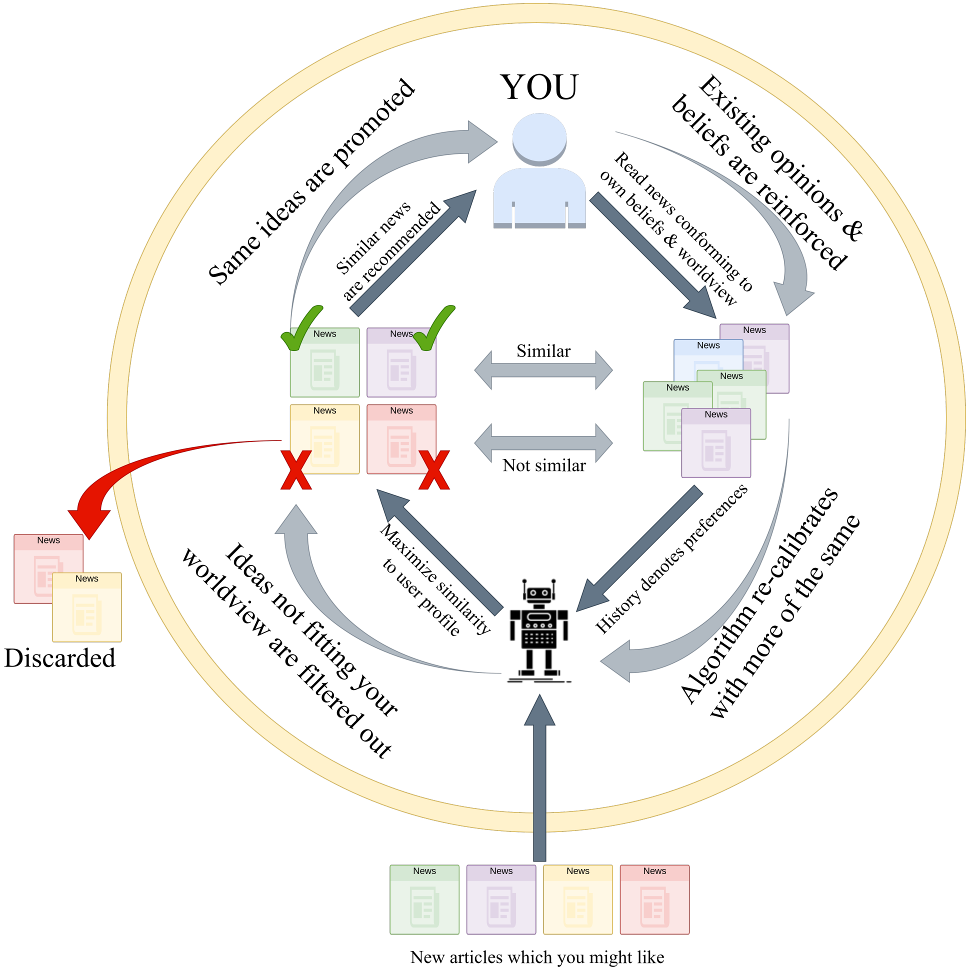
NeMig – A Bilingual News Collection and Knowledge Graph about Migration

Andreea Iana[1], Mehwish Alam[2], Alexander Grote[3], Nevena Nikolajevic[3], Katharina Ludwig[4], Philipp Müller[4], Christof Weinhardt[3], Heiko Paulheim[1]

[1] Data and Web Science Group, University of Mannheim
[2] Télécom Paris, Institut Polytechnique de Paris
[3] Karlsruhe Institute of Techonology
[4] Institute for Media and Communicaton Studies, University of Mannheim

News Recommendation



News Datasets

News recommendation benchmarks

- ✚ Large-scale datasets
- ✚ Rich user click behavior
- − General / less sensitive topics
- − Few news & user features, such as sentiments, stances, political info, ...

News data for media discourse analysis

- ✚ Rich information for fake news detection, news bias detection, ...
- − No user feedback data

News Datasets

News recommendation benchmarks

- ➕ Large-scale datasets
- ➕ Rich user click behavior
- ➖ General / less sensitive topics
- ➖ Few news & user features, such as sentiments, stances, political info, ...

News data for media discourse analysis

- ➕ Rich information for fake news detection, news bias detection, ...
- ➖ No user feedback data

NeMig: news data

- ➕ Polarizing topic
- ➕ Sentiment and political annotations

NeMig: user data

- ➕ Explicit click feedback
- ➕ Demographics & political preferences

Data Collection

Topic

- … on which most people have a somewhat **clear opinion**
- … which **divides partisans along party lines**

→ **refugees & migration**

News outlets

- ... feature a heterogeneous set of news articles on the topic

→ mix of **legacy & alternative** sources spanning **entire political spectrum**

Multiple languages

- **German & English**

Data Collection: English

Source	Inclusion Criteria	Exclusion Criteria
• **45 outlets** • Keywords: refugee*, asylum seeker*, migrant*, immigrant*, asylum applicant*, asylee*, person seeking asylum*, displaced person, displaced people, deportaton, immigration*	• Article should contain **at least 2 keywords**, separated by **min. 50 words** • Article length: **min. 150 words** • Published between **01.01.2021 - 01.07.2022**	• Paid articles • Non-English articles • Disclaimers, advertisements, buying options, reader comments, etc. • Announcements about publications, TV programs or movie or book recommendations etc.

Data Collection: German

Source	Inclusion Criteria	Exclusion Criteria

Source

- **40 outlets**
- Keywords: flüchtl*, geflücht*, asyl*, zuwander*, einwander*, immigrant*, immigration*, migration*, migrant*, ausländer, refug*, rapefug*, invasor*

Inclusion Criteria

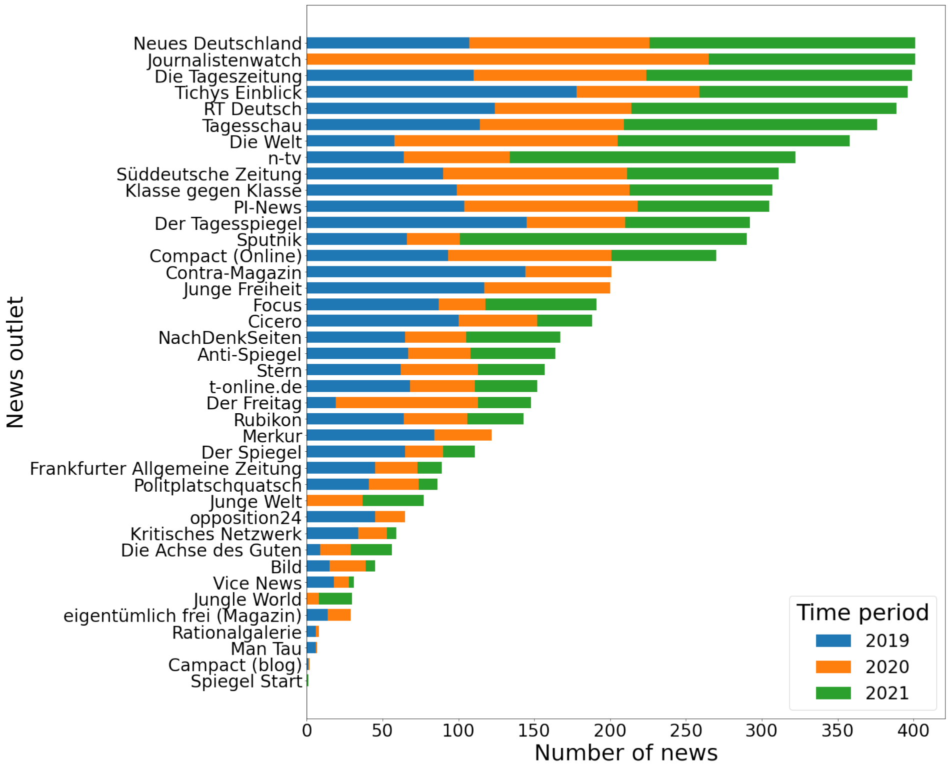
- Article should contain **at least 2 keywords**, separated by **min. 50 words**
- Article length: **min. 150 words**
- Published between **01.01.2019 - 31.12.2021**

Exclusion Criteria

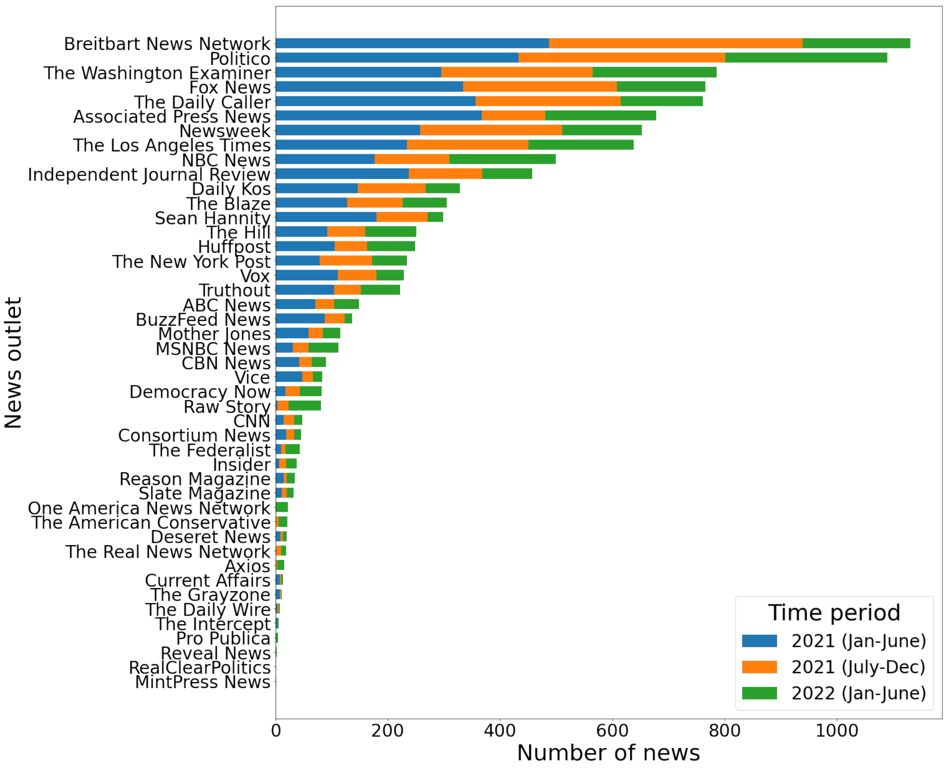
- Paid articles
- Non-German articles
- Disclaimers, advertisements, buying options, reader comments, etc.
- Announcements about publications, TV programs or movie or book recommendations etc.

Data Collection

German: **7,346 news**



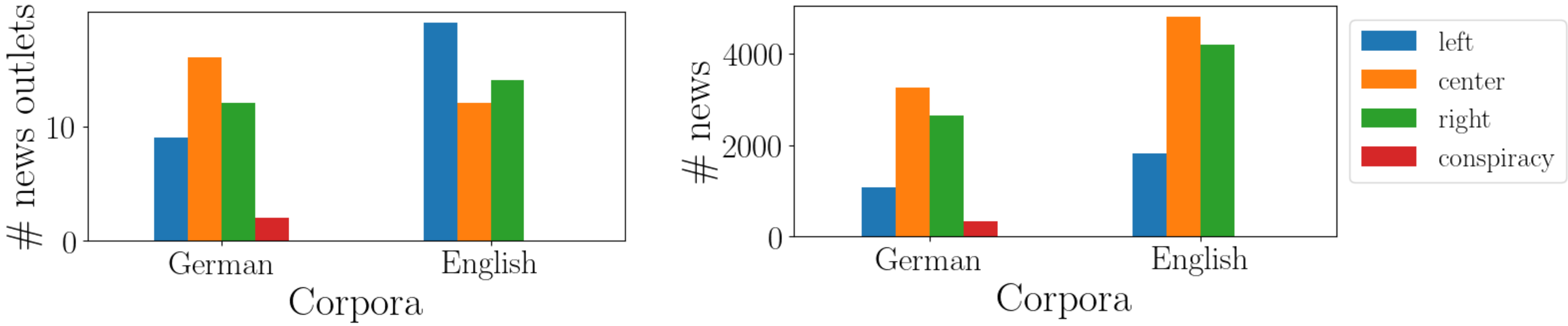
English: **10,814 news**



Data Annotation: Political Leaning

Political leaning of **news outlets** (~ proxy for articles)

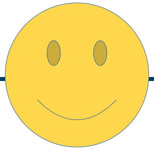
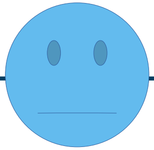
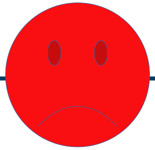
- English: classification using AllSides Media Bias Chart[1]
- German: no official categorization exists → classification by researcher team from media & communication domain



Data Annotation: Sentiment Analysis

Sentiment Classifier

- Multilingual XLM-R[2] trained on Tweets & fine-tuned for sentiment analysis
- Input: news title + abstract

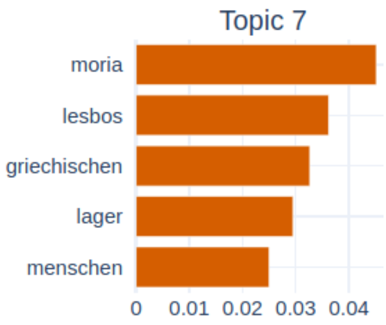
			
German	144	3,726	3,476
English	590	2,933	7,291

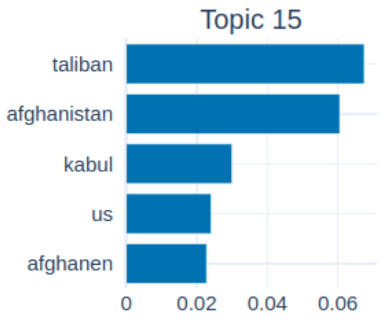
Data Annotation: Sub-topic Modelling

BERTopic[3]: create dense clusters of news embeddings

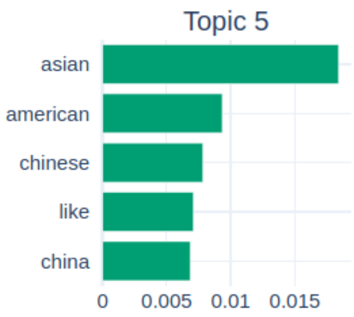
- Embeddings generated w/ pre-trained English Sentence Transformer[4] (English) or multilingual Sentence Transformer (German)
- Topic representations assigned w/ class-based TF-IDF approach

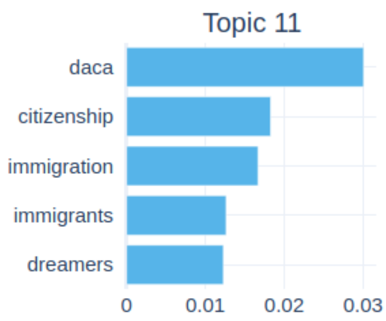
German: **25 sub-topics**





English: **40 sub-topics**





Data Annotation: Entity Recognition & Linking

Events described using named entities (in text & metadata)

Named Entity Recognition → Entity Linking → Entity Filtering

Pre-trained XLM-R fine-tuned on German / English CoNNL03[5]

Linkage to Wikidata[6] w/ multilingual seq2seq entity linking model[7]

Filter incorectly extracted or linked entities based on model's confidence & Wikidata properties

	German			English		
	Total	Linked	Not linked	Total	Linked	Not linked
Title	1,075	1,075	0	1,542	1,542	0
Abstract	2,383	2,383	0	2,365	2,365	0
Body	19,683	19,683	0	33,857	33,857	0
Publishers	40	34	6	45	44	1
Authors	490	193	297	1,242	386	856
Keywords	4,481	2,395	2,086	5,175	2,636	2,539

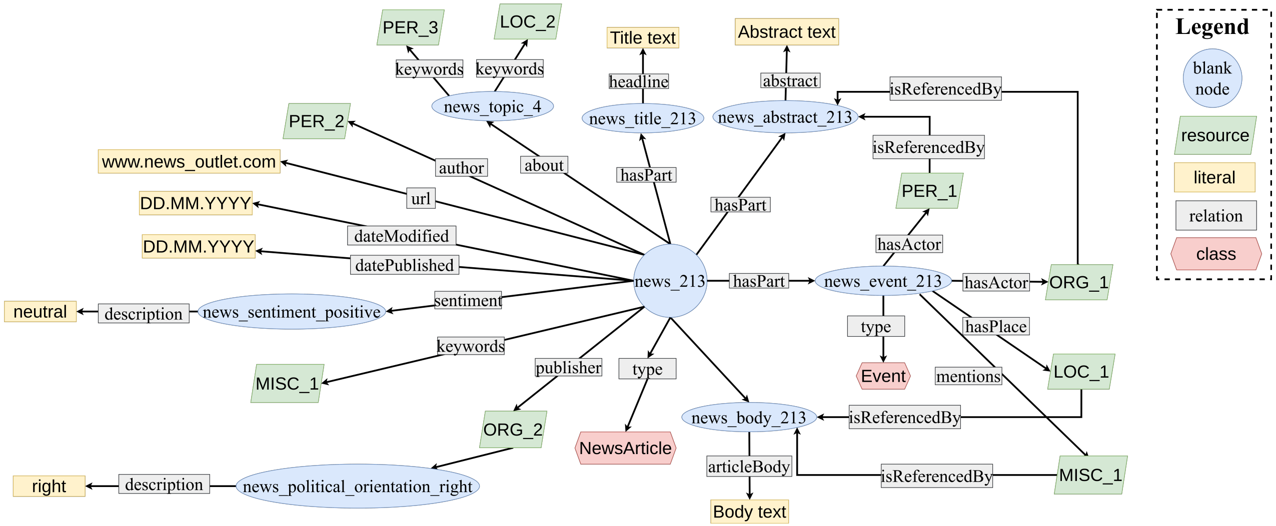
NeMigKG

Nodes

- **Literals**: titles, dates, …
- **Resources**: *Wikidata resources* (if linked) or *custom resources* (otherwise)

Relations: reusing *established* ontologies & schemas

Enrichment with external knowledge: up to **2-hop Wikidata neighbors** of linked entities

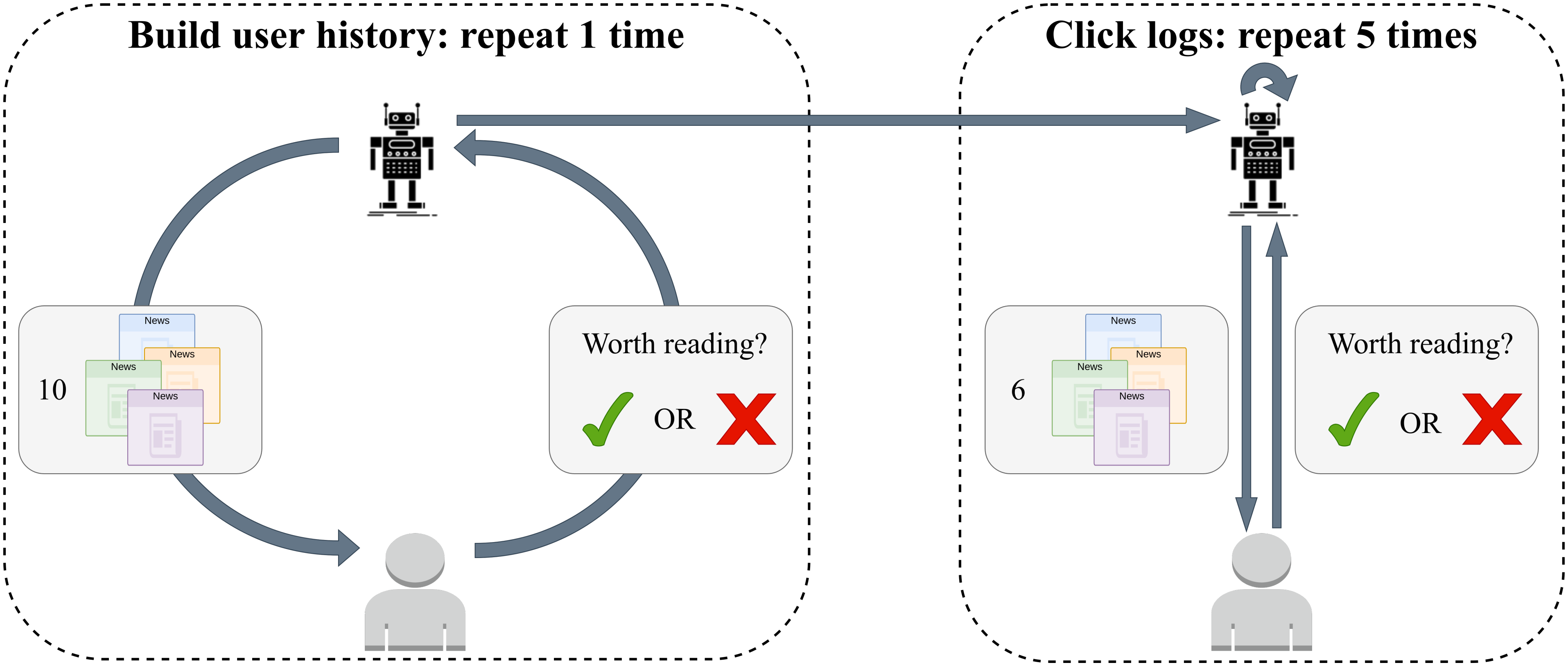


User Data Collection: Online Study

Goal: collect implicit & explicit user feedback

Intro	Stimulus	Outro
• Media use • Political standpoint • Political interest • Disenchantment with issues • Empathy	• Feedback collection	• Perceived filter bubble • Online political participation • Polarisation (affective, perceived, ideological) • Prosocial behavior • Demographics

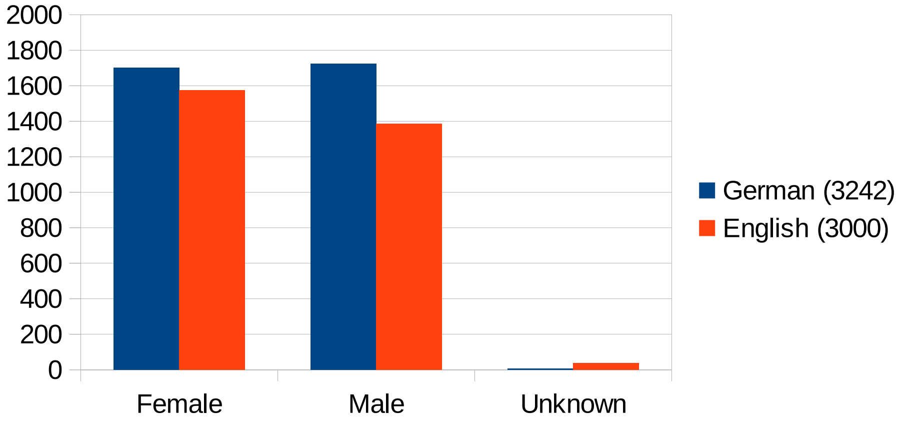
User Data Collection: Online Study



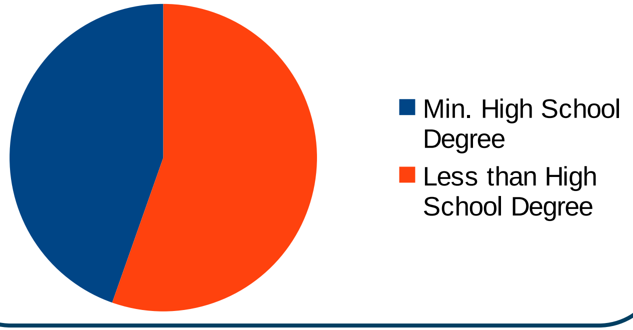
User Data Collection: Online Study

Partipants

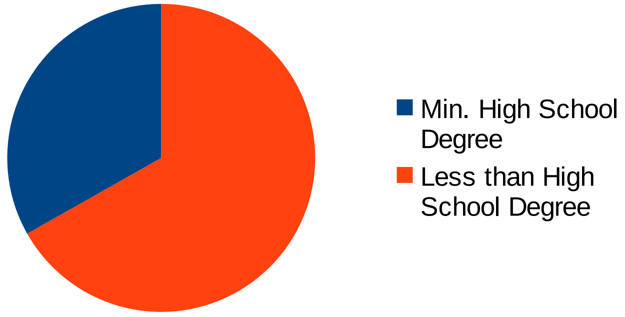
- Recruited through **online-access panels**
- Selected using a **quote procedure** to create a **representative sample**
- German & US Internet users
- 18 to 74 years old



German: 3242



English: 3000



Benchmarking: Recommendation Models

Variety of model architectures (news & user encoders, training objectives)

- Content personalization: NRMS[8]
- Content + aspect-based personalization: NAML[9], MINS[10], TANR[11]
- Candidate-aware user modeling: CAUM[12]
- Knowledge-aware model: DKN[13]
- Sentiment-debiasing objective: SentiDebias[14]

Benchmarking: Content Personalization

	German				English			
Model	AUC	MRR	nDCG@3	nDCG@6	AUC	MRR	nDCG@3	nDCG@6
NRMS	57.96±0.66	47.83±0.46	44.90±0.65	60.49±0.35	52.40±0.89	41.46±0.59	37.49±1.07	54.21±0.53
NAML	50.49±0.13	47.37±0.84	44.27±1.20	60.13±0.65	50.02±0.20	41.82±0.70	38.14±0.84	54.49±0.62
MINS	57.70±0.53	47.85±0.78	44.96±0.33	60.50±0.71	52.91±0.68	41.96±1.13	38.19±0.41	54.59±0.41
CAUM	58.18±0.85	48.17±1.08	45.27±0.60	60.74±0.55	53.10±0.78	41.83±0.93	38.16±1.42	54.50±0.32
DKN	**58.73±0.39**	**48.20±0.39**	**45.48±0.43**	**60.78±0.29**	**53.69±0.66**	**42.14±0.66**	**38.57±0.83**	**54.74±0.55**
TANR	50.62±0.14	47.28±0.44	44.22±0.51	60.06±0.33	50.17±0.10	41.82±0.12	38.26±0.33	54.49±0.17
SentiDebias	56.80±0.30	47.27±0.44	44.09±0.18	60.04±0.44	52.63±0.90	41.63±1.34	37.58±0.13	54.33±0.58

Models with similar architectures perform similarly regardless of training objectives

→ secondary optimization goals might have little effect on recommendation performance

External knowledge might benefit user modeling in setups with scarce training data

Quality of content personalization influenced by data sparsity

→ 0.02% sparsity for German vs. 9.27% sparsity for English

Benchmarking: Aspect-based Diversity

Aspect-based diversity[15]: level of uniformity of an aspect's distribution among recommendations

Evaluated w/ normalized entropy of aspect A_p's distribution in the recommendation list

class j of aspect A_p

$$D_{A_p}@k = -\sum_{j \in A_p} \frac{p(j) \log p(j)}{\log(|A_p|)}$$

#classes of aspect A_p

Benchmarking: Aspect-based Diversity

	German				English			
Model	nDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3	nDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3
NRMS	44.90±0.65	18.55±0.25	33.75±0.54	30.68±0.32	37.49±1.07	22.19±0.34	32.50±0.27	26.31±0.94
NAML	44.27±1.20	**19.31±0.21**	33.63±0.18	30.41±0.37	38.14±0.84	**23.89±0.65**	32.08±0.64	25.66±1.24
MINS	44.96±0.33	18.18±0.44	33.79±0.34	30.10±0.44	38.19±0.41	21.94±1.11	32.34±0.16	25.22±1.11
CAUM	45.27±0.60	18.09±0.33	33.13±0.43	30.42±0.33	38.16±1.42	21.90±0.63	32.09±0.26	25.90±0.63
DKN	**45.48±0.43**	18.46±0.16	33.79±0.44	30.92±0.59	**38.57±0.83**	22.04±0.40	32.46±0.66	**26.58±0.90**
TANR	44.22±0.51	18.52±0.18	33.70±0.36	30.39±0.38	38.26±0.33	21.59±0.19	32.41±0.68	26.47±0.82
SentiDebias	44.09±0.18	18.66±0.30	**33.84±0.34**	**30.99±0.30**	37.58±0.13	22.10±1.11	**33.26±0.34**	25.81±1.11

Sentiment debiasing model achieves highest sentiment-based diversity

Political diversification nearly identical for all models

Summary & Outlook

Status Quo

News recommenders control users' perception & access to information

Existing data resources lack user data or features for analyzing biases in news curation algorithms

NeMig

Datasets in German & English about refugees & migration

News data: textual content + metadata + sentiment + political orientation

User data: explicit click feedback + demographics + political information

Outlook

Analyze the multidimensional implications of news curation algorihms in monolingual & cross-lingual scenarios (i.e., filter bubbles, bias towards certain aspects, diversification of recommendations)

Generate synthethic user datasets w/ political information

Thank you!

andreea.iana@uni-mannheim.de

https://twitter.com/iana_andreea

https://zenodo.org/record/7908392

References

[1] https://www.allsides.com/media-bias/media-bias-chart

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[3] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[4] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[5] Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003)

[6] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[7] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics* 10 (2022), 274–290.

[8] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6389–6394.

[9] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3863–3869.

References

[10] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022. News recommendation via multi-interest news sequence modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7942–7946.

[11] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*. 1154–1159.

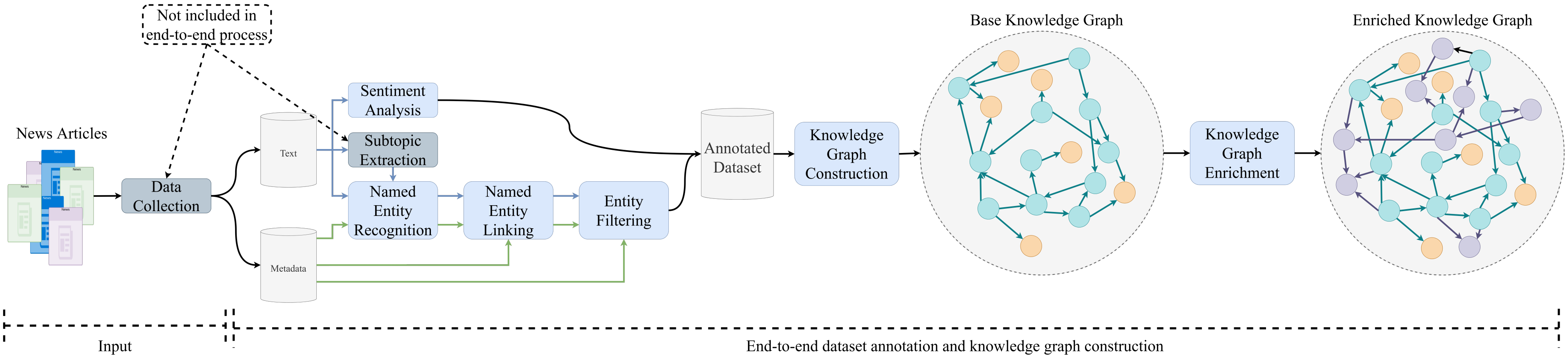
[12] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News recommendation with candidate-aware user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[13] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*. 1835–1844.

[14] Chuhan Wu, Fangzhao Wu, Tao Qi, Wei-Qiang Zhang, Xing Xie, and Yongfeng Huang. 2022. Removing AI's sentiment manipulation of personalized news delivery. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–9.

[15] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation. *arXiv:2307.16089*

Data Annotation



NeMigKG

Graph types

- **Base**: literals + entities
- **Entities**: only *resource nodes* from base KG
- **Enriched entities**: only *resource nodes* + *k-hop triples* from Wikidata
- **Complete**: enriched entities KG + literals

		German NeMigKG					English NeMigKG				
		Base	Entities	Enriched ($k = 1$)	Enriched ($k = 2$)	Complete	Base	Entities	Enriched ($k = 1$)	Enriched ($k = 2$)	Complete
Triples	Nodes	89,601	59,480	97,021	367,514	397,635	134,775	91,833	138,734	433,965	476,907
	Relations	19	11	768	1,172	1,180	19	11	828	1,195	1,203
	Triples	516,948	458,482	821,529	2,846,849	2,905,315	933,801	8 47,695	1,354,454	3,599,561	3,685,667
Nodes	% resources	66.38	100	100	100	92.42	68.14	100	100	100	91.00
	% literals	33.62	0	0	0	7.58	31.86	0	0	0	9.00
Resources	% blank	61.28	61.29	37.57	9.92	9.92	58.48	58.48	38.71	12.38	12.38
	% custom (not linked)	4.00	4.00	2.45	0.65	0.65	3.68	3.68	2.43	0.78	0.78
	% Wikidata (linked)	34.71	34.71	59.97	89.43	89.43	37.84	37.84	58.86	86.85	86.85

Benchmarking: Aspect-based Personalization

Aspect-based personalization[16]: level of homogeneity between a user's recommendations and history w.r.t. an aspect's distribution

Evaluated w/ generalized Jaccard similarity

$$PS_{A_p}@k = \frac{\sum_{j=1}^{|A_p|} \min(\mathcal{R}_j, \mathcal{H}_j)}{\sum_{j=1}^{|A_p|} \max(\mathcal{R}_j, \mathcal{H}_j)}$$

probability of news with class j of A_p to be in the recommendation list R

probability of news with class j of A_p to be in the user's history H

Benchmarking: Aspect-based Personalization

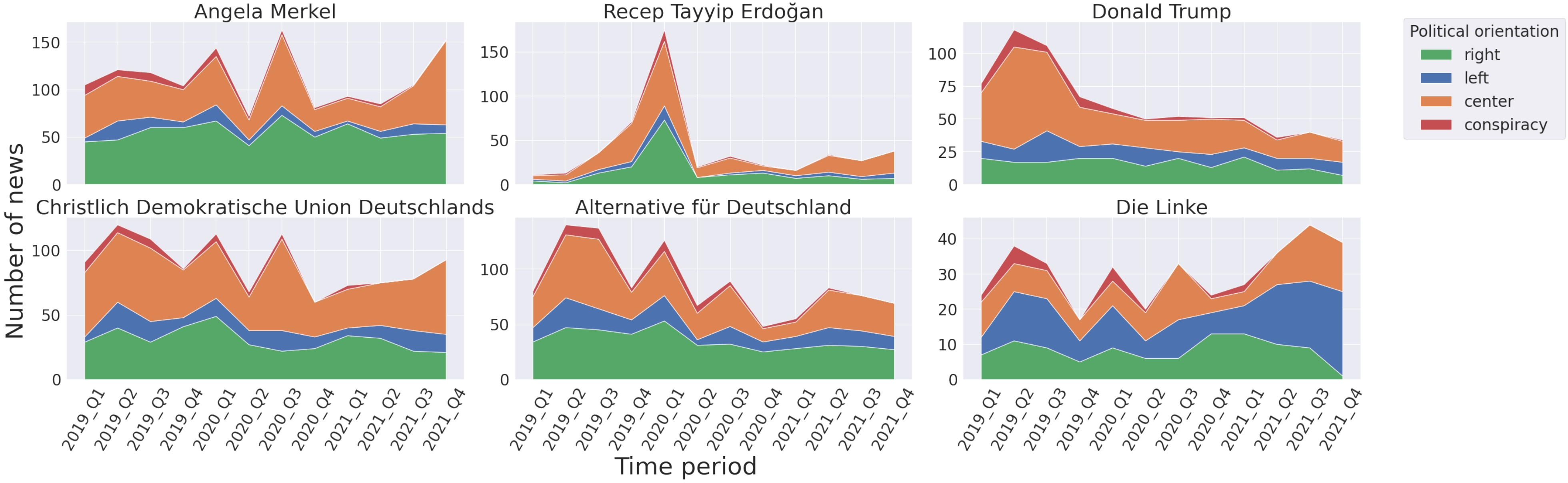
Model	German nDCG@3	German PS_{ctg}@3	German PS_{snt}@3	German PS_{pol}@3	English nDCG@3	English PS_{ctg}@3	English PS_{snt}@3	English PS_{pol}@3
NRMS	44.90±0.65	20.72±0.45	**42.66±0.30**	36.55±0.09	37.49±1.07	18.48±0.18	41.38±0.40	40.66±0.22
NAML	44.27±1.20	20.42±0.31	42.59±0.25	36.23±0.32	38.14±0.84	18.24±0.81	**41.43±0.19**	**41.17±0.30**
MINS	44.96±0.33	**21.10±0.60**	42.56±0.34	**36.74±0.27**	38.19±0.41	18.83±0.41	41.20±0.16	41.15±0.21
CAUM	45.27±0.60	20.85±0.79	42.48±0.43	36.59±0.36	38.16±1.42	18.67±1.45	41.41±0.26	40.95±0.41
DKN	**45.48±0.43**	20.95±0.35	42.51±0.36	36.08±0.24	**38.57±0.83**	18.56±0.35	41.25±0.14	40.70±0.15
TANR	44.22±0.51	20.88±0.24	42.64±0.19	36.42±0.34	38.26±0.33	**19.10±0.29**	41.27±0.24	40.98±0.20
SentiDebias	44.09±0.18	20.50±0.39	42.64±0.34	36.40±0.24	37.58±0.13	18.57±0.33	41.29±0.34	40.84±0.20

More diversity comes at the cost of personalization

Categorical & political personalization are more aligned with content personalization

News Trends Analysis: DE Dataset

Discover trends & correlations between entities mentioned & events



News Trends Analysis: EN Dataset

Discover trends & correlations between entities mentioned & events