gesis Leibniz-Institut für Sozialwissenschaften

hhu Heinrich Heine Universität Düsseldorf

Interlinking Social Science Research Data and Social Media Data

RQ: How can survey data be *interlinked* and *complemented* with data from different sources?

The collection of social science research data is expensive. Therefor, secondary research and exploitation of new information sources are important. This work presents two machine learning based approaches to help overcome current challenges.

Finding and reusing survey questions and items

- Interlinking of social science survey items accross surveys.
- Foster reuse of data by extending search applications and data catalogs with new features
- >100 000 suitable survey items from the GESIS archive
- Taxonomy and vocabulary for the description of survey items, so called question features
- Multiclass classification for the Information Type feature

Q: „How would you rate the current economic conditions in Germany?“

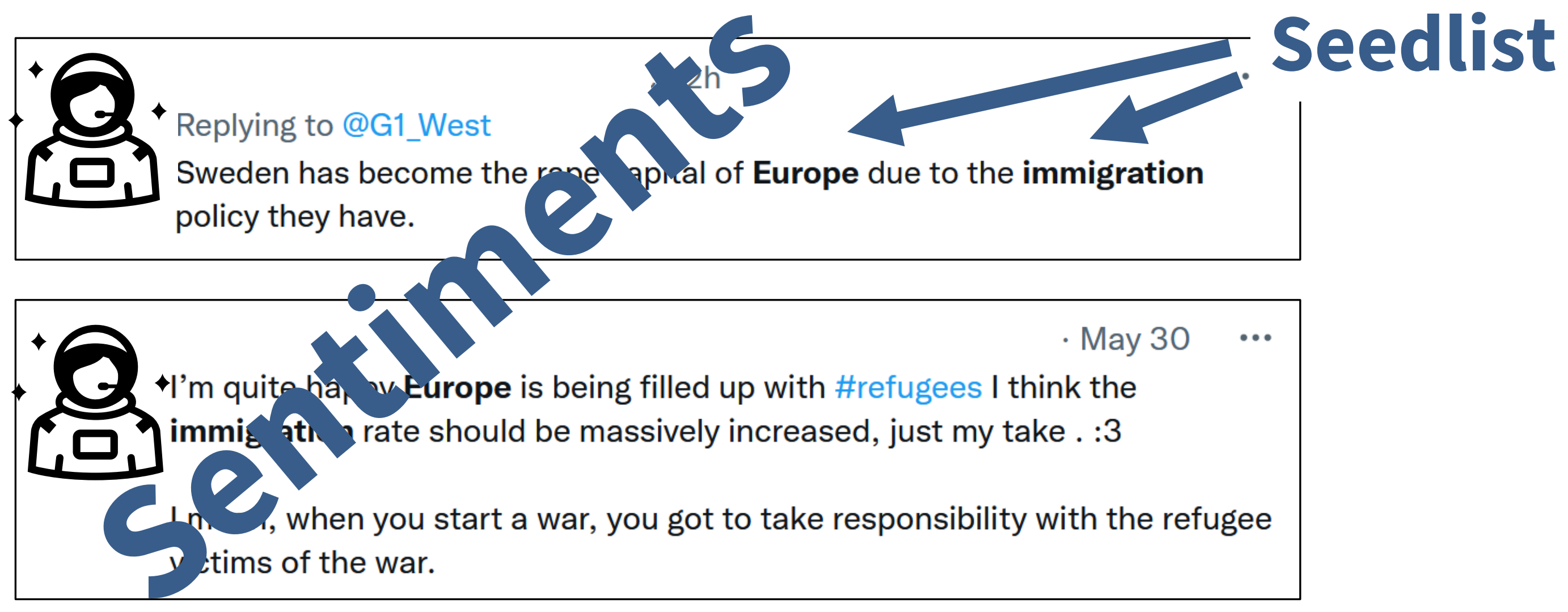
Information Type	[✓ Evaluation]	Fact	Cognition
Focus	Self-focus	Family-member	[✓ Object-focus]
Time Reference	Past	[✓ Present]	Future
Periodicity	[✓ Point in time]	Time span	Periodic p.
Relative Location	Apartment	Neighborhood	Country
Geo. Location	<Continent>	[✓ <Country>]	<City>

Example from ALLBUS '18

Approach	L1 (3 classes)		L2 (9 classes)	
	micro-f1	macro-f1	micro-f1	macro-f1
LSTM	**0.7644**	**0.7462**	**0.5764**	**0.5370**
SVC	0.5934	0.4593	0.3991	0.2097
RF	0.5482	0.5241	0.3636	0.2423
LogReg	0.5416	0.5232	0.3368	0.2682
MNB	0.5207	0.5135	0.3373	0.2563

Example from ALLBUS '18

Extension with social media discourse data



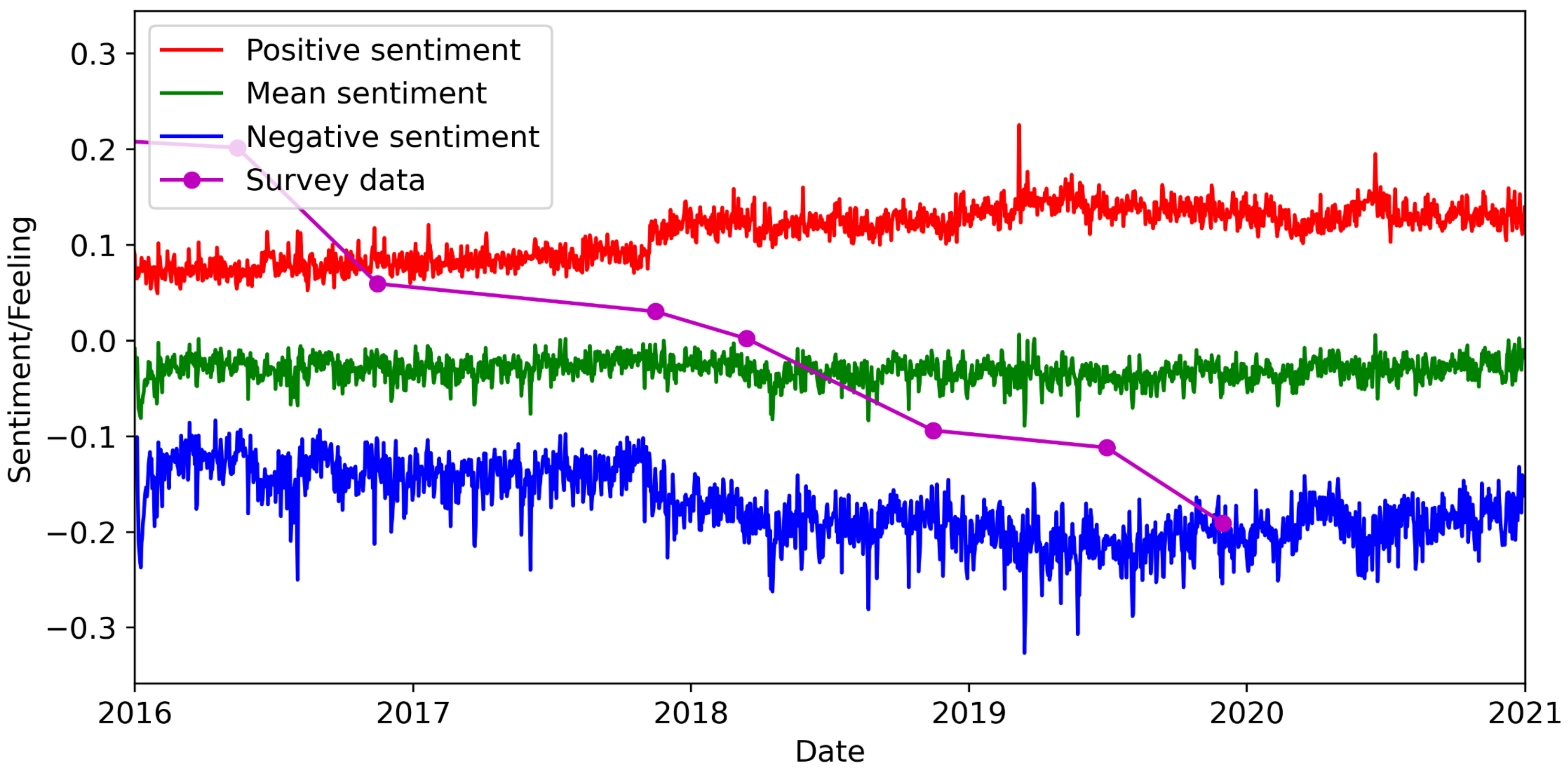
Replying to @G1_West
Sweden has become the rape capital of **Europe** due to the **immigration** policy they have.

· May 30 ···
I'm quite happy **Europe** is being filled up with #refugees I think the **immigration** rate should be massively increased, just my take . :3

I mean, when you start a war, you got to take responsibility with the refugee victims of the war.

Extraction of immigration related tweets based on seed terms

- Micropost retrieval from Twitter for a given topic
- Reduction of humanly-introduced bias (vocabulary mismatch)
- Systematic evaluation of methods in extraction pipeline
- Based on large scale data from the TweetsKB Twitter archive (10 billion tweets, since 2013)
- Generation of seedlists based on term frequency and semantic similarity (word embeddings)
- Operationalization through sentiment analysis
- Restriction to certain demographic groups



Sentiment of immigration tweets in the UK 2016-2021

Machine learning has many applications in the area of interlinking survey data. Thereby, a major challenge will be to preserve acceptance by social science researchers. Provenance and transparency are key aspects. Also, a comprehensive evaluation of the applied ML methods will be required.

Contact: Felix Bensmann (felix.bensmann@gesis.org)

Leibniz Gemeinschaft