Re-examining Sexism and Misogyny Classification with Annotator Attitudes

Aiqi Jiang[1,*] and **Nikolas Vitsakis**[1,*] and **Tanvi Dinkar**[1]
and **Gavin Abercrombie**[1] and **Ioannis Konstas**[1]
[1]The Interaction Lab, Heriot-Watt University
{a.jiang, nv2006, t.dinkar, g.abercrombie, i.konstas}@hw.ac.uk

Abstract

Gender-Based Violence (GBV) is an increasing problem online, but existing datasets fail to capture the plurality of possible annotator perspectives or ensure the representation of affected groups. We revisit two important stages in the moderation pipeline for GBV: (1) manual data labelling; and (2) automated classification.

For (1), we examine two datasets to investigate the relationship between annotator identities and attitudes and the responses they give to two GBV labelling tasks. To this end, we collect demographic and attitudinal information from crowd-sourced annotators using three validated surveys from Social Psychology. We find that higher Right Wing Authoritarianism scores are associated with a higher propensity to label text as sexist, while for Social Dominance Orientation and Neosexist Attitudes, higher scores are associated with a negative tendency to do so.

For (2), we conduct classification experiments using Large Language Models and five prompting strategies, including infusing prompts with annotator information. We find: (i) annotator attitudes affect the ability of classifiers to predict their labels; (ii) including attitudinal information can boost performance when we use well-structured brief annotator descriptions; and (iii) models struggle to reflect the increased complexity and imbalanced classes of the new label sets.[1]

Content Warning: This document includes examples of harmful and offensive language. These are found in the Appendices.

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; Parikh et al., 2019), resulting in women often feeling uncomfortable online (Stevens et al., 2024).

To counter this, there have been attempts to facilitate content moderation using natural language processing (NLP) methods to automatically identify misogynistic language. As a result, there now exist several datasets designed for supervised classification of various forms of GBV. However, Abercrombie et al. (2023) identified several 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, particularly of those with minoritised identities who are best placed to recognise GBV.

To fill this gap, we revisit the task of classifying online text following *strongly perspectivist* data practices (Basile et al., 2023; Cabitza et al., 2023), which aim to preserve labels provided by multiple annotators in the collection and modelling of data. We re-annotate two recent datasets, namely Explainable Detection of Sexism (EDOS) (Kirk et al., 2023), and Detection of Online Misogyny (DOM) (Guest et al., 2021), this time with (1) multiple ratings per item; and (2) demographic and attitudinal information about the annotators, which we maintain throughout the classification pipeline.

Prior work by Davani et al. (2024) that also collected attitudinal survey data from annotators attempts to capture morality via the Moral Foundations Questionnaire (MFQ) (Graham et al., 2013) as a predictor of the perception of offensiveness in toxic language. However, due to criticism of the MFQ regarding poor internal consistency (Kivikangas et al., 2021), we look towards other factors that influence individuals' responses, with evidence from social psychology and sociology pointing towards the constructs of *right wing authoritarianism* (RWA), *social dominance orientation* (SDO) and

[*] These authors contributed equally.

[1] Data and code are available at https://github.com/HWU-NLP/GBV-attitudes.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15103–15125
November 12-16, 2024 ©2024 Association for Computational Linguistics

Hostile Neosexism (HN) [2] (Altemeyer, 1983; Pratto et al., 1994; Chulvi et al., 2023) as potentially relevant towards understanding the link between attitudes and GBV-related behaviour.

We extend a pilot study by Abercrombie et al. (2024) to explore the link between these attitudes and annotating behaviours on our re-annotated dataset, and find that annotators with higher propensity toward RWA are more likely to label text as sexist, possibly due to its association with benevolent sexist attitudes (De Geus et al., 2022).[3] In contrast, we find that those with a higher propensity toward SDO and HN – both associated with hostile sexism (La Macchia and Radke, 2020; Chulvi et al., 2023) – are less likely to label items as sexist, possibly due to the text aligning with internalised beliefs.

While the datasets we re-annotated were originally conceived of for the classification of single 'gold standard' aggregated labels, we aim to represent diverse perspectives in predicting individual annotator labels (Leonardelli et al., 2023). To better study the effect of including annotator attitudes as input to a classification task, we conduct a large set of instruction-based zero-shot, few-shot (in-context learning; ICL), and fine-tuning experiments with four open-source Large Language Models (LLMs), namely `Flan-T5` (Chung et al., 2024), `Llama 2` (Touvron et al., 2023), `Llama 3` (Meta AI, 2024), and `Mistral` (Jiang et al., 2023). Following Fleisig et al. (2023) we experiment with different prompt templates to better incorporate the annotator information (shown in Figure 1). We find that ICL works 17% and 26% better than the majority baseline for majority vote and individual annotator tasks respectively, and fine-tuning LLMs performs 31% better when predicting individual labels per annotator. The best way to incorporate attitudinal data for annotators is to include well-structured brief annotator descriptions about demographics and attitudes but exclude demonstrations. Our experimental results also indicate that models are biased towards annotators' attitudes.

2 Background

The GBV Framework We follow Abercrombie et al. (2023) in adopting this framework and the term GBV as a class label. It 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.

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. This occurs by either hindering the ability of classifiers to recognise subtle and implicit forms of abuse, or by creating a prediction bias in the classifiers – e.g. in the form of harmful stereotypes – against historically minoritised voices (Davani et al., 2023). *Standpoint theory* (Harding, 1991) contends that only people with relevant lived experiences are able to recognise 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.

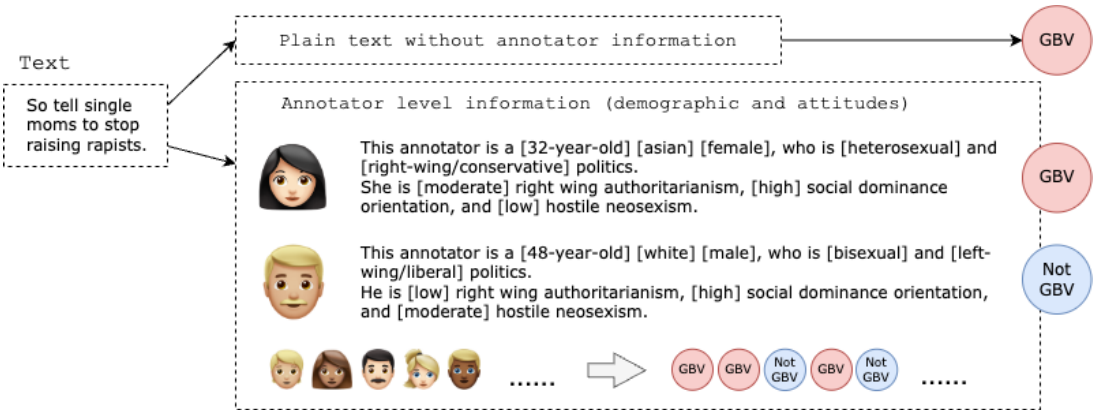
There is now a growing recognition of the need to collect, retain, and distribute labels provided by multiple annotators, and this has been adopted across a range of NLP tasks (for an extensive list, see Plank, 2022). This is particularly so for controversial tasks such as identification of abusive or 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).[4]

Beliefs and Attitudes We ground our approach to the analysis of annotator beliefs in the Dual Process Motivational Model of Ideology and Prejudice (Duckitt, 2001; Duckitt and Sibley, 2009). This links sociopolitical and ideological attitudes to prejudice captured by three related constructs: Right Wing Authoritarianism (RWA), Social Dominance Orientation (SDO), and Hostile Neosexism (HN). RWA explains propensity towards cultural conservatism and traditionalism-related beliefs (Altemeyer, 1983; Feather and McKee, 2012; Van Assche et al., 2019), while SDO explains favourable views towards social hierarchies of power, where

[2] For more details on the *RWA*, *SDO* and *HN* measures, please refer to section 2.

[3] I.e. "Attitudes towards women that seem subjectively positive but are actually discriminatory" (Chulvi et al., 2023).

[4] For further background, see the Perspectivist Data Manifesto at `https://pdai.info/`


Text
Plain text without annotator information
GBV
Annotator level information (demographic and attitudes)
GBV
Not GBV
......
GBV
GBV
Not GBV
GBV
Not GBV
......

So tell single moms to stop raising rapists.

This annotator is a [32-year-old] [asian] [female], who is [heterosexual] and [right-wing/conservative] politics.
She is [moderate] right wing authoritarianism, [high] social dominance orientation, and [low] hostile neosexism.

This annotator is a [48-year-old] [white] [male], who is [bisexual] and [left-wing/liberal] politics.
He is [low] right wing authoritarianism, [high] social dominance orientation, and [moderate] hostile neosexism.

Figure 1: Two prompting paradigms under which models may assign different labels to the same text: Items are entered as plain text, or the prompt is enriched with socio-demographic and attitudinal information about annotators.

inequality between groups is seen as inevitable or even natural (Christopher and Wojda, 2008; Pratto et al., 1994; Jagayat and Choma, 2021). HN is characterised by continued discrimination against women, denial of women's demands, opposition to policies aimed at improving women's social status, and the belief that feminist-driven changes unfairly disadvantage men (Tougas et al., 1995; Swim et al., 1995; Chulvi et al., 2023).

These constructs have been extensively assessed and found to be strongly related. They explain different forms of sexism and gender-based discrimination. RWA is linked to benevolent sexism, that is attitudes that force women into traditional predefined roles (e.g., being a mother) that seem superficially advantageous but are, in reality, marginalising and disempowering (De Geus et al., 2022). SDO correlates with hostile sexism, and pertains to beliefs in deterministic gender imbalances justifying male dominance through disparaging characterisations of women (De Geus et al., 2022; La Macchia and Radke, 2020). Finally, HN is primarily associated with hostile sexist and anti-feminist attitudes in particular (Chulvi et al., 2023; Off, 2023).

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

3 Related Work

GBV Datasets Abercrombie et al. (2023) systematically reviewed resources for automated detection of GBV, finding a small number of datasets that contain theoretical underpinnings (e.g. Samory et al., 2021; Jha and Mamidi, 2017). We select two of these datasets (Kirk et al., 2023; Guest et al., 2021) for reannotation.

Annotator Characteristics A number of NLP studies have attempted to use annotators' demographic characteristics as predictors of their responses to items (e.g. Akhtar et al., 2021; Dutta et al., 2023; Gordon et al., 2022; Goyal et al., 2022; Larimore et al., 2021; Pei and Jurgens, 2023). However, it has repeatedly been shown that demographic characteristics do not predict annotator behaviour at the individual level (Beck et al., 2024; Biester et al., 2022; Hwang et al., 2023; Orlikowski et al., 2023; Beck et al., 2024).

Recent studies have therefore attempted to uncover annotators' *social attitudes* and relate these to their responses. Sap et al. (2022) found that crowd workers with racist beliefs were less likely to consider anti-Black language as toxic. While they conducted two annotation experiments, one with many annotators but few items, and another with fewer annotators but more items, our data collection aims at both breadth and depth. Hettiachchi et al. (2023) measured crowd workers' responses to a misogynistic language labelling task, and surveyed their moral attitudes (in addition to demographic and personality-type information). They found that higher *moral integrity* and lower *benevolent sexism* scores correlated with label agreement with expert annotators. Davani et al. (2024) found that while cross-cultural differences exist, individual moral values significantly influence annotators' response to perceived offensiveness levels. Hence, we seek to explore the relationship between demographics, social attitudes, and crowd-sourced responses to GBV identification tasks.

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 annotator's responses to abusive language identification tasks, the latter using multi-task learning. The SEMEVAL 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. This drew several 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 the 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 ICL. Classification is turned into a single- or few-word generation task of the target label, merely by providing a few, or even no, specific examples as in input to the model in the form of an instruction or "prompt" (Plaza-del arco et al., 2023; Roy et al., 2023; Pendzel et al., 2023; Hartvigsen et al., 2022; Sen et al., 2023; Ziems et al., 2024). This is particularly appealing given the time, effort, and cost of collecting large-scale datasets with a large pool of annotators. To that end, we benchmark the new version of the dataset and its additional labels, and examine the ability of state-of-the-art systems to recognise GBV.

4 Data Collection

We selected the test sets and a subsection of the training sets of two previously published datasets: Explainable Detection of Sexism (EDOS) (Kirk et al., 2023), and Detection of Online Misogyny (DOM) (Guest et al., 2021). We chose these as (1) Abercrombie et al. (2023) had identified them as among the resources most thoroughly grounded in social science theory; (2) they are English language datasets, the language of our stakeholder partners, with whom we are co-designing GBV-mitigation tools under the framework of participatory design; and (3) the textual data is from two different platforms, providing an opportunity for cross-(sub-)domain comparison.

Pre-processing consisted solely of filtering out any items which included images. We leave annotations of multi-media items for future work. This left 3,896 items, of which we re-annotated a random selection of 1,000 from the test sets for evaluation and 600 from the training sets for fine-tuning.[5] Table 1 shows a comparison between the original and new label distributions, with the new labels determined by majority vote.

Dataset	Label	#Original	#New
EDOS	*Sexist*	299	406
	Not sexist	901	794
DOM	*Misogynistic*	47	97
	Nonmisogynistic	353	303
Ours	*GBV*	346	503
	Not GBV	1254	1097

Table 1: Label distributions in the datasets. "#Original" represents the distribution of labels from original data sources, and "#New" represents the distribution of our re-annotated labels, determined by majority vote.

As the Amazon Mechanical Turk (MTurk) crowd-sourcing platform is widely used to collect annotations and personal information for sensitive tasks (Sap et al., 2020; Kumar et al., 2021), we recruited 43 annotators on MTurk (19 women and 24 men with a mean age of 38, see Appendix A for a full Data Statement with detailed annotator information). To ensure attentive participation, we recruited only workers with ≥ 500 completed tasks and a $\geq 98\%$ approval rating. For comparison of the new labels with the original EDOS and DOM labels, we recruited people based in the same region as the original annotators, the United Kingdom. We further collected demographic information and responses to questions from three surveys designed to measure the attitudes of workers.

Measurement of Attitudes To measure the annotators' attitudes, we used survey questions from two verified scales widely used in social psychology to measure the constructs described in section 2: the *Very Short Authoritarianism* (VSA) (Bizumic et al., 2018) and *Short Social*

[5] We maintained the 3:1 size ratio between EDOS and DOM of the original datasets.

Dominance Orientation (SSDO) (Pratto et al., 2013) measuring RWA and SDO, respectively. We also collected responses to the five questions of the *Brief Hostile Neosexism Scale* (BHNS) to measure HN (Chulvi et al., 2023). As shown in Figure 2 of subsection B.4, overall attitudes show tendencies towards social dominance and neosexism, but not towards authoritarianism, although attitudes on all scales vary considerably among the annotator pool. See section 2 and Appendix B for more details.

Data Labelling Chulvi et al. (2023) have shown that the responses of around 12 annotators per item are sufficient to capture levels of disagreement for a similar sexist language labelling task. We collect up to 23 labels per item to enable investigation in this task. We provide annotators with the relevant parts of the original annotator instructions and guidelines from Kirk et al. (2023) and Guest et al. (2021). Instructions are provided in Appendix D.

Intra-Annotator Agreement We measure agreement between our recruited annotators as well as between the aggregated labels, decided by majority vote, and the original `EDOS` and `DOM` labels. We report raw percentage agreement and Krippendorf's α, which measures agreement between two or more raters and can handle missing values (Gwet, 2014).

Crowd workers	**Majority vote *v* Original labels**	
α	α	%
0.02	0.25	70.8

Table 2: Reliability measured by inter-annotator agreement (Krippendorf's α and percentage agreement (%)).

As shown in Table 2, agreement between the crowd-sourced annotators is low at only $\alpha = 0.02$, although aggregated labels are more similar to the original labels (also produced by majority vote).

5 Statistical Analysis

Our hypothesis is two-tailed and exploratory in nature: whether gender, SSDO, BHNS, or VSA scores are predictive of annotator behaviour in labelling items as sexist/misogynist.

Experimental Design Since we have multiple annotations per annotator, we employ a mixed effects regression model (Raudenbush, 1994). Our dependent variable is the binary label given by each annotator, while our predictors are gender (*male*/*female*), SSDO and BHNS scores (both aggregated into *High*, *Moderate* and *Low*), and VSA scores (aggregated into a five-point scale from *very low* to *very high*). Our model includes by-annotator random intercepts, as most individuals annotated multiple items, while we reject one participant who only provided a single annotation. All categorical variables are dummy-coded (see Appendix C for details).

To evaluate possible effects of pairwise comparisons, we employ a Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) due to its specific focus on false discovery rates in study designs with independent statistics, and smaller sample sizes (Benjamini and Hochberg, 2000; Thissen et al., 2002), such as our study.

	Estimate	SE	z	p	$p_{\text{adj.}}$
Intercept	0.85	0.58	1.48	0.140	0.251
Gender-Male	1.16	0.54	2.14	**0.032**	0.073
VSA - Low	–0.31	1.11	–0.28	0.783	0.951
VSA - High	2.61	0.96	2.73	**0.006**	**0.039**
VSA - V. High	0.73	1.21	0.60	0.548	0.822
SSDO - Low	0.06	0.82	0.08	0.937	0.951
SSDO - High	–1.71	0.69	–2.48	**0.013**	**0.039**
BHNS - Low	0.07	1.09	0.06	0.952	0.951
BHNS - High	–1.54	0.60	–2.59	**0.010**	**0.039**

Table 3: Regression model evaluative outcomes. $P_{\text{val.}}$ refers to significance of initial findings; $P_{\text{adj.}}$ refers to the adjusted $P_{\text{val.}}$ after a Benjamini & Hochberg correction.

Results The results of our regression analysis are shown in Table 3. Our post-hoc correction resulted in our initially significant result on the effects of gender being rejected. Nevertheless, we report a significant positive effect of the VSA-High condition on rating items as sexist (estimate = 2.61, SE = 0.96, z = 2.73, p= 0.006, $p_{\text{adj.}}$ = 0.039). We further report a significant negative effect of the SSDO-High condition in annotating items as sexist (estimate = -1.71, SE = 0.69, z = -2.48, p= 0.013, $p_{\text{adj.}}$ = 0.039). Finally, we report a strong negative effect of the BHNS-High condition on annotating items as sexist (estimate = -1.54, SE = 0.60, z = -2.59, p= 0.010, $p_{\text{adj.}}$ = 0.039).

Discussion Our findings echo prior work showing that demographics do not always influence annotation behaviour (Beck et al., 2024; Biester et al., 2022; Orlikowski et al., 2023). However, our results suggest a directional effect of the annotators' attitudes: higher VSA scores predict hypersensitivity in annotating sexism, indicating a positive propensity to label items as sexist. Conversely, higher scores along the SSDO and BHNS scales predict lower levels of annotations of sexism.

These findings are particularly interesting if placed within the context of the constructs that the scales themselves measure. Since RWA has been shown to be associated with benevolent sexism (De Geus et al., 2022), this could explain why annotators with higher VSA scores demonstrate a higher propensity to label items as sexist. We should note that our results show a significant effect only in the VSA-High condition, not the VSA-Very High condition, despite the trend of the effects being in the same direction. This suggests that while there is a significant difference in annotation behaviour between the high and moderate VSA groups, this effect does not extend consistently to the very high VSA group. We will address this inconsistency in our future work and conduct further analysis to explore potential factors, such as uneven sample sizes or demographic influences, that may have contributed to this unexpected result.

In contrast, with SDO and HN being linked to hostile sexism (La Macchia and Radke, 2020; Chulvi et al., 2023), the effect found by their respective scale might explain annotator leniency towards sexist items, aligning with prior work by Sap et al. (2022), i.e., on annotators with racist beliefs showing similar leniency towards racist language. This suggests that different dimensions of authoritarianism, social dominance, and neosexism can influence the nature of bias in annotations.

6 Classification Experiments

We benchmark the new dataset to explore i) whether a broader label set provides richer information and ii) how varying conceptualisations of GBV, annotator demographics and attitude information affect model performance in identifying GBV text online. To achieve this, we conduct two tasks to predict (a) majority labels per text, and (b) individual annotator labels with diverse label texts.

Dataset We use our re-annotated dataset introduced in section 4 to predict the majority label per text. Our dataset contains 1600 instances, including 1200 instances from `EDOS` subset and 400 from `DOM`. We also augment it with individual annotator labels, obtaining 6,000 and 23,000 instances for the fine-tuning and test sets respectively. Three different label texts are used: “GBV” as the aggregated label in our re-annotated dataset, “Sexist” from `EDOS`, and “Misogynistic” from `DOM` (see Table 1).

Prompt Design We experiment with five different prompt templates for our detection task.

(1) Label prompt: a simple prompt structure to give the label based on the text alone. The template is “`<text>` This text is classified as”.

(2) Task description (task): starts with an instruction describing the detection task, followed by the text to be classified. The instruction for the task description is “Classify the following text from a social media platform. It might contain a form of `<label>`. Output `<label>` if it contains `<label>`, or not `<label>` if not.”, and the template is “`<task description>` Text: `<text>` `<choices>` Answer:”.

(3) Few-shot demonstrations (demos): incorporates the task description and adds two examples (demonstrations) of texts with their corresponding labels before the text is classified. The template is “`<task description>` `<demonstration>` Text: `<text>` `<choices>` Answer:”.

(4) Annotator descriptions (anno): combines the task description with a description of the annotator’s demographic and attitude information before the text. The annotator description can be either a full description of all questions and answers from questionnaires (full) or a brief description of each scale (short), plus its corresponding range based on the compound score[6] for the annotator. The template is “`<task description>` `<annotator description>` Text: `<text>` `<choices>` Answer:”.

(5) Combined prompt: integrates the task description, few-shot demonstrations, and annotator description before the target text. For each demonstration, we add two annotators’ descriptions and labels. The template is “`<task description>` `<demonstration>` `<annotator description>` Text: `<text>` `<choices>` Answer:”.

We use the answer format from Gao et al. (2021). `<choices>` is described as “Choices: A. GBV or B. Not GBV.” for “GBV” label text, and corresponding changes are made for “Sexist” and “Misogynistic” labels (see Appendix E for more details).

[6] A compound score is a unified measure derived by aggregating individual responses to multiple questions in a questionnaire, enabling quantification and comparison. More details are provided in Appendix B.

Models We conduct two sets of experiments, namely ICL and fine-tuning.[7]

(1) RoBERTa$_{base}$, RoBERTa$_{hate}$: we perform ICL only to smaller encoder-only pre-trained LMs. The latter has been pre-trained on toxic language datasets making it a good candidate for the GBV classification task.

(2) FLAN-T5 : we perform both ICL and fine-tuning experiments with a (much larger) encoder-decoder instruction fine-tuned LLM. Fine-tuning enhances the model's reliability, while the subjectivity of GBV classification makes it difficult for an ICL model to capture the relationship between annotator attitudes and behaviours to labels with only few-shot demonstrations.

(3) LLaMA 2 7B, LLaMA 3 8B, Mistral 7B: We *fine-tune* only the base version of three decoder-only LLMs. LLaMA 2 and LLaMA 3 have been shown to adapt to new tasks with relatively few instructions (Milios et al., 2023), making them ideal for our low-resource setting (600 training instances), while Mistral exhibits significant performance, especially on text classification tasks.

Experimental Design To predict the majority labels for the GBV detection task, we apply the ICL experiment directly to the test set with 1000 instances, only using the "Label Prompt" template for inference. For the individual annotator label prediction task, we conduct three ICL experiments on FLAN-T5 with three different label texts used for the whole set, and another ICL for original labels from two subsets respectively (predicting instances from EDOS using "Sexist" and those from DOM using "Misogynistic"). Then fine-tuning experiments use only "GBV" label with all four LLMs. Both experiments test our augmented re-annotated dataset and its two subsets and utilise different prompts (i.e. prompt templates 2-5) under zero-shot and few-shot scenarios, to further investigate the influence of individual annotator's behaviours. Given skewed label distribution, we report the macro F1 score; for hyperparameter settings, see Appendix F.

Results and Analysis Table 4 shows classification results on majority labels via ICL. All three models outperform the majority-class baseline on both sets of annotations. However, RoBERTa$_{base}$ does so only marginally. Results from RoBERTa$_{hate}$ underline the strength of models tailored for a specific task, such as GBV detection here. FLAN-T5 outperforms all models, showcasing its superior capability when lacking annotated datasets.

Model	Original Annotation	Re-annotation
Majority class	44.54	40.72
RoBERTa$_{base}$	45.77	42.11
RoBERTa$_{hate}$	52.05	48.65
FLAN-T5	**61.40**	**57.55**

Table 4: Results of predicting majority labels via in-context learning for the GBV detection task.

Table 5 presents ICL results for FLAN-T5 using different label texts and input prompts on the augmented dataset. Among the four label settings, predicting "Sexist" on our full dataset consistently outperforms the others based on re-annotated labels, achieving the highest score of 65.62. Using the DOM label "Misogynistic" also performs better with short annotator descriptions. Regarding the datasets, better performance is generally achieved on the benchmark when compared to two subsets. Besides, adding annotator descriptions or demonstrations usually leads to superior performance, while combining both annotator information and demonstrations does not always enhance performance, highlighting the importance of proper prompt design.

Table 6 shows results of fine-tuning experiments on individual annotator labels using various input prompts for the GBV detection task. Unsurprisingly, FLAN-T5 outperforms the ICL variant (Table 5). Fine-tuning is used to improve the model's reliability for annotators. Given that GBV classification is a subjective task, it can be challenging for an ICL model to capture the relationship between nuanced annotator metadata (attitudes) and their behaviours (labels) with just few-shot demonstrations or definitions. See Appendix G for further results. Among the four LLMs, FLAN-T5 outperforms LLaMA 2, LLaMA 3, and Mistral, showing significant performances with short annotator description for the new label. Besides, the effectiveness of input prompts varies. Adding full annotator descriptions generally provides better results, particularly for LLaMA 2 and LLaMA 3, and inputs with short annotator descriptions also improve the results. The combined prompts with short annotator information and demonstrations, especially for Mistral, show great improvements. These results suggest that more comprehensive in-

[7] We use low-rank adaptation (LoRA) (Hu et al., 2022) for all models to reduce the number of trainable parameters.

Model: FLAN-T5	**Original**			**GBV**			**Sexist**	**Misogynistic**
	All	EDOS	DOM	All	EDOS	DOM	All	All
Majority class (single)	36.14	35.71	37.41	36.14	35.71	37.41	36.14	36.14
task	60.93	**64.87**	56.95	60.29	60.53	56.81	65.25	61.12
+demos	59.75	64.24	54.25	**62.60**	**62.92**	**58.49**	63.11	59.67
+anno (short)	**64.68**	64.28	**62.15**	61.13	61.35	57.05	**65.62**	**64.69**
+anno (short)+demos	62.42	63.72	59.32	59.91	60.50	54.80	63.43	62.32
+anno (full)	59.03	62.21	53.97	61.22	61.38	57.65	62.23	58.20
+anno (full)+demos	59.03	62.21	53.97	61.22	61.38	57.65	62.23	58.20

Table 5: Results of in-context learning on our re-annotated dataset using FLAN-T5 with different label texts: (i) "Original" uses the original labels, namely "Sexist" for EDOS subset and "Misogynistic" for DOM subset, (ii) "GBV" as the aggregated label for both subsets, (iii) "Sexist" and (iv) "Misogynistic" for both subsets. Six different input prompts are evaluated among three label texts. Best results are shown in bold by column.

New Label - GBV (maj. 36.14)	**FLAN-T5**	**LLaMA 2**	**LLaMA 3**	**Mistral**
task	63.78 ± 1.84	51.87 ± 1.76	50.32 ± 2.75	59.20 ± 2.10
+demos	65.12 ± 1.66	49.40 ± 1.79	**52.12 ± 1.05**	41.07 ± 1.18
+anno (short)	**65.79 ± 1.89**	51.17 ± 1.58	43.39 ± 1.47	52.56 ± 1.79
+anno (short)+demos	64.95 ± 1.03	41.16 ± 1.06	50.40 ± 0.53	**67.40 ± 1.55**
+anno (full)	64.50 ± 1.17	**53.21 ± 0.12**	51.70 ± 2.51	40.96 ± 2.97
+anno (full)+demos	62.23 ± 0.54	50.02 ± 1.10	43.31 ± 0.29	54.15 ± 0.46

Table 6: Results of fine-tuning LLMs on individual annotator labels using different input prompts for the GBV detection task. F1 score for the majority class (maj.) is 36.14. Best results are displayed in bold by column.

formation generally enhances model performance, but the effectiveness can vary by model.

Discussion We explore the GBV classification pipeline with an emphasis on the role of diverse annotator demographics and attitudes. Our classification experiments in section 6 reveal that detailed prompts, enriched with annotator bias, significantly improve LLM adaptability in identifying GBV content, even in low-resource scenarios. This highlights the importance of well-crafted input prompts in enhancing the model's ability to accurately interpret and respond to complex social phenomena involving variably interpretable elements.

However, a contrasting result is presented when we explore the use of annotator information with demonstrations in ICL and fine-tuning experiments. We observe a decrease in model performance when prompts are overloaded with additional information. This indicates that the quality of input prompt can influence the model's efficiency and additional information may hinder rather than help performance. Statistical findings in section 5 reveal that annotators' personal biases affect their labelling tendencies. Pre-trained models may lack perspectivist information, and adding a few demonstrations is insufficient for the model to learn these biases, potentially causing confusion. Besides, the Flan-T5 model is trained on samples with a maximum length of 1024 tokens. As it uses relative positional encoding, it is possible to train the model with far more tokens, but training with samples with a greater length could hurt model performance (Chung et al., 2024). It is also possibly because the models get negative biases from the particular demonstrations used, affecting their ability to accurately interpret GBV-related content.

In our analysis of the FLAN-T5 model's performance using different label texts, we found that "Sexist" leads to the best results, followed by "Misogynistic", with "GBV" performing the poorest, which might be attributed to several factors. These three labels have different conceptualizations of sexism. "Sexist" is relatively straightforward and less ambiguous, whereas "Misogynistic" is more specific, introducing stronger connotations to complicate classification. "GBV" is broader and more complex, covering violence and discrimination in various contexts and forms. It may increase cognitive load and interpretation difference. Additionally, the original labels from EDOS and DOM might align better with "Sexist" and "Misogynistic" classifications, respectively, compared to "GBV". Furthermore, "Sexist" and "Misogynistic" could be more understandable for models based on previous pre-trained resources. Therefore, it is essential to develop a clearer and more precise definition and taxonomy of gender-based violence to enhance the model's learning and classification abilities.

Furthermore, our analysis of various experiments on the newly annotated labels indicates a generally poorer performance compared to the original labels, which could be attributed to increased complexity, annotator bias, and class imbalance in the new annotation set. This emphasises the challenge of adapting automated classification models to variably interpretable and evolving social issues such as GBV detection, where language complexity and human perspectives intersect.

7 Conclusion

We have revisited the annotation and classification tasks for online GBV with a particular focus on the underlying attitudes of a broad range of annotators.

Through the re-annotation of two datasets, we collect demographic and attitudinal information about crowd-sourced annotators using validated surveys from Social Psychology, and incorporate a diverse range of annotator perspectives, exploring the relationship between these factors and the labels they provide. We find that annotators with stronger right-wing authoritarian traits show a higher propensity to label items as sexist, whereas those with more socially dominant and neosexist attitudes do the opposite. This suggests that people exhibiting right-wing authoritarian characteristics may be less attuned to subtle gender-related discourse.

We then conduct classification experiments on both aggregated and individual annotator labels using various prompting strategies and LLMs. Our findings indicate that models are biased by annotators' attitudes. While incorporating annotator information can enhance model capacity, but adding excessive information can be detrimental. This highlights the challenges of the increased complexity and imbalance of incorporating broader, *perspectivist* label sets, which adversely affect performance on all our experiments.

Limitations

This study concentrates on sexism and misogyny from the scope of the `EDOS` and `DOM` datasets. Future research directions require a broader GBV framework that captures the full spectrum of GBV-related issues and more inclusive dataset standards.

We recruit annotators only from MTurk, who might provide unreliable data on personal information (Huang et al., 2023) or sensitive topics such as GBV, raising potential data quality issues.

Due to the exploratory and experimental nature of this work, the statistical annotation analysis suffers from a number of important limitations, namely the relatively small sample size and unequal amount of annotations per annotator. Future studies should aim for a fixed number of annotations per annotator, while a larger sample size would lead to more generalisable results. There is also the need for further clarification and nuanced interpretation of the statistical results, especially in the context of very high VSA scores.

Lastly, our classification experiments and analysis are limited to open-source LLMs such as `Flan-T5`, `LLaMA 2`, `LLaMA 3`, and `Mistral`. The exclusion of other LLMs, such as the GPT family, limits the reproducibility and breadth of our work. The training resources for LLMs we used are in limited languages, which may not accurately capture GBV in multilingual contexts. It indicates the need for future work to explore a wider range of LLMs to model GBV online.

Although we account for annotator attitudes in our study to explore their impact on labelling, accurately measuring the effect of these attitudes remains a challenge. While this approach can provide insights into potential biases, it does not fully resolve the issue of maintaining objectivity in tasks like sexism and misogyny detection. Ensuring that annotations are both accurate and objective still requires a careful selection of annotators with relevant expertise and experience.

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of the School of Mathematical and Computer Sciences at Heriot-Watt University,[8] which reviewed our annotation methodologies and protocols to ensure compliance with ethical standards.

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.

[8] Project identification code is 5536.

- 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[9] and the Living Wage in our jurisdiction, which is considerably higher.

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

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.

Acknowledgements

We would like to thank the anonymous reviewers for all their valuable comments and suggestions. Aiqi Jiang, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas were supported by the EPSRC project "Equally Safe Online" (EP/W025493/1). This work was supported by the Heriot-Watt University high-performance computing facility (DMOG) and the Edinburgh International Data Facility (EIDF).

References

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.

Gavin Abercrombie, Nikolas Vitsakis, Aiqi Jiang, and Ioannis Konstas. 2024. Revisiting annotation of online gender-based violence. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 31–41, Torino, Italia. ELRA and ICCL.

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.

Bob Altemeyer. 1996. *The authoritarian specter*. Harvard University.

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.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Yoav Benjamini and Yosef Hochberg. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP

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

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)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2007–2021, New York, NY, USA. Association for Computing Machinery.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

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.

Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by mimicking annotator annotation) in toxic comment identification across diverse communities.

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.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Friedrich Funke. 2005. The dimensionality of right-wing authoritarianism: Lessons from the dilemma between theory and measurement. *Political Psychology*, 26(2):195–218.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

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).

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Peter Green and Catriona J. MacLeod. 2016. simr: an r package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498.

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.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

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.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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.

J Matias Kivikangas, Belén Fernández-Castilla, Simo Järvelä, Niklas Ravaja, and Jan-Erik Lönnqvist. 2021. Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin*, 147(1):55.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

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.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

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.

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

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.*

Meta AI. 2024. Meta llama 3. `https://llama.meta.com/llama3`. Accessed: 2024-06-12.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

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.

Gefjon Off. 2023. Complexities and Nuances in Radical Right Voters' (Anti)Feminism. *Social Politics: International Studies in Gender, State & Society*, 30(2):607–629.

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.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. 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.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Sagi Pendzel, Tomer Wullach, Amir Adler, and Einat Minkov. 2023. *Regulating Hate Speech Created by Generative AI*, chapter Generative AI for Hate Speech Detection: Evaluation and Findings. Auerbach Publications.

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.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stephen W Raudenbush. 1994. Random effects models. *The handbook of research synthesis*, 421(3.6).

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.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

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. In *ICWSM*, pages 573–584.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. 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.

David N Smith and Christopher W Gunn. 1999. Authoritarian aggression and social stratification: A research note. *Social thought & research*, pages 95–112.

Francesca Stevens, Florence E. Enock, Tvesha Sippy, Jonathan Bright, Miranda Cross, Pica Johansson, Judy Wajcman, and Helen Z. Margetts. 2024. Women are less comfortable expressing opinions online than men and report heightened fears for safety: Surveying gender differences in experiences of online harms.

Janet K Swim, Kathryn J Aikin, Wayne S Hall, and Barbara A Hunter. 1995. Sexism and racism: Old-fashioned and modern prejudices. *Journal of personality and social psychology*, 68(2):199.

David Thissen, Lynne Steinberg, and Daniel Kuang. 2002. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, 27(1):77–83.

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.

F. Tougas, R. Brown, A. M. Beaton, and S. Joly. 1995. Neosexism scale. *Personality and Social Psychology Bulletin*, 21(8):842–849.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

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.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A Data Statement

We provide a data statement to summarise the main features of the selected datasets, as recommended by McMillan-Major et al. (2023).

Curation rationale Textual data is from the test set of EDOS and DOM, selected for the reasons highlighted in section 4. For further details of the original data collection processes, see Kirk et al. (2023) and Guest et al. (2021).

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'. No information is available regarding authors of DOM texts.

Annotator demographics:

- Age: $24 - 56$, $m = 35.8$, $s = 7.9$
- Gender: Female: 19 (44.2%); Male: 24 (55.8%).
- Ethnicity: White: 35 (81.4%); Asian: 6 (14.0%); Black: 1 (2.3%); Other: 1, (2.3%).
- Sexual orientation: Heterosexual: 23 (53.5%); Bisexual: 18 (41.9%); Don't know: 1 (2.3%); Prefer not to say: 1 (2.3%).
- Political orientation: Left-wing/liberal: 7 (16.3%); Centre 16 (37.2%); Right-wing/conservative 16 (37.2.%); None/prefer not to say: 4 (9.3%).
- 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 provides 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 ensuring 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 consist of sub-dimensions, namely Conservativism, Authoritarianism, Traditionalism, Authoritarian Aggression 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)

With 6 questions and a 9-point scale, the score range for each question is 1 to 9. The compound score is calculated by summing the scores across these 6 items and adjusting for reverse scoring where applicable. This total score categorises respondents into five levels of RWA: very low, low, moderate, high, and very high, ranging from 6 to 54.

- Very low right wing authoritarianism: 6-15
- Low right wing authoritarianism: 16-25
- Moderate right wing authoritarianism: 26-35
- High right wing authoritarianism: 36-45
- Very high right wing authoritarianism: 46-54

While previous studies have used the scale with three breakpoints (low, moderate, and high), there is evidence to suggest that the moderate range might be concealing effects between the combinations of sub-dimensions that form the original RWA scale, and thus the VSA (Funke, 2005). To address this, we follow the original guidelines set by the creator of the RWA scale, from very low to very high (Altemeyer, 1996; Smith and Gunn, 1999), allowing for a finer distinction between the data that stays true to methodologies previously used to study the construct.

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.

With 4 questions and a 7-point scale, the score range for each question is 1 to 7. The compound score is calculated by summing the scores across these 4 items and adjusting for reverse scoring where applicable. This total score categorises respondents into three levels of SDO: low, moderate, and high, ranging from 4 to 28.

- Low social dominance orientation: 4-10
- Moderate social dominance orientation: 11-17
- High social dominance orientation: 18-28

B.3 Brief Hostile Neosexism Scale (BHNS)

Chulvi et al. (2023)'s scale is based on a 7-point Likert scale, ranging from *Strongly disagree* to *Strongly agree*. All emphasis in text was also present in the original neosexism scale. For all 6 items, higher numeric values indicate a higher level of hostile neosexism and are weighted higher.

- Some of the demands of the feminist movement seem to me to be a bit exaggerated.
- I sometimes feel that our society pays too much attention to the rights of certain minorities.
- In the name of equality, many women try to gain certain privileges.
- Many women interpret innocent comments and actions as sexist.
- Women are easily offended.
- Women exaggerate the problems they suffer because they are women.

With 6 questions and a 7-point scale, the score range for each question is 1 to 7. The compound score is calculated by summing the scores across these 6 items and adjusting for reverse scoring where applicable. This total score categorises respondents into three levels of hostile neosexism: low, moderate, and high, ranging from 6 to 42.

- Low hostile neosexism: 6-14
- Moderate hostile neosexism: 15-28
- High hostile neosexism: 29-42

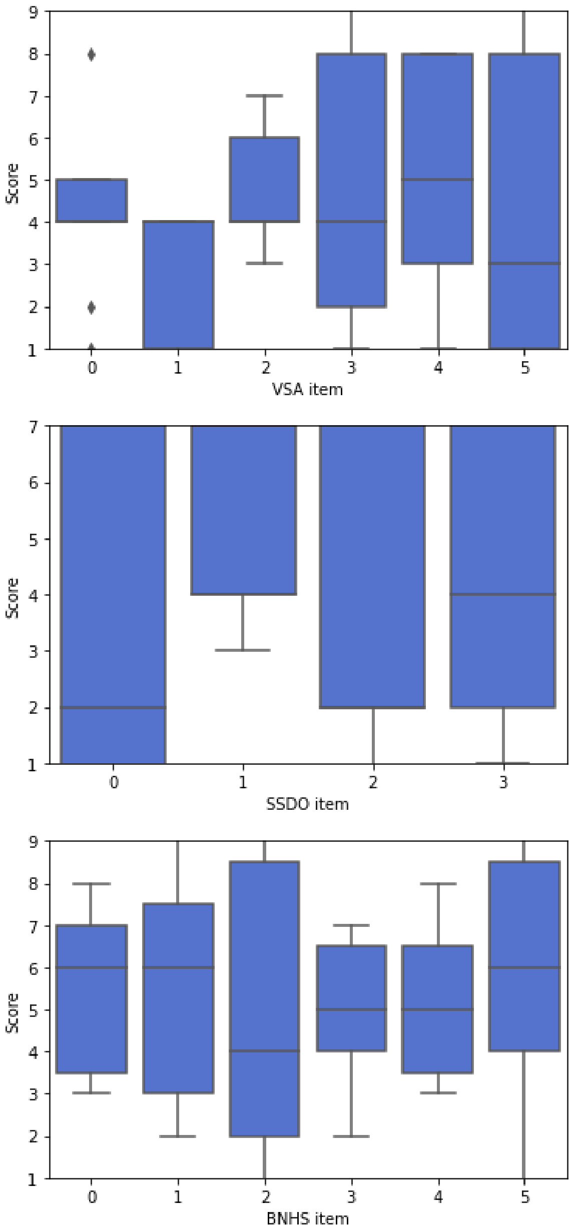


Figure 2: Responses to the six VSA, four SSDO, and six BHNS survey items on $[1-9]$, $[1-7]$, and $[1-7]$ scales, respectively.

B.4 Survey Responses

Annotator responses to the survey questions are presented in Figure 2.

We find that for VSA, the annotators tend towards the centre of the scale ($m = 4.78, s = 3.35$), while for SSDO, they are towards the more dominant end of the scale on average ($m = 6.14, s = 5.19$), and for BHNS, they tend towards hostile neosexist attitudes ($m = 4.78, s = 3.35$), as shown in Figure 2. There is, however, substantial variance on all three scales. To sum up, overall attitudes show tendencies towards social dominance and neosexism, but not towards authoritarianism, and attitudes on all scales vary considerably among the annotator pool.

C Statistical Testing

All statistical analyses were done in R (R Core Team, 2021), with the packages: 1.3.1 tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), simR 1.0.7 (Green and MacLeod, 2016).

The reference groups for our dummy coding were as follows:

- Gender: Women
- VSA: Moderate scores
- SSDO: Moderate scores
- BHNS: Moderate scores

We chose the moderate scale for all of our scales as a baseline as previous research has shown that the effects of social attitudes tend to influence behaviour when scoring alongside the edges of a scale (either low or high) Davani et al. (2024); Hettiachchi et al. (2023).

D Annotation Instructions

We provide an annotation instruction and the annotators must read and accept it before they start annotating texts. We give a content warning before the instruction: This research exposes participants to offensive language which may cause mental or physical stress to the reader. Please consider this before participating, you are under no obligation to take part and if you choose not to we thank you for considering taking part. Please do remember that you can withdraw from participating at any time.

Our full annotation instruction is described below:

Sexism This task defines sexism as: "Any abuse or negative sentiment that is directed towards women based on their gender, or on the combination of their gender with one or more other identity attributes (e.g. Black women, Muslim women, Trans women)"

An entry must be labelled Sexist if it meets both of the following conditions:

1. The entry refers to a woman, a group of women, women in general, or to supporters of feminism. For this task, "woman" refers to any person who identifies as a woman, irrespective of gender assigned at birth (i.e., include transgender women within this definition). In addition, explicitly threatening or inciting harm against individual women must also be included.

2. The entry expresses negative sentiment against its target on the basis of gender: for instance, it is derogatory, demonising, insulting, threatening, violent or prejudicial.

Your task is to label the entry rather than the speaker. Even in cases where the speaker could be sexist, please carefully consider whether the statement itself meets the above criteria.

Notes on quotes and jokes: Entries which make a quote (indicated by "") without any further comments should be taken at face value. If an entry contains a joke, please carefully consider its intention. If a joke meets the above criteria, it should be labelled as Sexist and put into the corresponding secondary category, even in cases where the tone is light-hearted or positive.

Not Sexist For each Not Sexist entry, you need to decide whether it contains abuse directed at another protected characteristic (i.e., a fundamental aspect of a person's identity) besides gender. Examples of other protected characteristics include, but are not limited to: race, ethnicity, immigration status, religion, age, sexuality, and disability status. If it does, write out the target of the abuse.

Common types of confusing Not Sexist content: Some entries that you should label Not Sexist may easily be confused with Sexist content. Please review the following examples, all of which should be labelled Not Sexist:

1. Entries which contain vulgar, inappropriate or offensive language, but do not specifically target women, e.g.,
   - "We're here at the bar, now suck my penis"
   - "Hahahaa you silly dickhead"

2. Entries that direct abuse against individuals without a gender-based attack, e.g.,
   - "I hate Hilary Clinton"
   - "She is so lame"
   - "Donald Trump is a bellend"

3. Entries abusive of other protected characteristics, but not gender, e.g.,

    - "Jews make me sick"
    - "White honkies gona dieeeeee if they cross me"

    NB: Abusive entries that attack gender with other characteristics (e.g., “I hate black women”), or contain gendered slurs (e.g., “Don’t be such a bitch”) should still be labelled Sexist

4. Entries that criticise feminism as a theoretical framework, ideology or practice, e.g.,

    - "I dont identify as a feminist. I just dont like the connotations, I try not to be political."
    - "Feminism isnt a well formulated theory, it’s not disprovable and so isnt a proper science."

    NB: Take care to distinguish between criticism of feminism as a theory, which by itself should be labelled Not Sexist, and abuse of feminists as people (e.g., "Feminists are such loony eyed old bags"), which should be labelled Sexist. However, entries which combine criticism of feminism with abuse of feminists should be labelled Sexist.

E Input Format and Examples

We provide examples for five prompt templates in Table 7. The example text “So tell single moms to stop raising rapists.” and its label is “A. GBV”.

F Experimental Details

Models We implement three models in section 6 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 `roberta-base`, `cardiffnlp/twitter-roberta-base-hate-latest`, `google/flan-t5-xl`, `meta-llama/Llama-2-7b-hf`, `meta-llama/Meta-Llama-3-8B`, and `mistralai/Mistral-7B-v0.1`.

Experimental Setting We randomly split our training set into training (4,800 samples) and validation (1,200 samples) sets by the ratio of 4:1 for fine-tuning. For ICL experiments, we use OpenPrompt (Ding et al., 2022), a standard framework for prompt learning over pre-trained language models. We use the default hyperparameters in Hugging Face. For fine-tuning LLMs, we prioritise several hyperparameters, using cross-entropy loss and the AdamW optimiser (Loshchilov and Hutter, 2019) with a $1e-3$ weight decay. We use the learning rate $1e-3$ for `FLAN-T5` and $3e-4$ for the others. We set the batch size to 128, the micro batch size to 4, the maximum sequence length to 512. We use 5 epochs for `LLaMA 2` and `LLaMA 3`, 4 epochs for `Mistral` and 2 epochs for `FLAN-T5`. For the LoRA hyperparameters, we set the attention dimension to 8, the alpha parameter to 16, and the dropout value to 0.05 for LoRA layers. We use the task type “SEQ_2_SEQ_LM” and modules q and v for `FLAN-T5`, and use “CAUSAL_LM” and modules q_proj and v_proj for the other decoder-only LLMs.

We save the checkpoint with the highest F1 score as the final model.

Computation All experiments are conducted on the Heriot-Watt University high-performance computing facility (DMOG) and the Edinburgh International Data Facility (EIDF).

G Experimental Results

We provide comprehensive results for ICL and fine-tuning experiments in the following tables. Table 8 shows more ICL results on two subsets based on Table 5. Table 9, Table 10, Table 11, and Table 12 present more fine-tuning results among LLMs tested on original labels by extending Table 6.

Label prompt:
<text> This text is classified as <label>
Example:
So tell single moms to stop raising rapists. This text is classified as

Task description:
<task description> Text: <text> <choices> Answer:
Example:
Classify the following message from a social media platform. It might contain a form of gender-based violence (GBV). Output A. GBV if it contains GBV, or B. Not GBV if not.
Text: So tell single moms to stop raising rapists.
Choices: A. GBV or B. Not GBV.
Answer:

Few-shot demonstrations:
<task description> <demonstration> Text: <text> <choices> Answer:
Example:
Classify the following message from a social media platform. It might contain a form of gender-based violence (GBV). Output A. GBV if it contains GBV, or B. Not GBV if not.
Text: Fuck the niggers and the jews. Both have a history of eating white people and fuck you for suggesting I should get along with people who want to eat me.
Choices: A. GBV or B. Not GBV.
Answer: B. Not GBV
Text: Well then good because someone has to knock her ass back in line when she acts like a woman and posts bullshit about her (((based minorities))). If someone is not white they are NOT one of us.
Choices: A. GBV or B. Not GBV.
Answer: A. GBV
Text: So tell single moms to stop raising rapists.
Choices: A. GBV or B. Not GBV.
Answer:

Annotator descriptions (short):
<task description> <annotator description> Text: <text> <choices> Answer:
Example:
Classify the following message from a social media platform. It might contain a form of gender-based violence (GBV). Output A. GBV if it contains GBV, or B. Not GBV if not.
This annotator is a 32-year-old asian female, who is heterosexual and right-wing/conservative politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism.
This worker is moderate right wing authoritarianism, moderate social dominance orientation, and moderate hostile neosexism.
Text: So tell single moms to stop raising rapists.
Choices: A. GBV or B. Not GBV.
Answer:

Annotator descriptions (full):
<task description> <annotator description> Text: <text> <choices> Answer:
Example:
Classify the following message from a social media platform. It might contain a form of gender-based violence (GBV). Output A. GBV if it contains GBV, or B. Not GBV if not.
This annotator is a 32-year-old asian female, who is heterosexual and right-wing/conservative politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism.
This worker is moderate right wing authoritarianism, moderate social dominance orientation, and moderate hostile neosexism. The following are questions
and annotator's answers
from each scale.
Scale 1: Very Short Authoritarianism Scale (VSA)
Statement 1: It2̌019s great that many young people today are prepared to defy authority.
Answer: Strongly agree
Statement 2: What our country needs most is discipline, with everyone following our leaders in unity.
Answer: Strongly agree
Statement 3: God2̌019s laws about abortion, pornography, and marriage must be strictly followed before it is too late.
Answer: Very strongly agree
Statement 4: There is nothing wrong with premarital sexual intercourse.
Answer: Unsure or neutral
Statement 5: Our society does NOT need tougher Government and stricter Laws.
Answer: Strongly agree
Statement 6: 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.
Answer: Strongly agree
Scale 2: Short Social Dominance Orientation Scale (SSDO)
Statement 1: In setting priorities, we must consider all societal groups.
Answer: Strongly disagree
Statement 2: We should not push for equality of societal groups.
Answer: Slightly disagree

[Continued Table]

Statement 3: The equality of societal groups should be our goal.
Answer: Strongly agree
Statement 4: Superior societal groups should dominate inferior groups.
Answer: Strongly disagree
Scale 3: Brief Hostile Neosexism Scale
Statement 1: Some of the demands of the feminist movement seem to me to be a bit exaggerated.
Answer: Somewhat agree
Statement 2: I sometimes feel that our society pays too much attention to the rights of certain minorities.
Answer: Strongly agree
Statement 3: In the name of equality, many women try to gain certain privileges.
Answer: Somewhat agree
Statement 4: Many women interpret innocent comments and actions as sexist.
Answer: Slightly disagree
Statement 5: Women are easily offended.
Answer: Somewhat disagree
Statement 6: Women exaggerate the problems they suffer because they are women.
Answer: Slightly disagree
Text: So tell single moms to stop raising rapists.
Choices: A. GBV or B. Not GBV.
Answer:

Combined prompt (short annotator description):
<task description> <demonstration> <annotator description> Text: <text> <choices>
Answer:
Example:
Classify the following message from a social media platform. It might contain a form of gender-based violence (GBV). Output A. GBV if it contains GBV, or B. Not GBV if not.
This annotator is a 32-year-old white female, who is bisexual and left-wing/liberal politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism. This worker is moderate right wing authoritarianism, high social dominance orientation, and high hostile neosexism.
Text: Fuck the niggers and the jews. Both have a history of eating white people and fuck you for suggesting I should get along with people who want to eat me.
Choices: A. GBV or B. Not GBV.
Answer: A. GBV
This annotator is a 48-year-old asian female, who is heterosexual and right-wing/conservative politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism. This worker is moderate right wing authoritarianism, moderate social dominance orientation, and high hostile neosexism.
Text: Fuck the niggers and the jews. Both have a history of eating white people and fuck you for suggesting I should get along with people who want to eat me.
Choices: A. GBV or B. Not GBV.
Answer: B. Not GBV
This annotator is a 35-year-old american male, who is bisexual and right-wing/conservative politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism. This worker is moderate right wing authoritarianism, moderate social dominance orientation, and high hostile neosexism.
Text: Well then good because someone has to knock her ass back in line when she acts like a woman and posts bullshit about her (((based minorities))). If someone is not white they are NOT one of us.
Choices: A. GBV or B. Not GBV.
Answer: A. GBV
This annotator is a 32-year-old asian female, who is heterosexual and right-wing/conservative politics.
Three scales are used to show the annotator's attitudes, namely the Very Short Authoritarianism (VSA) scale to measure Right Wing Authoritarianism (RWA), the Short Social Dominance Orientation (SSDO) scale to measure Social Dominance Orientation (SDO), and the Brief Hostile Neosexism scale to measure Hostile Neosexism. This worker is moderate right wing authoritarianism, moderate social dominance orientation, and moderate hostile neosexism.
Text: Well then good because someone has to knock her ass back in line when she acts like a woman and posts bullshit about her (((based minorities))). If someone is not white they are NOT one of us.
Choices: A. GBV or B. Not GBV.
Answer: B. Not GBV
Text: So tell single moms to stop raising rapists.
Choices: A. GBV or B. Not GBV.
Answer:

Table 7: Input examples with different prompt templates.

Model: FLAN-T5	**Original Label - GBV**			**Original Label - Sexist**			**Original Label - Misogynistic**		
	Our	EDOS	DOM	Our	EDOS	DOM	Our	EDOS	DOM
Majority class (single)	44.54	43.61	47.15	44.54	43.61	47.15	44.54	43.61	47.15
task	63.60	62.43	67.09	66.94	66.32	64.14	71.47	71.52	67.21
+demos	65.49	64.57	67.21	69.19	68.82	67.21	**72.84**	**73.16**	67.69
+anno (short)	65.72	64.84	67.09	63.40	62.28	62.89	67.63	66.15	**69.48**
+anno (short)+demos	65.86	65.69	63.62	**70.21**	**69.46**	**70.00**	70.45	70.32	67.81
+anno (full)	**67.20**	**66.51**	**67.78**	67.32	66.48	65.75	69.51	69.44	67.09
+anno (full)+demos	**67.20**	**66.51**	**67.78**	67.32	66.48	65.75	69.51	69.44	67.09
	New Label - GBV			**New Label - Sexist**			**New Label - Misogynistic**		
Majority class (single)	36.14	35.71	37.41	36.14	35.71	37.41	36.14	35.71	37.41
task	60.29	60.53	56.81	65.25	**64.87**	62.04	61.12	61.28	56.95
+demos	**62.60**	**62.92**	**58.49**	63.11	64.24	54.39	59.67	60.18	54.25
+anno (short)	61.13	61.35	57.05	**65.62**	64.28	**66.41**	**64.69**	**64.00**	**62.15**
+anno (short)+demos	59.91	60.50	54.80	63.43	63.72	53.88	62.32	62.29	59.32
+anno (full)	61.22	61.38	57.65	62.23	62.21	56.67	58.20	58.57	53.97
+anno (full)+demos	61.22	61.38	57.65	62.23	62.21	56.67	58.20	58.57	53.97

Table 8: Results of in-context learning on FLAN-T5 by using different label texts: (i) "GBV" as the aggregated label, (ii) "Sexist" from EDOS dataset, and (iii) "Misogynistic" from DOM dataset. Six different input prompts are evaluated among three label texts. The best results are shown in bold by column.

Model: FLAN-T5	**Original Label**		
	All	EDOS	DOM
Majority class (single)	44.54 ± 0.0	43.61 ± 0.0	47.15 ± 0.0
task	66.82 ± 1.03	64.94 ± 1.23	71.36 ± 2.54
+demos	67.06 ± 1.35	66.61 ± 2.16	67.68 ± 1.58
+anno (short)	68.19 ± 2.62	66.78 ± 2.41	71.31 ± 2.79
+anno (short)+demos	67.94 ± 1.77	64.98 ± 1.68	70.83 ± 0.65
+anno (full)	71.51 ± 1.47	70.24 ± 0.85	73.32 ± 1.74
+anno (full)+demos	70.04 ± 0.33	69.47 ± 0.69	72.64 ± 0.23
	New Label		
Majority class (single)	36.14 ± 0.0	35.71 ± 0.0	37.41 ± 0.0
task	63.78 ± 1.84	64.06 ± 1.73	62.96 ± 3.06
+demos	65.12 ± 1.66	65.33 ± 0.94	64.69 ± 2.44
+anno (short)	65.79 ± 1.89	66.53 ± 0.77	64.37 ± 2.11
+anno (short)+demos	64.95 ± 1.03	65.06 ± 0.45	64.02 ± 1.69
+anno (full)	64.50 ± 1.17	64.88 ± 1.73	63.11 ± 2.07
+anno (full)+demos	62.23 ± 0.54	63.07 ± 0.60	59.90 ± 1.76

Table 9: Results of fine-tuning FLAN-T5 on single annotator labels using different input prompts for the GBV detection task.

Model: LLaMA 2	**Original Label**		
	All	EDOS	DOM
Majority class (single)	44.54 ± 0.0	43.61 ± 0.0	47.15 ± 0.0
task	49.32 ± 2.57	49.53 ± 1.29	47.30 ± 7.95
+demos	47.09 ± 1.45	48.59 ± 1.92	42.46 ± 2.04
+anno (short)	52.06 ± 1.17	51.89 ± 2.00	51.63 ± 2.19
+anno (short)+demos	51.12 ± 1.98	51.29 ± 3.08	49.98 ± 1.78
+anno (full)	52.23 ± 1.76	52.05 ± 1.95	51.66 ± 1.08
+anno (full)+demos	52.09 ± 1.38	50.90 ± 2.12	55.96 ± 1.83
	New Label		
Majority class (single)	36.14 ± 0.0	35.71 ± 0.0	37.41 ± 0.0
task	51.87 ± 1.76	51.45 ± 2.04	52.43 ± 2.39
+demos	49.40 ± 1.79	50.36 ± 1.56	46.53 ± 2.39
+anno (short)	51.17 ± 1.58	50.20 ± 0.83	53.95 ± 4.11
+anno (short)+demos	41.16 ± 1.06	41.59 ± 1.29	37.13 ± 1.42
+anno (full)	53.21 ± 0.12	51.60 ± 1.07	57.76 ± 3.86
+anno (full)+demos	50.02 ± 1.10	47.18 ± 1.36	59.55 ± 0.24

Table 10: Results of fine-tuning LLaMA 2 on single annotator labels using different input prompts for the GBV detection task.

Model: LLaMA 3	**Original Label**		
	All	EDOS	DOM
Majority class (single)	44.54 ± 0.0	43.61 ± 0.0	47.15 ± 0.0
task	47.92 ± 0.84	48.55 ± 0.75	45.01 ± 1.72
+demos	50.04 ± 2.12	49.42 ± 1.74	51.27 ± 1.88
+anno (short)	47.23 ± 2.32	45.45 ± 1.66	53.91 ± 6.25
+anno (short)+demos	41.16 ± 1.06	41.59 ± 1.29	37.13 ± 1.42
+anno (full)	50.65 ± 0.27	50.33 ± 0.30	51.09 ± 0.68
+anno (full)+demos	45.07 ± 0.49	44.40 ± 0.68	47.27 ± 0.19
	New Label		
Majority class (single)	36.14 ± 0.0	35.71 ± 0.0	37.41 ± 0.0
task	50.32 ± 2.75	50.68 ± 2.53	48.65 ± 4.59
+demos	52.12 ± 1.05	52.07 ± 1.37	51.70 ± 1.26
+anno (short)	43.39 ± 1.47	41.57 ± 0.59	49.29 ± 6.43
+anno (short)+demos	50.40 ± 0.53	50.40 ± 0.27	48.59 ± 1.98
+anno (full)	51.70 ± 2.51	51.17 ± 2.36	53.12 ± 3.03
+anno (full)+demos	43.31 ± 0.29	40.70 ± 0.73	51.57 ± 1.16

Table 11: Results of fine-tuning LLaMA 3 on single annotator labels using different input prompts for the GBV detection task.

Model: Mistral	**Original Label**		
	All	EDOS	DOM
Majority class (single)	44.54 ± 0.0	43.61 ± 0.0	47.15 ± 0.0
task	58.75 ± 0.78	59.42 ± 0.96	54.86 ± 1.14
+demos	44.51 ± 0.94	43.61 ± 0.27	47.03 ± 0.59
+anno (short)	45.90 ± 1.33	47.07 ± 1.64	41.20 ± 1.81
+anno (short)+demos	64.31 ± 0.45	66.85 ± 0.31	54.51 ± 0.77
+anno (full)	44.41 ± 1.19	43.52 ± 2.05	46.92 ± 1.72
+anno (full)+demos	47.74 ± 0.66	48.15 ± 0.41	43.66 ± 0.63
	New Label		
Majority class (single)	36.14 ± 0.0	35.71 ± 0.0	37.41 ± 0.0
task	59.20 ± 2.10	57.45 ± 1.18	64.79 ± 1.54
+demos	41.07 ± 1.18	39.86 ± 1.11	44.92 ± 2.31
+anno (short)	52.56 ± 1.79	53.98 ± 0.69	57.61 ± 1.36
+anno (short)+demos	67.40 ± 1.55	67.53 ± 0.93	66.62 ± 1.10
+anno (full)	40.96 ± 2.97	40.23 ± 2.12	42.92 ± 1.87
+anno (full)+demos	54.15 ± 0.46	54.00 ± 0.77	52.83 ± 2.05

Table 12: Results of fine-tuning Mistral on single annotator labels using different input prompts for the GBV detection task.