Seminar

Text Data meets Econometrics

Carsten Jentsch, Niklas Benner & Jonas Rieger

Winter Term 2020/2021

technische universität dortmund

fakultät statistik

Text Data

Characteristics:

- weak- or unstructured data,
- dirty data,
- interest in the abstraction of (sentence) structures,
- high-dimensional data (documents x words).

Application Areas:

- comparison of the coverage of different media,
- analysis of journalistic channels (twitter, facebook, WhatsApp, ...),
- spam filters, search engines, user-specific advertising, ...

Possible Subject Areas

Preprocessing:

- Case Sensitivity, Tokenization, Stopwords, Stemming, Lemmatization,
- Bag-of-Words, N-Gram, tf-idf, ...

Classification & Clustering:

- (un)supervised categorization, recommendation tools, ...

Topic Modeling:

- LSA/LSI, pLSA/pLSI, LDA, CTM, NTM, STM, ...

Word Embeddings:

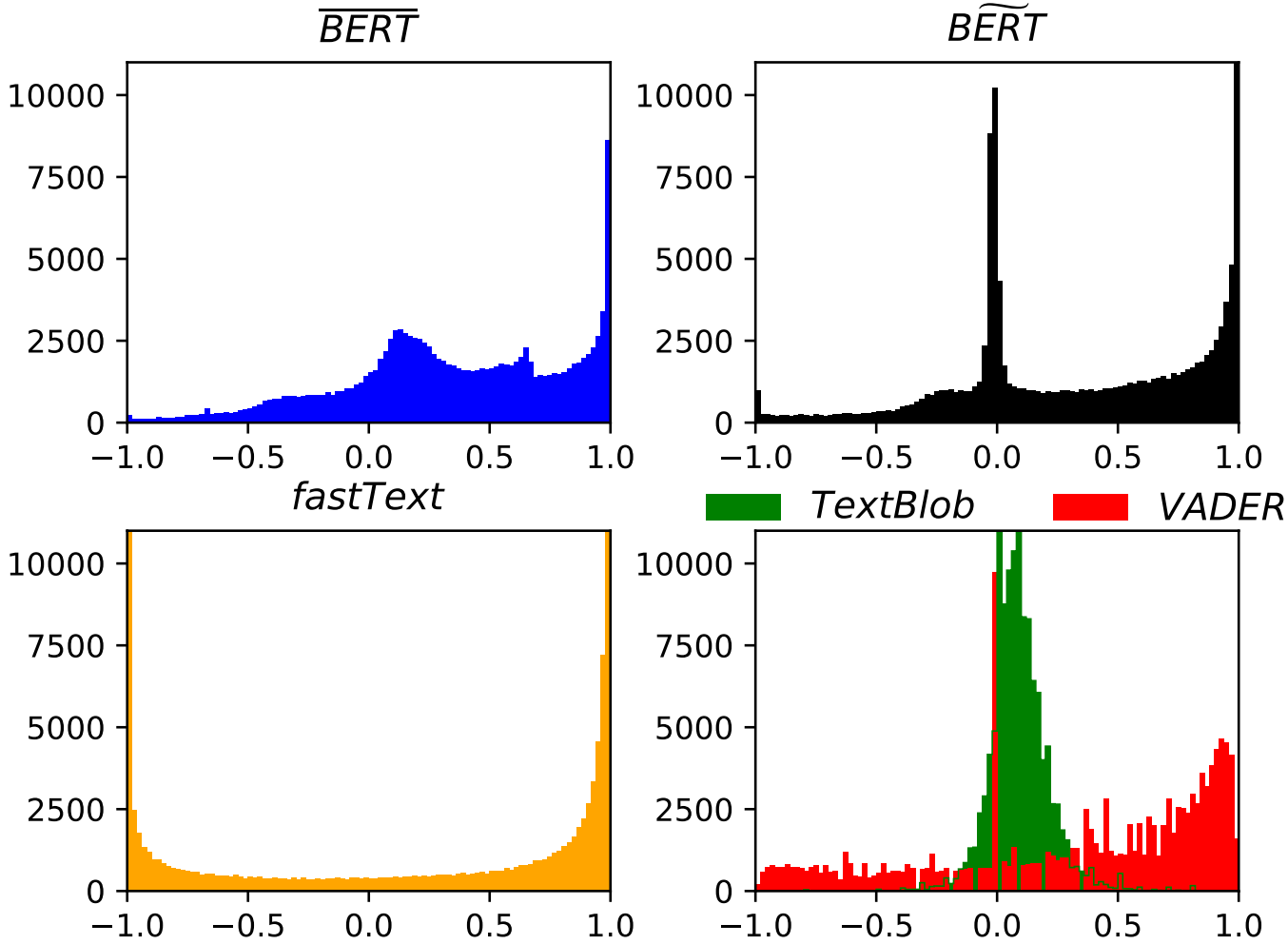
- Word2Vec, FastText, BERT, ...

(Party) Positions:

- Wordscore, Wordfish, ...

Example: Disagreement in ECB's Governing Council

- Idea: Dissent between central bankers weakens the effectiveness of monetary policy
- Analysis of 2131 (1625) speeches by ECB (NCB) representatives



Example: Disagreement in ECB's Governing Council

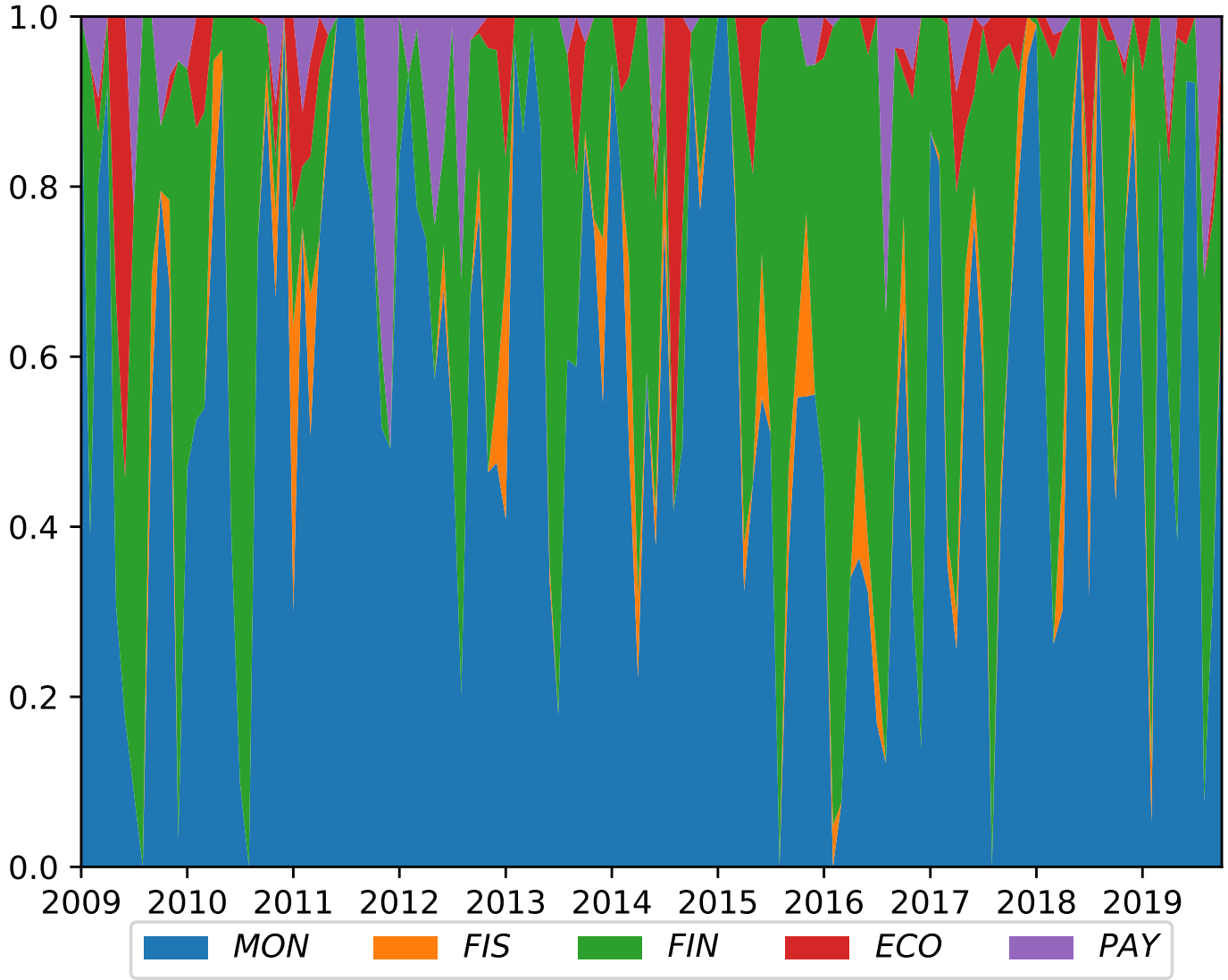


Figure: Share of the words assigned to the topics over time

Example: Disagreement in ECB's Governing Council

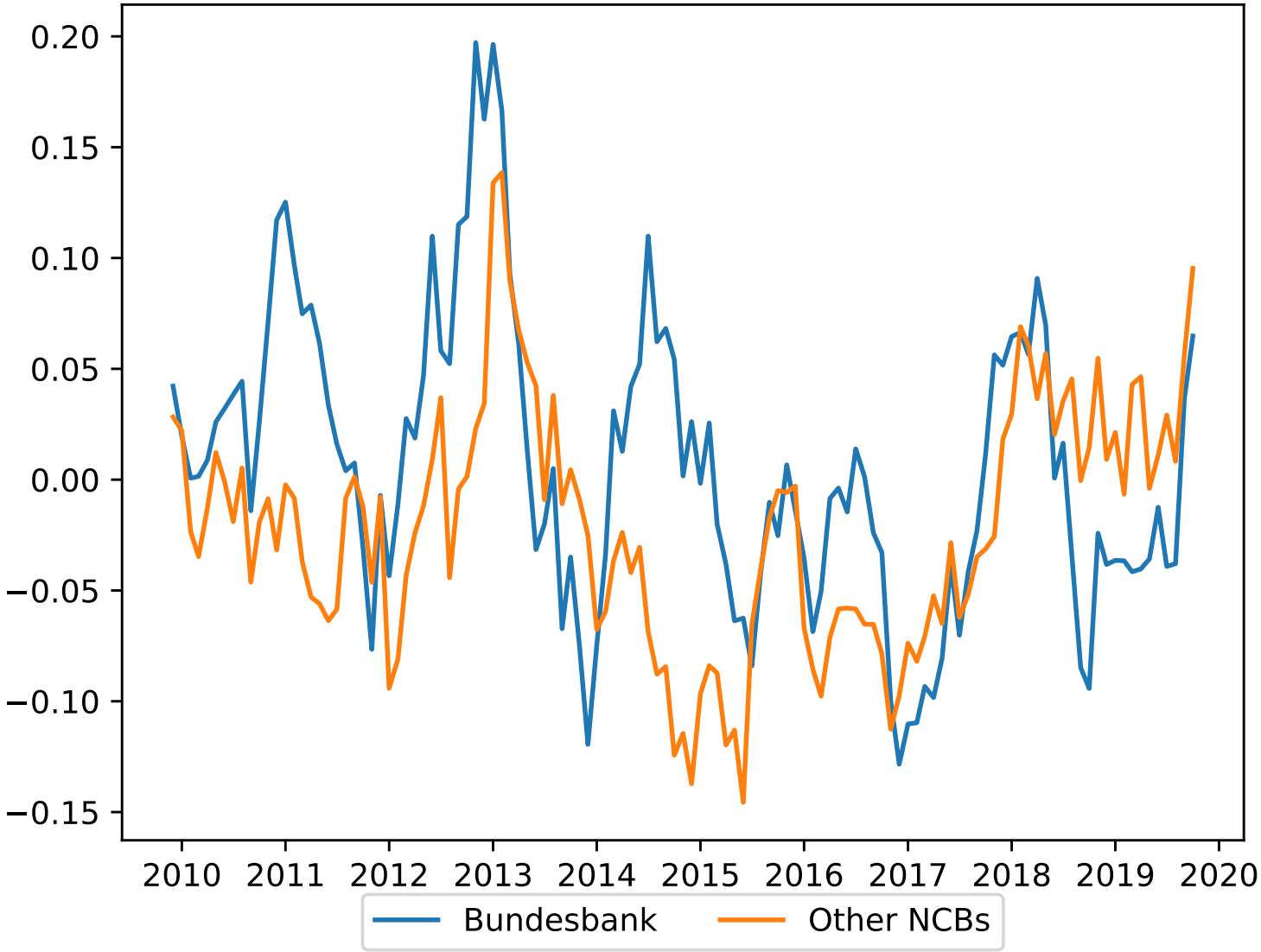


Figure: Sentiment difference to the ECB (moving average)

Demands

- presentation (slides: English, oral: English/German)
  - 30 minutes (Bachelor),
  - 45 minutes (Master),
- report of about 12 pages (English/German),
- participation in discussion and feedback,

- individual appointment to talk about the slides: at the latest one week before the day of the presentation itself,
- slides at least two days before our meeting.

Schedule

- appointment (in consultation with the participants) for a preliminary meeting to assign projects to participants,
- setting an individual schedule for presentations and discussions (in dependence of the number of participants),
- we prefer meetings of around three or four presentations per day,
- setting the deadline for the reports to a sufficiently late date at the end of the semester.

Application

For a non-binding application for the seminar please send an e-mail to

rieger@statistik.tu-dortmund.de

until

30.09.2020.