Revisiting Annotation of Online Gender-Based Violence

Gavin Abercrombie[♡], Nikolas Vitsakis[♡], Aiqi Jiang[♡♣], Ioannis Konstas[♡♠]
[♡]Interaction Lab, Heriot-Watt University, Edinburgh
[♣]Computational Linguistics Lab, Queen Mary University of London
[♠]Alana AI
{g.abercrombie, nv2006, a.jiang, i.konstas}@hw.ac.uk

Abstract

Online Gender-Based Violence (GBV) is an increasing problem, but existing datasets fail to capture the plurality of possible annotator perspectives or ensure representation of affected groups. In a pilot study, we revisit the annotation of a widely used dataset to investigate the relationship between annotator identities and underlying attitudes and the responses they give to a sexism labelling task. We collect demographic and attitudinal information about crowd-sourced annotators using two validated surveys from Social Psychology. While we do not find any correlation between underlying attitudes and annotation behaviour, ethnicity does appear to be related to annotator responses for this pool of crowd-workers. We also conduct initial classification experiments using Large Language Models, finding that a state-of-the-art model trained with human feedback benefits from our broad data collection to perform better on the new labels. This study represents the initial stages of a wider data collection project, in which we aim to develop a taxonomy of GBV in partnership with affected stakeholders.

Keywords: Gender-Based Violence, Misogyny, Sexism, Abusive language, Hate speech, Annotation, LLMs

1. Introduction

Gender-Based Violence (GBV) is an increasing problem in online spaces, affecting around half of all women and targeting those from marginalised groups in particular (Glitch UK and EVAW, 2020).

To counter this, there have been attempts to facilitate moderation of such content using natural language processing (NLP) methods to automatically identify misogynistic language. As a result, there now exist a number of datasets designed for supervised classification of various forms of GBV.

However, Abercrombie et al. (2023) identified a number of weaknesses in approaches to the creation of corpora for this task. One prominent shortcoming has been the lack of representation in the labelled data of people's different points of view, and particularly of people with the minoritised identites who are best placed to recognise GBV.

To fill this gap, we aim to revisit the task of annotating online text following *strongly perspectivist* data practices (Abercrombie et al., 2022; Basile et al., 2023; Cabitza et al., 2023) in the collection, modeling, and distribution of datasets, preserving the labels provided by multiple annotators. In this pilot study, we re-annotate a recently collected dataset, this time with (1) multiple ratings per item; and (2) demographic and attitudinal information about the annotators.

We make the following **research contributions**: (1) we collect a corpus of the responses of multiple annotators to each item in a subset of a widely used English language GBV dataset, along with demographic and attitudinal information about the annotators. We make this resource available to the research community at `https://github.com/GavinAbercrombie/EquallySafeOnline`. (2) We analyse this data to investigate the relationship between annotator demographics and attitudes and the labels that they apply to items. (3) We conduct benchmark experiments to investigate the capabilities of current state-of-the-art systems in identifying GBV in text.

2. Background

The GBV framework encompasses phenomena such as sexism, misogyny, and violence against women and girls—although it also recognises that people of all genders are affected by GBV.[1] It was first introduced by the United Nations (UN General Assembly, 1993; United Nations, 2021). For further details of the theoretical foundation of this framework and motivation for its application to the field of NLP, see Abercrombie et al. (2023).

Annotator Variability and Perspectivist Data Practices While labels collected for supervised classification have traditionally been aggregated to a single 'gold' or 'ground truth' label for each item, recent work has recognised that this can lead to the erasure of minoritised voices, and can subsequently hinder the ability of classifiers to recognise subtle and implicit forms of abuse. *Standpoint theory* (Harding, 1991) contends that only people with

[1]For example, men face pressure to conform to masculine gender role norms (European Institute for Gender Equality, 2021).

NLPerspectives 2024@LREC-COLING 2024, pages 31–41
21 May, 2024.

relevant lived experience are capable of recognising subtle, implicit abuse such as stereotypes and micro-aggressions. According to the *matrix of domination* Collins (2002), this experience likely results from sharing intersectional social categorisations with the intended targets of the abuse. With label aggregation, the labels provided by people with such identities and experiences are often erased.

There is now a growing recognition of the need to collect, retain, and distribute labels provided by multiple annotators, and this has been adopted across a range of NLP tasks (Plank, 2022). This is particularly so for controversial tasks such as identification of abusive or toxic language, in which annotator variation may be caused by differences of opinion or ideology (e.g. Akhtar et al., 2021; Almanea and Poesio, 2022; Cercas Curry et al., 2021; Leonardelli et al., 2021). *Strong Perspectivism* aims to preserve this variation through modelling, classification, and evaluation (Cabitza et al., 2023). For further background, see the Perspectivist Data Manifesto at `https://pdai.info/`.

Beliefs and attitudes We ground our theoretical approach in the Dual Process Motivational Model of Ideology and Prejudice (Duckitt and Sibley, 2009; Duckitt, 2001), specifically, the differential effect hypothesis aspect of the model. This hypothesis explains that sociopolitical and ideological attitudes linked to prejudice can be adequately captured by two distinct but often related constructs, Right Wing Authoritarianism (RWA) and Social Dominance Orientation (SDO) related attitudes. The former explains propensity towards cultural conservativism and traditionalism related beliefs (Altemeyer, 1983; Feather and McKee, 2012; Van Assche et al., 2019), while the latter explains favourable views towards social hierarchies of power, where inequality between groups is seen as inevitable or even natural (Christopher and Wojda, 2008; Pratto et al., 1994; Jagayat and Choma, 2021).

Both of these constructs have been extensively assessed and found to be strongly related and to explain different forms of sexism and gender based discrimination. RWA has been found to be a good predictor of 'benevolent sexism', that is attitudes that force women into traditional predefined roles (i.e., being a mother) that seem subjectively advantageous but are, in reality, marginalising and disempowering (De Geus et al., 2022). SDO pertains towards beliefs towards deterministic gender imbalances justifies male dominance through a disparaging charactrisation of women (La Macchia and Radke, 2020; De Geus et al., 2022).

Taken as a whole, these constructs have been widely used to explain gender based discrimination, through both offline (Perez-Arche and Miller, 2021; Christopher and Wojda, 2008; Patev et al., 2019) and online (Jagayat and Choma, 2021) contexts, and have been validated across cultures (Çetiner and Van Assche, 2021; De Geus et al., 2022), while also being used to explain that such beliefs transcend demographic identities (Renström, 2023).

3. Related Work

Annotator Characteristics A number of NLP studies have attempted to group annotators according to their demographic characteristics and use these factors as predictors of their responses to items (e.g. Akhtar et al., 2021; Gordon et al., 2022; Goyal et al., 2022). However, it has repeatedly been shown that demographic characteristics do not predict annotator behaviour at the individual level (Beck et al., 2023; Biester et al., 2022; Chulvi et al., 2023; Orlikowski et al., 2023).

Several recent studies have therefore attempted to uncover the *social attitudes* of annotators and relate the results to the responses they produce. Sap et al. (2022) surveyed crowd workers, and found that those with racist beliefs were less likely to consider anti-Black language to be toxic. While they conducted two annotation experiments, one with many annotators but few items and the other with fewer annotators but more items, our data collection aims at both breadth and depth.

Hettiachchi et al. (2023) measured the responses of crowd workers to a misogynistic language labelling task, as well as their moral attitudes (in addition to demographic and personality-type information), which they obtained through survey questions. They found that higher *moral integrity* and lower *benevolent sexism* scores correlated with label agreement with expert annotators.

It is in this vein that we seek to discover the relationship between the demographics, social attitudes, and responses to GBV identification tasks provided by crowd-sourced annotators.

Modelling multiple perspectives Previously, research on modelling with label variation focused on using disagreements to inform improved prediction of a single aggregated label (see Uma et al., 2021, for a survey). More recent work has attempted to preserve these variations at inference. For example, Cercas Curry et al. (2021) and Mostafazadeh Davani et al. (2022) predicted each annotators' responses to abusive language identification tasks, the latter using multi-task learning. The SEMEVAL 2023 shared task on learning with disagreement (Le-Wi-Di) (Leonardelli et al., 2023) explicitly attempted to focus the field on attention to levels of disagreement between annotators when labelling text for toxicity. This drew a number of approaches including that of Vitsakis et al. (2023), who focused on preserving the full range of points of view at inference at the expense of overall classification performance.

Toxic language detection with LLMs With the recent explosion in the use of LLMs, there has been a paradigm shift in approaches to identification of phenomena such as toxic language as researchers have shifted from training models from scratch (e.g. Davidson et al., 2017; Jiang et al., 2022) or fine-tuning pre-trained models (e.g. Caselli et al., 2020; Cercas Curry et al., 2021) to harnessing the power of the new models to classify items with few, or even no, specific examples.

To benchmark the new version of the dataset, we present the results of initial experiments using a recent open-source LLM (see §5).

4. Data Collection and Analysis

4.1. Datasets

We selected the test set of a previously published dataset: Explainable Sexism (EDOS[2]), (Kirk et al., 2023), which we chose as (1) Abercrombie et al. (2023) had identified it as among the resources most thoroughly grounded in social science theory; and (2) it is English language, the language of our stakeholder partners, with whom we are co-designing GBV-mitigation tools.

Pre-processing of the data consisted of filtering out any items which include images. We leave annotation of multi-media items for future work. This left 3,896 items, of which we randomly selected 400 for re-annotation. We will release all code for implementation of the data collection and processing procedure on acceptance.

4.2. Annotators

We recruited 41 annotators on the Amazon Mechanical Turk crowd-sourcing platform. To ensure attentive participation, we recruited only workers with at least 500 completed tasks and a $\geq 98\%$ approval rating. For comparison with the original EDOS labels, which were labelled by annotators from the United Kingdom, we also limited recruitment to workers based in the UK. Prior to annotation, in a separate task batch (i.e. at an earlier time and date), we collected demographic information and responses to questions from two surveys designed to measure the attitudes of the workers.

Demographic information The annotators self-reported as 16 women, 24 men, and one other. We supply a full Data Statement in Appendix A.

[2]Language resource: (Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul, 2023)

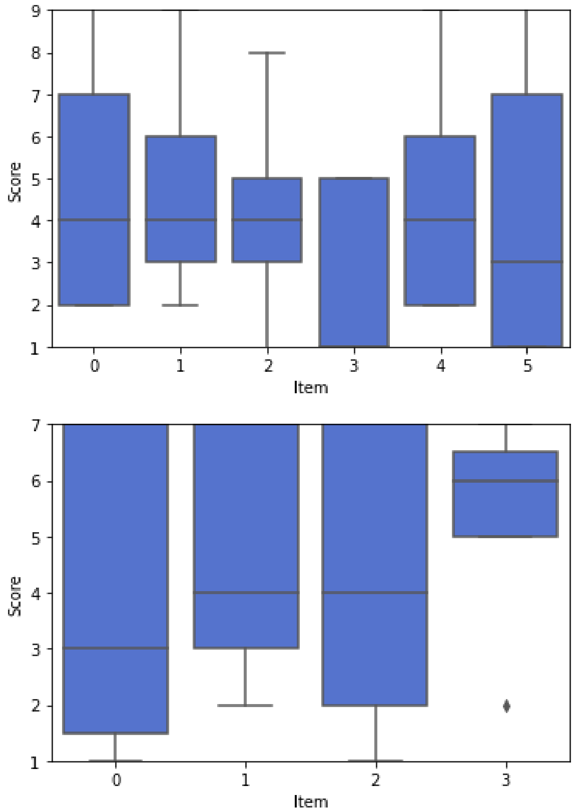

Figure 1: Responses to the six VSA and four SSDO items on $[1-9]$ and $[1-7]$ scales, respectively.

Attitudes To measure the annotators attitudes, we used survey questions from two verified scales widely used in social psychology: the Very Short Authoritarianism (VSA) scale (Bizumic et al., 2018) and the Short Social Dominance Orientation (SSDO) (Pratto et al., 2013) scales to measure Right Wing Authoritarianism (RWA) and Social Dominance Orientation (SDO) respectively. Further details of these scales are provided in Appendix B.

We find that for VSA, the annotators tend slightly towards the centre of the scale ($m = 4.55, s = 3.26$), while dor SSDO, they are somewhat towards the more dominant end of the scale on average ($m = 5.36, s = 3.79$), as shown in Figure 1. Overall, the annotators display a mix of more to less authoritarian and dominant attitudes.

4.3. Data Labelling

Annotators were provided with the original instructions from EDOS. We collected up to ten responses from different annotators per item, which we examine here.

Intra-Annotator Agreement We measure the levels of agreement between our recruited annotators as well as between the aggregated labels, decided by majority vote, and the original EDOS labels.

We report raw percentage agreement and Krippendorf's alpha, which can measure agreement between two or more raters and also handle missing values (Gwet, 2014).

Crowd workers		Majority vote *v.* Original labels	
α	%	α	%
0.11	56.7	0.37	73.2

Table 1: Reliability as measured by inter-annotator agreement (Krippendorf's α and Cohen's κ and raw percentage agreement (%)). Cohen's κ for multiple annotators is calculated pairwise.

As shown in Table 1, agreement between the crowd-sourced annotators is low. In fact, they only agree unanimously on five items in the dataset (0.0125%). Although the aggregated labels are somewhat closer to the original labels (also produced by majority vote), agreement is still quite poor at only $\kappa = 0.37$. Where the aggregated label doesn't agree with the original, we find discord among the new annotators in 100 per cent of cases. A comparison of the original and new test set labels is presented in Table 2, where we can see that the crowd-workers consider more items to be sexist than the original annotators. In the following paragraphs, we investigate whether information about annotators can explain the observed variations.

Original		New	
Sexist	*Not sexist*	*Sexist*	*Not sexist*
108	292	127	273

Table 2: Aggregated classes of the two label sets.

Group Responses: Demographics We examine the correlations between annotators' demographic characteristics and their propensity to label items as 'sexist'. Aside from age, which is continuous, we binarised each variable as the majority category versus the others, such that *gender* becomes *female*/*non-female* etc.[3] As shown in Table 3, only *white* ethnicity correlated with labelling behaviour to a statistically significant degree ($p < 0.05$).

Group Responses: Social Attitudes We now turn to the attitude scale scores (see Table 4). We find no correlation between responses to the VSA scale and annotation behaviour. Although higher scores on the SSDO do correlate with annotators propensity to label items as *sexist*, this result is not statistically significant at $p = 0.14$.

[3]We recognise that the resulting binary categories, e.g. *bi-sexual/not bi-sexual* may not be representative of the underlying population.

Demographic variable	**Correlation** Spearman's r	**Significance** p-value
Age	0.12	0.61
Gender: *female*	-0.40	0.08
Ethnicity: *white*	0.51	0.02
Sexuality: *bi-.*	0.54	0.15
Politics: *right*	-0.21	0.39

Table 3: Correlations between characteristics and the percentages of items labelled as 'sexist'.

Attitude scale	**Correlation** Spearman's r	**Significance** p-value
VSA	0.08	0.78
SSDO	0.42	0.14

Table 4: Correlations between attitudinal survey scores and percentage of items labelled as 'sexist'.

5. Initial classification experiments

To investigate whether our broader label collection provides richer information for automated classifiers, we benchmark the new data and compare with performance on the original labels. For this, we aggregate the labels by majority vote.

We select three pre-trained models as our baselines for the experiments. `Llama2` represents the recent trend of LLMs developed using Reinforcement Learning with Human Feedback (Touvron et al., 2023). `DeBERTaV3` (He et al., 2023) are widely used `BERT`-based architectures with high performances across NLP benchmarks. Antypas and Camacho-Collados (2023) provide a fine-tuned version of the twitter-based pre-trained model (Loureiro et al., 2023) based on 13 different hate speech datasets in English.

We fine-tune the models on the two sets of labels separately, and compare performance against the majority class of the original labels (*not sexist*). As we have somewhat unbalanced classes, we report macro F1, as well as accuracy scores.

Model	**Original Label** mF1	Acc	**New Label** mF1	Acc
Majority Vote	42.26	73.18	40.56	68.25
$\texttt{DeBERTa}_{base}$	42.91	70.43	40.63	68.42
$\texttt{RoBERTa}_{hate}$	65.22	71.68	62.39	67.92
`Llama2`	50.60	54.64	51.79	55.39

Table 5: Results on the sexist text detection task.

Table 5 shows classification results. All three models demonstrate better performance (as measured by F1 score). However, $\texttt{DeBERTa}_{base}$ only does marginally better. Results from $\texttt{RoBERTa}_{hate}$ underline the strength of models tailored for a specific task, such as sexism detection in this case. While the performance of `Llama2` lies between these two, it is the only model that performs better on the newly collected labels than the originals.

6. Discussion and Conclusion

This paper presents an initial foray into revisiting the annotation of GBV with the aim of capturing diverse perspectives and ensuring the presence of affected voices throughout the classification pipeline.

Low agreement rates show that annotators interpret many of the itmes differently, and while our experiments with capturing the annotators' underlying attitudes do not yield any significant correlations, we do find a potential link between the reported ethnicity of these annotators and their responses. In future work we aim to expand data collection to achieve greater statistical power and further examine these potential links between annotators' underlying attitides and the perspectives they apply to the GBV labelling task.

Initial classification results using `Llama2` suggest some promise that sophisticated models that incorporate human feedback may be able to exploit the rich information that comes from broader data collection practices. Future experiments will therefore focus on modelling the plurality of perspectives represented in the multi-label data, and exploring ways to ensure that minoritised voices are not subsumed by the majority.

Limitations

We recognise that our annotator pool for this pilot study is relatively small, and may not be representative of the population of workers on the crowdworking platform. Future work will aim to explore these factors further with (1) a larger sample; (2) other GBV datasets, such as Detection of Online Mysogyny (Guest et al., 2021). Although these datasets are among the most solidly theory-driven available, they still have several shortcomings with regards to the tenets of (i) perspectivist data practices, (ii) participatory design and design justice theory, and (iii) the GBV framework. Ultimately, we need new taxonomies and annotation schema, and the collection of new datasets. We hope that these initial efforts will inform future work in this area.

Ethical Considerations

IRB approval This study was approved by the institutional review board (IRB) of our Heriot-Watt University as project $2023 - 5536 - 8232$.

Annotator welfare and compensation As annotators were exposed to potentially upsetting language, we took the following mitigation measures:

- Participants were warned about the content (1) before accepting the task on the recruitment platform, (2) in the Information Sheet provided at the start of the task, and (3) in the Consent Form where they acknowledged the potential risks.
- Participants were required to give their consent to participation.
- They were able to leave the study at any time on the understanding that they would be paid for any completed work.
- The task was kept short (all participants completed each round in under 30 minutes) to avoid lengthy exposure to upsetting material.

Following the advice of Shmueli et al. (2021) we paid participants at a rate that was above both Prolific's current recommendation of at least £9.00 GBP/$12.00 USD[4] and the Living Wage in our jurisdiction, which is considerably higher.

We follow the recommendations of Kirk et al. (2022) on presenting harmful text both to annotators and to the readers of this document.

Annotator identities Due to the size of our annotation pool, for this study, analysis of annotators' demographic characteristics was limited to individual features. We recognise that responses to GBV are influenced by complex intersectional identities that we have been unable to capture here, but which will be the focus of future data collection and analysis.

Author positionality Tackling abusive language is an inherently political task, in which every decision made by researchers and developers (consciously or by default) has potential ramifications for affected stakeholders. We approach this topic through the prism of design justice (Costanza-Chock, 2020), and are actively working with experts from relevant NGOs to co-design technical solutions to online GBV. We therefore reject status quo practices that do not centre those most affected by GBV. However, while the design and engineering aspects of this work are based on feminist thought and theory, this does not affect the experiments and statistical analyses we conduct, which follow standard scientific practice.

Acknowledgements

Gavin Abercrombie, Aiqi Jiang, and Ioannis Konstas were supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1). We thank the NLPerspectives reviewers for their helpful comments and feedback.

[4] `https://www.prolific.co/blog/how-much-should-you-pay-research-participants`

Bibliographical References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

Julian Aichholzer and Clemens M Lechner. 2021. Refining the short social dominance orientation scale (SSDO): A validation in seven European countries. *Journal of Social and Political Psychology*, 9(2):475–489.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Bob Altemeyer. 1983. *Right-wing authoritarianism*. Univ. of Manitoba Press.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Flavio Azevedo, John T Jost, Tobias Rothmund, and Joanna Sterling. 2019. Neoliberal ideology and the justification of inequality in capitalist societies: Why social and economic dimensions of ideology are intertwined. *Journal of Social Issues*, 75(1):49–88.

Valerio Basile, Gavin Abercrombie, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, Elisa Leonardelli, and Sara Tonelli, editors. 2023. *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP (and Beyond) @ECAI2023*. CEUR, Krakow, Poland.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Boris Bizumic, John Duckitt, et al. 2018. Investigating right wing authoritarianism with a very short authoritarianism scale. *Journal of Social and Political Psychology*, 6:129–150.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Şeyda Dilşat Çetiner and Jasper Van Assche. 2021. Prejudice in Turkey and Belgium: The cross-cultural comparison of correlations of right-wing authoritarianism and social dominance orientation with sexism, homophobia, and racism. *Analyses of Social Issues and Public Policy*, 21(1):1167–1183.

Andrew N Christopher and Mark R Wojda. 2008. Social dominance orientation, right-wing authoritarianism, sexism, and prejudice toward women in the workforce. *Psychology of Women Quarterly*, 32(1):65–73.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the Second Workshop on Perspectivist Approaches to NLP (NLPerspectives)*.

Patricia Hill Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Sasha Costanza-Chock. 2020. *Design Justice Community-Led Practices to Build the Worlds We Need*. MIT Press.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Roosmarijn De Geus, Elizabeth Ralph-Morrow, and Rosalind Shorrocks. 2022. Understanding ambivalent sexism and its relationship with electoral choice in britain. *British Journal of Political Science*, 52(4):1564–1583.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.

John Duckitt. 2001. A dual-process cognitive-motivational theory of ideology and prejudice. In *Advances in experimental social psychology*, volume 33, pages 41–113. Elsevier.

John Duckitt and Chris G Sibley. 2009. A dual-process motivational model of ideology, politics, and prejudice. *Psychological inquiry*, 20(2-3):98–109.

European Institute for Gender Equality. 2021. Traditional norms of masculinity. Available at `https://eige.europa.eu/publications-resources/toolkits-guides/gender-equality-index-2021-report/traditional-norms-masculinity?language_content_entity=en`.

Norman T Feather and Ian R McKee. 2012. Values, right-wing authoritarianism, social dominance orientation, and ambivalent attitudes toward women. *Journal of Applied Social Psychology*, 42(10):2479–2504.

Glitch UK and EVAW. 2020. The ripple effect: COVID-19 and the epidemic of online abuse.

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.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Kilem L Gwet. 2014. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, LLC.

Sandra Harding. 1991. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets.

Arvin Jagayat and Becky L Choma. 2021. Cyber-aggression towards women: Measurement and psychological predictors in gaming communities. *Computers in human behavior*, 120:106753.

Andrew T Jebb, Vincent Ng, and Louis Tay. 2021. A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

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

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.

Stephen T La Macchia and Helena RM Radke. 2020. Social dominance orientation and social dominance theory. *Encyclopedia of personality and individual differences*, pages 5028–5036.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. 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. 2020. RoBERTa: A robustly optimized BERT pretraining approach. In *The Eleventh International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2023. Tweet insights: A visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*.

Orla McBride, Jamie Murphy, Mark Shevlin, Jilly Gibson-Miller, Todd K Hartman, Philip Hyland, Liat Levita, Liam Mason, Anton P Martinez, Ryan McKay, et al. 2021. Monitoring the psychological, social, and economic impact of the COVID-19 pandemic in the population: Context, design and conduct of the longitudinal COVID-19 psychological research consortium (C19PRC) study. *International journal of methods in psychiatric research*, 30(1):e1861.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.*

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.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Alison J Patev, Calvin J Hall, Chelsie E Dunn, Ashlynn D Bell, Bianca D Owens, and Kristina B Hood. 2019. Hostile sexism and right-wing authoritarianism as mediators of the relationship between sexual disgust and abortion stigmatizing attitudes. *Personality and individual differences*, 151:109528.

Haley Perez-Arche and Deborah J Miller. 2021. What predicts attitudes toward transgender and nonbinary people? An exploration of gender, authoritarianism, social dominance, and gender ideology. *Sex Roles*, 85(3-4):172–189.

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.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Felicia Pratto, Atilla Çidam, Andrew L Stewart, Fouad Bou Zeineddine, María Aranda, Antonio Aiello, Xenia Chryssochoou, Aleksandra Cichocka, J Christopher Cohrs, Kevin Durrheim, et al. 2013. Social dominance in context and in individuals: Contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. *Social Psychological and Personality Science*, 4(5):587–599.

Felicia Pratto, Jim Sidanius, Lisa M Stallworth, and Bertram F Malle. 1994. Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology*, 67(4):741.

Emma A Renström. 2023. Exploring the role of entitlement, social dominance orientation, right-wing authoritarianism, and the moderating role of being single on misogynistic attitudes. *Nordic Psychology*, pages 1–17.

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.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Mirjana Tonković, Francesca Dumančić, Margareta Jelić, and Dinka Čorkalo Biruški. 2021. Who believes in COVID-19 conspiracy theories in Croatia? prevalence and predictors of conspiracy beliefs. *Frontiers in psychology*, 12:643568.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

UN General Assembly. 1993. Declaration on the elimination of violence against women. UN General Assembly resolution 48/104 assembly. Resolution, United Nations.

United Nations. 2021. 'Endemic violence against women cannot be stopped with a vaccine' says WHO chief. `https://news.un.org/en/story/2021/03/1086812`. Accessed: 2023-06-07.

Jasper Van Assche, Yasin Koç, and Arne Roets. 2019. Religiosity or ideology? On the individual differences predictors of sexism. *Personality and Individual Differences*, 139:191–197.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

A. Data Statement

We provide a data statement, as recommended by McMillan-Major et al. (2023).

Curation rationale Textual data is from the test set of EDOS (Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul, 2023), selected for the reasons highlighted in subsection 4.1. For further details of the original data collection process, see Kirk et al. (2023).

Language variety: en. English, as written in comments on internet forums on the Gab and Reddit platforms.

Author demographics: According to Kirk et al. (2023), post authors are 'are likely male, western and right-leaning, and hold extreme or far-right views about women, gender issues and feminism'.

Annotator demographics:

- Age: $24 - 57$, $m = 36.4$, $s = 9.3$
- Gender: Female: 16 (39.0%); Male: 24 (58.5%); Genderfluid: 1 (2.4%).
- Ethnicity: White: 33 (84.8%); Asian: 4 (9.8%); Black: 2 (4.9%); Arab: 1, (2.4%); Mixed: 1 (2.4%).
- Sexual orientation: Heterosexual: 29 (70.7%); Bisexual: 12 (29.3%).
- Political orientation: Left-wing/liberal: 9 (22.0%); Centre 15 (36.6%); Right-wing/conservative 7 (17.1%); None/prefer not to say: 10 (24.4%).
- Training in relevant disciplines: Unknown

Text production situation:

- Time and place: August 2016 to October 2018; Gab and Reddit.
- Modality: Text.
- Intended audience: Internet forum users.

Text characteristics The posts were taken from forums known to attract misogynistic rhetoric: Gab, an extreme-right leaning forum and subreddits labelled as 'Incels', 'Men Going Their Own Way', 'Men's Rights Activists', and 'Pick Up Artists'. Kirk et al. (2023) also provide a full data statement.

B. Measuring Social Attitudes

The VSA scale (Bizumic et al., 2018) is a modified version of the original RWA Altemeyer (1983), which reduced the original 30-item questionnaire into 6 items, while the SSDO scale is a modified version of the original SDO developed by Pratto et al. (1994), which reduced the original 16-item scale into 4 items. Both scales have been verified towards both internal and external validity while ensuing that all elements of the original subscales are adequately captured (Altemeyer, 1983; Pratto et al., 1994).

Furthermore, both the VSA and the SSDO scales have been verified through a variety of cultures and contexts (Aichholzer and Lechner, 2021; Pratto et al., 2013; McBride et al., 2021; Azevedo et al., 2019; Tonković et al., 2021). Each participant answered through the full battery of questions present in each questionnaire, as removing a subsection of items can invalidate the questionnaire responses (Jebb et al., 2021). The full lists of items are presented below.

B.1. Very Short Authoritarianism Scale (VSA)

The scale reporting was based on a 9-point Likert scale, ranging from Very strongly disagree to Very strongly agree. The scale is consisted of subdimensions, namely Conservativism, Authoritarianism, Traditionalism, Authoritarian Agression and Authoritarian Submission. Letter R indicates that the item is reverse scored.

- It's great that many young people today are prepared to defy authority. (Conservatism or Authoritarian Submission)- (**R**)
- What our country needs most is discipline, with everyone following our leaders in unity (Conservatism or Authoritarian Submission)
- God's laws about abortion, pornography , and marriage must be strictly followed before it is too late. (Traditionalism or Conventionalism)
- There is nothing wrong with premarital sexual intercourse. (Traditionalism or Conventionalism) (**R**)
- Our society does NOT need tougher Government and stricter Laws. (Authoritarianism or Authoritarian Aggression) (**R**)
- The facts on crime and the recent public disorders show we have to crack down harder on troublemakers, if we are going to preserve law and order. (Authoritarianism or Authoritarian Aggression)

B.2. Short Social Dominance Orientation Scale (SSDO)

The scale reporting was based on a 7-point Likert scale, ranging from Strongly disagree to Strongly agree. All emphasis in text was also present in the original SSDO scale. For items 2 and 4, higher numeric values indicate a higher level of SSDO and are weighted higher.

- In setting priorities, we must consider all *societal* groups.
- We should not push for equality of *societal* groups.
- The equality of *societal* groups should be our goal.
- Superior *societal* groups should dominate inferior groups.

C. Language Resource References

Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul. 2023. *Explainable Detection of Online Sexism*. Codalab.

A. Experimental Details

Models We implement three models in §5 based on the Python library Transformers provided by Hugging Face (Wolf et al., 2020). These models are pre-trained and available in Hugging Face models, namely `microsoft/deberta-v3-base`, `cardiffnlp/twitter-roberta-base-hate-latest`, and `meta-llama/Llama-2-7b-hf`.

Experimental Setting We randomly split our dataset into training and validation sets by the ratio of 4:1 for fine-tuning. We prioritise several hyper-parameters for all models, where they use cross-entropy loss and the AdamW optimiser (Loshchilov and Hutter, 2019) with a $1e-5$ learning rate and $1e-3$ weight decay. We set the batch size to 128, the micro batch size to 4, and the maximum sequence length to 256. We do training for 10 epochs and 5 epochs separately for five `BERT`-based models and `Llama2`, all with warmup steps of 30. We save the checkpoint with the highest F1 score as the final model.

Computation All experiments are conducted on high-performance computing (HPC) facility at our institution. Further details on acceptance.