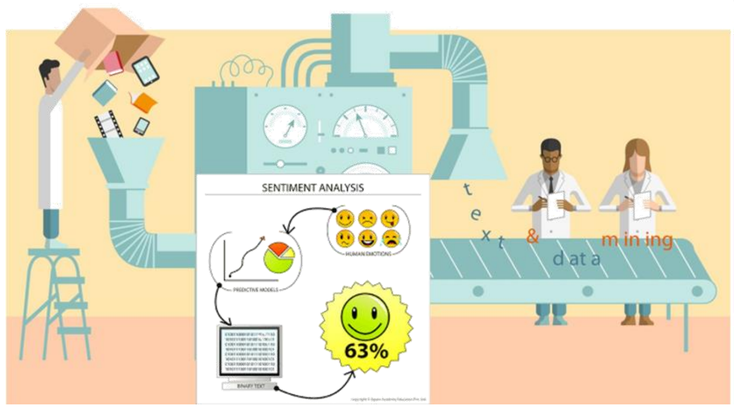
Text-based indexation and monitoring of corporate reputation using the R package 'sentometrics'

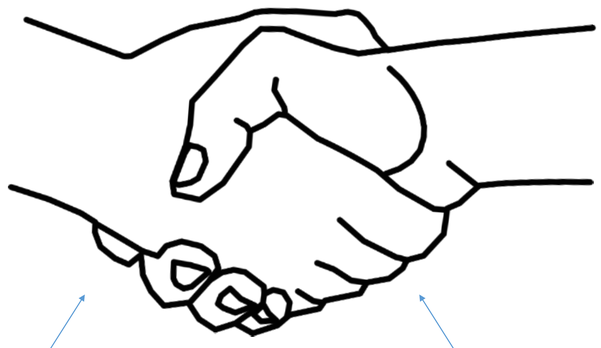
Andres Algaba, David Ardia, Keven Bluteau, Samuel Borms and Kris Boudt

The Sentometrics research project



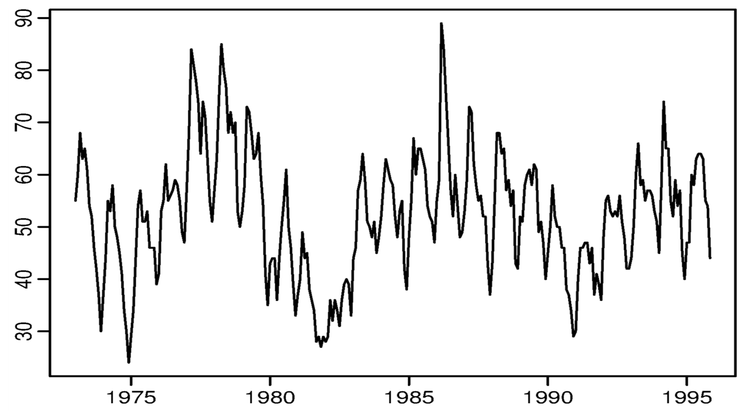
textual sentiment analysis

Cross-disciplinary
Cross-university
Econometrics expertise
Many applications



sentometrics

research ⟷ R package



time series econometrics

Investment analysis | Reputation monitoring | ...

Event detection | Macroeconomic forecasting

The R package 'sentometrics'

A framework that integrates (i) the qualification of sentiment from texts, (ii) the aggregation into different sentiment measures, and (iii) the optimized prediction based on these measures.

(STEP 1) Build a corpus of texts with quantifiable metadata ("features") $\in [0, 1]$

ID	DATE	TEXTS	FEAT. 1	FEAT. 2	...
1	1995-01-02	Text 1	1	0	...
2	1995-01-05	Text 2	0	1	...
...	...	...	...	...	...

(STEP 2) Pick lexicons and compute textual sentiment

Lexicon-based sentiment analysis augmented with valence shifters (negation, amplification, downtoners)

Across-document aggregation | document-level sentiment

Weighting schemes: counts, proportional, tf-idf

(STEP 3) Aggregate document-level sentiment scores into time series (daily, weekly, monthly, yearly)

Within-document aggregation | a time series

Across-time aggregation | a *smoothed* time series

```
        date LM_eng--wsj--equal_weight LM_eng--wapo--equal_weight
1: 1995-12-01             -0.03038392              -0.03096058
2: 1996-01-01             -0.03074413              -0.03262021
3: 1996-02-01             -0.03349817              -0.03567584
4: 1996-03-01             -0.03106851              -0.03681972
5: 1996-04-01             -0.02889475              -0.03420715
6: 1996-05-01             -0.02873871              -0.03299130
```

lexicon | feature | time aggregation scheme | P time series

Weighting schemes: equal, proportional, exponential, Almon polynomial, linear

(STEP 4) Estimate a sentiment-based prediction model (linear, logistic)

$$y_{u+h} = \delta + \boldsymbol{\gamma}^\top \boldsymbol{x}_u + \beta_1 s_u^1 + \ldots + \beta_p s_u^p + \ldots + \beta_P s_u^P + \epsilon_{u+h}$$

target | other | sentiment

Elastic net penalized regression because typically $P \gg N$

(STEP 5) Evaluate model performance and sentiment attribution

Out-of-sample errors analysis

Time-varying attribution of sentiment measures to predictions across lexicons, features and weighting schemes

Model confidence set (Hansen et al., 2011)

Full application: "*Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values*" (Ardia, Bluteau & Boudt, 2017)

Illustration: reputational sentiment time series

Swiss-based corpus

GDELT database to get urls ('CHE' as country actor code)
Scraped using Python's 'newspaper' library; 211,000 articles
April 2013 to December 2017

Seven reputation dimensions

Products, innovation, citizenship, workplace, governance, leadership & performance (Fombrun et al., 2015)
Features are a number of characteristic keywords (e.g., CEO, R&D or profitability) per dimension

Company features

Credit Suisse, UBS, Novartis & Roche
Full name detection in summary

Lexicons

Harvard General Inquirer
Loughran & McDonald (2011)
Henry (2008)

Time series construction

Weekly aggregation
Linear smoothing
26-week time lag
Averaging across dimensions

Next: validation of reputational indices and its dimensions w.r.t. a set of reputation proxies. Hard because reputation is *latent*.

We perform STEP 1, STEP 2 and STEP 3.