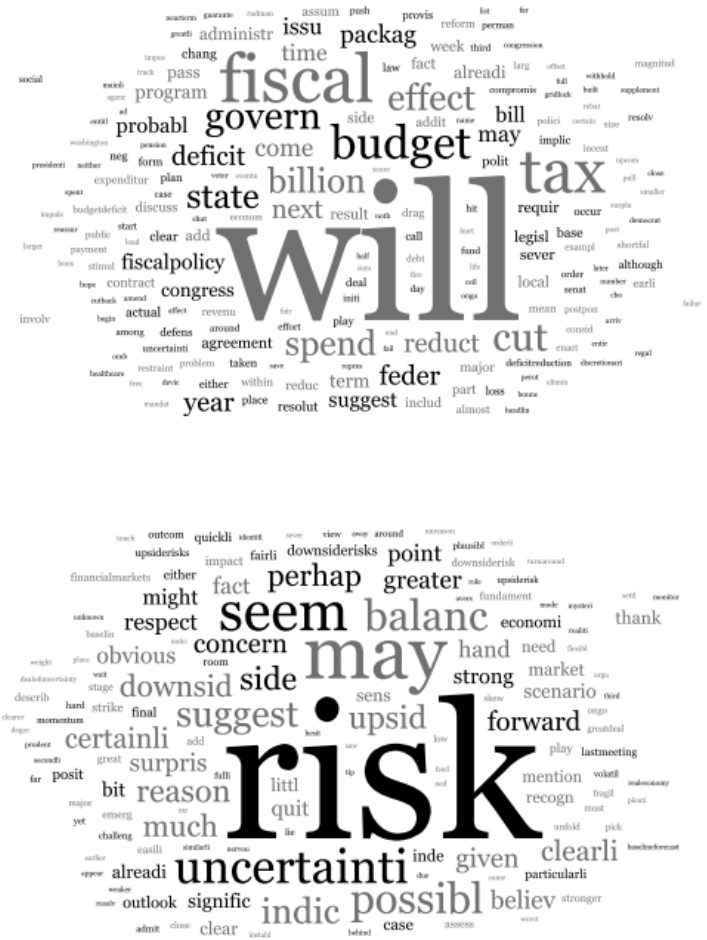
- Release lags
- Cost of acquisition
- Static
- Measurement errors

The traditional approach also omits a large source of data… texts!

MOTIVATION

Topics matter

Overall sentiment measure should be a weighted combination of several topic–based sentiment indices

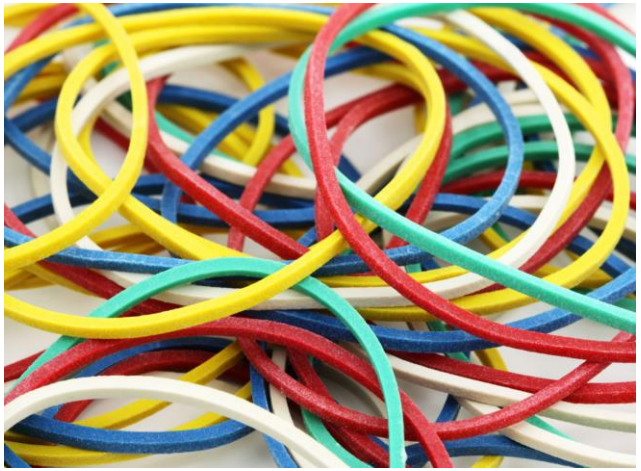


Taken from Hansen et al. (2018)

CONTRIBUTION

Can we improve growth forecasts with traditional variables by adding **text–based variables**?

Propose an elastic approach to optimize aggregation from thousand of sentiment indices



<u>Yes</u>, texts complement traditional variables, typically for longer–term horizons (above six months)

STEPS

1. **Classify texts by topic** and choose a subset of topics to select relevant texts
2. **Compute** for each text n of corpus at time t the **sentiment** using $\boldsymbol{L}$ **methods**
3. For each corpus at time t and method l, obtain $\boldsymbol{K}$ **topic–based sentiments**
4. For each topic k and method l, obtain $\boldsymbol{B}$ **time series aggregated values**
5. **Optimize combination** of variables
6. **Forecast** (out–of–sample)
7. Evaluate forecasting power
8. Measure attribution

NEWS DATA

All English articles from major US newspapers in the **LexisNexis** database with reference to the US

Dates range from September, 1994 to March, 2016

Filters:

- Geographic location (US only)
- Filter out non–economy related topics
- Need to have a major reference to the topic to be included (relevance score equal or above 85 on LexisNexis)

STEP 1 – CLASSIFY TEXT BY TOPICS

But what if no topics?

- **Latent topic models** (e.g., Latent Dirichlet Allocation), unsupervised inference via a generative statistical model

- Machine learning **classification** (support vector machines, neural network, etc.)

- **Keywords**: set of keywords define a topic

STEP 2 – SENTIMENT CALCULATION

Lexicon approach + Valence shifting

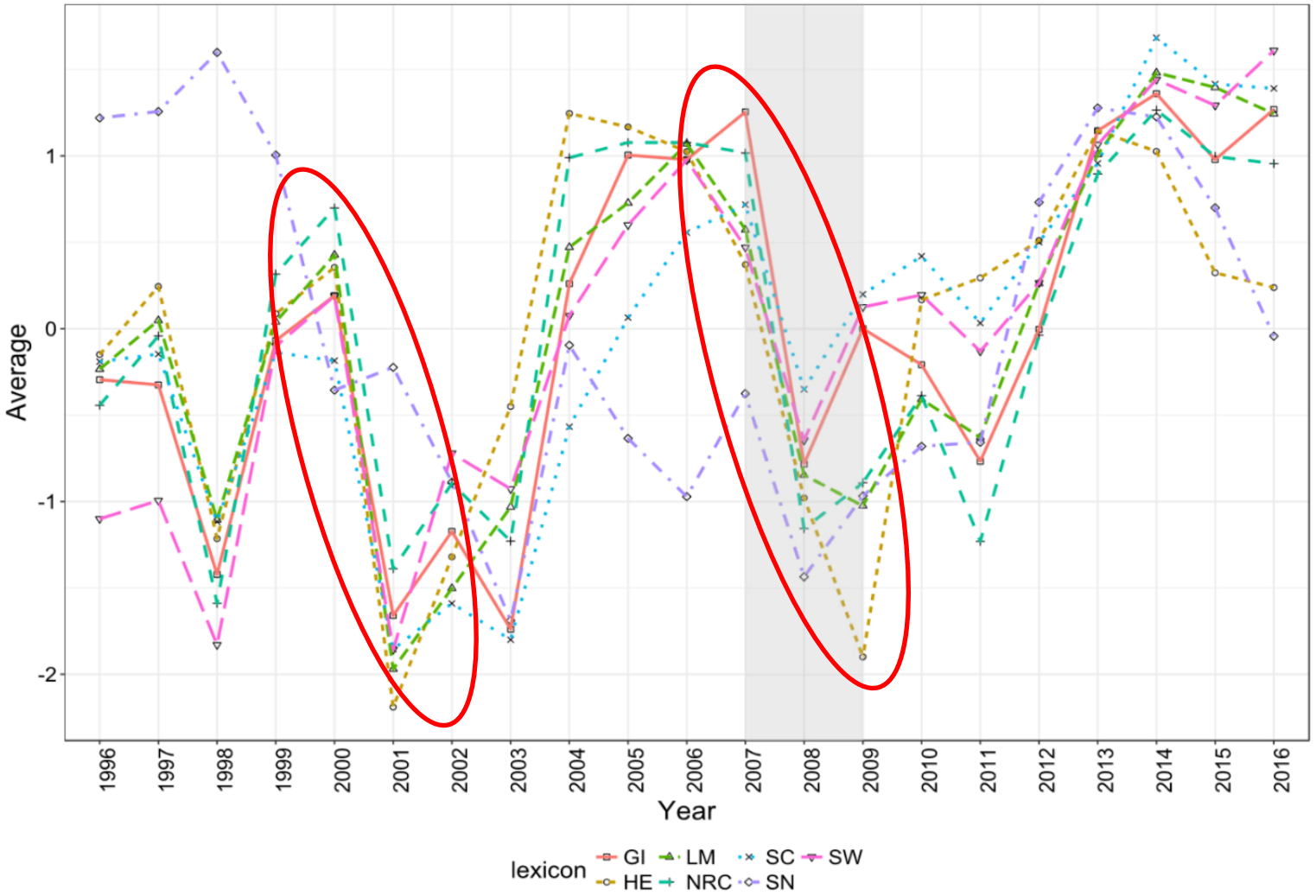
- **Harvard General Inquirer**
- **Henry** (2008)
- **Loughran & McDonald** (2011)
- **SentiWordNet** of Baccianella et al. (2016)
- **SenticNet** of Cambria et al. (2016)
- **SO–CAL** of Taboada et al. (2011)
- **NRC** of Mohammad and Turney (2010)

Net sentiment measure:

$$s_{n,t,l} \equiv \frac{N^{+}_{n,t,l} - N^{-}_{n,t,l}}{N^{+}_{n,t,l} + N^{-}_{n,t,l} + N^{0}_{n,t,l}}$$

SENTIMENT INDICES

Yearly average of the individual news article sentiments for each lexicon



STEP 3 – TOPIC AGGREGATION

We have for each day t, L vectors $\boldsymbol{s}_{t,l} \equiv (s_{1,t,l}, \ldots, s_{N_t,t,l})'$

Text–to–topic aggregation matrix $\boldsymbol{W}_t$ $(K\ x\ N_t)$, equally weighted (but could also be built from LDA probabilities)

L vectors that capture the **daily sentiment for each of the $\boldsymbol{K}$ topics.** They are generated by taking $\boldsymbol{W}_t \boldsymbol{s}_{t,l}$ $(K\ x\ 1)$

STEP 4 – TIME-SERIES AGGREGATION

Maximum **time aggregation** lag τ ($0 \leq \tau < T$), stack the vectors column–by–column into $K \times (\tau + 1)$ matrices

$$\boldsymbol{V}_{t,l} \equiv \begin{bmatrix} | & & | \\ \boldsymbol{W}_{t-\tau}\boldsymbol{s}_{t-\tau,l} & \cdots & \boldsymbol{W}_{t}\boldsymbol{s}_{t,l} \\ | & & | \end{bmatrix}$$

We do this for $l = 1, \ldots, L$, and stack matrices row–by–row

$$\boldsymbol{V}_{t} \equiv \begin{bmatrix} \boldsymbol{V}_{t,1} \\ \vdots \\ \boldsymbol{V}_{t,L} \end{bmatrix}$$

Given $\boldsymbol{V}_t$ and a suitable time aggregation (e.g., Beta weighting) matrix $\boldsymbol{B}$ of size $(\tau + 1) \times B$, we then construct the **final vector of size $\boldsymbol{LKB} \times \boldsymbol{1}$ of textual sentiment predictors**

SENTIMENT INDICES

STEP 5 – MODELS

Calibration:

- **Aggregate** all textual sentiment indices optimally **given a variable of interest**

- We use a **penalized least squares criterion** to regularize the estimation of the high-dimensional parameters

- Requires calibration of the elastic net penalty parameters, where we follow Zou et al. (2007) and minimize a BIC–like criterion

STEP 5 – MODELS

Dependent variable is h–month log–**growth** of the **US industrial production**:

$$IP_t(h): \log(IP_{t+h}) - \log(IP_t)$$

Horizons: $h = 1, 3, 6, 9, 12$ months

Benchmark against dataset of "**traditional variables**":

- Vintage dataset of 105-128 (depending on the period) **variables** from the **FRED–MD** database (McCracken, 2016)
- 16 **financial metrics** such as dividend ratios, long/short term yields, stock variances (Goyal and Welch, 2008)
- CBE's forward–looking volatility index (**VIX**)
- Media–attention **EPU** index (Baker & al., 2016)
- Six survey–based **Conference Board indices** (CB)

STEP 5 – MODELS

We study the following models:

Models $\mathbf{M_{1a}}$ (without sentiment $\boldsymbol{s}$), $\mathbf{M_{1b}}$

$$IP_t(h) = cst + \alpha IP_{t-h}(h) + \boldsymbol{\gamma}'\boldsymbol{x}_t + \boldsymbol{\beta}'\boldsymbol{s}_t + \epsilon_t$$

Models $\mathbf{M_{2a}}$ (without sentiment $\boldsymbol{s}$), $\mathbf{M_{2b}}$

$$IP_t(h) = cst + \alpha IP_{t-h}(h) + \boldsymbol{\gamma}'\boldsymbol{f}_t + \boldsymbol{\beta}'\boldsymbol{s}_t + \epsilon_t$$

$\boldsymbol{x}$: Merged macro, financial, survey dataset

$\boldsymbol{f}$: factors using IC_{p1} information criteria of Bai and Ng (2002) computed with the merged dataset

$\boldsymbol{s}$: 4,928 sentiment values (LKB = 7 x 44 x 16)

STEP 6 – FORECASTING

Rolling forward **out–of–sample** forecasting

The full out-of-sample period is from January 2001 (January 2003 for $h = 12$) to December 2016 (192 observations for $h = 1$ and 168 for $h = 12$) .

Separation into **three periods**: pre–crisis, crisis and post–crisis

STEP 7 – FORECAST EVALUATION

		RMSFE				MAFE			
Period	h	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$
Full sample	1	0.68	0.70	0.64	0.70	0.49	0.49	0.45	0.49
	3	1.52	1.54	1.59	1.52	0.96	1.01	1.02	1.01
	6	4.86	3.93	5.01	3.14	2.36	2.35	2.85	2.14
	9	7.01	4.95	8.36	4.58	3.71	3.28	4.89	3.19
	12	6.39	5.19	8.69	5.14	4.25	3.41	6.03	3.32
Pre–crisis	1	0.55	0.57	0.56	0.56	0.43	0.42	0.43	0.44
	3	0.99	0.93	1.21	0.93	0.72	0.70	0.87	0.70
	6	1.67	1.65	2.62	1.62	1.31	1.36	1.80	1.32
	9	2.41	2.42	4.67	2.53	1.96	1.93	3.00	1.98
	12	3.27	2.00	6.07	1.90	2.72	1.67	3.73	1.57
Crisis	1	1.19	1.27	1.08	1.27	0.81	0.87	0.69	0.88
	3	3.20	3.19	3.17	3.04	2.46	2.52	2.31	2.29
	6	11.30	8.54	10.64	6.20	7.63	6.44	7.45	4.99
	9	8.58	7.94	9.92	7.94	6.67	6.20	7.67	6.20
	12	10.43	10.14	9.42	10.12	8.34	7.84	7.49	7.70
Post–crisis	1	0.53	0.50	0.49	0.50	0.42	0.40	0.40	0.41
	3	0.78	0.93	0.89	1.03	0.62	0.74	0.70	0.82
	6	1.72	2.26	2.86	2.32	1.32	1.68	2.05	1.77
	9	8.47	4.93	9.72	4.07	3.93	3.22	5.27	3.00
	12	6.02	3.85	9.81	3.78	3.80	2.98	7.01	2.90

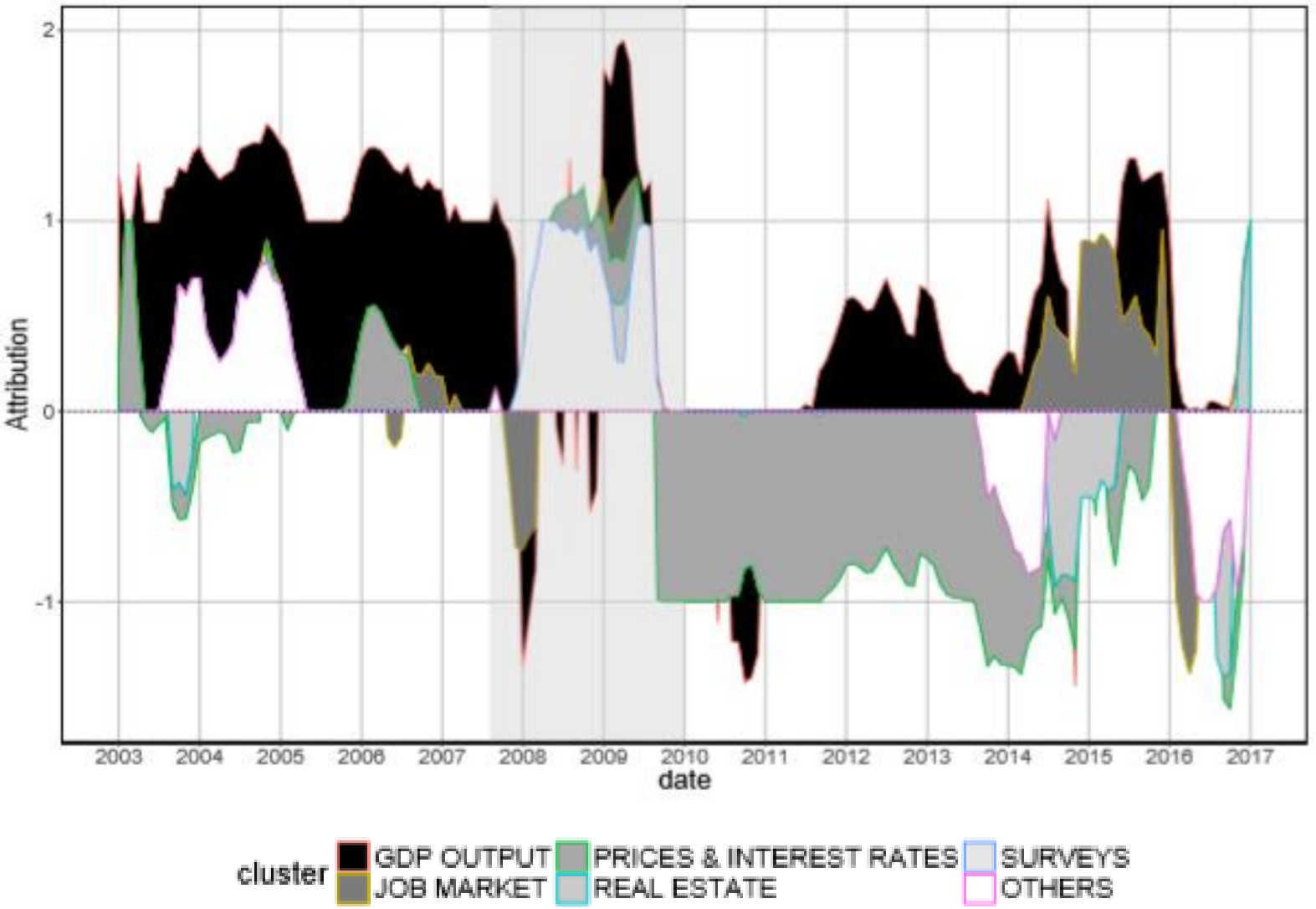
STEP 7 – FORECAST EVALUATION

What if you omit one dimension? Which dimension is most important?

		RMSFE					MAFE				
	h	$\mathcal{M}$	LEX	TOPIC	TIME	ALL	$\mathcal{M}$	LEX	TOPIC	TIME	ALL
$\mathcal{M}_{1b}$	1	0.70	0.69	0.68	0.68	0.64	0.49	0.48	0.48	0.49	0.46
	3	1.54	1.50	1.41	1.52	1.58	1.01	0.98	0.93	0.96	0.99
	6	3.93	4.51	4.52	4.86	5.24	2.35	2.42	2.32	2.36	2.55
	9	4.95	5.91	5.57	7.01	8.37	3.28	3.43	3.28	3.71	4.17
	12	5.19	5.85	6.11	6.39	8.24	3.41	4.01	4.09	4.25	5.02
$\mathcal{M}_{2b}$	1	0.70	0.69	0.68	0.65	0.68	0.49	0.49	0.48	0.46	0.48
	3	1.52	1.53	1.50	1.39	1.31	1.01	1.06	1.07	0.95	0.92
	6	3.14	3.72	3.23	3.62	3.34	2.14	2.39	2.17	2.25	2.20
	9	4.58	5.65	5.36	6.99	6.23	3.19	3.74	3.42	4.16	4.06
	12	5.14	6.79	7.14	8.18	7.82	3.32	4.81	5.04	5.30	5.34

Dark gray: Simplified model statistically superior
Light gray: Full model statistically superior

STEP 8 – ATTRIBUTION ANALYSIS



OUTLINE

1. What makes Sentometrics interesting from a methodological viewpoint?
2. Forecasting US economic growth
3. **The R package sentometrics**

SENTOMETRICS R PACKAGE

Package 'sentometrics'

December 18, 2018

Type Package

Title An Integrated Framework for Textual Sentiment Time Series Aggregation and Prediction

Version 0.5.6

Author David Ardia [aut],
Keven Bluteau [aut],
Samuel Borms [aut, cre],
Kris Boudt [aut]

Maintainer Samuel Borms <samuel.borms@unine.ch>

Description Optimized prediction based on textual sentiment, accounting for the intrinsic challenge that sentiment can be computed and pooled across texts and time in various ways. See Ardia et al. (2018) <doi:10.2139/ssrn.3067734>.

Depends R (>= 3.3.0), data.table

In continuous development (e.g., Google Summer of Code 2019)

FUNCTIONALITIES

Functionality	Functions	Output
1. Corpus management		
(a) Creation	`sento_corpus()`	*sentocorpus*
(b) Manipulation	**quanteda** corpus functions (*e.g.*, `docvars()`, `corpus_sample()`, or `corpus_subset()`)	
(c) Conversion	`to_sentocorpus()`	
(d) Features generation	`add_features()`	

Creation of a **corpus**

Generation of metadata **features** ("topics")

Selection of texts

EXAMPLE

```
R> library("sentometrics")
```

```
R> data("usnews", package = "sentometrics")
```

```
R> uscorpus <- sento_corpus(usnews)
```

```
R> uscorpus <- add_features(uscorpus, keywords = keywords,
+    do.binary = FALSE, do.regex = FALSE)
```

FUNCTIONALITIES

Functionality	Functions	Output
2. Sentiment computation		
(a) Lexicon management	`sento_lexicons()`	*sentolexicons*
(b) Computation	`compute_sentiment()`	*sentiment*
(c) Manipulation	`sentiment_bind()`, `to_sentiment()`	
(d) Summarization	`peakdocs()`	
3. Sentiment aggregation		
(a) Specification	`ctr_agg()`	
(b) Aggregation	`sento_measures()`, `aggregate.sentiment()`	*sentomeasures*
(c) Manipulation	`measures_delete()`, `measures_fill()`, `measures_global()`, `measures_merge()`, `measures_select()`, `measures_subset()`, `diff()`, `scale()`, `get_measures()`	
(d) Visualization	`plot.sentomeasures()`	
(e) Summarization	`summary()`, `peakdates()`, `nobs()`, `nmeasures()`, `get_dimensions()`, `get_dates()`	

Computation of sentiment using lexicons and valence shifters

Aggregation of sentiment into time series

Plotting and summarization of time series

EXAMPLE

```
R> data("list_lexicons", package = "sentometrics")
R> data("list_valence_shifters", package = "sentometrics")
R> lexiconsIn <- c(list_lexicons[c("LM_en", "HENRY_en", "GI_en")],
+    list(NRC = lexicon::hash_sentiment_nrc,
+         HULIU = lexicon::hash_sentiment_huliu,
+         SENTIWORD = lexicon::hash_sentiment_sentiword,
+         JOCKERS = lexicon::hash_sentiment_jockers,
+         SENTICNET = lexicon::hash_sentiment_senticnet,
+         SOCAL = lexicon::hash_sentiment_socal_google))
R> lex <- sento_lexicons(lexiconsIn = lexiconsIn,
+    valenceIn = list_valence_shifters[["en"]])


R> ctrAggPred <- ctr_agg(howWithin = "proportionalPol",
+    howDocs = "equal_weight", howTime = "beta",
+    by = "day", fill = "latest", lag = 270, aBeta = 1:3, bBeta = 1:2)
R> sentMeasPred <- sento_measures(uscorpus, lexicons = lex, ctr = ctrAggPred)
```

EXAMPLE

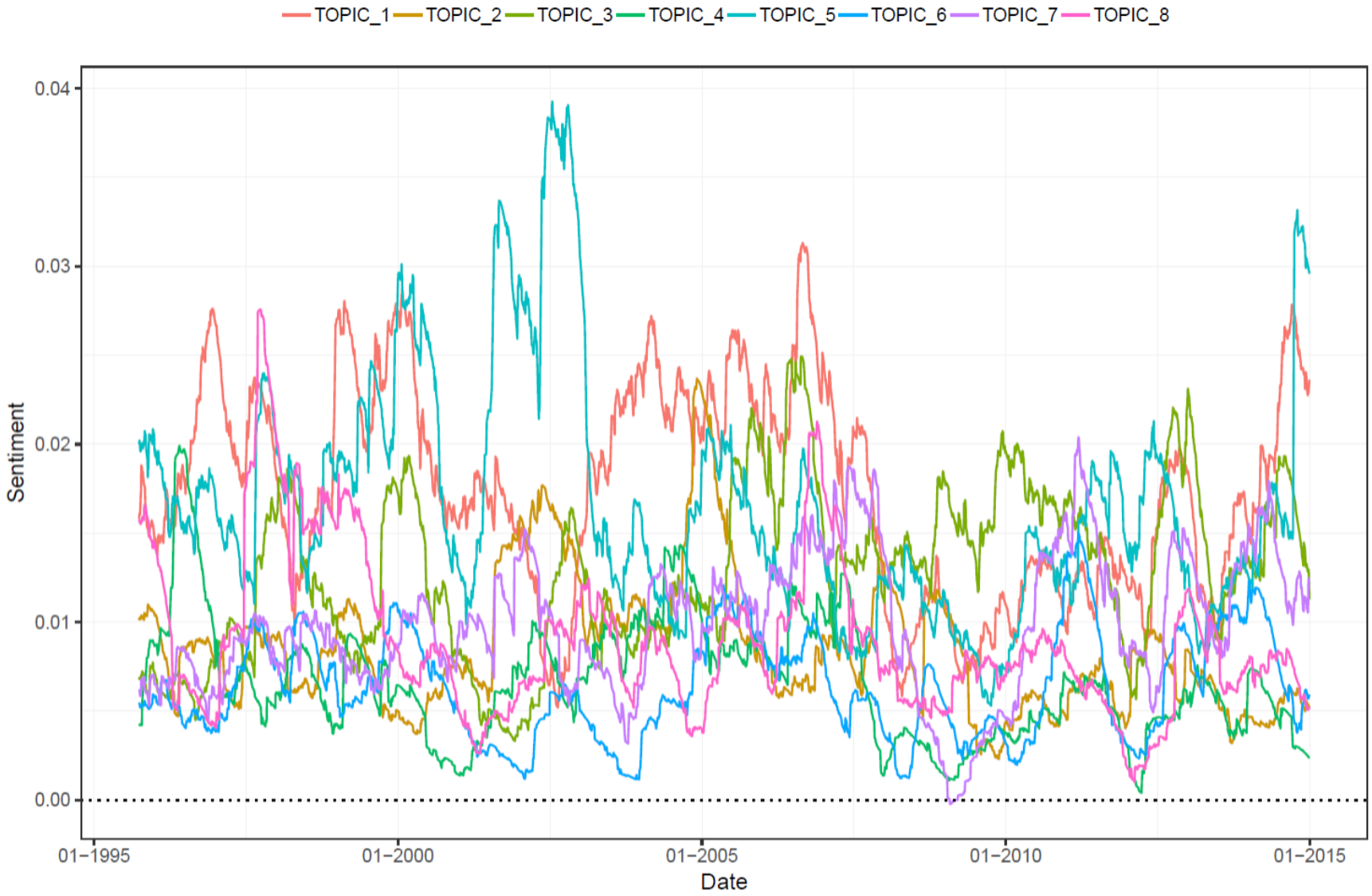


Figure 3: Textual sentiment time series across latent topic features.

FUNCTIONALITIES

Functionality	Functions	Output
4. Modelling		
(a) Specification	`ctr_model()`	
(b) Estimation	`sento_model()`	*sentomodel*, *sentomodeliter*
(c) Prediction	`predict.sentomodel()`	
(d) Diagnostics	`summary()`, `get_loss_data()`, `attributions()`	*attributions*
(e) Visualization	`plot.sentomodeliter()`, `plot.attributions()`	

Elastic net **estimation** and (out-of-sample) **prediction**

Attribution analysis

EXAMPLE

```
R> ctrIter <- ctr_model(model = "gaussian",
+     type = "BIC", h = h, alphas = c(0, 0.1, 0.3, 0.5, 0.7, 0.9, 1),
+     do.iter = TRUE, oos = oos, nSample = M, nCore = 1)
R> out <- sento_model(sentMeasIn, x = x[, "lag", drop = FALSE], y = y,
+     ctr = ctrIter)
```

```
R> attrFit <- attributions(fit, sentMeas)
R> head(attrFit[["features"]])
```

FUNCTIONALITIES

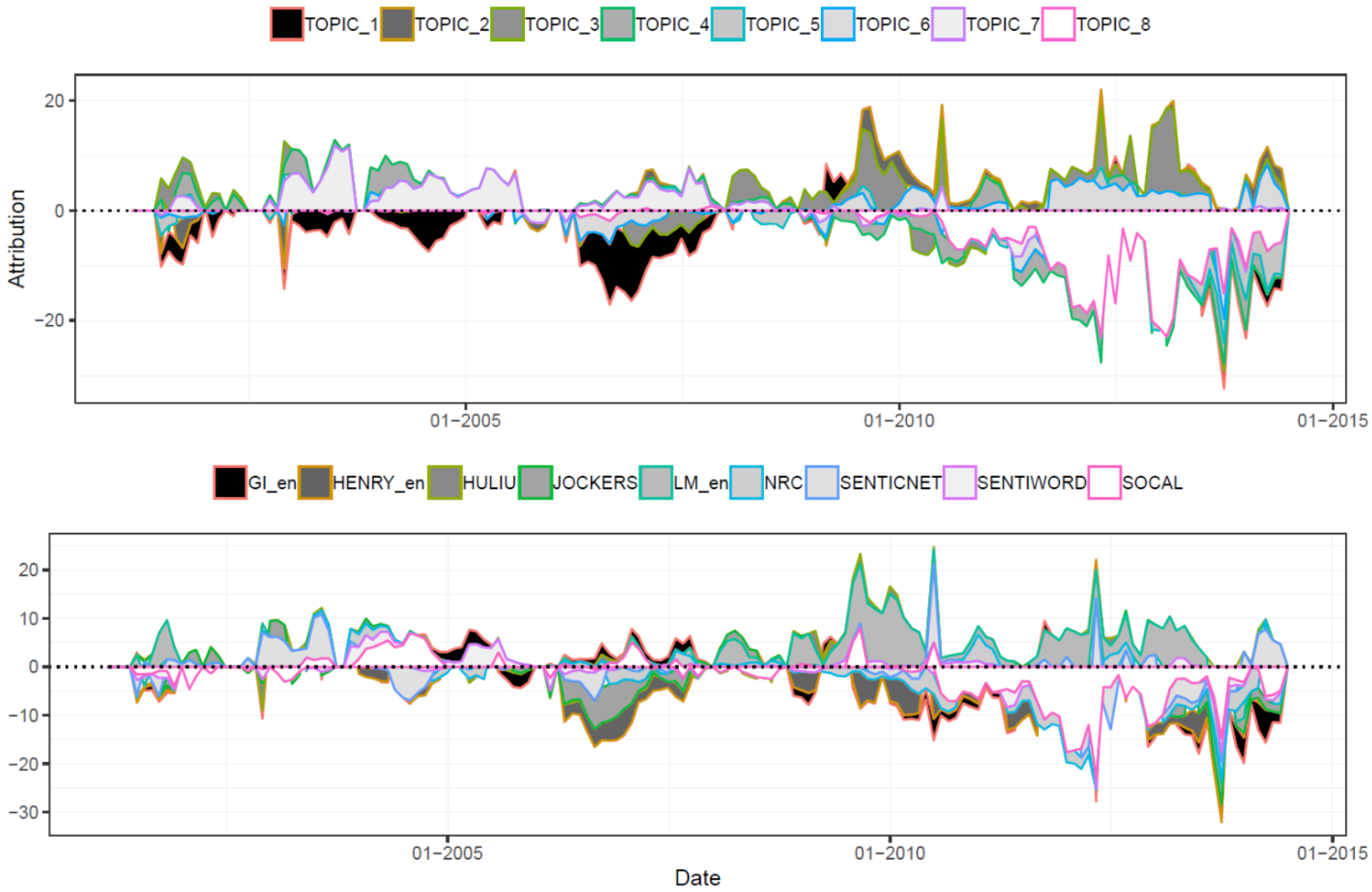


Figure 5: Attribution to features (top) and lexicons (below).

SENTOMETRICS R PACKAGE

The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment

David Ardia
University of Neuchâtel
HEC Montréal

Keven Bluteau
University of Neuchâtel
Vrije Universiteit Brussel

Samuel Borms
University of Neuchâtel
Vrije Universiteit Brussel

Kris Boudt
Ghent University
Vrije Universiteit Brussel
Vrije Universiteit Amsterdam

Abstract

We provide a hands–on introduction to optimized textual sentiment indexation using the R package **sentometrics**. Textual sentiment analysis is increasingly used to unlock the potential information value of textual data. The **sentometrics** package implements an intuitive framework to efficiently compute sentiment scores of numerous texts, to aggregate the scores into multiple time series, and to use these time series to predict other variables. The workflow of the package is illustrated with a built–in corpus of news articles from two major U.S. journals to forecast the VIX index.

Keywords: Penalized Regression, Prediction, R, **sentometrics**, Textual Sentiment, Time Series.

CONCLUSION

Sentometrics is about econometrics meeting sentiment

Transformation of qualitative sentiment data into quantitative sentiment variables

Application in econometric analysis of relationships between sentiment and other (economic) variables

Software package sentometrics