CENTER FOR
CIVIL SOCIETY
RESEARCH

Automated protest event analysis

Sophia Hunger

University of Bremen & WZB Berlin Social Science Center

Exciting times

- previously: data are **rare, hard to collect, expensive**
- humans produce **a lot of text**
- available **computational power**

What do we mean by “text-as-data”?

- Texts are used to communicate → contain information
- understanding texts as data, not main intention of the communication
- necessary: simplification
- imposing abstraction

On the shoulders of giants

- *"if I have seen further [than others], it is by standing on the shoulders of giants."*
- *need* to build on decades long research traditions and theories
- *provide* a toolkit to further develop established methods
- CSS in itself as *cumulative process*
- for instance: **Protest Event Analysis**

Protest Event Analysis

- modern societies as protest societies
- protest as non-institutional arena for attitudinal polarization and political conflict

Protest Event Analysis

- modern societies as protest societies
- protest as non-institutional arena for attitudinal polarization and political conflict

Protest Event Analysis

- standard methodology for the **systematic collection of protest events** among political scienstist sociologists
- often based on **media reporting**, i.e. newspaper articles, and - more recently - on social media
- human coding → **labour- and resource-intensive**

A more computational take on PEA: Automatisation

- protest event analysis based on **automated event extraction**
- **classifiers** to pre-select relevant texts from newspaper articles
- standard approaches and cutting-edge methods, such as transformer models
- **geo-referencing** based on Named-Entity-Recognition and GoogleMaps API
- combined with reduced human coding

The German Political Protest and Radicalization Monitoring - MOTRA

The German Political Protest and Radicalization Monitoring - MOTRA

Goals

- Systematic gathering of protest events across in length and depth across political phenomena, space, and time in Germany → demonstration, confrontational, violence
- Embedding of political protest and radicalization in public debates
- Understand radicalization of individuals and bridge to organized actors

Three data collection instruments

- Protest Event Analysis (PEA)
- Public debates in mass and social media
- Biographic profiles of radicalized protest actors

The German Political Protest and Radicalization Monitoring - MOTRA

Goals

- **Systematic gathering of protest events across in length and depth across political phenomena, space, and time →** demonstration, confrontational, violence
- Embedding of political protest and radicalization in public debates
- Understand radicalization of individuals and bridge to organized actors

Three data collection instruments

- **Protest Event Analysis (PEA) → AUTOMATIZATION**
- Public debates in mass and social media
- Biographic profiles of radicalized protest actors

Nationwide Protest Monitoring in Depth and Length - PEA

Research items of interest

- When, where, who, with which claims does protest occur?
- To what degree do we see politically motivated violence in the protest arena?
- Which spatial foci can we link with radicalized protest?

Nationwide Protest Monitoring in Depth and Length - PEA

Research items of interest

- When, where, who, with which claims does protest occur?
- To what degree do we see politically motivated violence in the protest arena?
- Which spatial foci can we link with radicalized protest?

Main challenge

- collection of protest event data from newspapers (Süddeutsche Zeitung)
- combine with ProDat data (1950-2000)
- many articles
- planned expansion with regional newspapers

Nationwide Protest Monitoring in Depth and Length - PEA

Data gathering

- Identification of protest events in print news media
  - Start and basis of the process: full corpus of **Süddeutsche Zeitung**
  - Extensions: taz, Die WELT, regional press, Junge Freiheit + police press releases

Nationwide Protest Monitoring in Depth and Length - PEA

Data gathering

- Identification of protest events in print news media
    - Start and basis of the process: full corpus of **Süddeutsche Zeitung**
    - Extensions: taz, Die WELT, regional press, Junge Freiheit + police press releases
- **Automatized classification** of articles covering protest events based on hand-coded training data → allowing study in length by pre-identifying relevant news

Nationwide Protest Monitoring in Depth and Length - PEA

Data gathering

- Identification of protest events in print news media
  - Start and basis of the process: full corpus of **Süddeutsche Zeitung**
  - Extensions: taz, Die WELT, regional press, Junge Freiheit + police press releases
- **Automatized classification** of articles covering protest events based on hand-coded training data → allowing study in length by pre-identifying relevant news
- Post-classification: **human coding** of protest form, actors' constellation, addressees and political claims

Protest Monitoring - PEA

Structure of pipeline

- collecting articles
- pre-filtering
- coding training data (needs to be balanced)
- training classifier
- apply classifier
- human coding
- de-depulication and cleaning

Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**

Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**
- looking for the needle in a haystack

Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**
- looking for the needle in a haystack
- **classifier:** an algorithm that automatically orders or categorizes data into one or more of a set of “classes.”

Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**
- looking for the needle in a haystack
- **classifier:** an algorithm that automatically orders or categorizes data into one or more of a set of “classes.”
- rules used by machines to classify data

Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**
- looking for the needle in a haystack
- **classifier:** an algorithm that automatically orders or categorizes data into one or more of a set of “classes.”
- rules used by machines to classify data
- binary or multi-class
- supervised and unsupervised classifiers: need for training data

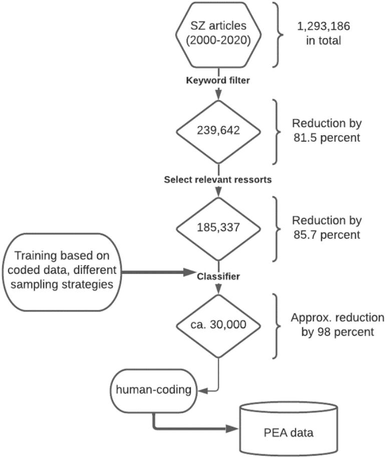
Automatization of PEA

The “Haystack”

- many documents → electronic news databases and computational tools
- first (important!) step: **Which documents contain protest?**
- looking for the needle in a haystack
- **classifier:** an algorithm that automatically orders or categorizes data into one or more of a set of “classes.”
- rules used by machines to classify data
- binary or multi-class
- supervised and unsupervised classifiers: need for training data
- example: Spam, Netflix suggestions

Protest Event Data generating process

Figure 1: PEA pipeline


SZ articles (2000-2020)
1,293,186 in total
Keyword filter
239,642
Reduction by 81.5 percent
Select relevant ressorts
185,337
Reduction by 85.7 percent
Training based on coded data, different sampling strategies
Classifier
ca. 30,000
Approx. reduction by 98 percent
human-coding
PEA data

•

Variables of interest

Included Variables:

- Location of protest
  - Categorical: in Germany or abroad
  - If in Germany: state and city
- Protest offline or online
- Protest event: selection of all relevant sentences (strings)
- Action form: categorical and exact wording as string
- Claim, Actors, Addressee
- Date of protest event: date and exact wording as string

Variables of interest

Included Variables:

- Location of protest
  - Categorical: in Germany or abroad
  - If in Germany: state and city
- Protest offline or online
- Protest event: selection of all relevant sentences (strings)
- Action form: categorical and exact wording as string
- Claim, Actors, Addressee
- Date of protest event: date and exact wording as string

Carried out by a team of ca. 4 RAs in the PolDem Coding-Tool

Pre-filtering

Protest keywords

- e.g. *protest*, politically motivated*, (extreme left*, extreme right*, racist*, islamist*, antisemiti*) AND (motivated*, background, criminal*), confessor*, Staatsschutz*, demonstr*, *demo, *demos, *manifestation*, torch parade*, *march*, human chain*, *bomb*, Molotow*, graffiti*, arson*, *attack*, *attacks*, *graffiti*, graffiti*, hostage, *attack*, *terror*, *assault*, *attack*, death threat*, hate message*, hate mail*, threatening mail*, threatening letter*, etc.

Pre-Filtering

Excluded sections

- Forum & Leserbriefe, Geld & Technik, Hobby, Immobilien, JETZT.DE, jetzt.muenchen, Jugend, Schule, Berg- und Ski-Journal, Kinder- und Jugendliteratur, Kinder- und Jugendmedien, Kinderseiten, Kunstmarkt, Literatur, Literaturbeilage, Meinungsseite, Mietmarkt, Mobiles Leben, Mode, Reise, Zeitung in der Schule, Zeitvertreib, Sport

Balancing pipeline

▷ With a list of German *places* from Wikipedia we balance the training towards domestic events: mentioning one German location and being reported from Germany

Balancing pipeline

▷ With a list of German *places* from Wikipedia we balance the training towards domestic events: mentioning one German location and being reported from Germany

Location Entity Recognition & GoogleMaps API

- Identify locations with Name Entity Recognition-NER in spacyr:: to benefit from a Python-based pre-trained model (de_core_news_sm)
- Send request through GoogleMaps API to obtain latitude and longitude
- Is the "place" in Germany?

Protest Event Data training samples

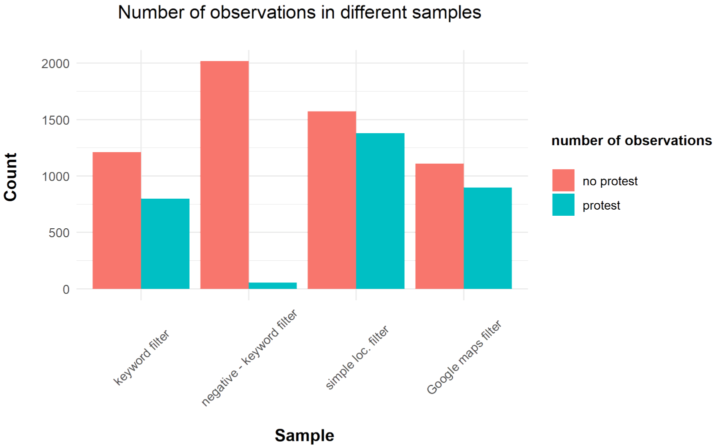
Table 1: Sampling and Balancing Strategy over Rounds

Sample	Sampling selection strategy	N	
1	Ressorts + Protest Keywords	2712	
2	Negative Cases	2107	
3	1 + List of German Locations	2045	
4	1 + Location Entity Recognition & GoogleMaps & narrative differentiation	1979	
		Total=	8661

▷ Goal: looking for the needle in the haystack

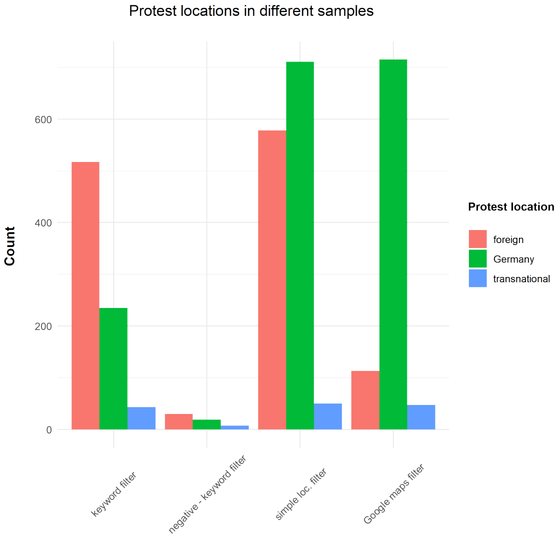
Sample balancing - validation

Figure 2: Observations by sample



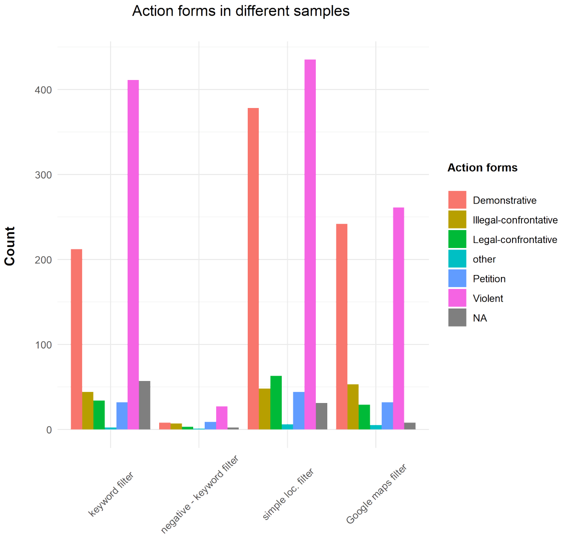
Sample balancing - validation

Figure 3: Protest location by sample



Sample balancing - validation

Figure 4: Action form by sample



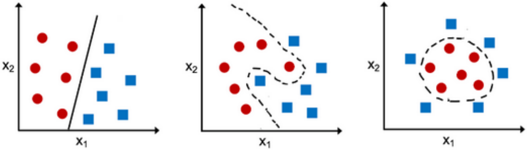
Special additional variable

Protest narratives

- Beyond differentiating from *concrete* protest events and just articles containing protest keywords (e.g. "MP Weidel protested in the parliamentary session..."), we introduce "narratives"
- **Narratives**: actual content on protest dynamics, however, only mentioning references without talking about concrete events, e.g. "The 1970s protests changed politics" or "Terrorist threats remain latent in France..."
- Expectation: generate a fine-grained differentiation in semantics by discriminating between concrete "acts" and references; the latter being able to contaminate negative articles including *other* protest wording

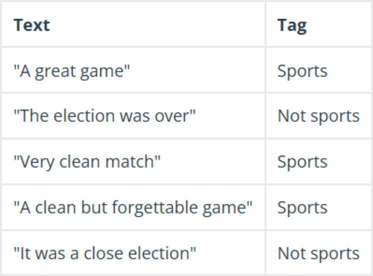
Machine learning classifiers

Figure 5: Binary classification problem



Machine learning classifiers

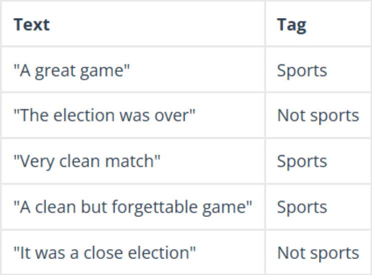
Figure 6: Naive Bayes



Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Machine learning classifiers

Figure 6: Naive Bayes



Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Figure 7: What about this?

$$P(a\,very\,close\,game) = P(a) \times P(very) \times P(close) \times P(game)$$

Automatized classification of German events in news: machine learning predictions

Machine learning models

- Different performance across classic quanteda.textmodels and tidymodels in R, and Python models with nltk and skitlearn
- Overall tested: Random Forest, SVM, Logistic Regressions, Naive Bayes, Lasso regressions

Automatized classification of German events in news: machine learning predictions

Machine learning models

- Different performance across classic quanteda.textmodels and tidymodels in R, and Python models with nltk and skitlearn
- Overall tested: Random Forest, SVM, Logistic Regressions, Naive Bayes, Lasso regressions

Pre-processing of new text corpus

- Removal of numbers, symbols, punctuation
- Removal of German stop-words based on quanteda + extended list by Götze (2019) + German-based lemmatization

Automatized classification of German events in news: machine learning predictions

Machine learning models

- Different performance across classic quanteda.textmodels and tidymodels in R, and Python models with nltk and skitlearn
- Overall tested: Random Forest, SVM, Logistic Regressions, Naive Bayes, Lasso regressions

Pre-processing of new text corpus

- Removal of numbers, symbols, punctuation
- Removal of German stop-words based on quanteda + extended list by Götze (2019) + German-based lemmatization

Best performance: lemmatized Naive Bayes classifier with quanteda (multinomial distribution specification) reaching specificity of 0.74 and precision of 0.36

Post-classification coding

Human coding including

- Actors
- Number of participants
- Demands/claims
  - based on updated Prodat issue list
  - directional
- Addressees

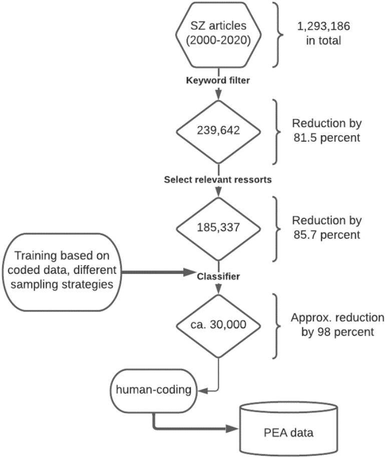
Post-classification coding

Human coding including

- Actors
- Number of participants
- Demands/claims
  - based on updated Prodat issue list
  - directional
- Addressees

Protest Event Data generating process

Figure 8: PEA pipeline


SZ articles (2000-2020)
1,293,186 in total
Keyword filter
239,642
Reduction by 81.5 percent
Select relevant ressorts
185,337
Reduction by 85.7 percent
Training based on coded data, different sampling strategies
Classifier
ca. 30,000
Approx. reduction by 98 percent
human-coding
PEA data

What to do with this?

The bridging power of text-as-data

- highlights the value of description
- leaves room for discovery
- text-based automated analyses can connect quantitative and qualitative scholars

Protest Event Analysis

Research interests 2020/2021

- What was the influence of the pandemic on protests in Germany?
- Covid protests: Formation of a new movement? Radicalization?
- Was the new "Querdenken" movement able to dominate German street protests?
- Other movements: Continuous constraints or adaptation to a "new normal"?

Main findings

Development and radicalization of the German protest landscape

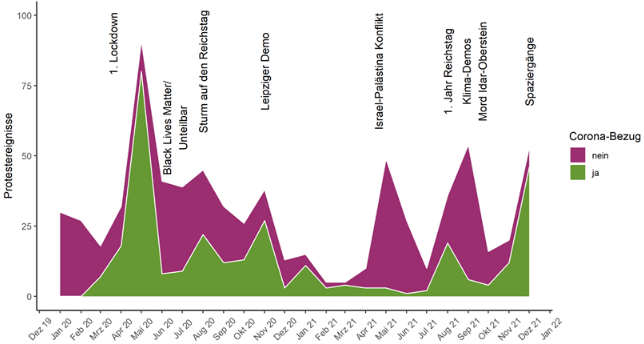
- **Re-mobilization** of the protest arena in the second year of the pandemic - still marked by anti-containment protests
- parallel **radicalization of the action repertoire** - "Querdenker" and radical right as main drivers of street radicalization

Core findings

- **Street protest:** still the **central means** of collective expression of opinion during the pandemic
- **Radicalization** of protest dynamics during the second year of the pandemic
- Dominance of Covid protests mainly in 2020
    - Institutionalization of ”Querdenken”
    - but no complete displacement of other movements
- Relative **re-mobilization** of the protest landscape in 2021
    - fewer restrictions and politicization in the context of the German federal elections

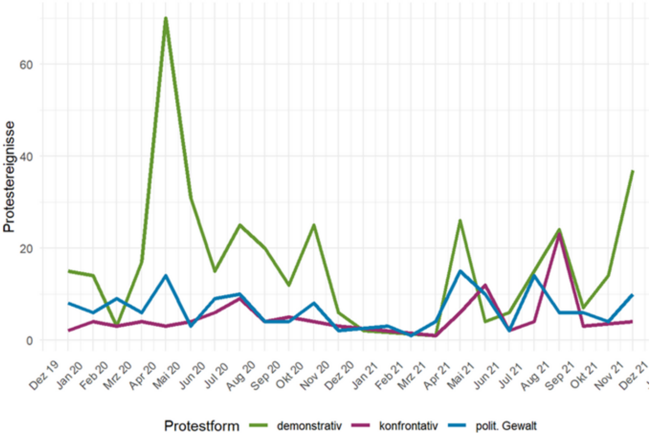
2020 and 2021: Covid and other protests in comparison

Figure 9: PEs in Germany



Forms of protest in comparison

Figure 10: Protest forms



Forms of protest in comparison

- still (mainly in 2020) dominance of demonstrations
- 2021: relative radicalization of the action repertoire
- driven by which issues?

Main issues

Dominance of Covid/"Querdenken" protest (43 percent)

- Mobilization in waves
- Mass events but also increasingly confrontational and violent

Environment and mobility

- Less confrontational

Integration, migration and racism

- Mobilization on both sides
- Often violent: Right-wing extremism and Islamism

Combination with ProDat

Combination with ProDat

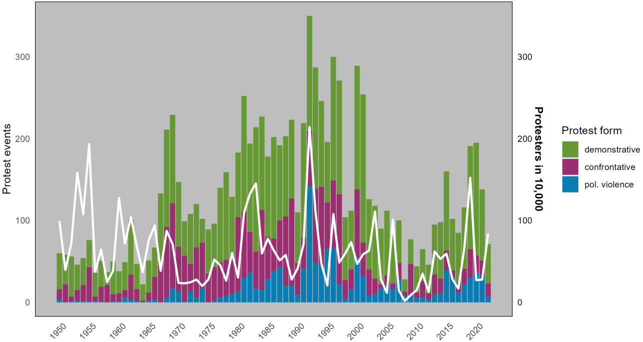


Figure 11: Protest 1950 - 2022

Many thanks for your attention!

sophia.hunger@uni-bremen.de