

Talking about Europe: Exploring 70 years of news archives

Enrico Bergamini¹ Emmanuel Mourlon-Druol²

¹PhD candidate in Economics, University of Turin

²Professor of Economic History, University of Glasgow

EADH 2021

Introduction

- Bergamini, E.; Mourlon-Druol, E.; *"Talking about Europe: Exploring 70 years of news archives"* 2021, Bruegel Working Paper.
- Papadia, F.; Bergamini, E.; Mourlon-Druol, E.; Porcaro, G.; *"How interested are European citizens in European matters?"* 2021, Bruegel Policy Brief.
- **Motivation:** Long-term investigation of news coverage about "Europe", in "public opinion" eyes.
- Archival and methodological interest in working with longer-spanning databases compared with the related literature

In a nutshell

What have we done so far?

- Scraping full-text large news archives
- **Systematization** of unstructured archives
- **Methodology:** isolating the topic "Europe" via keywords matching combined with NLP-based topic recognition
- Event detection on derived time-series (E/T)

Literature streams

- Digitised textual sources increased application possibilities for digital humanities (DH). Machine Learning (ML) applications in public opinion, journalism, and DH studies increased the effectiveness of "distant reading" vis-à-vis "close reading". Broersma and Harbers (2018)
- Media coverage of the European Union, public opinion and support:
 - Barth and Bijsmans (2018); Diez Medrano (2003); Herzer (2019); Baisnée (2003); Abel (2013); Bellamy (2013)
- Quantitative analysis of large-textual information:
 - Vliegenthart et al. (2008); Greene and Cross (2017); Müller et al. (2018); Pukallus (2016); European Commission and Directorate General for Research (2007); Baker et al. (2016); Müller and Hornig (2020); Baker et al. (2020)

Scraping archives

- Nationally relevant, long-spanning, public archives (1945-2019)

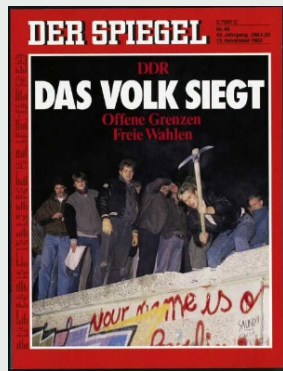
La Stampa - 9.9M



LeMonde - 2.8M



Spiegel - 300K



Scraping

- Data-collection of roughly 13M articles OCR ready

ARCHIVES

L'ALLEMAGNE A CAPITULÉ

Publié le 08 mai 1945 à 00h00 - Mis à jour le 08 mai 1945 à 00h00



Article réservé aux abonnés

Londres, 7 mai, 15 heures.

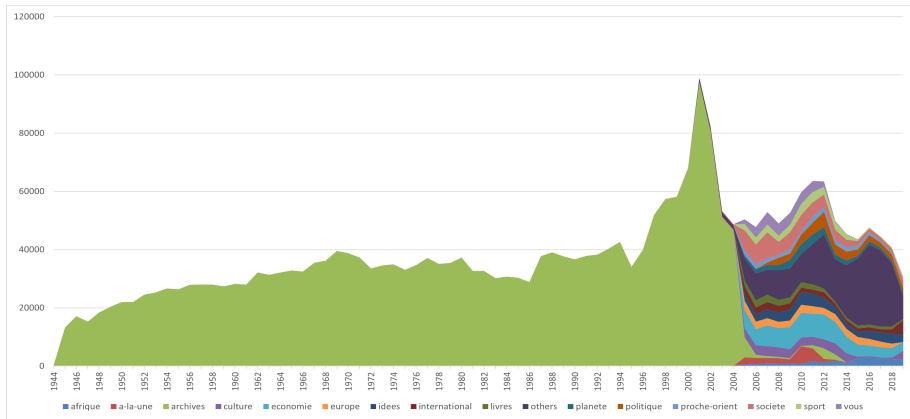
Selon l'agence Reuter, la radio allemande de Flensburg (seul poste émetteur officiel encore en service) annonce que l'amiral Dœnitz a ordonné la capitulation sans condition de l'ensemble des forces allemandes.

Exploring and cleaning the archives

- Large, undocumented, unstructured, (and cumbersome) data.
- These novel dataset contains daily (weekly in Der Spiegel) full-text articles, and (often) incomplete information about the sections:
 - Need for specific exploratory data analysis for each archive: **samples' consistency** was staple for deriving frequencies.
 - Consulting archivists.

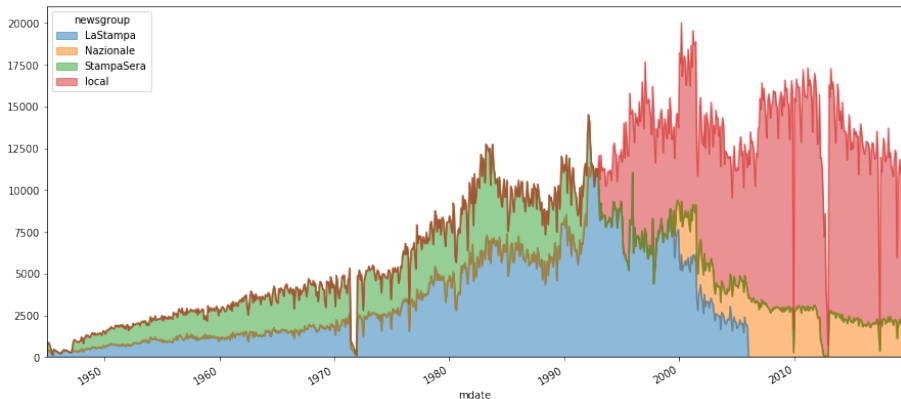
Exploring and cleaning the archives

Figure: Le Monde



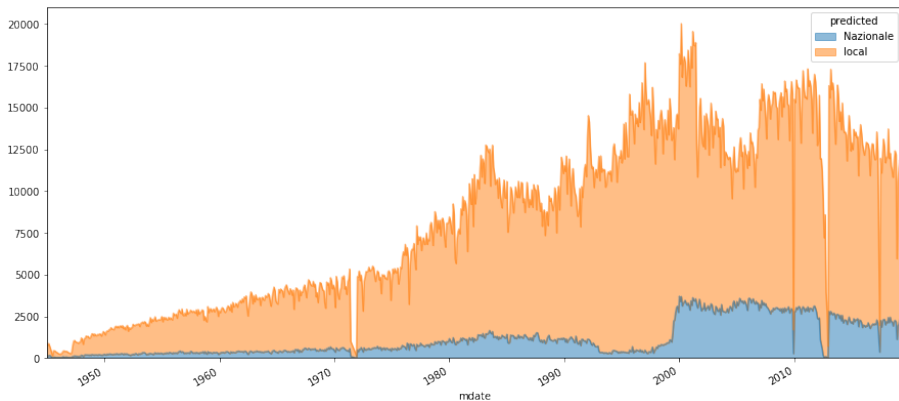
Exploring and cleaning the archives

Figure: La Stampa



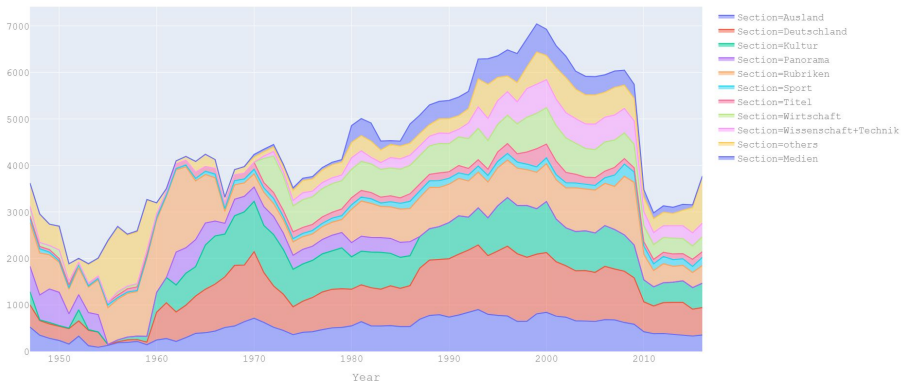
Exploring and cleaning the archives

Figure: La Stampa



Exploring and cleaning the archives

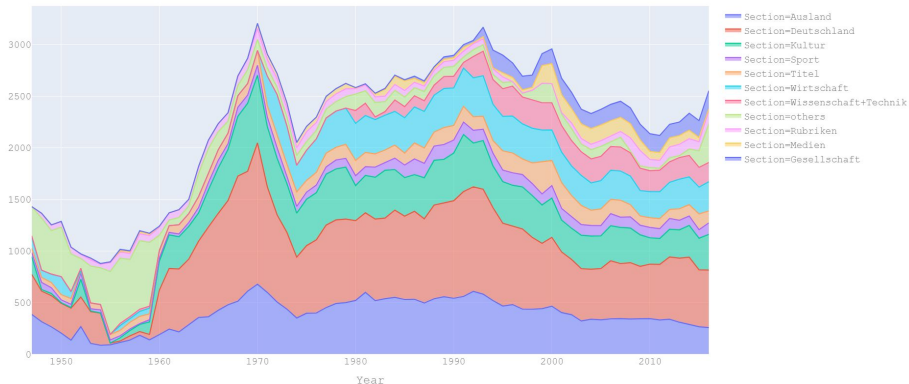
Figure: Spiegel



Exploring and cleaning the archives

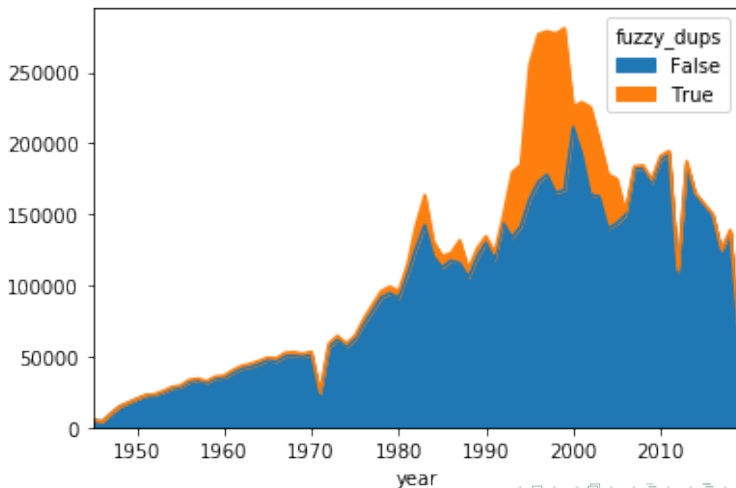
Figure: Spiegel

Spiegel



Exploring and cleaning the archives

Figure: Duplicates detection (fuzzy-match)



Text pre-processing

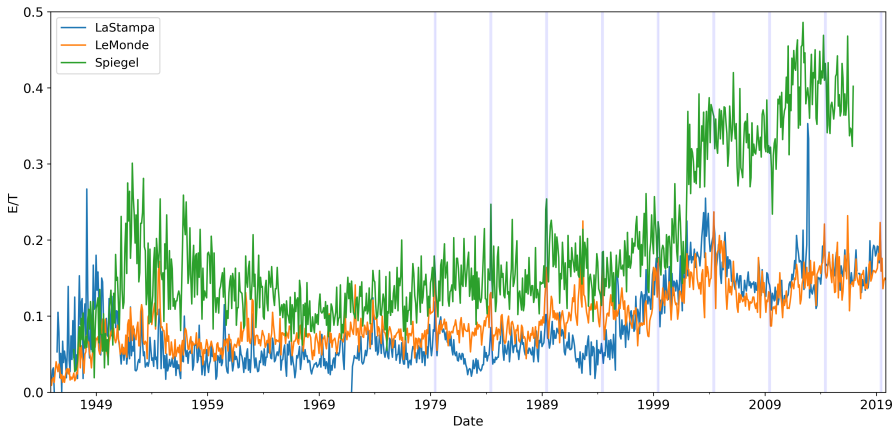
- We transform the textual information into bag-of-words model
- Removal of stopwords
- Lemmatisation (playing, played, plays → play)
- Bigrams and trigrams (european_union)

Topic extraction: keywords

- We define a first set of keywords, quite broad, for all the three languages (including acronyms).
 - *Europe, Union européenne, UE, Communauté économique européenne, Communauté européenne, CEE, Commission européenne, Parlement européen, CECA, CEEA / Euratom, Schengen, Euro, Maastricht, BCE, Banque centrale européenne, Bruxelles, Strasbourg, européen, Marché Commun, instances européennes, Commission de Bruxelles, ...*
- We create a count variable summing up each appearance of these keywords.
- We deflate "euro" and "european" to 0.5, and choose a threshold of 3, in order to match an article as "European".

Topic extraction: keywords

Figure: E/T 1945-2019 (keywords)



Topic extraction: keywords

- Assuming our keywords capture a wide enough sample, we need to select a more precise sub-sample and deflate the E/T ratio:
 - False positives (currency, sports)
 - Definitional issues: international politics, national politics, which we are *partly* interested in.

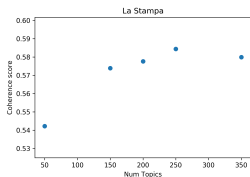
Topic extraction: topic modelling methodology

- At this stage, we have a sub-corpus of matched articles (and their relative "intensity", as count variable).
- We run a topic modelling exercise on this sub-corpus, in order to extract topics.
- Topic modelling infers latent topics from a collection of texts.
- Latent Dirichlet Allocation (David M. Blei, Andrew Ng, Michael I. Jordan, 2006)

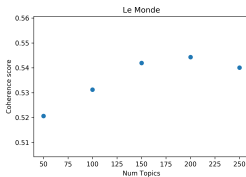
Topic extraction: topic modelling methodology

- We implement three language-specific LDA models, relying on McCallum (2002) framework for large topic modelling (Gibbs sampling).
- to-be implemented: STDM (Roberts et al. (2019)), gLDA (Jagarlamudi et al 2012), nnMF Greene and Cross (2017)
- Hyper-parameter optimisation: maximise coherence score values.

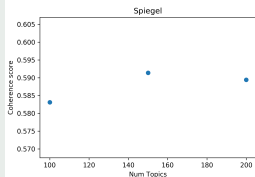
La Stampa



LeMonde

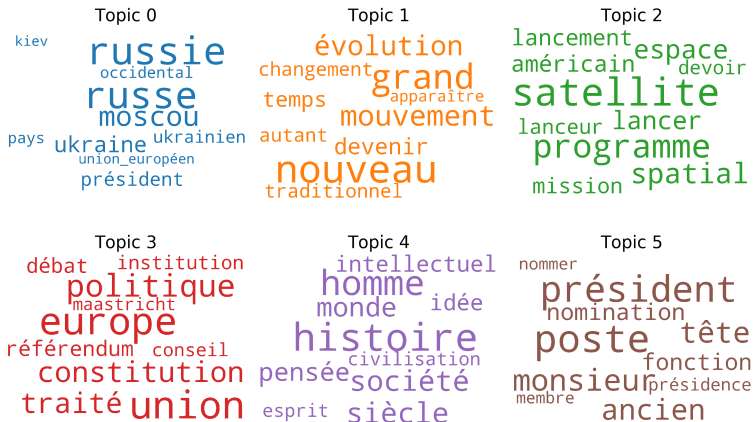


Spiegel



Topic extraction

Figure: Topics - LeMonde



Topic extraction: deflating matched articles

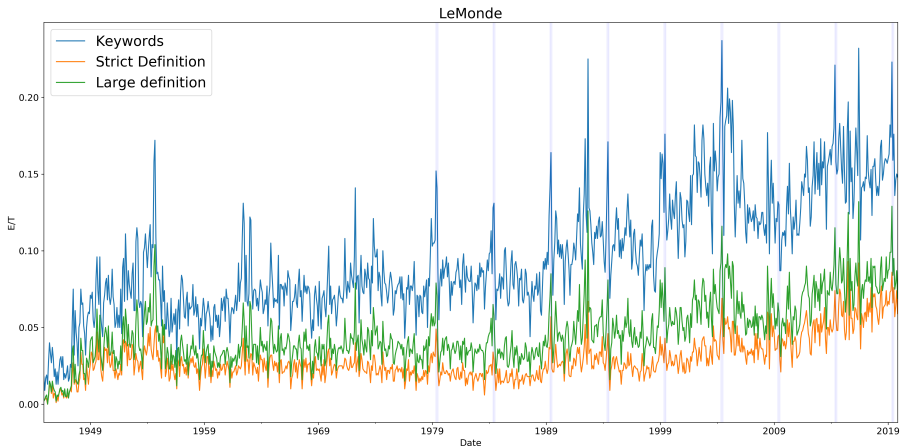
- We assign each article to a topic, with traditional highest probability allocation (how limiting?).
- We flag topics by their degree of certainty:
 - ① Certainly regarding Europe
 - ② Potentially regarding Europe
 - ③ Certainly not regarding Europe
- We include articles belonging to (1) and exclude those belonging to (3).
- For the articles belonging to topics (2), we only include articles with a keyword score higher than the average of that topic.

Topic extraction: deflating matched articles

Label	Topic ID	Number of Articles	Average Keywords per Article	Keywords
European	198	22986	5.28	président, angela_merkel, nicolas_sarkozy, dirigeant, chef_etat, bruxelles, juncker, vouloir, françois_hollande, monsieur_sarkozy
European	189	4021	4.98	commercial, accord, etat_unis, commerce, négociation, pays, omc, gatt, libre_échange, marché
European	193	27963	4.78	système, exemple, cas, principe, forme, technique, méthode, existe, formule, pratique
European	168	2505	4.52	ue, pays, union_européen, union, membre, etat_membre, adhésion, élargissement, bruxelles, candidat
European	192	19786	4.10	conseil, comité, conférence, membre, réunion, travail, ministre, réunir, organisation, représentant
International Politics	172	622	4.03	allemand, allemagne, bonn, français, gouvernement, france, chancelier, accord, traité, république_fédéral
Uncertain	171	2834	3.98	pouvoir, mesure, cas, contrôle, prévoir, système, condition, devoir, permettre, disposition
Uncertain	180	9247	3.92	grand, nouveau, année, important, devoir, pouvoir, permettre, partie, venir, également
International Politics	185	2328	3.89	serbe, kosovo, yougoslavie, belgrade, serbie, bosnie, yougoslave, international, guerre, albanais
Uncertain	170	2366	3.88	banque, financier, crédit, bancaire, établissement, prêt, marché, dette, risque, milliard_euro

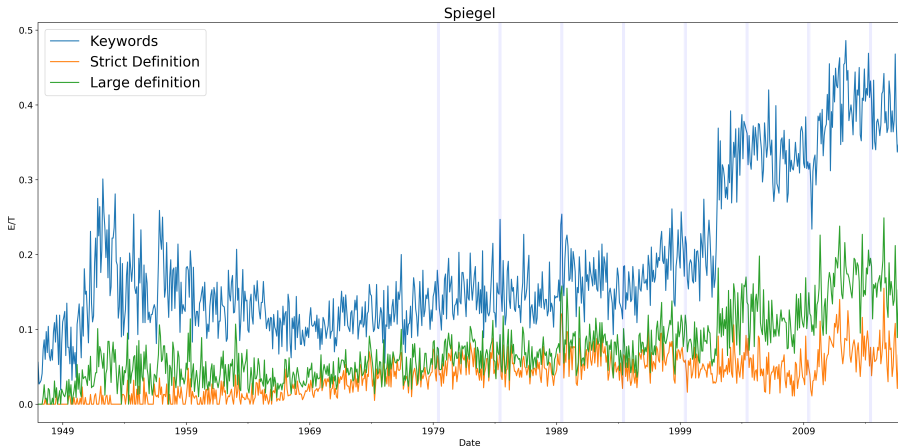
Topic extraction: deflating matched articles

Figure: LeMonde - compared methodologies - [link](#)



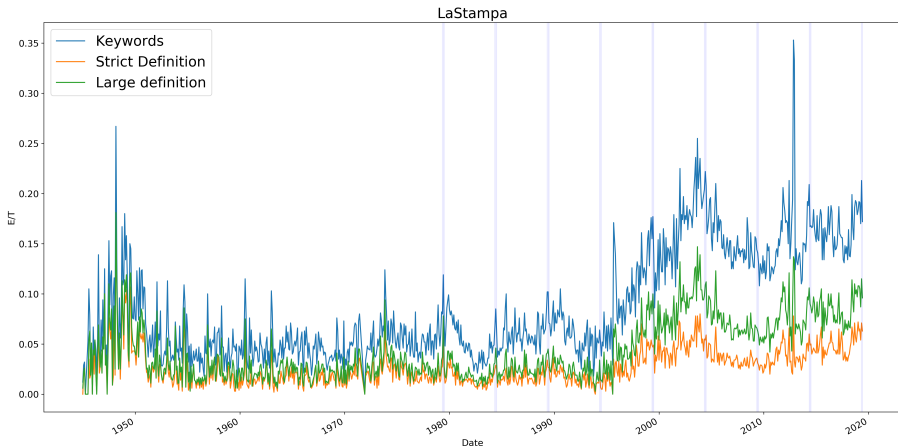
Topic extraction: deflating matched articles

Figure: Spiegel - compared methodologies - [link](#)



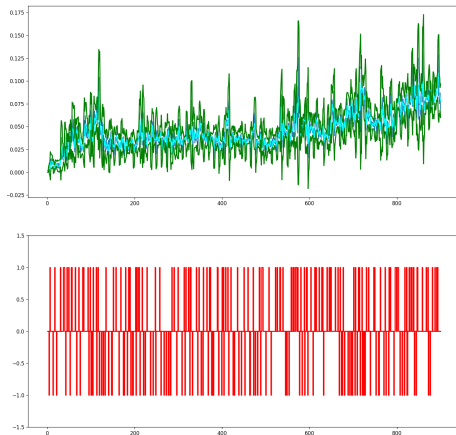
Topic extraction: deflating matched articles

Figure: La Stampa - compared methodologies - [link](#)



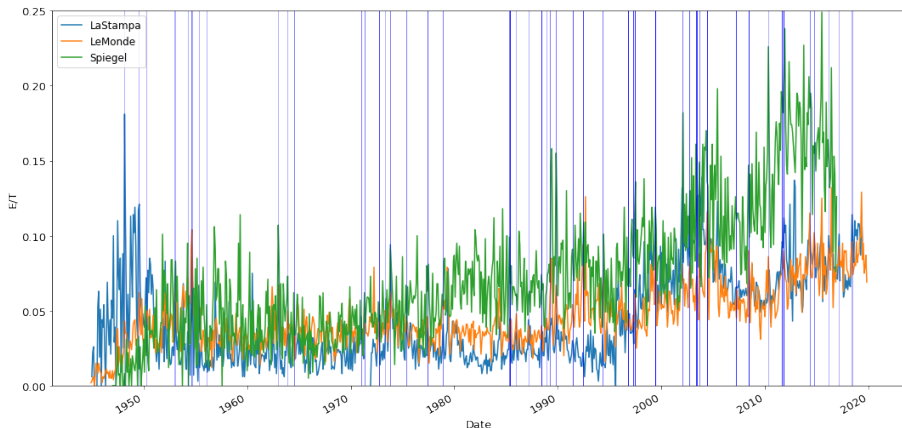
Event detection

Figure: Z-score event detection Brakel (2020)



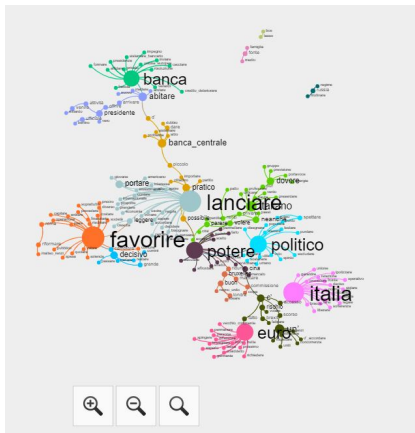
Frequency of European media coverage

Figure: Results - [link](#)



Where next?

Figure: Geotemporal networks - [link](#)



Conclusions

- We collect and systemize a novel, large web-scraped dataset to investigate media coverage, across three countries.
- We propose a methodology to isolate the topic "Europe" via keywords matching combined with NLP-based topic recognition.
- We identify a level increase in European news coverage corresponding to the 1990s wave of European integration.
- We perform event detection on derived time-series (E/T), matching key events in the history of European integration.

Comments and questions

Thank you!

References I

- Abel, R. (2013). The Pleasures and Perils of Big Data in Digitized Newspapers. *Film History*, 25(1-2):1–10. Publisher: Indiana University Press.
- Baisnée, O. (2003). *La production de l'actualité communautaire. Eléments d'une sociologie du corps de presse accrédité auprès de l'Union européenne (France/Grande-Bretagne)*. PhD thesis, Institut d'Etudes Politiques de Rennes.
- Baker, S., Bloom, N., Davis, S., and Terry, S. (2020). COVID-induced economic uncertainty and its consequences. Library Catalog: VoxEU.
- Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636. Publisher: Oxford Academic.
- Barth, C. and Bijsmans, P. (2018). The Maastricht Treaty and public debates about European integration: the emergence of a European public sphere? *Journal of Contemporary European Studies*, 26(2):215–231.

References II

- Bellamy, R. (2013). 'An Ever Closer Union Among the Peoples of Europe': Republican Intergovernmentalism and Democratic Representation within the EU. *Journal of European Integration*, 35(5):499–516. Publisher: Routledge
_eprint: <https://doi.org/10.1080/07036337.2013.799936>.
- Brakel, J.-P. (2020). Peak signal detection in realtime timeseries data - Smoothed Z-score Algorithm (peak detection with robust threshold).
- Broersma, M. and Harbers, F. (2018). Exploring machine learning to study the long-term transformation of news. *Digital Journalism*, 6(9):1150–1164.
- Diez Medrano, J. (2003). *Framing Europe: Attitudes to European Integration in Germany, Spain, and the United Kingdom*. Princeton N.J., princeton university press edition.
- European Commission and Directorate General for Research (2007). *The transformation of political mobilisation and communication in European public spheres*. Publications Office, Luxembourg. OCLC: 1044364964.

References III

- Greene, D. and Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1):77–94. Publisher: Cambridge University Press.
- Herzer, M. (2019). *The Media, European Integration and the Rise of Euro-journalism, 1950s–1970s*. Palgrave Macmillan.
- McCallum, A. K. (2002). MALLET: A machine learning for language toolkit.
- Müller, H. and Hornig, N. (2020). A new Uncertainty Perception Indicator (UPI) – concept and first results. *Dortmund Center for data-based Media Analysis (DoCMA)*, (1):39.
- Müller, H., Porcaro, G., and von Nordheim, G. (2018). Tales from a crisis: diverging narratives of the euro area | Bruegel. Technical Report 3, Bruegel.
- Pukallus, S. (2016). *Representations of European citizenship since 1951*. Palgrave Macmillan.
- Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2).

References IV

Vliegenthart, R., Schuck, A. R. T., Boomgaarden, H. G., and De Vreese, C. H. (2008). News Coverage and Support for European Integration, 1990–2006. *International Journal of Public Opinion Research*, 20(4):415–439. Publisher: Oxford Academic.