Comparative Survey of German Hate Speech Datasets: Background, Characteristics and Biases

Markus Bertram, **Johannes Schäfer**[1], Thomas Mandl[1]

University of Hildesheim
[1]Institute for Information Science
and Natural Language Processing (IwiSt)

{johannes.schaefer,mandl}@uni-hildesheim.de

October 11, 2023

Hate Speech Datasets

Empirical research on hate speech

- Different data sources (social media platforms)
- Different filtering techniques (rare phenomena)
- **Different concepts/definitions**
  (toxicity, abusive/offensive language, profanity, (illegal) hate speech)

⇒ Characteristics of datasets and biases?

Basis: Bias and comparison framework for English abusive language datasets (Wich et al., 2022)

→ Our work: Survey of **German** datasets

Agenda

Bias and comparison framework for abusive language datasets

Wich et al. (2022)

Goal: Identify characteristics and biases of datasets

1. Latent Semantic Indexing (LSI) to measure the **intra-dataset similarity between classes**
2. Embedding-based similarity:
   Inter-dataset similarity and **intra-dataset** similarity between classes
3. MI-based word rankings: Most **prominent words** for the hate speech (HS) class in each dataset, inter-dataset comparison
4. Cross-dataset topic model: Clear **HS topic(s)** or different topics more prominent?
5. Shapley values: Identify important **features** for HS classifiers

Overview of German hate speech datasets

Four shared task datasets – Mostly Twitter data (collected in 2017-2020) – Rather few manually labeled samples (even fewer abusive cases), motivates combining multiple datasets

Dataset name	Source	# of labeled samples	# of unlabeled samples	% abusive of labeled data	Inter-rater agreement
Covid2021	Twitter	4,960	0	22%	$\alpha = .92$
De-reddit-corpus	Reddit	0	2,992,835	-	-
Germeval2018	Twitter	8,541	0	34%	$\alpha = .78$
Germeval2019	Twitter	9,862	0	52%	$\kappa = .59$
Hasoc2019	Facebook, Twitter	4,669	0	12%	$\kappa = .88$
Hasoc2020	Twitter	3,400	0	29%	$\kappa = .83$
iHS	Twitter	1,249	275,022	40%	$\kappa = .44 - .55$
IWG Hate. pub.	Twitter	469	0	23%	$\alpha = .38$
Telegram	Telegram	1,149	5,421,845	16%	$\alpha = .74$

Challenges in preparing these datasets

German hate speech datasets

- Different **concepts** annotated: Binary vs. fine-grained classes or sub-classes; automatic annotation in De-reddit-corpus $\longrightarrow$ are datasets even comparable?
- Including different **sources**: Most available datasets contain only Twitter data
- Partial overlap: Dataset iHS includes some Germeval data
- Different dataset sizes: Downsampling of larger datasets?

Latent Semantic Indexing (LSI) to measure the intra-dataset similarity between classes

(A = abusive, N = neutral)

Dataset	$A \rightarrow A$	$A \rightarrow N$, $N \rightarrow A$	$N \rightarrow N$
Covid2021	.70	.71	.72
De-reddit-corpus	.29	.26	.24
Germeval2018	.39	.41	.44
Germeval2019	.41	.40	.36
Hasoc2019	.53	.57	.61
Hasoc2020	.48	.50	.56
iHS	.47	.49	.51
IWG Hatespeech public	.28	.17	.21
Telegram	.34	.37	.44

Key results

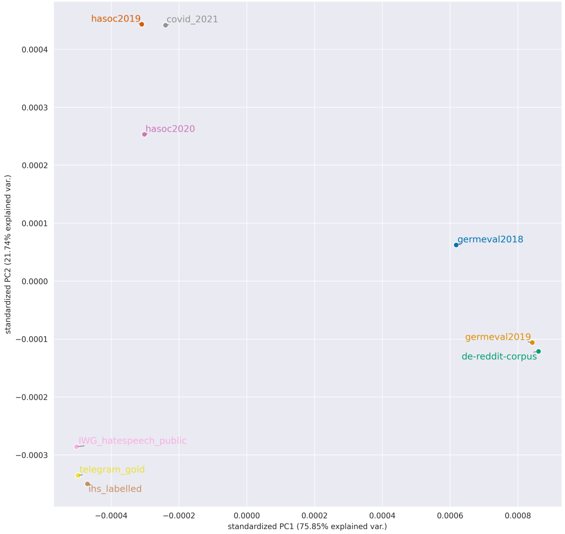
- Differences between classes in each dataset rather small

$\rightarrow$ According to this analysis: classes difficult to distinguish ($A \rightarrow N$ value not lower)

$\rightarrow$ High intra-dataset similarity

$\Rightarrow$ Hate speech detection task is difficult in each dataset

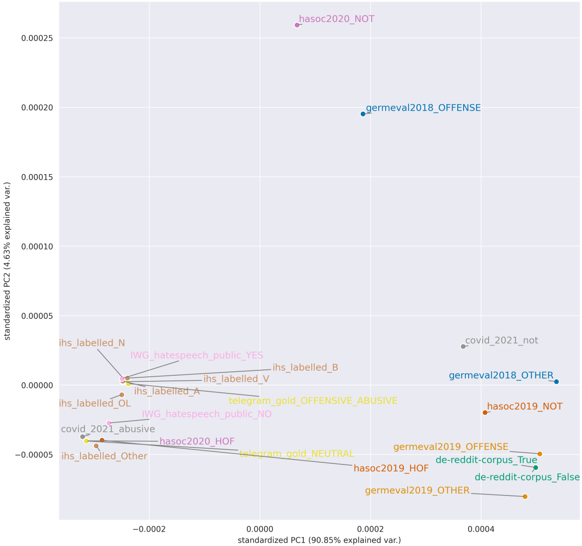
Embedding-based inter-dataset similarities



Key results

- 2D PCA projection (limited informative value)
- HASOC19/20 and Germeval18/19 each close together
- Germeval2019 closer to De-reddit-corpus than to Germeval2018
- No Twitter vs. Telegram/reddit separation
- Covid data close to pre-Covid data

Embedding-based similarity: separated classes



Key results

- Embedding centroids for individual classes in each dataset
- No clear clusters for abusive vs. neutral
  (only possibly for HASOC19/20 and Covid data)

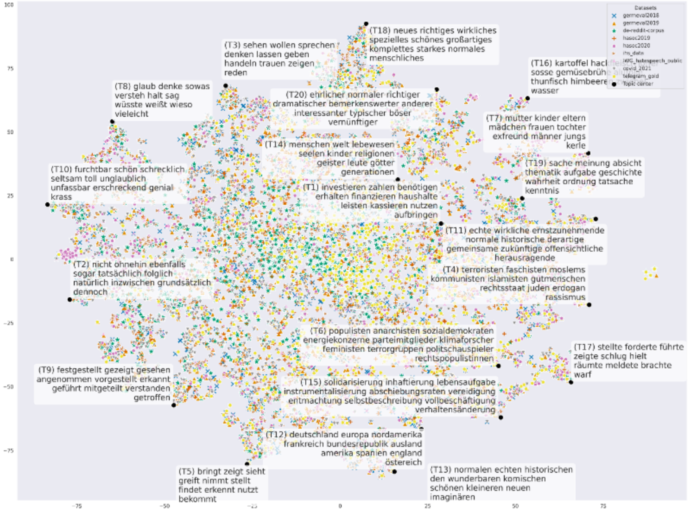
MI-based word rankings

Dataset	MI-based word rankings for the hate speech class
Covid2021	corona, dumm, merkel, mensch, virus, geben, glauben, anderer, idiot, einfach
De-reddit-corpus	einfach, geben, halt, anderer, sehen, leute, sagen, mensch, finden, eigentlich
Germeval2018	merkel, frau, deutsch, deutschland, dumm, geben, grüne, sehen, deutsche, land
Germeval2019	merkel, frau, deutschland, deutsch, dumm, sehen, land, geben, spd, deutsche
Hasoc2019	alias, loch, deutschland, papa, merkel, capitol, land, frau, sagen, sehen
Hasoc2020	arsch, hurensohn, scheiß, porno, dumm, deutsch, gratis, frau, ficken, halt
iHS	fuck, arsch, scheiße, ficken, nutte, dumm, idiot, abschaum, hure, einfach
IWG Hatespeech public	flüchtling, kind, frau, absagen, vergewaltigen, finden, schwimmbad, menschenwürde, verstoß, sexuell
Telegram	kind, geben, volk, mensch, deutsch, deutschland, anderer, bringen, krank, sehen

Key results

- Most terms indicating insult/profanity
- Identity terms (bias!)

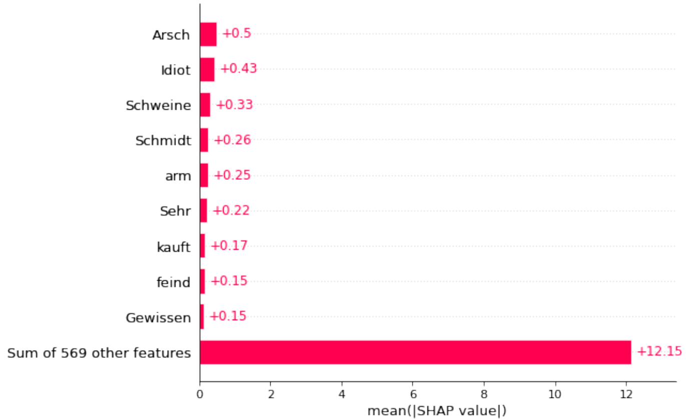
Cross-dataset topic model



Key results

- Most topics not relevant to HS; possible exceptions:
  - T4 (*terroristen*, *faschisten*, *moslems*, etc.)
  - T6 (*feministen*, *terrorgruppen*)
  - T15 (*inhaftierung*, *abschieberaten*, etc.)
- Some topics include identity terms (often targets of HS)
- No clear clustering of datasets to specific topics (e.g. no COVID-19 topic)

Feature importance using Shapley values



Key results

- Most important tokens to detect HS class
- Here displayed for dataset iHS
- Inter-dataset comparison: Vast majority of features are different

Comparative Survey of German Hate Speech Datasets

Conclusion

- Distinction of abusive vs. neutral class is difficult in these datasets
- Combination of (rather small) datasets seems to be important to cover wider range of hate speech phenomena
- Datasets cover many topics
- Biases to certain identity terms

Related publications

- Bias Mitigation for Capturing Potentially Illegal Hate Speech (dataset iHS) Schäfer (2023)
- HS-EMO: Analyzing Emotions in Hate Speech Schäfer and Kistner (2023)

- M. Wich, T. Eder, H. Kuwatly and G. Groh. (2022). Bias and comparison framework for abusive language datasets, AI and Ethics 2 1–23. http://dx.doi.org/1.1007/s43681-021-00081-0.
- Johannes Schäfer. (2023). Bias Mitigation for Capturing Potentially Illegal Hate Speech. In: Datenbank-Spektrum. https://doi.org/10.1007/s13222-023-00439-0.
- J. Schäfer and E. Kistner. (2023). HS-EMO: Analyzing Emotions in Hate Speech. In: Proceedings of KONVENS 2023.
  Data/Code: https://github.com/Johannes-Schaefer/HS-EMO.