Exploring the Influence of Label Aggregation on Minority Voices: Implications for Dataset Bias and Model Training

Anonymous ACL submission

Abstract

Resolving disagreement in manual annotation typically consists of removing unreliable annotators and using a label aggregation strategy such as majority vote or expert opinion to resolve disagreement. These may have the side-effect of silencing or under-representing minority but equally valid opinions. In this paper, we study the impact of standard label aggregation strategies on minority opinion representation. We investigate the quality and value of minority annotations, and then examine their effect on the class distributions in gold labels, as well as how this affects behaviour of models trained on the resulting datasets. Finally, we discuss the potential bias introduced by each method and how they can be amplified by the models.

1 Introduction

Disagreements among annotators are common for any annotation task, especially for subjective tasks such as sexism detection. They arise for two main reasons: carelessness or misunderstanding of the task; and differing yet equally valid opinions (e.g. men and women may disagree on what constitutes a sexist comment) (Beck et al., 2020).

Although disagreement information has recently been incorporated into training (Plank et al., 2014) and evaluation (Mostafazadeh Davani et al., 2022; Gordon et al., 2022), the predominant approach relies on label aggregation to generate gold labels (Zampieri et al., 2019). Disagreements are typically resolved by taking the majority vote or relying on an expert (Davidson et al., 2021; Founta et al., 2018; Waseem and Hovy, 2016). However, these strategies can introduce bias into the dataset, as they tend to disregard differing annotations that represent the views of a minority of the annotators, stemming from different cultural and social contexts (Jiang et al., 2021; Aroyo and Welty, 2013). It is potentially important to represent these minority voices in both data and downstream models, but there has been limited systematic investigation into how label aggregation strategies may introduce bias in this respect. The potential implications of this on models trained with such data also remain largely unexplored.

We address this gap by investigating different label aggregation strategies and their impact on minority opinion inclusion, as well as testing the downstream effects on model behaviour on 2 datasets for sexism detection. We also conduct a qualitative analysis to understand the type of disagreements between annotators.

Our results show that majority aggregation captures opinions well when sexism is clear and there is little difference of opinion. However, this produces fewer occurrences of fine-grained but important classes such as dehumanisation, and mistreatment of women, which may be important to recognise specifically, due to their potentially more harmful nature (Posetti et al., 2021). This effect is amplified in the predictions of models that are trained on the aggregated data. Overall, our investigation highlights the need for consideration of the most suitable strategy according to the data and objective of the task.

2 Related Work

A common approach for dataset creation is to employ multiple annotators to label input, and subsequently to derive a gold label using an aggregation strategy such as majority voting, averaging, or expert intervention (Sabou et al., 2014; Waseem and Hovy, 2016). There are two types of annotations that deviate from the gold label: noisy labels and labels influenced by differing experiences, beliefs, and opinions, including minority viewpoints that may not align with the majority or expert opinions. Efforts have been made to identify and address noisy labels (Kazai et al., 2013; Rottger et al., 2022; Sap et al., 2022; Beck et al., 2020).

Recently, studies have drawn attention to the under-representation of minority opinions (Prabhakaran et al., 2021), and researchers have criticised the notion of a single, definitive gold label (Mieleszczenko-Kowszewicz et al., 2023; Aroyo and Welty, 2015; Plank, 2022; Baan et al., 2022), as these approaches tend to prioritise the majority viewpoint to achieve state-of-the-art (SOTA) performance scores. Nevertheless, SOTA systems often rely heavily on gold labels for achieving superior performance on standard evaluation metrics (Vitsakis et al., 2023; Gordon et al., 2021). In response, methods have been explored to enhance gold label generation. Ren et al. (2023) proposes label confidence-based noise corrections and incorporates new data, while Wu et al. (2023) identifies labelling biases and trends (e.g., favouring a particular class) and implements adjustments. Moreover, Wei et al. (2023) examines the downstream impact of aggregating and utilising multiple labels.

In this study, we investigate various label aggregation strategies employed to generate a single gold label, assessing their impact on minority annotations and analysing their downstream effects on model behaviour.

3 Datasets

We have selected two publicly available datasets, both providing individual annotations and gold labels obtained after resolving disagreements.

Dataset	EDOS	GE
Lang	English	English
Posts	20k	13.6k
Platform	Gab, Reddit	Twitter
Annotators	19	105
Ann/post	3	5
Ann type	expert	MTurk
Selection	expert	convert & maj
Number of labels	T1-2, T2-4,T3-11	2

Table 1: Details of datasets

Call me sexist, but: This dataset (GE) (Samory et al., 2021) combines multiple datasets to produce sexism classes. Each post is annotated for the sexist nature of its content and the incivility of the phrasing. The sexism label is derived by combining the two via a scoring system.[1] Although the dataset contains 13.6k posts, we only use the 5.7k for which annotations from individual annotators are available.

[1]Details of binary label conversion in Appendix B

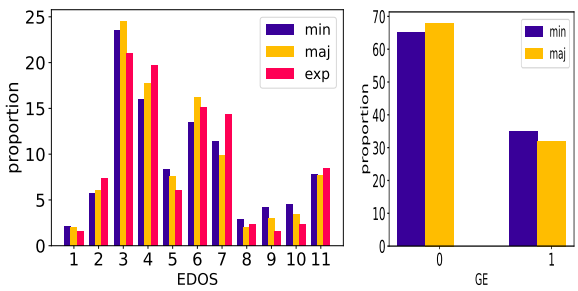

Figure 1: Class distributions of data for EDOS and GE

Explainable Detection of Online Sexism: EDOS is used in SemEval 2023 task 10 (Kirk et al., 2023), featuring a hierarchical class structure: Tier 1 - sexist and not sexist; Tier 2 - threats; derogation; animosity; and prejudice. Tier 3 is divided into: threats of harm (1); incitement and encouragement of harm (2); descriptive attacks (3); aggressive and emotive attacks (4); dehumanisation and sexual objectification (5); casual use of gendered slurs, profanities and insults (6); immutable gender differences (7); backhanded gender compliments (8); condescending explanations or unwelcome advice (9); supporting mistreatment of individual women (10); and supporting systemic discrimination against women (11).[2]

4 Investigating Aggregation on Data

We first investigate the impact of various label aggregation strategies on the distribution of labels within the dataset. In addition to majority vote and expert opinion[3], we also explore minority aggregation, where the gold label is determined by selecting the minority vote. We present the distribution of different labels for the EDOS and GE datasets in Figure 1.

We observe that **(a)** in GE, by using minority aggregation, more inputs are labelled as sexist, suggesting that some sexist posts may be overlooked with a majority vote; and **(b)** in EDOS, the ratio of certain labels, including threats of harm (1), dehumanisation and sexual objectification (5), backhanded gender compliments (8), condescending explanations or unwelcome advice (9), and supporting mistreatment of individual women (10), which are already not very frequent, would be further reduced in EDOS when using the majority or expert opinion.

[2]All analysis for EDOS T1 & T2 in Appendix A

[3]Only for EDOS - there is no expert annotation available for the GE dataset.

Looking at these distributions, minority aggregation captures the more nuanced forms of sexism, while those chosen more frequently by the majority tend to be easy to identify. Expert aggregation also captures sexism more widely. However, the risk of using this strategy is that the gold labels will then have the expert's biases. Another problem is that only a single label can be chosen in spite of multiple viable options, e.g., a post could be both dehumanising and contain gendered slurs. In this case, most annotators choose the more easily identifiable label.

5 Investigating Aggregation on Models

To understand the effect of different label aggregation strategies on model performance, we fine-tuned BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) on each.[4] Figure 2 shows the comparative class distributions for both models for GE and EDOS. Table 2 shows the class distributions of the test data for each strategy.

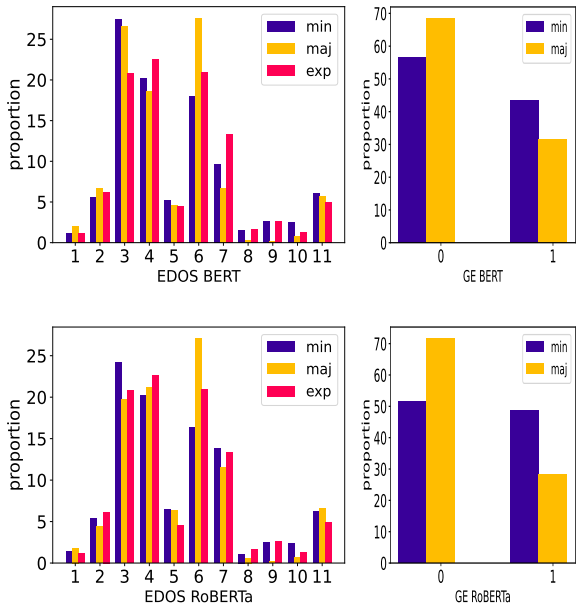

Figure 2: Class distributions from model predictions

We observe that **(a)** for both BERT and RoBERTa, there is a significant difference between majority and minority aggregation for GE. More posts are identified as sexist when minority aggregation is used. Interestingly, the gap between the sexist and non-sexist classes also increases in the model prediction: for sexist, the difference between minority and majority in the data is 2.86%, but in model prediction, it is between 9% and 20% points, indicating that these biases can be amplified by the model. **(b)** For both models, in EDOS, majority aggregation has a higher proportion of threats of harm (1), while it has very low counts for condescending explanations (9), backhanded compliments (8), and mistreatment of individual women (10). Minority aggregation has the highest proportion of descriptive attacks (2), dehumanisation and sexual objectification (5), condescending explanations (9), and supporting systemic discrimination of women (11). Unlike GE, the bias amplification is less significant. This could be due to data sparsity for some labels, particularly those representing the minority, and so the model does not learn it.

We see that the class distribution of the model predictions differs according to the label aggregation strategy used, with the effects of the data amplified in the model. Taking minority opinions into consideration can lead to over-predicting classes related to sexism. However, using minority aggregation and expert aggregation helps in classifying posts into the more subtle classes that are harder to identify, as well as those which constitute more harm to the receiver.

We suggest that rather than a single label aggregation strategy being ideal, it should be task-dependent, based on factors such as the objective of the task. For example, content moderation on a Twitch stream where no sexism is tolerated aims to capture all nuances of sexism, while content moderation policies on X or Facebook might be more lenient, where one might want to be more certain that a post is considered sexist by the majority of users. Considering whether a post might be viewed as sexist by the receiver depends on their personal sensitivities but may be more useful than considering the intent of the sender.

6 Minority Opinions Or Noise

In this work, we have considered labels selected by the minority label aggregation as minority opinions. However, as mentioned previously, some deviations from the majority or expert annotations may come from noisy annotations. To investigate this, we select 100 posts with discrepancies between annotators from each dataset. Subsequently, we manually categorize them as either differing opinions or noise.[5]

[4] Model parameters and F1 scores in Appendix C.

[5] Data provided as additional file. Text rephrased using a fine-tuned llama2 model (https://huggingface.co/

		Data			BERT			RoBERTa		
dataset	class	majority	minority	expert	majority	minority	expert	majority	minority	expert
GE	Not sexist	68.54	65.68 $_{\downarrow 2.86}$	-	68.54	56.63 $_{\downarrow 9.91}$	-	71.71	51.43 $_{\downarrow 20.28}$	-
	Sexist	31.46	34.32 $_{\uparrow 2.86}$	-	31.46	43.37 $_{\uparrow 9.91}$	-	28.29	48.57 $_{\uparrow 20.28}$	-
EDOS	1. Threats	1.75	2.10 $_{\uparrow 0.35}$	1.75	1.98 $_{\uparrow 0.84}$	1.22 $_{\uparrow 0.04}$	1.14	1.82 $_{\uparrow 0.23}$	1.42 $_{\downarrow 0.17}$	1.59
	2. Incitement	5.65 $_{\downarrow 1.10}$	5.53 $_{\downarrow 1.22}$	6.75	6.69 $_{\uparrow 0.54}$	5.62 $_{\downarrow 0.53}$	6.15	4.39 $_{\downarrow 3.00}$	5.39 $_{\downarrow 2.00}$	7.39
	3. Descriptive	25.37 $_{\uparrow 4.70}$	23.99 $_{\uparrow 3.32}$	20.67	26.60 $_{\uparrow 5.79}$	27.43 $_{\uparrow 6.62}$	20.81	19.68 $_{\uparrow 3.50}$	24.18 $_{\uparrow 8.00}$	16.18
	4. Agg & emo	18.64 $_{\downarrow 3.41}$	15.68 $_{\downarrow 6.37}$	22.05	18.66 $_{\downarrow 3.95}$	20.21 $_{\downarrow 2.40}$	22.61	21.20 $_{\downarrow 0.60}$	20.15 $_{\downarrow 1.65}$	21.80
	5. Dehumanise	8.14 $_{\uparrow 1.98}$	8.68 $_{\uparrow 2.52}$	6.16	4.63 $_{\uparrow 0.09}$	5.21 $_{\uparrow 0.67}$	4.54	6.28 $_{\uparrow 1.79}$	6.48 $_{\uparrow 1.99}$	4.49
	6 Profanity	15.65 $_{\uparrow 1.04}$	13.57 $_{\downarrow 1.04}$	14.61	27.63 $_{\uparrow 6.72}$	17.98 $_{\downarrow 2.93}$	20.91	27.10 $_{\uparrow 4.74}$	16.39 $_{\downarrow 5.97}$	22.36
	7. Immutable diff	9.32 $_{\downarrow 3.91}$	11.52 $_{\downarrow 1.71}$	13.23	6.76 $_{\downarrow 6.67}$	9.60 $_{\downarrow 3.74}$	13.34	11.51 $_{\downarrow 4.11}$	13.83 $_{\downarrow 1.79}$	15.62
	8. Backhand comp	2.20 $_{\downarrow 0.93}$	3.43 $_{0.30}$	3.13	0.37 $_{\downarrow 1.24}$	1.49 $_{\downarrow 0.12}$	1.61	0.61 $_{\downarrow 1.64}$	1.09 $_{\downarrow 1.16}$	2.25
	9. Condescend	3.28 $_{\uparrow 1.10}$	4.11 $_{\uparrow 1.93}$	2.18	0.15 $_{\downarrow 1.32}$	2.60 $_{\uparrow 1.13}$	1.47	0.15 $_{\downarrow 0.32}$	2.45 $_{\uparrow 1.98}$	0.47
	10. Mistreat	2.82 $_{\uparrow 0.69}$	3.93 $_{1.80}$	2.13	0.81 $_{\downarrow 1.84}$	2.49 $_{\downarrow 0.16}$	2.65	0.68 $_{\downarrow 0.35}$	2.34 $_{\uparrow 1.31}$	1.03
	11. Discrimination	7.18 $_{\downarrow 0.15}$	7.45 $_{\uparrow 0.12}$	7.33	5.73 $_{\uparrow 0.81}$	6.10 $_{\uparrow 1.18}$	4.92	6.59 $_{\downarrow 0.24}$	6.26 $_{\downarrow 0.57}$	6.83

Table 2: Class distributions % from test data & model predictions & difference from the original gold labels

We find that (a) for GE, 69% of the disagreements are 'valid', occurring for two main reasons. First, gender-based compliments (e.g. *"women are beautiful like flowers. Delicate and deserving of tender loving care"*) may be considered positive or sexist depending on the annotator's context. Second, gendered slurs and profanity (such as 'bitch' or 'whore') used casually seem to cause disagreement. The reason for this is not always clear, though sometimes it may be due to misunderstood sarcasm (e.g. "I stay up till 2 smoke dope play video games, but she's the bitch for walking out on me"). (b) for EDOS, 76% of the disagreements are 'valid', mostly arising from annotators choosing one of many viable options. They frequently struggle to choose between derogation and 2 other classes - animosity and prejudice, and similarly between aggressive and emotive attacks (4) and slurs, profanities and insults (6). Again, sarcasm seems to be one of the main causes of noisy disagreements. In this dataset, we also found quoted tweets to be problematic since the comments were of the opposite point of view to the original message with which they were concatenated. For example, in the quoted tweet *"women should be happy when men leer at them and catcall. It means they're attractive na na na YEAH SURE!"*, the capitalised comment at the end is disagreeing with the original message.

Following this analysis, we conclude that for EDOS and GE, most disagreements are indeed 'valid', and thus, minority aggregation is generally a viable strategy for investigating the class distribution of minority voices in the data and comparing the label aggregation strategies. However, this depends on the quality of the dataset, and if the annotations are too noisy, we recommend first using a de-noising technique.

7 Conclusion

We have investigated the impact of 3 label aggregation strategies on the representation of minority opinions, and the effect of using the gold labels generated by each on downstream model behaviour. We also verified via a qualitative study that most of the disagreements in our 2 sexism detection datasets arise from genuine differing opinions rather than noisy annotations, though this might not be the case in other datasets.

We conclude that each label aggregation strategy has its own bias and leads to different label distributions and model behaviour. Using majority aggregation will lead to a dataset and model that may exclude certain opinions, particularly those that are more sensitive to sexism. Using expert aggregation is more sensitive to sexism than majority aggregation) (i.e. it will label more posts as sexist), but runs the risk of biasing the dataset towards a single expert's opinion, so is also less inclusive. Incorporating minority opinions gives the most diverse and inclusive dataset but can lead the trained model to over-classify posts as sexist.

The label aggregation strategy should, therefore, be chosen carefully, keeping in mind the objective of the task and use case. Moreover, when choosing a label aggregation strategy, one should be mindful of minority opinions. Where feasible, we suggest performing at least an analysis of how the label distribution changes with label aggregation strategy, and comparing it with the minority label aggregation to ensure that the chosen strategy does not introduce substantial biases against minority voices.

georgesung/llama2_7b_chat_uncensored)

8 Limitations

Working with only 2 datasets limits the generalisability of our results, especially for datasets that contain many noisy annotations. In those cases, an additional noise removal step is required. Differing opinions and data ambiguity are general problems for all datasets, but fine-grained labels should be used cautiously. They are useful only if easily distinguishable (by humans as well as machines); otherwise, they unnecessarily contribute to disagreements which are less related to differing opinions and more about being forced to make a single choice, which can skew results to some extent. One should also be mindful about the choice of expert used in the case of expert arbitration for disagreements: in our study, only one dataset used an expert and therefore, the results of comparing an expert selection strategy with the other strategies might differ in other datasets.

9 Ethics

Our work is part of ongoing research into bias and fairness sexism detection. While it deals with sensitive and offensive topics and language, it furthers the study of inclusivity and diversification of NLP to better serve and accommodate people from varying demographic groups. The datasets we use contain no sensitive or personal information about the annotators - any such information has already been either anonymised or not made publicly available by the original authors of the datasets. The work has undergone ethics approval from the authors' institution.

References

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics.

Sam Davidson, Jordan Hosier, Yu Zhou, and Vijay Gurbani. 2021. Improved named entity recognition for noisy call center transcripts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 361–370, Online. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):1–22.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.*, 16(2):138–178.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Wozniak, Ewa Dziecioł, Przemysław Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP. CEUR Workshop Proceedings*.

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.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Julie Posetti, Nabeelah Shabbir, Diana Maynard, Kalina Bontcheva, and Nermine Aboulez. 2021. The chilling: Global trends in online violence against women journalists. *New York: United Nations International Children's Emergency Fund (UNICEF)*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lijuan Ren, Liangxiao Jiang, and Chaoqun Li. 2023. Label confidence-based noise correction for crowdsourcing. *Engineering Applications of Artificial Intelligence*, 117:105624.

Paul Rottger, 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, Seattle, United States. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669, Toronto, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. 2023. To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2523–2535.

Ming Wu, Qianmu Li, Fei Yang, Jing Zhang, Victor S Sheng, and Jun Hou. 2023. Learning from biased crowdsourced labeling with deep clustering. *Expert Systems with Applications*, 211:118608.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A EDOS T1 and T2

A.1 Investigating label aggregation strategies on data

From Table 3 we observe that in T1, majority aggregation and expert aggregation are similar, also with fewer posts labelled sexist than minority aggregation. In T2, majority aggregation labels the most posts as sexist for class derogation, while minority aggregation has the highest proportion for class prejudice. expert aggregation is highest for classes threats and animosity.

dataset	class	majority	minority	expert
EDOS T1	not sexist	75.62	72.11	75.73
EDOS T1	sexist	24.38	27.89	24.27
EDOS T2	threats	8.46	8.12	9.22
EDOS T2	derogation	48.92	44.20	44.11
EDOS T2	animosity	30.99	33.88	35.10
EDOS T2	prejudice	11.63	13.80	11.58

Table 3: Class distributions for EDOS T1 and T2

A.2 Investigating label aggregation strategies on Models

From Figure 3 and Table 4 in T1 majority aggregation has the lowest proportion of class sexist and minority aggregation the highest. In T2, majority aggregation and expert aggregation have a higher proportion of class threats and animosity than minority aggregation. Minority aggregation has a higher proportion for class prejudice.

dataset	class	BERT		
		majority	minority	expert
EDOS T1	not sexist	77.36	69.77	74.63
EDOS T1	sexist	22.64	30.23	25.37
EDOS T2	threats	8.20	5.45	8.48
EDOS T2	derogation	49.44	49.74	48.32
EDOS T2	animosity	35.78	34.11	37.18
EDOS T2	prejudice	6.58	10.70	6.02
		RoBERTa		
EDOS T1	not sexist	80.54	75.00	80.76
EDOS T1	sexist	19.46	25.00	19.24
EDOS T2	threats	8.20	5.45	8.48
EDOS T2	derogation	49.44	49.74	48.32
EDOS T2	animosity	35.78	34.11	37.18
EDOS T2	prejudice	6.58	10.70	6.02

Table 4: Class distributions of model predictions

B Dataset Information

All sensitive information is anonymised or not made publicly available by the original authors of the datasets.

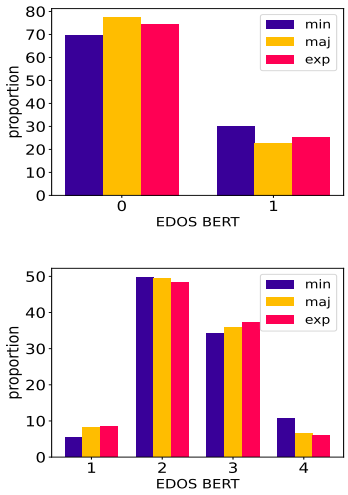

Figure 3: Class distributions of model predictions for EDOS T3 and GE, based on different label aggregation strategies

B.1 EDOS Class Structure

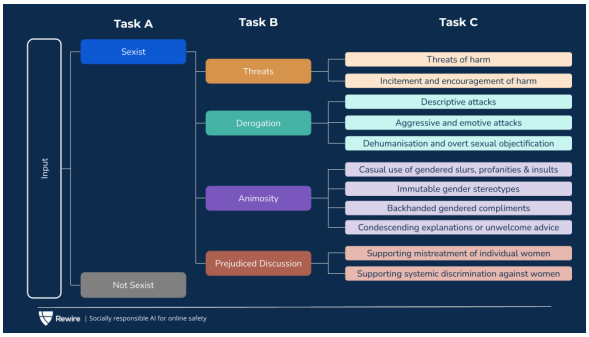

Figure 4: EDOS dataset class structure. Figure from Kirk et al. (2023)

B.2 GE Class Structure

The sexist nature of content classes C1 to C6 and the incivility of phrasing classes P1 to P3 from Samory et al. (2021) are explained below.

Content

1. Behavioural expectations
2. Stereotypes and comparative opinions
3. Endorsement of inequality
4. Denying inequality and rejecting feminism
5. Maybe sexist: can't tell without context
6. Not sexist: not a direct statement of sexist opinion

Phrasing

1. Uncivil and sexist: attacks, foul language, or derogatory depictions directed towards individuals because of their gender

2. Uncivil but not sexist: offensive phrasing or message that is not clearly directed towards the target‘s gender

3. Civil (can be sexist or not sexist): neutral phrasing that does not contain offenses or slurs

Each annotator annotates a post for C and P. If they choose classes C1 to C4 in C or P1 in P, their annotation for a post is sexist. If not, it is not sexist. A majority vote of the 5 annotators‘ sexism choices is used to generate the gold label.

C Model Details

Model parameters for EDOS and GE in Table 6. The code is run on an NVIDIA GeForce RTX 3090 GPU. F1 scores in Table 5.

dataset	BERT		
	majority	minority	expert
GE	77.6	75.3	-
EDOS T1	77.3	69.8	78.3
EDOS T2	80.2	78.4	82.2
EDOS T3	79.2	76.6	80.2
	RoBERTa		
GE	76	68.7	-
EDOS 1	71.9	64.3	69
EDOS 2	76.2	75.9	76.2
EDOS 3	72.2	60.6	77.8

Table 5: F1 scores for models

parameter	GE	EDOS T1	EDOS T2	EDOS T3
num labels	2	2	5	11
max num epochs	4	10		
model	bert-base-uncased, roberta-base for Sequence Classification			
max sequence length	max_length			
learning rate	5e-5			
batch size	16			
optimizer	AdamW			
seed	constant (for reproducibility)			
loss	default for Sequence Classification by transformers library			

Table 6: Model parameters