Sentiment and Econometrics: Toward A Unified Framework of Textual Sentiment Analysis for Economic and Financial Applications

Samuel Borms
Université de Neuchâtel
Vrije Universiteit Brussel

August 17, 2020

Public PhD defense

Motivation

Alternative data

You read the papers. You listen to the radio. You watch TV. Thank you.

Alternative data are “qualitative sentiment data.” Information value?!

Motivation of my work

Focus on **texts**.

$\mathbf{f}$(questions, methods, traditional data, textual data) $\rightarrow$ better answers.

But... application-specific textual data transformation is hard.

My thesis attempts to define $\mathbf{f}()$. In my version, econometrics meets sentiment meets econometrics.

User personas – for who is this useful?

The aim is to provide a **gateway** for specialists and non-specialists willing to create and use **textual sentiment** data.

Researcher

Data Scientist

Asset Manager

Macroeconomist

Contribution #1: Formalization of a methodological frame of thought for applied textual (sentiment) analysis

Generalization of "sentiment"

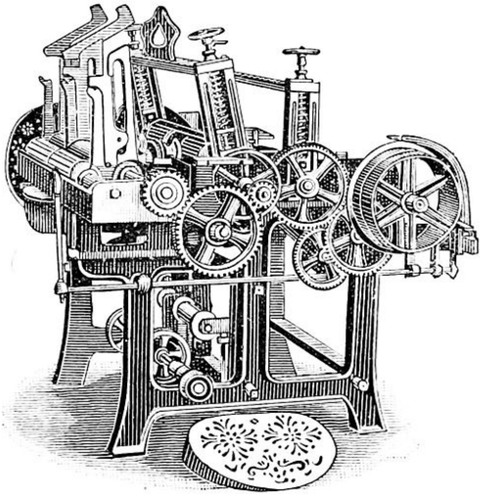
Step away from the limiting view on sentiment in most literature.

Definition. Sentiment is the disposition of an entity toward an entity, expressed via a certain medium.

$\Longrightarrow$ Use the medium... (e.g. press data)
... to extract the expressed disposition... (e.g. positivity or bias)
... to measure something about one or more entities (e.g. the economy).

Fill in the details along the analysis.

A framework for solving problems with sentiment data



From problem to data transformation to modeling to concluding and back.

Integrates numerous challenges likely to face.

sentiment

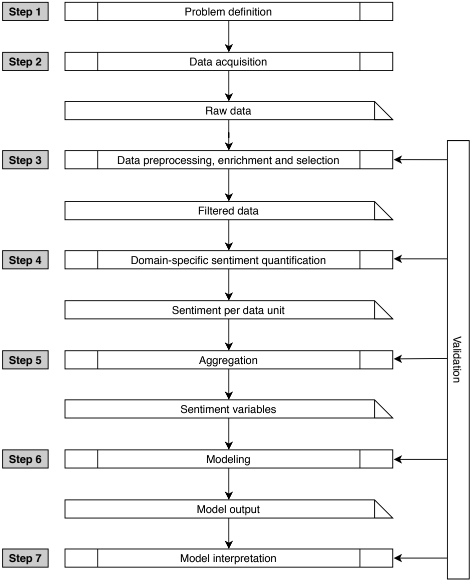
$+$

econ**ometrics**

$\Longrightarrow$

sentometrics

The sentometrics analysis cycle



The "joint hypothesis problem" and validation

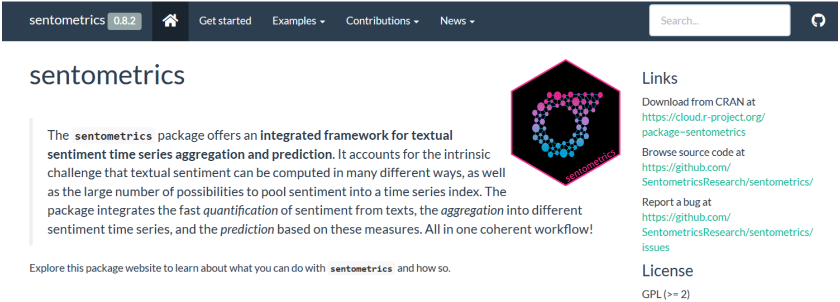
Main challenge is "joint hypothesis problem"-like. You need to validate both the data transformation and the answer to your research question.

Cyclical validation approach required.

Contribution #2: Formalization of the methodological framework's core into a computational one

R software package **sentometrics**

Development, release and thorough documentation of open-source R software package **sentometrics**. Free to install and use!



sentometrics

> The `sentometrics` package offers an **integrated framework for textual sentiment time series aggregation and prediction**. It accounts for the intrinsic challenge that textual sentiment can be computed in many different ways, as well as the large number of possibilities to pool sentiment into a time series index. The package integrates the fast *quantification* of sentiment from texts, the *aggregation* into different sentiment time series, and the *prediction* based on these measures. All in one coherent workflow!

Explore this package website to learn about what you can do with `sentometrics` and how so.

Links

Download from CRAN at https://cloud.r-project.org/package=sentometrics

Browse source code at https://github.com/SentometricsResearch/sentometrics/

Report a bug at https://github.com/SentometricsResearch/sentometrics/issues

License

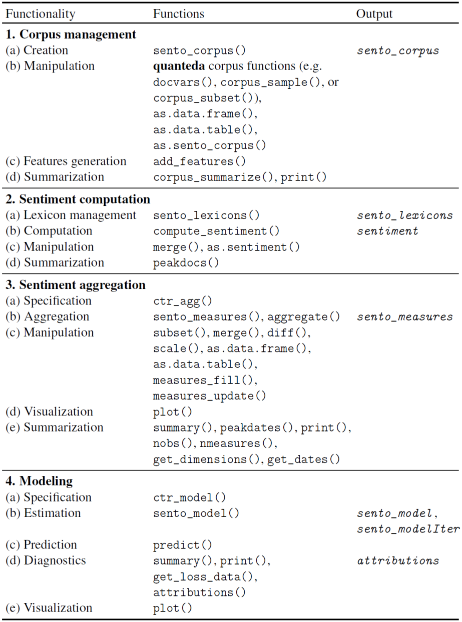
GPL (>= 2)

Continuous improvements and additions going forward. See dedicated package website sentometricsresearch.github.io/sentometrics.

Functionalities

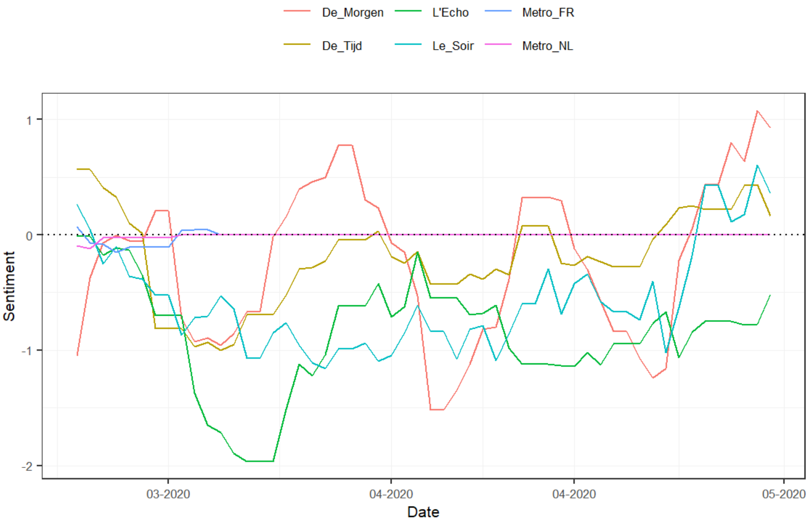
Maps to methodological framework in that it covers at least one functionality for all steps possible (3–7).

Unique when it comes to flexible aggregation into sentiment time series.



Functionality	Functions	Output
1. Corpus management		
(a) Creation	`sento_corpus()`	*sento_corpus*
(b) Manipulation	**quanteda** corpus functions (e.g. `docvars()`, `corpus_sample()`, or `corpus_subset()`), `as.data.frame()`, `as.data.table()`, `as.sento_corpus()`	
(c) Features generation	`add_features()`	
(d) Summarization	`corpus_summarize()`, `print()`	
2. Sentiment computation		
(a) Lexicon management	`sento_lexicons()`	*sento_lexicons*
(b) Computation	`compute_sentiment()`	*sentiment*
(c) Manipulation	`merge()`, `as.sentiment()`	
(d) Summarization	`peakdocs()`	
3. Sentiment aggregation		
(a) Specification	`ctr_agg()`	
(b) Aggregation	`sento_measures()`, `aggregate()`	*sento_measures*
(c) Manipulation	`subset()`, `merge()`, `diff()`, `scale()`, `as.data.frame()`, `as.data.table()`, `measures_fill()`, `measures_update()`	
(d) Visualization	`plot()`	
(e) Summarization	`summary()`, `peakdates()`, `print()`, `nobs()`, `nmeasures()`, `get_dimensions()`, `get_dates()`	
4. Modeling		
(a) Specification	`ctr_model()`	
(b) Estimation	`sento_model()`	*sento_model*, *sento_modelIter*
(c) Prediction	`predict()`	
(d) Diagnostics	`summary()`, `print()`, `get_loss_data()`, `attributions()`	*attributions*
(e) Visualization	`plot()`	

Litmus test – example sentiment time series variables



Contribution #3: Structured news-based measurement of firm-level sustainability and of economic uncertainty

Application #1: News-based firm-level sustainability

Increasing interest to invest in companies who do well environmentally, socially and governance wise (ESG).

Sustainable asset managers use external ESG ratings and in-house research to screen the investment universe.

I add news to the mix through creation of daily firm-level indices capturing frequency and sentiment of news reporting about ESG issues.

Empirical analysis

291 European stocks, Dutch news from Belga, 1999–2018, Sustainalytics.

Use of specific keywords to track the news relevant to ESG **coverage**.
More negative **sentiment** word list ("controversies").

Querying, cleaning, selection, aggregation, validation.

Keywords

A **semi-supervised** approach: expertise + models. Man & machine.

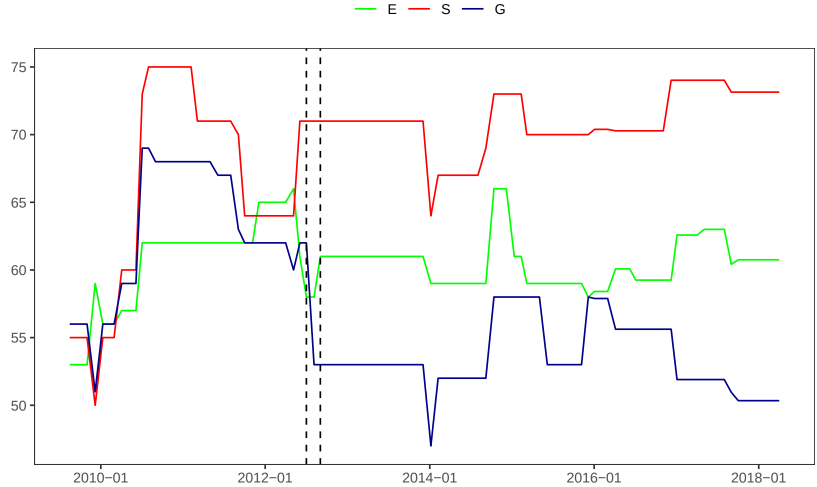
Expert: pick some important words (klimaat, mobiliteit, ecologie, etc.).

Model: from >100000 words, tell me which words are semantically related based on estimated **word embeddings**.

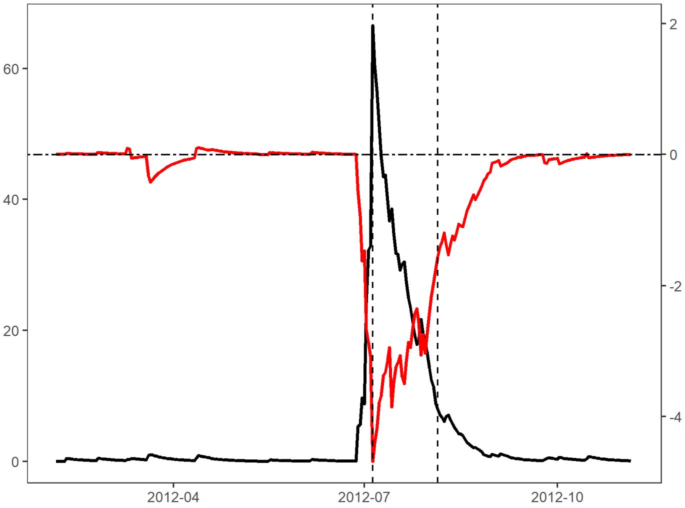
Expert: check if the most related words are useful.

Algorithm: follow up the news that use these keywords.

Barclays Sustainalytics ratings (monthly)



Barclays news coverage & sentiment during LIBOR scandal



Short-term news **signals**. View of the risk-averse doctor analyzing patients.

Stock and sector screening

Monthly rebalanced portfolios based on the news-based indicators perform at least as well as portfolios based on external ESG ratings.

News coverage indicators more informative than pure sentiment ones.

Sector rotation most promising.

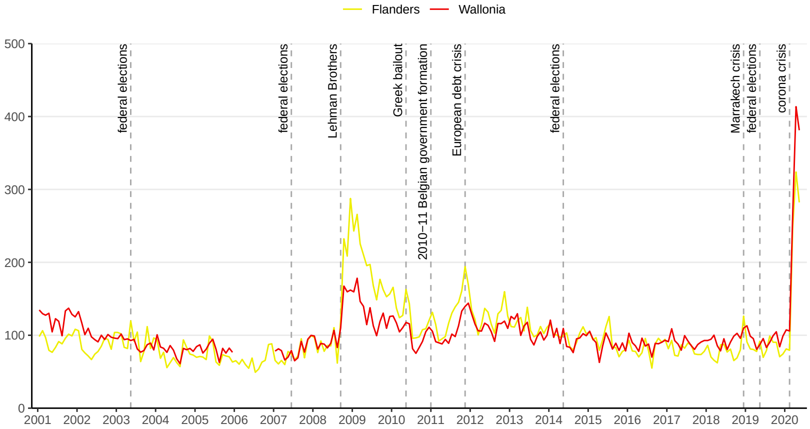
Possible extensions: international news, event study, factor portfolios.

Application #2: An EPU index for Belgium

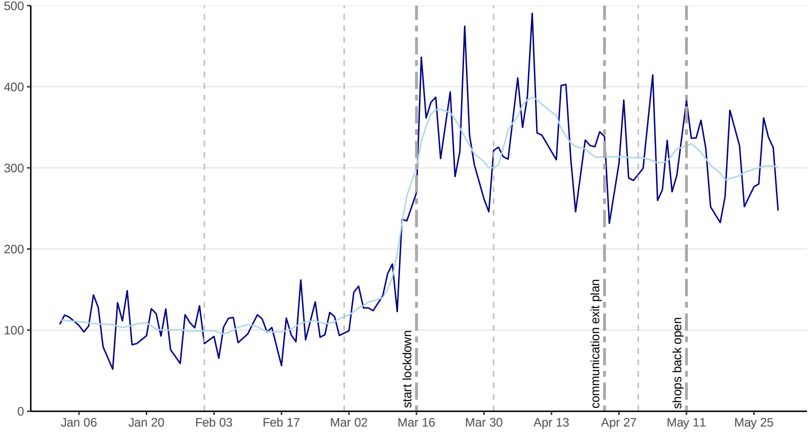
Application to the case of Belgium of common methodology to measure economic policy uncertainty (EPU) from news articles.

Same word embedding approach to keywords definition.

Monthly evolution of news-based EPU in Belgium

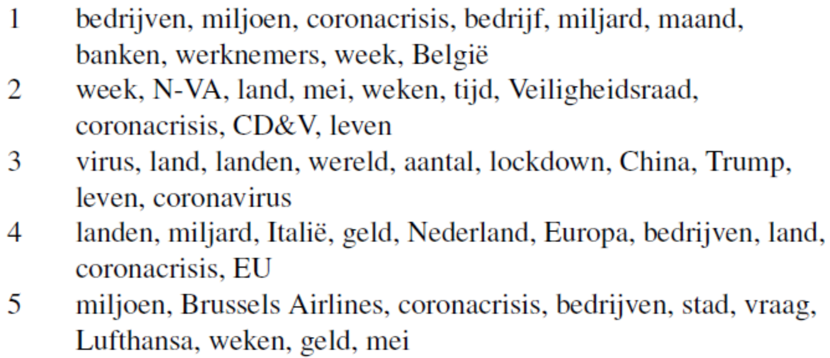


Daily zoom-in during 2020



Explaining peaks

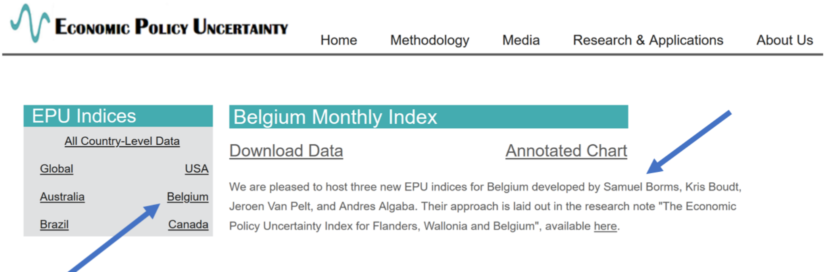
Additional validation in the form of automated qualitative “news reader.”



1	bedrijven, miljoen, coronacrisis, bedrijf, miljard, maand, banken, werknemers, week, België
2	week, N-VA, land, mei, weken, tijd, Veiligheidsraad, coronacrisis, CD&V, leven
3	virus, land, landen, wereld, aantal, lockdown, China, Trump, leven, coronavirus
4	landen, miljard, Italië, geld, Nederland, Europa, bedrijven, land, coronacrisis, EU
5	miljoen, Brussels Airlines, coronacrisis, bedrijven, stad, vraag, Lufthansa, weken, geld, mei

5 clusters of news in April 2020, all related to COVID-19.

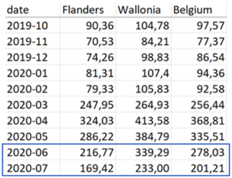
www.policyuncertainty.com



Belgium Monthly Index

Download Data   Annotated Chart

We are pleased to host three new EPU indices for Belgium developed by Samuel Borms, Kris Boudt, Jeroen Van Pelt, and Andres Algaba. Their approach is laid out in the research note "The Economic Policy Uncertainty Index for Flanders, Wallonia and Belgium", available here.



date	Flanders	Wallonia	Belgium
2019-10	90,36	104,78	97,57
2019-11	70,53	84,21	77,37
2019-12	74,26	98,83	86,54
2020-01	81,31	107,4	94,36
2020-02	79,33	105,83	92,58
2020-03	247,95	264,93	256,44
2020-04	324,03	413,58	368,81
2020-05	286,22	384,79	335,51
2020-06	216,77	339,29	278,03
2020-07	169,42	233,00	201,21

But what in August 2020? And so on...

Conclusion

Contributions wide and large

A systematic approach to use textual data in applied research.

A computational toolbox allowing to do so quickly and efficiently.

A catalyst to more effectively use textual data.

An inspiration to formalize similar frameworks for audio and video data.

A structured application to following ESG-related news.

A monthly index to track EPU in Belgium.

Exciting future research

Better **validation** tools.

More **applications** based on the framework.

Integrated **theoretical developments** jointly accounting for various steps.

Intraday textual sentiment analysis.

Multimodal sentiment analysis.

Wouldn't it be nice if we could collaborate cross-disciplinary?

github.com/SentometricsResearch (in progress)

PhD papers

- Algaba, Ardia, Bluteau Borms & Boudt (2020). "**Econometrics meets sentiment: An overview of methodology and applications**". The Journal of Economic Surveys 34 (3), 512-547.
- Ardia, Bluteau, Borms & Boudt (2020). "**The R package *sentometrics* to compute, aggregate and predict with textual sentiment**". The Journal of Statistical Software, forthcoming.
- Borms, Boudt, Van Holle & Willems (2020). "**Semi-supervised text mining for monitoring the news about the ESG performance of companies**". Data Science for Economics and Finance: Methodologies and Applications (Springer), forthcoming.
- Algaba, Borms, Boudt & Van Pelt (2020). "**The Economic Policy Uncertainty index for Flanders, Wallonia and Belgium**". Bank- en Financiewezen digitaal 2020/6.

Thanks

Many thanks for your attention!

Looking forward to taking questions.