RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets

Supplementary Material

Dennis Assenmacher[*]
University of Münster
GESIS

Marco Niemann
University of Münster

Kilian Müller
University of Münster

Moritz V. Seiler
University of Münster

Dennis M. Riehle
University of
Koblenz-Landau

Heike Trautmann
University of Münster
University of Twente

[*]Corresponding author. E-Mail: dennis.assenmacher@uni-muenster.de

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

A Appendix

A.1 Instructions for Crowdworkers

A.1.1 Briefing (English Translation)

Task

Please imagine you are taking the role of a moderator (e.g. a newspaper editor). Please evaluate all comments based on whether they could be published under a news article without further review (unproblematic). For problematic comments, you will be asked to determine why, from one perspective, the comment should be subject to further review before publication. To do this, you must select one or more of the following categories: Insult, Threat, Sexism, Racism, Profane Language, Advertising, Meta/Organisational.

Please note: Your submitted data will only be used for the purposes of the research project at the University of Münster.

Quality assurance

To ensure that the questions are answered in a meaningful way, we have built in hidden time and content checks.

Forum

Please use the forum if you have questions about the project. Here the project manager will answer you and the other gurus can also learn from your questions. And always remember: There are no stupid questions.

A.1.2 Comment Rating (English Translation)

In the Figures 1 and 2 you can find two mock-ups of the interface used by our service provider `Crowd Guru` for the annotation of the comments. We resort to mock-ups as the crowdworking job was carried out in German (both comments and interface).

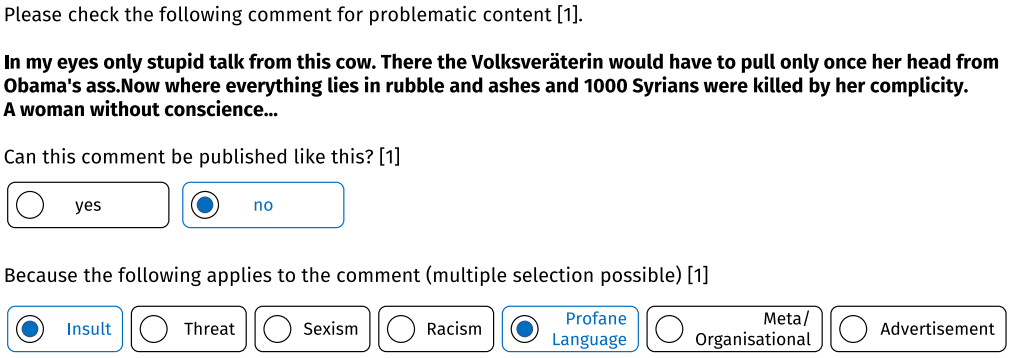

Figure 1: Comment Classification with Extended Label Selection

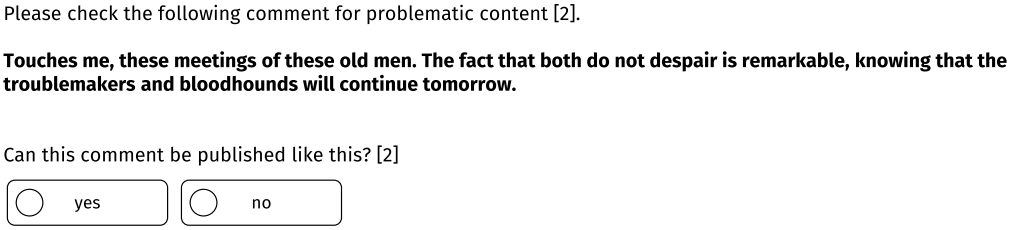

Figure 2: Comment Classification Initial State

A.1.3 Annotation Procedure

The following BPMN model gives an overview of how the annotation procedure has been organised.

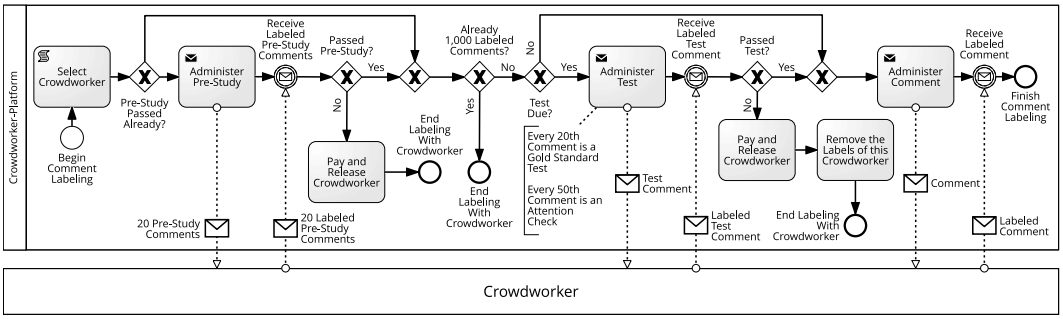

Figure 3: Annotation Process

A.2 Experimental Details

All of the conducted experiments can be easily reproduced by executing the corresponding jupyter notebooks. We split up the data creation, training and evaluation process in separate files and sub-folders. All of the notebooks are made publicly available at `https://github.com/Dennis1989/RP-Mod-RP-Crowd`.

For hyperparameter optimisation we used Bayesian optimisation (skopt). Table 1 summarises the corresponding parameter ranges for each baseline classifier.

Table 1: Configuration space for all baseline algorithms

Algorithm	Configuration Space
Naive Bayes	$\alpha \in$ [0,1] `prior` $\in \{0,1\}$
Gaussian Bayes	`smoothing` $\in [1e^{-9},1]$
Logistic Regression	$C \in [1e^{-6},1e^{6}]$ `solver` $\in$ {liblinear, saga, lbfgs}
Gradient Boosted Trees	`max_depth` $\in \{1,..,20\}$ `learning rate` $\in (10^{-}5,0)$ `min_samples_split` $\in$ (2,100) `min_samples_leaf` $\in$ (2,100)

To create AutoML pipelines we rely on `autosklearn`. We restrict the training time to 36.000 seconds and individual pipeline training to 800 seconds. Similar to our hyperparameter optimisation approach, we utilise internal 10-fold cross-validation. Additionally, we refit the final pipeline on the complete training data. Both BERT variants and their respective configurations (`Single-Task` and `Multi-Task`) are described in detail in the paper.

A.3 Experimental Results

In this section we provide additional insights into the global results of our training process. Table 2 displays the average results and corresponding standard deviation of our baseline trails, whereas Table 3 displays the results of our AutoML runs. As described in the paper we experienced model instability during the training process of our `single-task` BERT models. Figure 5 shows some characteristics of these failed models. Overall, we experience similar patterns as described recently in [2].

Table 2: Average validation accuracy for our baseline experiments with standard deviations

		`tf-idf`	`fasttext`
Naive (Gaussian Bayes)	`RP-Crowd-2`	0.7024 ±0.0000	0.5671 ±0.0002
	`RP-Crowd-3`	0.7261 ±0.0000	0.5725 ±0.0003
	`RP-Mod`	0.6666 ±0.0000	0.5798 ±0.0000
Logistic Regression	`RP-Crowd-2`	0.6995 ±0.0004	0.7200 ±0.0015
	`RP-Crowd-3`	0.7259 ±0.0020	0.7707 ±0.0007
	`RP-Mod`	0.6750 ±0.0006	0.6750 ±0.0009
Gradient Boosted Trees	`RP-Crowd-2`	0.6934 ±0.0026	0.7227 ±0.0046
	`RP-Crowd-3`	0.6935 ±0.0027	0.6782 ±0.0090
	`RP-Mod`	0.6453 ±0.0060	0.6704 ±0.0036

Table 3: Validation accuracy for our AutoML experiments

		`tf-idf`	`fasttext`
AutoML	`RP-Crowd-2`	0.709	0.727
	`RP-Crowd-3`	0.737	0.782
	`RP-Mod`	0.680	0.685

Table 4: Validation accuracy on all trained language models

		Single-Task	Multi-Task		
			$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
$\texttt{BERT}_{Base}$	`RP-Crowd-2`	0.728	0.789	0.786	0.781
	`RP-Crowd-3`	0.806	0.815	0.825	0.823
	`RP-Mod`	0.5	0.719	0.711	0.706
$\texttt{BERT}_{Hate}$	`RP-Crowd-2`	0.5	0.788	0.782	0.791
	`RP-Crowd-3`	0.836	0.798	0.815	0.808
	`RP-Mod`	0.5	0.717	0.707	0.709
$\texttt{GBERT}_{Base}$	`RP-Crowd-2`	0.665	0.811	0.811	0.814
	`RP-Crowd-3`	0.821	0.833	0.827	0.84
	`RP-Mod`	0.696	0.727	0.727	0.721

Table 5: Mean validation accuracy and standard deviation for $\texttt{GBERT}_{Base}$

		Multi-Task		
		$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
$\texttt{GBERT}_{Base}$	`RP-Crowd-2`	0.8120 ±0.0021	0.8116 ±0.0035	0.8124 ±0.0023
	`RP-Crowd-3`	0.8267 ±0.0057	0.8316 ±0.0034	0.8357 ±0.0036
	`RP-Mod`	0.7228 ±0.0020	0.7203 ±0.0053	0.7225 ±0.0016

Table 6: Validation F1 on all trained language models

		Single-Task	Multi-Task		
			$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
$\texttt{BERT}_{Base}$	RP-Crowd-2	0.768	0.794	0.795	0.779
	RP-Crowd-3	0.804	0.827	0.826	0.825
	RP-Mod	0.0	0.708	0.7	0.699
$\texttt{BERT}_{Hate}$	RP-Crowd-2	0.0	0.787	0.771	0.795
	RP-Crowd-3	0.841	0.805	0.819	0.814
	RP-Mod	0.0	0.721	0.721	0.71
$\texttt{GBERT}_{Base}$	RP-Crowd-2	0.744	0.814	0.809	0.811
	RP-Crowd-3	0.826	0.833	0.823	0.845
	RP-Mod	0.675	0.737	0.732	0.72

Table 7: Validation AUC on all trained language models

		Single-Task	Multi-Task		
			$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
$\texttt{BERT}_{Base}$	RP-Crowd-2	0.84	0.868	0.869	0.864
	RP-Crowd-3	0.879	0.899	0.899	0.903
	RP-Mod	0.498	0.793	0.785	0.774
$\texttt{BERT}_{Hate}$	RP-Crowd-2	0.499	0.867	0.869	0.871
	RP-Crowd-3	0.906	0.888	0.898	0.894
	RP-Mod	0.5	0.785	0.774	0.774
$\texttt{GBERT}_{Base}$	RP-Crowd-2	0.766	0.892	0.889	0.889
	RP-Crowd-3	0.899	0.914	0.912	0.914
	RP-Mod	0.77	0.79	0.791	0.791

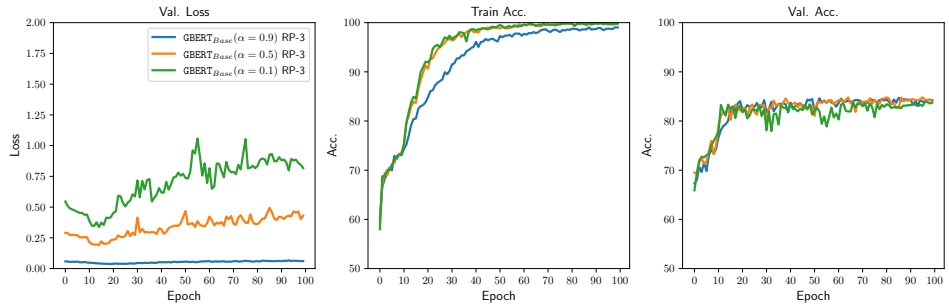

Figure 4: Validation Loss, Training Accuracy and Validation Accuracy of $\texttt{GBERT}_{Base}$ with different α values during the training process on RP-Crowd-3

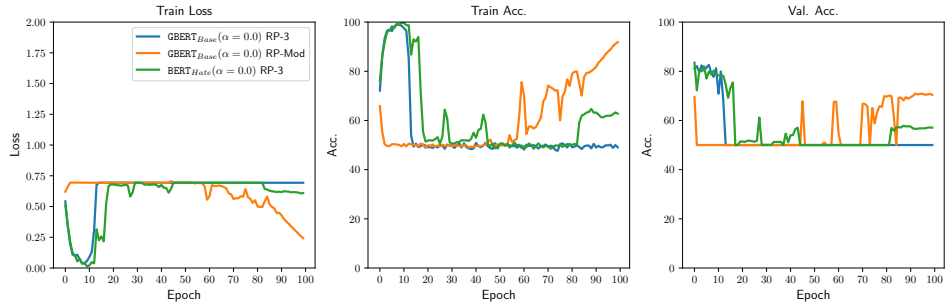

Figure 5: Three different single-task models that failed during training. In the multi-task setting these issues were not present.

A.4 Datasheet

A.4.1 Motivation

For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The RP-Mod and RP-Crowd dataset was created in the context of the MODERAT! project and is aimed at training machine learning models that allow to detect abusive language and to further classify identified instances into abusiveness categories. Another secondary use case will be the generation of explainer models for the previously generated machine learning models. Lastly, we account for the lack of large-scale German news comments datasets and provide the largest (to date) known instance of such a dataset. Given the platform character of our project, we hope to provide a basis for the generation of better machine learning models for German abusive comment detection.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organisation)?
The RP-Mod and RP-Crowd dataset was created at the University of Muenster (WWU) (Germany) at the Department for Information Systems in collaboration with the Rheinische Post (RP). Throughout the creation of the dataset, the following people have been involved (order alphabetically): Dennis Assenmacher (WWU), Jörg Becker (WWU), Jens Brunk (WWU), Hannah Monderkamp (RP), Kilian Müller (WWU), Marco Niemann (WWU), Dennis M. Riehle[2] (UKL), and Heike Trautmann (WWU).

Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

The research leading to these results received funding from the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014-2020), Project: MODERAT! (No. CM-2-2-036a).

Any other comments?

No further comments.

A.4.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

Each instance of the RP-Mod and RP-Crowd dataset represents a textual comment submitted to our partner RP, complemented by the moderation decision of their community managers, and crowdworker assigned detail labels. A schematic representation of each instance is provided through Table 8.

Table 8: Dataset Structure

Column	Description	Datatype	Ranges	RP-Mod	RP-Crowd
ID	Unique identifier	int	-	*	*
Text	Text of the comment	text	-	*	*
Reject Newspaper	comment is rejected by moderators	bool	{0,1}	*	
Reject Crowd	comment is rejected by crowdworkers (majority decision)	bool	{0,1}		*
Rejection Count Crowd	total number of rejections by crowd	int	[0,5]		*
Sexism Count Crowd	# rejects by crowdworkers based on sexism	int	[0,5]		*
Racism Count Crowd	# rejects by crowdworkers based on racism	int	[0,5]		*
Threat Count Crowd	# rejects by crowdworkers based on threats	int	[0,5]		*
Insult Count Crowd	# rejects by crowdworkers based on insult	int	[0,5]		*
Profanity Count Crowd	# rejects by crowdworkers based on profanity	int	[0,5]		*
Meta Count Crowd	# rejects by crowdworkers based on meta	int	[0,5]		*
Advertisement Count Crowd	# rejects by crowdworkers based on advertisement	int	[0,5]		*

[2]Dennis M. Riehle received his doctorate at the WWU and is now an Assistant Professor at the University of Koblenz-Landau (UKL)

How many instances are there in total (of each type, if appropriate)?

There are 85,000 comments in the `RP-Mod` and `RP-Crowd` dataset. Each instance has been labelled once by a professional community manager and five times by crowdworkers.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

As described in Section A.4.3, the presented `RP-Mod` and `RP-Crowd` dataset is a sub-sample of all comments submitted to the `RP` between November 2018 and June 2020. The sampling is conducted as outlined in Section A.4.3 (oversampling potentially abusive comments to account for the common imbalance in most published abusive language datasets; otherwise, random sampling).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? *In either case, please provide a description.*

Aside of the filtering described in Section A.4.3 the provided comment data is "raw". The `Rejection` and `Count` columns (see Table 8) contain aggregates of the individual decisions.

Is there a label or target associated with each instance? *If so, please provide a description.*

The `RP-Mod` and `RP-Crowd` dataset contains moderator- and crowdworker-assigned labels. Table 8 describes the labels; further semantic description for the individual labels can be found in Table 9.

Table 9: Full Labelling Schema translated from and based on [3]

	Label	Explanation
Theory-Deduced	sexism	Attacks on people based on their gender (identity), often with a focus on women.
	racism	Attacks on people based on their origin, ethnicity, nation (typ. meant to incite hatred).
	threats	Announcements of the violation of the physical integrity of the victim.
	insults	Denigrating, insolvent or contemptuous statements (left without further specification).
	profane language	Usage of sexually explicit and inappropriate language.
Orga.-Specific	meta / organisational	Organisational content such as request on why specific posts or commenters have been blocked. Not abusive per se, but heat up conversations providing little to no contribution.
	advertisement	Comments advertising unrelated services or products or linking to such. Not abusive per se; but lead to a deterioration of discourse sentiment and quality.

Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

Meta-information originally provided by the `RP` were removed (especially user-related information). This was necessary to minimise any potential privacy conflicts. Due to a technical problem, there are 389 instances in which a crowdworker classified a comment as abusive but did not provide a label. This results in 348 comments missing at least one of the five possible labels assigned by the crowdworkers.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? *If so, please describe how these relationships are made explicit.*

Not applicable.

Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.*

Depending on the exact task at hand, case-specific sampling will be meaningful (to filter for specific text lengths, filter specific rejection reasons, ...). Hence, ideas of potential splits can be found in the NeurIPS publication linked to this dataset (however, we leave the sampling to the user):

> Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz V. Seiler, Dennis M. Riehle, and Heike Trautmann. 2021. "`RP-Mod & RP-Crowd`: Moderator- and Crowd-Annotated German News Comment Datasets." In *Proceedings of the 35th Conference on Neural Information Processing Systems*. NeurIPS 2021. Virtual Event.

Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.*

As described above, there are 389 instances in which a crowdworker classified a comment as abusive but did not provide a label. Apart from these instances there are no further known errors or cases of redundancies in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.*

No, the `RP-Mod` and `RP-Crowd` dataset does not contain any confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.*

Yes, as the `RP-Mod` and `RP-Crowd` dataset provides textual data for the detection of abusiveness in comments, the dataset does include instances that might be offensive, insulting, threatening, or might otherwise cause anxiety. We want to express that the included comments do not represent the opinions of the authors of this paper and dataset. Furthermore, we ask people who might be susceptible to anxiety or offences to handle the data with care.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

The `RP-Mod` and `RP-Crowd` dataset does not directly relate to people. As all comments have been written, moderated, and labelled by natural persons, an indirect link exists. However, the dataset does not contain links to this personal information.

With the dataset we provide additional, anonymised demographic information for further analysis. Based on the provided information it is not possible to link back to the natural person.

Does the dataset identify any subpopulations (e.g., by age, gender)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

As the `RP-Mod` and `RP-Crowd` dataset is released without identifiers regarding the original commenters, the identification of subpopulations is not feasible.

Based on the additional demographic information, subpopulations among the crowdworkers can be identified. However, it is not possible to link the individuals or groups back to natural persons/groups.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.*

Searching comments from the RP-Mod and RP-Crowd dataset via search engines may allow to retrieve the original comment on the website of the RP. However, in this case, the comment has always been available online. Furthermore, a natural person can only be identified if the person registered with his/her real name (or any other identifiable information). The RP-Mod and RP-Crowd dataset itself contains no personal information so that the direct identification of individuals is not possible.

The crowdworkers' data has been fully anonymised. Hence, identifying individuals is not possible.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.*

Yes, the RP-Mod and RP-Crowd dataset may contain sensitive information (religious or political beliefs, racial or sexual identifiers, . . .). However, the dataset contains no link to the person behind the comment. If the person is identifiable via a search for this comment, the individuum previously decided to publicly release this information on the RP website. Hence, no previously unknown information about an individual is released.

Any other comments?

No further comments.

A.4.3 Collection Process

How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The comments in the RP-Mod and RP-Crowd dataset have been submitted to our partner RP by the individual commenters. Moderator decisions are assigned by the RP's community managers; crowd labels by crowdworkers as outlined in Section A.4.4.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?*

The RP-Mod and RP-Crowd dataset contains comments that were directly submitted to the RP by the commenters. Comments and moderation decisions have been provided to us as a direct export from their comment moderation system.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Yes, the the RP-Mod and RP-Crowd dataset were sampled from a larger dataset. To enable crowd annotation, we sampled only those comments having a length of 500 characters or less. Furthermore, we oversampled comments rejected by the moderators (RP-Mod), so all 7,141 originally rejected comments are included in the full RP-Mod and RP-Crowd dataset. The remaining 77,859 are randomly sampled from comments accepted by the moderators.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

As the moderation of comments is part of their daily business, the moderators of the RP were paid according to their contracts by the RP. The crowdworkers were paid 9,35€ (German minimum wage per hour).

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? *If not, please describe the timeframe in which the data associated with the instances was created.*

The RP-Mod and RP-Crowd dataset contains comments that have been submitted to the RP in the timeframe from November 2018 and June 2020.

Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

The terms of service allowing for the reuse of comments have been checked by the judicial departments of both institutions involved in the MODERAT! project (https://moderat.nrw). Furthermore, German procurement law required us to pay all crowdworkers the German minimum wage to prevent unfair labour conditions and exploitation.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

The RP-Mod and RP-Crowd dataset does not directly relate to people. As all comments have been written, moderated, and labelled by natural persons, an indirect link exists. However, the dataset does not contain links to this personal information.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was acquired from our project partner RP. Included are the original comments as submitted by the commenters, including the moderation decisions of the RP community managers.

Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

The Terms of Service of the RP were adjusted prior to the start of the data collection. Use and publication of the submitted data (comments) for research purposes have been explicitly listed.

Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Commenting users of RP were informed about the updated terms of services and had to consent to them.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

No, the comments were submitted under new Terms of Service, which do not include a mechanism to revoke consent.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

No formal analysis has been conducted. However, comments have either been publicly available before or cannot be traced back to the creating user (all links to the comment authors have been removed).

Any other comments?

No further comments.

A.4.4 Preprocessing/Cleaning/Labelling

Was any preprocessing/cleaning/labelling of the data done (e.g., discretisation or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

No preprocessing was done to the comment texts included in the `RP-Mod` and `RP-Crowd` dataset. The only applied restriction is the maximum length of 500 characters per comment. Crowd labels included in `RP-Crowd` were obtained through the crowdworking platform `Crowd Guru`[3].

For the experiments in the accompanying NeurIPS paper (see below) various preprocessing steps have been conducted. Preprocessing is strongly recommended for future users of the dataset.

> D. Assenmacher, M. Niemann, K. Müller, M. V. Seiler, D. M. Riehle, and H. Trautmann. "`RP-Mod & RP-Crowd`: Moderator- and Crowd-Annotated German News Comment Datasets." In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, NeurIPS 2021, Virtual Event, 2021.

Was the "raw" data saved in addition to the preprocessed/cleaned-labelled data (e.g., to support unanticipated future uses)? *If so, please provide a link or other access point to the "raw" data.*

Not applicable, as the `RP-Mod` and `RP-Crowd` dataset contains the unedited comments and moderator/crowd labels.

Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point.*

No software has been used for preprocessing, cleaning or labelling the data. All required and conducted preprocessing, cleaning, and labelling steps are outlined in this datasheet and the original publication.

Any other comments?

No further comments.

A.4.5 Uses

Has the dataset been used for any tasks already? *If so, please provide a description.*

At the time of publication, the `RP-Mod` and `RP-Crowd` dataset has been used in the experiments outlined in Section 5 of the paper "`RP-Mod & RP-Crowd`: Moderator- and Crowd-Annotated German News Comment Datasets" [1]. Furthermore, it has been used for internal experimentation in the context of the research project MODERAT! (`https://moderat.nrw`).

Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point.*

[3] `https://www.crowdguru.de/`

We will provide links to all published papers through the zenodo repository of the RP-Mod and RP-Crowd dataset. Furthermore, we will provide a subpage on our project homepage https://moderat.nrw that links to the zenodo repository and all associated papers.

What (other) tasks could the dataset be used for?

The RP-Mod and RP-Crowd dataset can be used for various purposes aside of the ones presented in the original paper "RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets" [1]. Potential use-cases include creating German abusive comment word-embeddings, training of binary, multi-class, or even multi-label classifiers, and corresponding explainer models. Furthermore, applications such as benchmarks against other datasets or more in-depth linguistic analysis of (non-)abusive comments are feasible.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labelled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Future users should—depending on the chosen application—be careful about aspects such as temporal degradation. As language and topics tend to be "living constructs" subject to constant changes, classifiers trained on the dataset might show deteriorating performance for texts released long after this dataset.

Are there tasks for which the dataset should not be used? *If so, please provide a description.*

There are no general limitations to the use of the data. However, we caution to use the RP-Mod and RP-Crowd dataset for the creation of fully automated detection systems. If classified texts originate from a different domain, their style and features might diverge from the ones present in the presented dataset. Furthermore, freedom of expression issues could arise from fully-automated detection systems.

Any other comments?

No further comments.

A.4.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organisation) on behalf of which the dataset was created? *If so, please provide a description.*

The dataset will only be distributed as stated below.

How will the dataset will be distributed (e.g., tarball on website, API, *GitHub*) *Does the dataset have a digital object identifier (DOI)?*

The RP-Mod and RP-Crowd dataset is available on zenodo with the DOI 10.5281/zenodo.5242915. (https://doi.org/10.5281/zenodo.5242915).

The dataset is distributed in the CSV format. An overview of all files is provided in Table 10.

When will the dataset be distributed?

The RP-Mod and RP-Crowd dataset are available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and*

Table 10: Files of the RP-Mod and RP-Crowd datasets

Filename	Description
RP-Mod-Crowd.csv	The full RP-Mod and RP-Crowd datasets including all labels stated in Table 8. Contains $85,000$ comments with 12 variables.
RP-Crowd-1.csv	Balanced dataset (equal quantities of accepted and rejected comments) with the crowd rejection threshold set to 1. Contains $57,410$ comments with 3 variables.
RP-Crowd-2.csv	Balanced dataset (equal quantities of accepted and rejected comments) with the crowd rejection threshold set to 2. Contains $17,368$ comments with 3 variables.
RP-Crowd-3.csv	Balanced dataset (equal quantities of accepted and rejected comments) with the crowd rejection threshold set to 3. Contains $6,302$ comments with 3 variables.
RP-Crowd-4.csv	Balanced dataset (equal quantities of accepted and rejected comments) with the crowd rejection threshold set to 4. Contains $1,974$ comments with 3 variables.
RP-Crowd-5.csv	Balanced dataset (equal quantities of accepted and rejected comments) with the crowd rejection threshold set to 5. Contains 424 comments with 3 variables.
RP-Mod.csv	Balanced dataset (equal quantities of accepted and rejected comments) containing all moderator rejected comments. Contains $14,282$ comments with 2 variables.
CrowdGuru-Ratings.xlsx	The raw labels as returned by our service provider Crowd Guru (one line per annotator and comment; incl. time to make the decision).
CrowdGuru-Demographic.xlsx	The anonymised demographical data of the crowdworkers involved in our annotation study (incl. age, gender, educational level, job, country of residence).

provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The RP-Mod and RP-Crowd dataset are distributed under the CC BY-NC-SA (https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode) license. A human-readable version of this agreement can be found at https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en. We further ask users of the dataset to cite this paper:

> D. Assenmacher, M. Niemann, K. Müller, M. V. Seiler, D. M. Riehle, and H. Trautmann. "RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets." In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, NeurIPS 2021, Virtual Event, 2021.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

Unknown.

Any other comments?

We hereby confirm that we have the rights to publish the datasets RP-Mod and RP-Crowd for non-commercial (i.e., research) purposes. All affected persons (commenters and crowdworkers) have been informed about this usage of their data and explicitly agreed on this.

Should, despite all precautions, the dataset (or any parts of it) violate any rights, the authors bear all responsibility.

The dataset will be published under the CC BY-NC-SA 4.0 license which can be found online under https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.

A.4.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

All authors of this paper act as maintainers for the RP-Mod and RP-Crowd dataset. To account for potential changes of institutions, please contact them via their corresponding GitHub accounts:

- Dennis Assenmacher (https://github.com/Dennis1989)
- Marco Niemann (https://github.com/MarcoNiemann)
- Kilian Müller (https://github.com/MuellerKilian)
- Moritz V. Seiler (https://github.com/mvseiler)
- Dennis M. Riehle (https://github.com/driehle)
- Heike Trautmann (https://github.com/trautm)

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Please contact us via zenodo or use the above-stated GitHub accounts to contact the contributing authors.

Is there an erratum? *If so, please provide a link or other access point.*

No. This is the initial publication of the RP-Mod and RP-Crowd dataset. Hence, no errata have been discovered so far.

Will the dataset be updated (e.g., to correct labelling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

All incoming update requests will be handled by the contributing authors listed in A.4.7. Any updates will be posted on the zenodo repository page as new versions.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.*

The data does not directly relate to people. No retention periods were stated.

Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

We will use versioning for any updates released. Hence, old versions will be kept for reference and consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

The dataset can be extended in accordance with the limitations imposed by the license. In any case, the authors, respectively maintainers of the `RP-Mod` and `RP-Crowd` datasets, should be contacted for the incorporation of potential fixes and extensions.

Any other comments?

No further comments.

References

[1] D. Assenmacher, M. Niemann, K. Müller, M. V. Seiler, D. M. Riehle, and H. Trautmann. `RP-Mod & RP-Crowd`: Moderator- and Crowd-Annotated German News Comment Datasets. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, NeurIPS 2021, Virtual Event, 2021.

[2] M. Mosbach, M. Andriushchenko, and D. Klakow. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *Proc. Ninth Int. Conf. Learn. Represent.*, ICLR 2021, pages 1–19, Virtual Event, 2021.

[3] M. Niemann, D. M. Riehle, J. Brunk, and J. Becker. What Is Abusive Language? Integrating Different Views on Abusive Language for Machine Learning. In *Proceedings of the 1st Multidisciplinary International Symposium on Disinformation in Open Online Media*, MISDOOM 2019, pages 59–73, Hamburg, Germany, 2020. Springer.