GermEval2024 Shared Task: GerMS-Detect – Sexism Detection in German Online News Fora

Stephanie Gross, Johann Petrak, Louisa Venhoff, Brigitte Krenn
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, 1010 Vienna, Austria
firstname.lastname@ofai.at

Abstract

We present an overview of the GermEval2024 shared task: GerMS-Detect on the detection of sexism and misogyny in the German language comments of online news fora. The data were annotated by a varying number of human annotators with regard to whether or not the comment is sexist or misogynist in a way that could discourage women from participating in the discussion. Ambiguous comments or comments which may contain more subtle forms of misogyny have often been judged and annotated differently by the human annotators. For this task, rather than assuming the existence of one "true" label for each comment, we accept that judgements on the presence or strength of misogyny can be highly subjective and encourage the development of models which can be used to reflect the potential disagreement for some of the comments. For this reason, the shared task was divided into two subtasks, where subtask 1 focused on classification models capable of detecting binary or ordinal levels of misogyny derived in different ways from the labels provided by the human annotators as well on predicting whether or not there is disagreement between the annotators of the comment. Subtask 2 was concerned with directly approximating the distribution of labels a group of specific annotators is likely to assign to a specific comment. Seven teams participated in subtask 1 and six participated in subtask 2. Of these, five teams contributed a paper for the workshop.

Content warning: We show illustrative examples of sexist and misogynous language.

1 Introduction and Motivation

Sexist and misogynist comments in online social media or other online fora can be harmful and be an important factor why women refrain from participating in online discussions. This effect of silencing women in online fora may get caused also by comments with subtle or implicit forms of misogyny. This calls for the deployment of tools to identify sexist content to support content moderation and monitoring. However, identifying sexist content is also a challenging task for humans because they often refer to some implied context which is not available or is formulated in a subtle way, avoiding strong or outright offensive language. Therefore the manually annotated datasets on which classifiers are trained on potentially contain high human annotator variation for the same content. The up to date prevalent approach is to unify diverging annotator opinions, assuming a ground truth, e.g., by employing majority vote, subsequent consensus by the annotators, or a decision by a meta reviewer. Plank (2022) emphasizes that human label variation needs more attention in machine learning research, as it impacts data, modeling and the evaluation of machine learning systems. Although the interest in preserving annotator variation is increasing (e.g., Pavlick and Kwiatkowski, 2019; Uma et al., 2021b; Plank, 2022; Davani et al., 2022), and relevant workshops and shared tasks were already organized in recent years (e.g., Abercrombie et al., 2022; Uma et al., 2021a; Ojha et al., 2023), multi-perspective approaches are still in their infancy.

We organized the GermEval2024 Shared Task: GerMS-Detect – Sexism Detection in German Online News Fora, with the goal to contribute to this line of research. This shared task also follows from success of previous shared tasks on sexism detection (such as Fersini et al., 2018; Basile et al., 2019; Kirk et al., 2023).

The corpus used in this shared task was collected from comments in the fora of a large Austrian online news site (derStandard.at[1]). The annotations reflect the newspaper's forum moderation policy regarding sexism and misogyny. Moreover, the definition of sexism reflected in the annotation guide-

[1] https://www.derstandard.at

lines is based on the definition given in Encyclopedia Britannica which defines **sexism** as "prejudice or discrimination based on sex or gender, especially against women and girls", and **misogyny** as "the extreme form of sexist ideology" which they state is "the hatred of women".[2] The corpus was then annotated by multiple annotators with labels ranging from 0 (no sexism/misogyny) over 1 (slight) to 4 (extreme sexism). Annotator judgements tend to differ, especially when the comment lacks context, or is worded in a subtle or deliberately ambiguous way. The goal of the shared task was to explore how different opinions from different annotators can be utilized and reflected in models trained on this corpus, rather than assuming that one label is the correct one to reflect the "true" sexism present in the comment and considering diverging labels as mistakes or noise. See Table 1 for sample *sexist* comments in the data.

Sample:	"Ich mag keine Kampflesben, die sollte man mal allesamt wegsperren"
EN:	"I do not like combat lesbians, they should all be locked away"
Sample:	"Bei aller Tragik und Ernsthaftigkeit.... wir haben schon a fesche Justizministerin"
EN:	"With all tragedy and seriousness.... we definitely have a dashing Minister of Justice"

Table 1: Sample comments from the dataset.

The datasets provided in the shared task contain, in the training set, the individual annotations from each annotator (identified by an annonymized annotator id) and, in the test set, the list of annotators (but not their labels) for which the trained models have to make predictions. The shared task is divided into two subtasks: in subtask 1, binarized and multi-class labels derived from the individual annotator-assigned labels as well as an indicator of annotator disagreement on the presence of sexism are derived from the set of annotations and have to be predicted by the model for the test set. In subtask 2, the distribution of binarized and multi-class labels for the given set of annotators has to be predicted by the model. For both subtasks there was a closed and an open track, where the closed track required that no additional training data or pretrained models which may have been trained for sexism or misogyny detection was allowed and all contributions had to be open source. For the open track, any approach, including proprietary data or models, including large language models, was allowed.

Characteristics which make our shared task unique are: The dataset (i) is in German language, (ii) it includes a high number of expert labels, (iii) it was collected with the goal to provide a more welcoming and safer climate of discussion in online newspaper fora, especially for female users, and (iv) it allows for experimenting with classifier training based on hard and soft labels. Uma et al. (2021b), for example, has shown that with datasets annotated by a high number of expert coders, training directly with soft labels achieved better results than training from aggregated or gold labels.

2 Dataset

Data collection The data stem from fora of a large Austrian online newspaper in German language and consist of 7984 user comments on newspaper articles.[3] They include (i) selected comments which were reported as problematic by forum users, (ii) randomly sampled comments, (iii) comments pre-classified as potentially sexist by a sexism classifier trained on an early subset of the annotated data, and (iv) comments from 24 article fora which were manually identified by forum moderators to contain an above-average number of comments considered as *sexist*. (For more details, see Krenn et al., 2024). The length of the comments ranges from one to 173 words, with a mean of 32 words per comment. The original newline and whitespace characters were preserved.

Data preprocessing For anonymization, (i) URLs were replaced with the placeholder `{URL}`, (ii) At-mentions (e.g. `@name`) were replaced with `{USER}`, (iii) comments were scanned for email addresses, but none were present in the texts, and (iv) each comment was manually checked for potential mentions of user names or nick names by three annotators and systematically replaced with the placeholder `{USER}`.

Further means for privacy protection were that all information indicating the position of a comment within a certain thread was excluded, as well as all information which would allow a comment to

[2] Source: https://www.britannica.com/topic/sexism (Accessed: 2024-07-30). As the dataset employed in the shared task comprises comments which are either sexist or misogynist or both, we use *sexism* or *sexist* to refer to sexist or misogynous comments in this paper.

[3] After the end of the shared task competition phase, all data was made publicly available, see https://huggingface.co/datasets/ofai/GerMS-AT and https://ofai.github.io/GermEval2024-GerMS/download.html.

be associated with a particular article forum. These privacy detection means influences the annotation process, as there is no further context available but only the individual comment text when annotators decide whether a comment is *sexist* or not and to what extent.

Data annotation Goal of the corpus annotation was to learn from moderator judgements in their everyday work. Therefore the majority of annotators who manually labelled the comments were experienced forum moderators (7 out of 10). However, the other three annotators were experienced in corpus annotation. There were 3 annotators who self-identified as male, 7 who self-identified as female, and all annotators were native speakers of German.

The annotators were provided with detailed annotation guidelines including a list of criteria determining what should be classified as *sexist*, covering the newspaper's gender policy. The criteria to judge a comment as sexist referred to in the annotation guidelines are:

- Generalizing stereotypes, i.e., attributions to groups of women, including role stereotypes (e.g., women are better suited for housework) and attribute stereotypes (e.g., women can not think logically)
- Reduction of a person to her appearance
- Women as sexual objects
- Female connoted insult
- Denigration of women, their performance and women's issues, e.g., denial of the existence of gender differences in salary
- Downplay of sexual violence and sexual harassment against women
- Whataboutism, e.g., claiming that men are much more likely to be affected by violence
- Abortion, e.g., abortion is equated with murder
- Misandry: Given a *sexist* utterance against men, can the male referent be replaced by a female referent and does the resulting utterance fall under one of the above categories?

For more details on the annotation guidelines, see (Krenn et al., 2024).

In addition, the annotators were asked to label those comments they have classified as *sexist* on a scale from 1 to 4 according to their personal perception of the severity of sexism expressed in the comment ("How uncomfortable do I feel reading this comment?"). While Röttger et al. (2022) argue to follow either a descriptive or prescriptive annotation paradigm when annotating a dataset, we aimed for a combination of detailed guidelines on what should be considered as sexist (prescriptive paradigm) and the subjective assessment of how sexist a user comment is (descriptive paradigm). This allowed us to create a dataset which captures gradations in the assessment of sexist utterances with twofold use: (i) for the training of binary classifiers (*sexist* versus *non-sexist*); (ii) in machine learning research for how to make models aware of more or less disagreement on labels (e.g., Uma et al., 2021b; Plank, 2022). These two use-cases are reflected in subtask 1 and subtask 2 of the GermEval2024 Shared Task GerMS-Detect, respectively.

Comments were annotated by assigning one of 5 possible labels ($0-4$), where 0 is the absence of *sexism* and $1-4$ express the levels of subjective severity of the expressed *sexism* as perceived by the individual annotators (1 = mild, 2 = present, 3 = strong, 4 = extreme). Each comment was annotated by 3–10 individual annotators (3 labels: 325 comments, 4 labels: 1073 comments, 5 labels: 6481 comments, 7 labels: 6 comments, 10 labels: 999 comments).

Annotator Agreement and Corpus Analysis Krippendorff Alpha over all annotations was 0.64 (ordinal scale), and for the binary data (sexist vs. not sexist) it was 0.59. According to Hayes and Krippendorff (2007), values over 0.667 are considered to be good. The lower values in the present dataset might be due to the highly subjective nature of what is considered sexist and the assessment of its severity. A Shapiro-Wilk Test showed significant results for all annotators ($p < 0.001$) indicating that the data are not normally distributed. Therefore a Kruskal Wallis H Test was calculated to check for overall significant differences between the means of the annotators. This test was significant with $H = 477.04$, $p < 0.001$. A Dunn-Bonferroni post-hoc test was conducted to compare the individual annotators. This test revealed significant differences ($p <= 0.05$), see Figure 1.

	A001	A002	A003	A004	A005	A007	A008	A009	A010	A012
A001	1.000	0.001	0.353	1.000	1.000	1.000	0.102	0.024	1.000	1.000
A002	0.001	1.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	0.000
A003	0.353	0.000	1.000	0.039	0.029	0.136	0.000	0.000	0.000	0.000
A004	1.000	0.000	0.039	1.000	1.000	1.000	0.124	0.022	1.000	1.000
A005	1.000	0.000	0.029	1.000	1.000	1.000	0.097	0.014	1.000	1.000
A007	1.000	0.000	0.136	1.000	1.000	1.000	0.062	0.010	1.000	1.000
A008	0.102	1.000	0.000	0.124	0.097	0.062	1.000	1.000	0.806	0.019
A009	0.024	1.000	0.000	0.022	0.014	0.010	1.000	1.000	0.083	0.000
A010	1.000	0.000	0.000	1.000	1.000	1.000	0.806	0.083	1.000	1.000
A012	1.000	0.000	0.000	1.000	1.000	1.000	0.019	0.000	1.000	1.000

Figure 1: Results of a Dunn-Bonferroni post-hoc test comparing individual annotators. Significant differences ($p <= 0.05$) are marked in red.

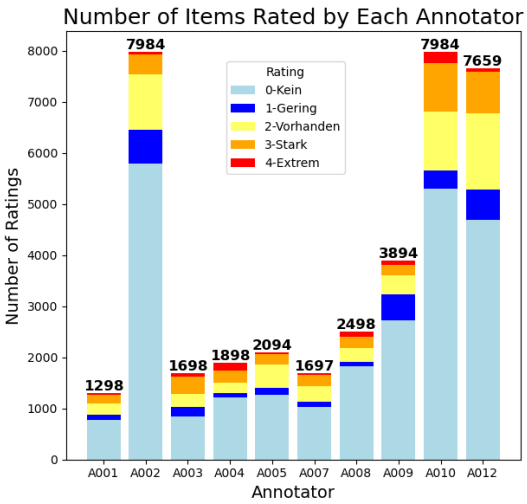

Figure 2: Number of items rated by each of the 10 individual annotators.

These low p-values might be due to different reasons:

- Systematic Differences: Individuals may have a systematic difference in how they rate items. For example, one rater consistently gives higher or lower ratings than the other one.
- Rating Bias: One or both raters might have a bias in their ratings, such as always rating items on the higher or lower end of the scale, leading to a significant difference when compared to other raters.
- Consistency in Rating: If one rater is very consistent in their ratings (e.g., always giving the same rating for similar items) while another rater is less consistent or varies more in their ratings, this can lead to significant differences in the distribution of ratings.
- Sample Size: If the number of items rated by each rater is large, even small differences in the average ratings can become statistically significant, leading to very low p-values.
- Scale of Measurement: The scale of ratings (0-4) might accentuate differences, especially if the differences between raters are consistent across many items.

Figure 2 shows the number of items rated by each annotator and their respective ratings. While three annotators labelled 95% of the data or more, the other 7 labelled 16–49%. Also, differences in the subjective assessment of the severity of a user comment are visible. Figure 3 shows the means and distributions of the ratings per annotator.

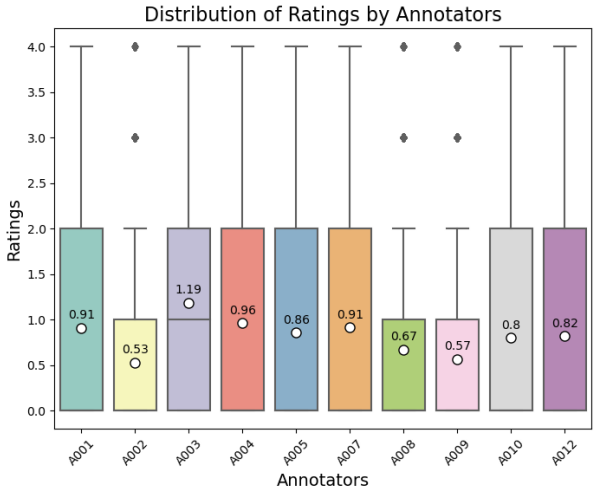

Figure 3: Annotator distributions and means of annotator ratings.

When comparing the significant pairs in the Dunn's test with the plots it becomes clear why some pairs are more significant than others, depending on the amount they rated, their inconsistencies in ratings and their differing means and distributions.

The higher level of subjectivity and the inconsistent number of labels per comment raise challenges for subtask 1 and subtask 2. However, dealing with annotator variation due to subjective assessment and developing robust models based on the dataset with certain inconsistencies are topics we wanted to target in the shared task.

In order to gain more insight in annotator variation, we propose a qualitative analysis of comments

with significant disagreement (e.g., a deviation of 3 * the standard deviation), such as qualitative content analysis and inductive category development (see Mayring, 2014). In the dataset discussed in this paper, for example irony, sexism against men, and non-sexist insults might play a role in comments with significant disagreement. However, this is ongoing work and needs further investigation. Additionally, deductive category application might be useful for further analysing significant disagreement in these types of dataset, e.g. the categories proposed by Sandri et al. (2023).

3 Task Description

3.1 Task Definition and Evaluation Metrics

Subtask 1: Classification In subtask 1 the goal was to predict labels for each text in a dataset where the labels are derived from the original labels assigned by several human annotators in several different ways:

- `bin_maj`: predict 1 if a majority of annotators assigned a label other than 0, predict 0 if a majority of annotators assigned a label 0. If there was no majority, then both the label 1 and 0 will count as correct in the evaluation.
- `bin_one`: predict 1 if at least one annotator assigned a label other than 0, 0 otherwise.
- `bin_all`: predict 1 if all annotators assigned labels other than 0, predict 0 otherwise.
- `multi_maj`: predict the majority label if there is one, if there is no majority label, any of the labels assigned is counted as a correct prediction for evaluation.
- `disagree_bin`: predict 1 if there is disagreement between annotators on 0 versus all other labels and predict 0 otherwise.

System performance on all five predicted labels was evaluated using F1 macro score over all classes. The final score which was used for ranking the submissions was calculated as the unweighted average over all 5 scores.

Subtask 2: Label distribution prediction In subtask 2 the goal was to predict the distribution for each text in a dataset where the target distribution is derived from the original distribution of labels assigned by several human annotators. The annotators assigned (according to the annotation guidelines) the strength of misogyny/sexism present in the given text via the labels 0 (for no sexism present) to 4 (extreme sexism). From the set of assigned labels, two target distributions where derived: a binarized version, specifying the fractions of annotators who assigned 0 and who assigned non-0 labels, and another distribution with the fractions of annotators who assigned labels 0 to 4. The participants had to submit a dataset which contained, for each example in the test set, the predicted fractions for both distributions.

For the evaluation of subtask 2, the Jensen-Shannon (JS) divergence between the target distribution and the predicted distribution was calculated and averaged for each of the binary and the multiclass distributions in the test set and the two JS divergences where then averaged to obtain the final score. The JS-divergence was chosen as it is a true metric and bounded, and it is therefore well suited to be used and combined into the final score.

Closed versus open tracks For each subtask, there was a closed and an open track. In the closed track, neither additional data labelled for sexism or misogyny, nor language models or embeddings which might have been pre-trained or instruction-finetuned with sexism/misogyny specific data were allowed to enhance reproducability.[4] For the open track, participants were encouraged to use whatever approach they preferred. However, only the closed track counted towards the competition of the shared task and a closed track submission was required for the submission of a paper.

3.2 Task Organisation

The GermEval2024 Shared Task GerMS-Detect was run on Codabench and organised in four different competitions: subtask 1 – closed track[5], subtask 1 – open track[6], subtask 2 – closed track[7], and subtask 2 – open track[8]. Reason for this was to keep the leader boards and the evaluation metrics separate. The task was organised in three phases: a trial phase, a development phase, and a competition phase (which ended on 2024-06-28). In the trial phase, an initial set of 1000 labeled training examples and 500 unlabeled test examples was available, in the development phase 4486 labeled training examples and 1512 unlabeled test examples were available and in the competition phase, 5998 labeled training examples and 1986 unlabeled test ex-

[4] For more details, see https://ofai.github.io/GermEval2024-GerMS/closed-track.html

[5] https://www.codabench.org/competitions/2744/

[6] https://www.codabench.org/competitions/2745/

[7] https://www.codabench.org/competitions/2746/

[8] https://www.codabench.org/competitions/2747/

amples were available. Each training set contained the labeled test data from the previous phase. All data is available from the GermEval2004 GerMS-Detect web site[9]. The training/test data splits were carried out in a way that simultanously stratified the distribution of annotator ids, class labels, and original annotation rounds (i.e., source of the data) as much as possible. The code used for evaluation is available from the GermEval GerMS-Detect Github repository[10].

4 Participant Systems and Results

Per subtask and track, one submission account was allowed per team. 13 teams registered for the shared task, of these, during the competition phase, 7 submitted to subtask 1 – closed track, 3 submitted to subtask 1 – open track, 6 submitted to subtask 2 – closed track and 2 submitted to subtask 2 – open track. 5 teams submitted papers describing their approaches and results, which will be discussed in the following chapters.

4.1 Leader Board Results

In the closed track, the 5 teams which submitted a paper were also the ones which achieved the highest results on the leader board, see Table 2 for a summary of their results.

Team	ST1-c F1 macro	ST1-o F1 macro	ST2-c JS	ST2-o JS
THAugs	0.642	-	-	-
ficode	0.641	-	0.354	-
Quabynar77	0.611	0.452	0.292	0.409
Team GDA	0.597	0.586	0.301	-
pd2904	0.483	-	0.388	-

Table 2: Top ranked leaderboard results and summary statistics for subtask 1 (ST1) and subtask 2 (ST2), the open track (o) and the closed track (c) of the 5 teams who submitted a paper. The best submission is marked in red.

Subtask 1 All five teams developed systems for subtask 1 - closed. The scores obtained by their best submissions are shown in Figure 4 with their p=0.05 confidence intervals[11]. At p=0.05 the best two submissions were not significantly different. For both submissions an ensemble method fine-tuning Deepset's `gbert-large`[12] (teams THAugs and ficode) was employed. The third best submission by team Quabynar fine-tuned Deepset's `gbert-base`[13]. The fourth best submission by team GDA employed a Support Vector Machine (SVM) classifier on top of `mE5-large` embeddings[14]. The fifth best submission by team pd2904 followed a more traditional approach by applying a combination of Random Forests, Light Gradient-Boosting, Extreme Gradient Boosting, SVM, and CatBoost models.

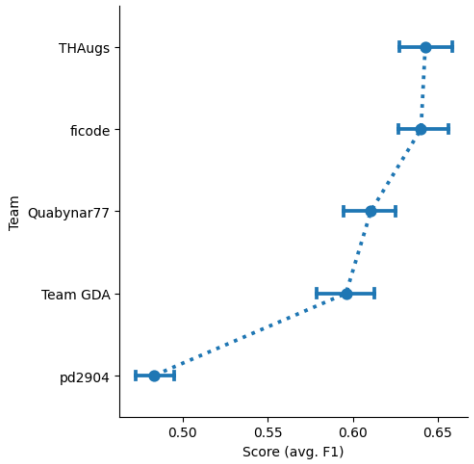

Figure 4: Comparison of results showing $p = 0.05$ confidence intervals of the teams who participated in subtask 1 closed.

Two teams additionally submitted results for the open track of subtask 1 (team GDA and Quabynar77). However, only team Quabynar described their approach for the open track in their paper. They applied few-shot learning on OpenAI's GPT 3.5 Turbo by selecting only the top 5 comments iteratively for each annotator, achieving an F1 macro score of 0.452.

Subtask 2 Four teams submitted results for subtask 2, see Figure 5 for an overview of their results. The top submission by team Quabynar77 fine-tuned Google's `bert-base-german-cased`[15]. The second best approach by team GDA employed a Support Vector Machine classifier with `gbert-large-pc` embeddings[16]. An

[9] https://ofai.github.io/GermEval2024-GerMS/

[10] https://github.com/OFAI/GermEval2024-GerMS/tree/main/python

[11] Calculated via bootstrapping of 500 samples using the CompStats (Nava-Muñoz et al., 2024) package, see https://compstats.readthedocs.io/en/latest/

[12] https://huggingface.co/deepset/gbert-large

[13] https://huggingface.co/deepset/gbert-base

[14] https://huggingface.co/intfloat/multilingual-e5-large

[15] https://huggingface.co/google-bert/bert-base-german-cased

[16] https://huggingface.co/deutsche-telekom/gbert-large-paraphrase-cosine

interesting difference between the two top submissions is that the approach fine-tuning `bert-base-german-cased` achieved the same result for the multi score distribution as the SVN classifier with `gbert-large-pc` embeddings, but performed better on the binary distribution: JS divergence = 0.248 (team Quabynar77) vs. 0.267 (team GDA).

Team ficode used the same ensemble method as in subtask 1, fine-tuning `gbert-large`. Team pd2904 also employed a similar approach as in subtask 1 by training the same types of traditional models for each annotator.

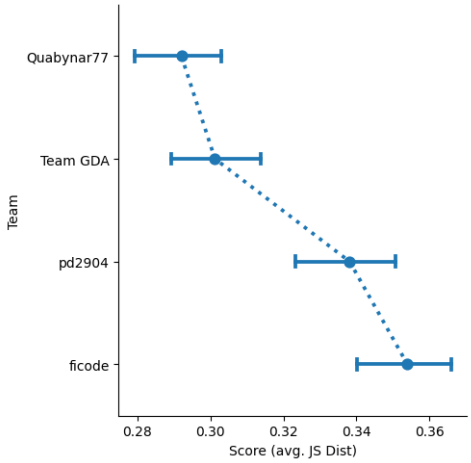

Figure 5: Comparison of results showing $p = 0.05$ confidence intervals of the teams who participated in subtask 2 closed.

Team Quabynar was the only team participating in the open track of subtask 2 who described their results in their paper. They applied the same approach as for the open track of subtask 1, using few-shot learning on OpenAI's GPT 3.5 Turbo (iteratively selecting the top 5 comments for each annotator), performing worse than all submissions of the closed track of subtask 2, i.e., achieving a higher score for the JS divergence.

5 Conclusion

In this paper, we presented an extended dataset based on Krenn et al. (2024) with a higher number of expert annotations, which allows training a classifier directly on the individual labels.

Further, we analysed and discussed annotator variation in detail and proposed a qualitative method to gain further insights in reasons for annotator variation, which might also be relevant for other datasets with significant annotator disagreement.

Additionally, we summarized the systems of teams submitting a paper to describe their approach. Four out of five teams used transformer architectures. German versions of BERT were the most popular models, but also multilingual-e5-large embeddings were employed. However, also submitted were results from a transformer based approach combined with a SVM classifier on top, as well as an approach based on traditional models (Random Forests, Light Gradient-Boosting, Extreme Gradient Boosting, SVM, and CatBoost models).

6 Limitations

Even though there are more *not sexist* comments in the dataset than *sexist* comments, the dataset has a selection bias towards sexist comments (see the description of the data collection in section 2), which makes the proportion of sexist comments much higher than in the news fora. Therefore a classifier trained on that data might label a comment as *sexist* with a higher probability. However, if the proportion of *sexist* comments would have reflected the proportion in the real data, a much larger amount of data would have needed labelling in order to span such a broad range of topics.

The comments in the dataset are annotated without further context, e.g., the article a forum is related to, or the thread a comment is part of. Therefore some *sexist* comments might be missed due to the lack of context. Also, ironic comments responding to a *sexist* comment might be misinterpreted as *sexist*.

The specific newspaper's forum moderation policy influenced the annotation guidelines and also the majority of the annotators were employed as forum moderators for that specific newspaper. In other contexts, other criteria for identifying sexism or misogyny might be relevant.

A limitation of the shared task is that submissions to open tracks did not count towards the competition ranking and closed track submissions were required for a paper submission. We only received one description of an approach for the open track, which is not sufficient for a proper comparison between the closed and the open track. However, the reason for emphasizing on the closed task was reproducibility.

7 Ethical Considerations

The foremost goal of the dataset collection was to train classifiers that support content moderators of an Austrian German language online newspaper with regards to identifying sexist and misogynous comments. In the forum of this online newspaper, 20K to 50K comments are made per day (with rising tendency), making solely manual monitoring for human moderators impossible. Therefore, support by automatic monitoring of classifiers is a precondition for moderators to intervene in a timely manner.

There is risk of harm to annotators by repeated exposure to sexist and misogynist utterances. Even though annotators are either professional forum moderators used to handling sexist and misogynous comments, or experts in corpus annotation, regular monitoring is necessary to watch for negative effects of excessive exposure to harmful content on individuals. Researchers and developers might be affected by the exposure to harmful content, as well as readers of the paper. The exposure to such harmful content may also lead to prejudiced discussions and the reproduction or reinforcement of harmful representation stereotypes. Therefore, content warnings are placed at the beginning of the paper before examples for *sexist* comments are presented, cf. (Kirk et al., 2022).

Violation of privacy is a risk which may concern forum users who are mentioned in the comments or whose comments are part of the dataset. As a countermeasure, all potential user names, at-mentions, URLs, email addresses were deleted.

A datasheet was published together with the dataset on huggingface to offer detailed information on the capacities and limitations of the dataset. The advantage of making the dataset publicly available is that fellow researchers can take up and further extend the work. We strongly recommend to publish a model card (Mitchell et al., 2019) with each model trained on the dataset. Still, misuse of the dataset can not be completely ruled out.

Acknowledgments

This work was conducted as part of the projects FemDwell[17] supported through FemPower IKT 2018[18] and EKIP – A Platform for Ethical AI Application[19] supported by the Austrian Research Promotion Agency (FFG)[20]. We thank the forum moderators at derStandard.at for their contributions to developing the annotation guidelines and their efforts in annotating the dataset.

[17] https://ofai.github.io/femdwell/

[18] https://austrianstartups.com/event/call-fempower-ikt-2018

[19] https://ekip.ai/

[20] https://www.ffg.at/en

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma. 2022. Proceedings of the 1st workshop on perspectivist approaches to nlp@ lrec2022. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of SemEval-2023*, pages 2193–2210, Toronto, Canada.

Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. Germs-at: A sexism/misogyny dataset of forum comments from an austrian online newspaper. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739.

Philipp Mayring. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Klagenfurt, AUT.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Sergio Nava-Muñoz, Mario Graff, and Hugo Jair Escalante. 2024. Analysis of systems' performance in natural language processing competitions. *Pattern Recognition Letters*.

Atul Kr Ojha, A Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. 2023. Proceedings of the 17th international workshop on semantic evaluation (semeval-2023). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. Semeval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.