SemEval-2023 Task 10: Explainable Detection of Online Sexism

Hannah Rose Kirk[1,2*], **Wenjie Yin**[1,3*], **Bertie Vidgen**[1], and **Paul Röttger**[1,2]

[1]Rewire
[2]University of Oxford
[3]Queen Mary University of London

Abstract

Online sexism is a widespread and harmful phenomenon. Automated tools can assist the detection of sexism at scale. Binary detection, however, disregards the diversity of sexist content, and fails to provide clear explanations for why something is sexist. To address this issue, we introduce SemEval Task 10 on the Explainable Detection of Online Sexism (EDOS). We make three main contributions: i) a novel hierarchical taxonomy of sexist content, which includes granular vectors of sexism to aid explainability; ii) a new dataset of 20,000 social media comments with fine-grained labels, along with larger unlabelled datasets for model adaptation; and iii) baseline models as well as an analysis of the methods, results and errors for participant submissions to our task.

Content warning: We show illustrative examples of sexist language to describe the taxonomy and analyse error types.

1 Introduction

Online sexism can inflict harm on women who are targeted, make online spaces inaccessible and unwelcoming, and perpetuate social asymmetries and injustices. Automated tools are now widely deployed to find sexist content at scale, supporting content moderation, monitoring and research. However, most automated classifiers do not give further explanations beyond generic, high-level categories such as 'toxicity', 'abuse' or 'sexism'. Opaque classification without explanations can cause confusion, anger and mistrust among moderated users, and make it harder to challenge moderation mistakes; moderators who use automated tools may not trust or fully understand the labels, reducing their efficiency; and it is difficult to assess model weaknesses in the absence of granular labels, which hinders the development of better models.

We organised the SemEval 2023 Task *Explainable Detection of Online Sexism (EDOS)*, which follows from success of previous shared tasks on abuse and hate detection (Zampieri et al., 2019; Ribeiro and Silva, 2019; Pavlopoulos et al., 2021; Basile et al., 2019; Fersini et al., 2018a,b). Our task proposes and applies a taxonomy with three hierarchical tasks for detecting sexist content. **Task A** is a binary task to detect whether content is sexist. For sexist content, **Task B** distinguishes between four distinct categories of sexism, and **Task C** identifies one of 11 fine-grained sexism vectors.

There are three factors that make our EDOS task unique: First, **data diversity**: we sample data from two large social media platforms (Reddit and Gab) using an ensemble of filtering methods. The entries are more diverse than sampling from keywords alone or a single platform. Second, **annotation quality**: we recruit highly-trained annotators who all self-identify as women. We make this decision to mitigate implicit biases in labelling and to evoke a participatory approach to AI where the communities primarily harmed from specific forms of content are included in the data development process (Birhane et al., 2022; Zytko et al., 2022). To improve consistency of labels with our detailed guidelines, all annotators were trained in multiple pilot tasks, and disagreements resolved by experts. Third, **task granularity**: we ground our dataset in a carefully-constructed taxonomy such that the multiple categories in Tasks B and C are both meaningful from a social scientific perspective, aid explainability in predictions and present a challenging machine learning task for multi-class classification.

In releasing the dataset and analysing the accompanying SemEval task submissions, we contribute to efforts in building automated sexism detection systems which are both more *accurate* and *explainable* via fine-grained labels.[1] These contributions mark a step towards a safer internet for women.

[*]Equal contribution

[1]All task data, individual annotations and guidelines are available at github.com/rewire-online/edos. A data statement is provided in Appendix A.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2193–2210

2 Taxonomy Design

2.1 Existing Taxonomies

We conducted a broad review of research on sexism and misogyny. Within this literature, we identified recent articles that propose taxonomies of sexist or misogynist content, in particular Jha and Mamidi (2017); Samory et al. (2021); Farrell et al. (2019); Zeinert et al. (2021); Guest et al. (2021); Rodríguez-Sánchez et al. (2021) and Parikh et al. (2021). Each taxonomy is unique, but they generally differ along several dimensions, including differences in (i) construction (whether they are built from theory or empirics); (ii) scope (how the top-level category of sexism or misogyny is defined and which subcategories are included); and (iii) structure (whether subcategories are hierarchically ordered). We review these details in Appendix B.

2.2 Our Taxonomy

Starting from this literature review, we created a first draft of the taxonomy. The taxonomy was further refined using a grounded theory approach (Glaser and Strauss, 2017) with empirical entries from our dataset (see §3) to merge or adjust the schema. It comprises three subtasks:

Task A: Binary Sexism The first level of our taxonomy makes a binary distinction between sexist and non-sexist content. We define *sexist content* as any abuse, implicit or explicit, 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 or Muslim women). Our taxonomy focuses on *sexism*, rather than *misogyny*. Misogyny refers to "expressions of hate towards women" (Ussher, 2016), while sexism also covers more subtle implicit forms of abuse and prejudice that can still substantially harm women.[2]

Task B: Category of Sexism The second level of our taxonomy disaggregates sexist content into four conceptually and analytically distinct categories. We consciously chose not to separate out categories by the supposed effect on the recipient, or supposed motivation of the speaker, because the harm caused by sexist content is idiosyncratic; and speaker intent is difficult to gauge, especially without broader context. The four categories are **(1) Threats, plans to harm & incitement:** Language where an individual expresses intent and/or encourages others to take action against women which inflicts or incites serious harm and violence against them. It includes threats of physical, sexual or privacy harm. **(2) Derogation:** Language which explicitly derogates, dehumanises, demeans or insults women. It includes negative descriptions and stereotypes about women, objectification of their bodies, strong negative emotive statements, and dehumanising comparisons. It covers negative statements directed at a specific women and women in general. **(3) Animosity:** Language which expresses implicit or subtle sexism, stereotypes or descriptive statements. It includes benevolent sexism, i.e., framed as a compliment. **(4) Prejudiced Discussion:** Language which denies the existence of discrimination, and justifies sexist treatment. It includes denial and justification of gender inequality, excusing women's mistreatment, and the ideology of male victimhood.

Task C: Fine-Grained Vectors The third level of our taxonomy disaggregates each *category* of sexism into fine-grained sexism *vectors*. We seek vectors which are mutually exclusive (i.e., each vector is distinct) and collectively exhaustive (i.e., all sexist content can be assigned to a vector). Tab. 1 gives an overview of the 11 fine-grained vectors in our taxonomy, along with their definitions and examples taken from prior literature then edited to clearly demonstrate each distinct vector.

3 Dataset Construction

3.1 Choice of Platform

A large portion of online harms research gathers data from Twitter (Vidgen and Derczynski, 2020), resulting in lacking community-based diversity in the severity and form of sexist content. Thus, in order to produce a diverse dataset with coverage of our fine-grained vectors of sexism, we opt to use two social media platforms. First, **Gab**, which is an 'alt-tech' social networking site established in 2016. It has positioned itself as a rival to mainstream sites, such as Twitter and Facebook, and explicitly aims to protect free (and far-right) speech (Zannettou et al., 2018; Lima et al., 2018; Jasser et al., 2021). Second, **Reddit**, which is a network of topic-based forums ('subreddits'), where users can share, view and comment on content related

[2] Abuse primarily targeted at groups other than women, like racism or antisemitism, is not covered by our taxonomy. However, we tasked annotators with flagging such content using an "other target" label, which, by majority vote, applied to 2,098 out of 15,146 non-sexist entries.

Category	Vector	Definition	Example
1. Threats, plans to harm and incitement	**1.1** Threats of harm	Expressing intent, willingness or desire to harm an individual woman or group of women. This could include, but is not limited to: physical, sexual, emotional or privacy-based forms of harm.	*I'll kill any women that talks back to me*
	1.2 Incitement and encouragement of harm	Inciting or encouraging an individual, group, or general audience to harm a woman or group of women. It includes language where the author seeks to rationalise and/or justify harming women to another person.	*Raping her would put her in her place*
2. Derogation	**2.1** Descriptive attacks	Characterising or describing women in a derogatory manner. This could include, but not limited to: negative generalisations about women's abilities, appearance, sexual behaviour, intellect, character, or morals.	*Women's football is so shit, they're so slow and clumsy*
	2.2 Aggressive and emotive attacks	Expressing strong negative sentiment against women, such as disgust or hatred. This can be through direct description of the speaker's subjective emotions, baseless accusations, or the use of gendered slurs, gender-based profanities and gender-based insults.	*I hate women*
	2.3 Dehumanising attacks and overt sexual objectification	Derogating women by comparing them to non-human entities such as vermin, disease or refuse, or overtly reducing them to sexual objects.	*Women are pigs*
3. Animosity	**3.1** Causal use of gendered slurs, profanities and insults	Using gendered slurs, gender-based profanities and insults, but not to intentionally attack women. Only terms that traditionally describe women are in scope (e.g. 'b*tch', 'sl*t').	*Stop being such a little bitch*
	3.2 Immutable gender differences and gender stereotypes	Asserting immutable, natural or otherwise essential differences between men and women. In some cases, this could be in the form of using women's traits to attack men. Most sexist jokes will fall into this category.	*Men and women's brains are wired different bro, that's just how it is*
	3.3 Backhanded gendered compliments	Ostensibly complimenting women, but actually belittling or implying their inferiority. This could include, but is not limited to: reduction of women's value to their attractiveness or sexual desirability, or implication that women are innately frail, helpless or weak.	*Women are delicate flowers who need to be cherished*
	3.4 Condescending explanations or unwelcome advice	Offering unsolicited or patronising advice to women on topics and issues they know more about (known as 'mansplaining').	*My gf always complains about period pains but she just doesn't understand the medical science for eliminating them!*
4. Prejudiced Discussion	**4.1** Supporting mistreatment of individual women	Expressing support for mistreatment of women as individuals. Support can be shown by denying, understating, or seeking to justify such mistreatment.	*Women shouldnt show that much skin, it's their own fault if they get raped*
	4.2 Supporting systemic discrimination against women as a group	Expressing support for systemic discrimination of women as a group. Support can be shown by denying, understating, or seeking to justify such discrimination.	*The leadership of men in boardrooms is a necessary evil—corporations need to be efficiently run*

Table 1: Taxonomy of sexism categories (Task B) and fine-grained vectors (Task C).

to their interests. Several studies have shown that Reddit is home to many sexist and anti-feminist communities, described as the 'manosphere' (Ging, 2017; Zuckerberg, 2018; Farrell et al., 2019; Ging et al., 2020; Ribeiro et al., 2021).

We source equal amounts of data from Reddit and Gab. For each platform, we first create a pool of 1M entries (see §3.2). We then sample 10k entries from each pool for labelling (see §3.3).

3.2 Data Collection

Gab We collect 34M publicly available Gab posts from August 2016 to October 2018.[3] This data has been widely used in academic studies (e.g., Kennedy et al., 2018; Cinelli et al., 2021). We randomly sample 1M entries to create the pool.

Reddit First, we compile a list of 81 subreddits which are likely to contain sexist content, based on previous work (Guest et al., 2021; Farrell et al., 2019; Zuckerberg, 2018; Jones et al., 2020; Qian et al., 2019; Ging, 2017; Ribeiro et al., 2021) and online wikis.[4] Then, we collect all comments from August 2016 to October 2018 in these subreddits using the Reddit API.[5] We manually review each subreddit and assign it to one of four categories (Incels, Men Going Their Own Way, Men's Rights Activists, Pick Up Artists) based on prior work by Lilly (2016).[6] To ensure that our dataset is not overly biased towards niche linguistic expressions and topics, we restrict our sampling to the 24 subreddits with at least 100k comments, resulting in a dataset of 42M comments. Finally, we randomly sample 250k comments from each of the four subreddit categories to create the pool.

[3] files.pushshift.io/gab

[4] rationalwiki.org/wiki and incels.wiki

[5] We only collect comments as our early analysis showed that many posts are single URLs, images or videos and thus harder to parse in isolation for labelling. We specify the date range to match the Gab data.

[6] We describe these categories in Appendix C.

		train		dev		test	
Task A							
not sexist		10,602	76%	1,514	76%	3,030	76%
sexist		3,398	24%	486	24%	970	24%
	total	14,000	100%	2,000	100%	4,000	100%
Task B							
1. threats, plans to harm and incitement		310	9%	44	9%	89	9%
2. derogation		1,590	47%	227	47%	454	47%
3. animosity		1,165	34%	167	34%	333	34%
4. prejudiced discussion		333	10%	48	10%	94	10%
	total	3,398	100%	486	100%	970	100%
Task C							
1.1 threats of harm		56	2%	8	2%	16	2%
1.2 incitement and encouragement of harm		254	7%	36	7%	73	8%
2.1 descriptive attacks		717	21%	102	21%	205	21%
2.2 aggressive and emotive attacks		673	20%	96	20%	192	20%
2.3 dehumanising attacks and overt sexual objectification		200	6%	29	6%	57	6%
3.1 casual use of gendered slurs, profanities, and insults		637	19%	91	19%	182	19%
3.2 immutable gender differences and gender stereotypes		417	12%	60	12%	119	12%
3.3 backhanded gendered compliments		64	2%	9	2%	18	2%
3.4 condescending explanations or unwelcome advice		47	1%	7	1%	14	1%
4.1 supporting mistreatment of individual women		75	2%	11	2%	21	2%
4.2 supporting systemic discrimination against women as a group		258	8%	37	8%	73	8%
	total	3,398	100%	486	100%	970	100%

Table 2: Distribution of class labels across tasks and data splits.

3.3 Data Preparation and Sampling

Cleaning We cleaned the text in the Gab and Reddit pools by: (1) replacing URLs and usernames with generic tokens; (2) dropping empty entries; (3) dropping entries that only contain URLs or emoji; (4) dropping non-English language entries; and (5) dropping duplicates. After these cleaning steps were completed, we apply our sampling techniques.

Boosted Sampling The prevalence of abusive content 'in the wild' is difficult to estimate reliably, but could be as low as 0.1% or 1% (Vidgen et al., 2019). Sexism represents only a subset of all abuse; So, random sampling from online platforms will lead to a dataset with a large class imbalance, which impedes the training of AI systems. A range of sampling techniques have been used in prior work to boost the proportion of abusive content in datasets, including keyword search (ElSherief et al., 2017) or lexicons (Farrell et al., 2019), and community-based (Guest et al., 2021; Vidgen et al., 2021a) or user-based sampling (Vidgen et al., 2019). For privacy concerns, we do not store user-based information. Instead, we apply a mix of community-based sampling (on Reddit), with a carefully-designed ensemble of varied sampling methods.

Ensemble of Sampling Methods Relying on a single sampling method makes the sampled data more prone to biases (Yin and Zubiaga, 2022). Following extensive pilots, we settle on six different techniques to sample 10k cleaned entries from each of Gab and Reddit pools (total $n = 20{,}000$), to ensure coverage of the 11 fine-grained sexism vectors and introduce lexical and topical diversity. The six techniques sample (1) 1,000 entries with at least one sexist keyword (e.g. "c*nt", "b*tch"); (2) 1,000 entries with at least one sexist keyword and one topical keyword (e.g. "she", "girl"); (3) 100 entries from each decile of toxicity scores from the Perspective model (total 1,000)[7]; (4) 100 entries from each decile of scores from a bespoke classifier for sexism detection which we trained on seven open source sexism/misogyny datasets (total 1,000)[8]; (5) 1,000 entries from cases where the score from Perspective's Toxicity model differed from the score of our custom classifier by at least $|0.8|$; and (6) 5,000 entries sampled using a combination of topical keywords and scores from other attributes of Perspective (e.g., Identity Attacks + "girl" or Sexually Explicit + "she").

3.4 Data Annotation

We follow the 'prescriptive paradigm' for data annotation, in that we want the annotators to apply our comprehensive annotation guidelines rather than applying their own subjective beliefs (Röttger et al., 2022). Our annotation guidelines contain definitions, clarifying notes, exemplars, edge cases, and general guidance.

Annotator Recruitment To mitigate the risk of implicit bias and to encourage participation from affected communities, we worked with trained anno-

[7] perspectiveapi.com

[8] For training details, see Appendix D.

tators who all self-identify as women. We required that all annotators pass a challenging 200-entry screening task that covered all 11 sexism vectors in our data. In total, we recruited and trained 19 annotators who passed the screening. We opted for expert annotation over crowdwork because pilot experiments demonstrated an marked difference in quality and consistency of labels.[9] The demographics of annotators and their experiences with online sexism are documented in Appendix A.

Annotation Process Three annotators labelled each entry. To further ensure label quality, we rely on expert adjudication for disagreements. Experts were called upon to give labels for (i) cases with less than 3/3 agreement (unanimous) in Task A, and (ii) cases with less than 2/3 agreement in Tasks B and C. The expert team consisted of two women authors of this paper, and two of the most experienced workers from the annotation team. Data was assigned to annotators in batches every two weeks over the course of two months. Throughout the process, we worked closely with annotators so that their feedback could be incorporated into our guidelines and so that their welfare could be continuously monitored and protected.

3.5 Dataset Distribution

The data was split into training, development, and test sets in the ratio of 70:10:20. Only sexist instances are included in Task B and C. Label distributions of all splits are shown in Tab. 2.

4 Task Description

4.1 Task Definition

Our SemEval task consists of three subtasks, reflecting our hierarchical taxonomy (§2.2). Task A is a binary classification task (sexist vs. not sexist). Task B and C are multi-class classification tasks, with four and 11 categories, respectively.

4.2 Task Organisation

We ran our SemEval task on CodaLab.[10] There were two primary phases: (i) the Training and Development Phase which ran from September 2022 to January 2023, and (ii) the Test Phase, which began on 10th January 2023 and ended on 31st January 2023. During the Training and Development Phases, we released entries and labels from the train and dev splits, as well as an additional *starting kit* including 1M unlabelled Reddit entries and 1M unlabelled Gab entries. During the Test Phase, we staggered the release of test entries for Task A from Tasks B and C, since the latter provides information on the correct labels of the former.

4.3 Evaluation Metrics and Baselines

We evaluate all systems with macro-average F1 score to account for imbalance between classes. We supply 7 baseline models (see Tab. 3) to benchmark a range of simple to complex systems. The simplest baselines are always predicting the most frequent class (*B0*) and uniformly predicting each class (*B1*). We train one traditional machine learning model (*B2*) where the data is first vectorized with TF-IDF and then fitted with an XGBoost model. The remaining baselines use DistilBERT (Sanh et al., 2019) and DeBERTa-v3-base (He et al., 2021a). For each model, we fine-tune on the training set (*B3, B5*), and do continued pre-training on the 2M unlabelled entries from Reddit and Gab combined (*B4, B6*). Across all tasks, DeBERTa with continued pre-training sets the highest baseline. The best baseline for Task A (F1=0.8235) is more saturated than for Tasks B and C (F1=0.5926; 0.3171), making this an interesting SemEval challenge with varying degrees of difficulty and remaining headroom for performance improvements.

5 Participant Systems and Results

5.1 Participant Overview

EDOS was a very popular SemEval task, with 599 accounts signing-up for dataset access.[11] In the Development Phase, 128 submissions were made for Task A, 71 for Task B and 52 for Task C. In the Test Phase, 134 submissions were made for Task A, 87 for Task B and 81 for Task C.

Validation Process To ensure a fair competition in the Test Phase, we allowed one submission account per team and up to two test submissions of task predictions, to allow for one upload mistake.[12] We issued a short survey to confirm compliance with our T&Cs and collect system information. Our final leaderboard includes those who responded to this survey: 84 to Task A, 69 to Task B and 63 to

[9] Results are presented in Appendix E.

[10] codalab.lisn.upsaclay.fr/competitions/7124

[11] Number of accounts signed up as of 25/04/2023.

[12] We unfortunately encountered some suspicious activity with multiple accounts using generic emails `ABCD1234@domain`, or making >10 test set submissions.

	Model	Continued Pre-training	Macro-F1 Task A	Task B	Task C
Baselines					
B0	MostFrequent	✗	0.4310	0.1594	0.0317
B1	Uniform	✗	0.4509	0.2413	0.0629
B2	XGBoost	✗	0.4933	0.2297	0.0881
B3	DistilBERT	✗	0.7621	0.5531	0.2935
B4		✓	0.7804	0.5367	0.3140
B5	DeBERTa-v3-base	✗	0.8235	0.4790	0.1517
B6		✓	**0.8235**	**0.5926**	**0.3171**
Top Ranked Systems					
PingAnLifeInsurance	DeBERTa-v3-large, twHIN-BERT-large	✓	**0.8746**		
stce	RoBERTa-large, ELECTRA	✓	0.8740	0.7203	0.5487
FiRC-NLP	DeBERTa (ensemble)	✗	0.8740		
JUAGE	PaLM (ensemble)	✗		**0.7326**	
PASSTeam	RoBERTa, HateBERT	✓		0.7212	0.5412
PALI	DeBERTa, RoBERTa	✓			**0.5606**
Summary Statistics					
		System count	84	69	63
		Q1	0.7994	0.5730	0.3153
		Mean	0.8095	0.5899	0.3829
		Median	0.8322	0.6191	0.4230
		Std	0.0746	0.1065	0.1274
		Q3	0.8537	0.6501	0.4758
		Percentage of systems which beat *B0*	100%	100%	100%
		Percentage of systems which beat *B6*	55%	60%	73%

Table 3: Baselines, top ranked leaderboard results and summary statistics per Task. The best baseline is **bolded**. The best participant submission is **bolded in red**.

Task C. 59 teams submitted to all three tasks, 10 submitted to two, and 20 submitted to just one.

5.2 Leaderboard Results

The top three submissions for each task are presented in Tab. 3.[13] The top three systems in all tasks use multiple models or an ensemble, and many top systems apply further pre-training and multi-task learning. Notably, *stce* and *PASSTeam* achieved top results on two or more tasks. Both used multi-task learning and further pre-training on the unlabelled data in the starter kit: *stce* use RoBERTa-large (Liu et al., 2019) and ELECTRA (Clark et al., 2020); while *PASSTeam* use a multi-task learning strategy (Yu et al., 2020) with fine-tuned RoBERTa and HateBERT (Caselli et al., 2021). In Task A, *PingAnLifeInsurance* also adopt a multi-task DNN structure (Liu et al., 2020) with further pre-training of DeBERTa-v3 (He et al., 2021a) and TwHIN-BERT (Zhang et al., 2022c) on the starter kit unlabelled data and an additional Kaggle dataset. *FiRC-NLP* use an ensemble of various DeBERTa models fine-tuned only on the labelled task data. In Task B, *JUAGE* was one of the few systems relying on prompt-based learning, but achieved first place using an instruction-tuned PaLM model (Chowdhery et al., 2022) with a parameter-efficient prompt tuned on the task data and majority voting over six iterations. In Task C, *PALI* further pre-trained DeBERTa-v3 using the starter kit unlabelled data and applied a second loss term (normalized temperature-scaled cross entropy).

Scores for Tasks B and C were substantially lower in mean and higher in variance than Task A (see Tab. 3). All participant systems beat our simplest baseline predicting the most frequent class, while a majority still beat our more complex baseline of DeBERTa-v3 with continued pre-training.

5.3 Popular Methods

Architecture For all tasks, the large majority of participants (∼90%) used a transformers architecture (Fig. 1). The most popular transformer-based models include RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021b,a), BERT (Devlin et al., 2019), BERTweet (Nguyen et al., 2020) and DistilBERT (Sanh et al., 2019). Several submissions use prompted language models such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022) or OPT (Zhang et al., 2022b). Other approaches include traditional machine learning methods (∼8%) and non-transformer deep neural networks (∼6%), which are often combined with other methods.

Additional Training and Data The majority of participants applied fine-tuning only on the target task (>90%) but some also apply further pre-training (∼30%), fine-tuning on an auxiliary task

[13] Full leaderboards are available on github.com/rewire-online/edos

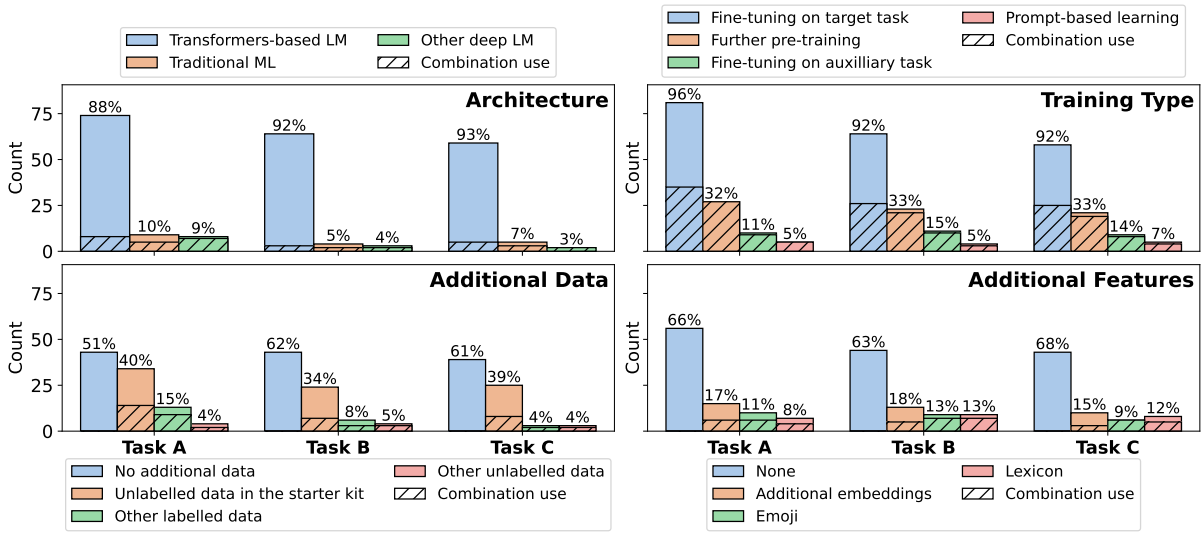

Figure 1: Participant methodologies by popularity (count and percentage of submissions using the method). Each subplot shows a different breakdown of system paper method decisions. The first subplot (top left) shows that the majority of participants used transformer-based language models. The second subplot (top right) shows fine-tuning on the target task was the most popular training method, but a substantial proportion of participants also applied further pre-training. The third subplot (bottom left) similarly shows that while the majority only used the labelled data provided in the task, many participants also used the in-domain unlabelled data in our starter kit. Finally, the last subplot (bottom right) shows the majority did not use additional features.

(∼15%) or prompt-based learning (∼5%). Around 40% of participants used the unlabelled data in the starter kit. Participants also sometimes used resources outside our task, both labelled (∼10%) and unlabelled data (∼5%). A number of innovative methods were applied including data augmentation, active learning or data cartography (Swayamdipta et al., 2020).

Additional Features The majority of participants did not use additional features (∼66%). If additional features were used, they were predominantly used in a combination of additional embeddings (∼18%), emoji (∼12%) or lexicons (∼12%).

5.4 Most Effective Methods

The distribution of participant F1 scores across methodologies is more tightly clustered for Task A than Tasks B and C (see Fig. 2). Across methodology variables, **architecture** choices lead to the most different distributions of performances, with transformers-based and other deep language models providing a more competitive average F1 than traditional ML systems across tasks. In contrast, **training type** seems to have the least influence on the distribution of F1 scores, with the most overlap between conditions, even in Task C. The highly related **additional data** condition has slightly more effect on performance, with the use of unlabelled data in the starter kit resulting in the highest average. Systems that did not use any **additional features** had higher average F1 scores in all tasks.

6 Error Analyses

We base our error analyses on the 10 best performing systems for each task as their errors indicate the remaining difficulty within our dataset.

6.1 Quantitative Error Analysis

Confusion Between Binary Labels Out of 4,000 test instances in Task A, 162 were misclassified by all top 10 systems, where 38.3% were false positives (not sexist content predicted as sexist) and 61.7% were false negatives (sexist content predicted as not sexist). Of the false negatives, 41% are from the *Animosity* category, 34% from *Derogation*, 16% from Prejudiced Discussion and 9% from Threats. This pattern is expected because *Animosity* is often contains implicit language, while *Threats* are commonly conveyed in explicit language.

Confusion Between Categories and Vectors Among the categories, *Threats* is least often misclassified as other categories and the least common mistake for instances with the true label of any other category (see Fig. 3). *Animosity* and *Deroga-*

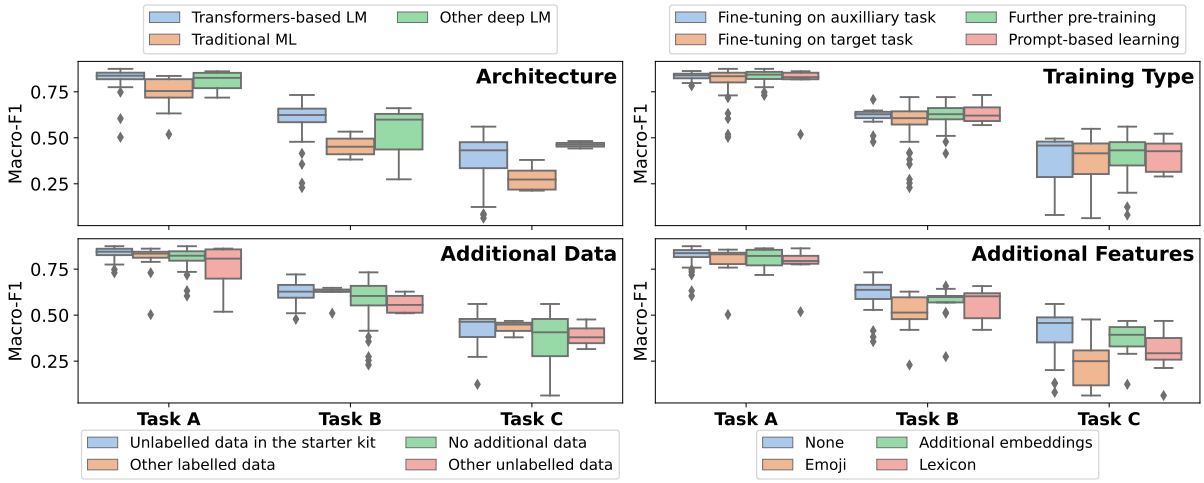

Figure 2: Participant methodologies by Macro-F1 score (across submissions using the method in isolation or combination). Each subplot shows a different breakdown of system paper method decisions and the average F1 score by task. F1 scores on Task A are generally higher and more tightly distributed than on Tasks B and C.

tion are often mistaken for each other. Instances of *Prejudiced Discussion* are also commonly misclassified as *Derogation* or *Animosity*. The confusion across vectors in Task C has a similar pattern (Appendix F). Some errors reflect class imbalance: the largest class (2.1) and the smallest class (3.4) are the most and least frequent confusions, respectively. Others errors reveal inherent similarity between certain vectors: 2.1 and 3.2 are commonly confused, with both referring to gendered stereotypes, but 2.1 is explicitly negative in sentiment. The same is true for 2.2 and 3.1, which share gendered slurs, except that 3.1 is causal (not targeted) usage.

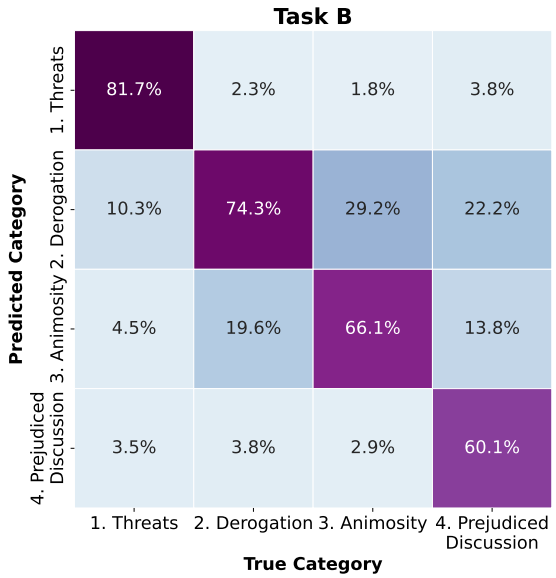

Figure 3: Confusion matrix for categories of sexism (Task B) across the top 10 performing systems.

6.2 Manual Error Inspection

Two self-identifying woman authors manually inspected 100 instances from the Task A test set that were misclassified by all top 10 teams. Out of the 100 entries, we identified 76 as genuine system misclassifications, while 24 were annotation errors.[14]

System Errors Of the 76 system errors, 52 were false negatives (68%) and 24 were false positives (32%). Of the **false negatives**, there were several common themes of misclassified examples. These themes include (i) expressions that are seemingly positive in sentiment or backhanded gendered compliments (e.g., "*STEM ain't hard, wamans can do it too!*"); (ii) gender stereotypes encoded in a joke (e.g., "*I just bought a smart car. It doesn't turn on when the wife is in the driving seat.*"); (iii) slang use (e.g., "*Tha Brothas want dem PAWGs.*") or rarer gendered insults (such as "hag" or "witch"); and (iv) irony (e.g., "*either you worship women or you're a misogynist.*"). Some explicit examples, paired with rhetorical devices such as questions, were also misclassified as not sexist (e.g., "*How does one learn how to choke a women? I assume there is an art to it? Is it that you're depriving blood to the brain more so than choking the windpipe?*").

Of the **false positives**, there were also a number of common patterns. These include (i) counter-speech (e.g., "*It's fucked up how women aren't*

[14]Note that the annotation error rate is likely higher in this subset relative to the dataset as whole, because it includes challenging examples that both automated systems and humans struggle to classify.

allowed to even be neutral. Always have to put on a fucking show to entertain men and keep them from attacking you."); (ii) matter-of-fact discussion or descriptions of groups associated with sexism (e.g., "*The whole point of MGTOW is, men are going their own way....which doesn't mean we've changed from desiring women to desiring men. It just means we don't allow women to decide what it is to be a man - or to be the main focal point of our lives.*"); or (iii) criticisms of an individual woman's actions (e.g., "*I know a woman who slept with one off her best friends on off boyfriend. Dropped her entire social circle for a jerk, not even a Chad. She was an, 8 as well. SMH. She had choices, chose this guy. Women can be dumb.*").

Annotation Errors Out of the 24 mislabelled examples, 17 were false negatives (29%) and 7 were false positives (71%). In general, annotators struggled with the distinction between individual-related and generalised statements. False negatives occur when gender stereotypes are expressed as speculations rather than characteristic descriptions (e.g., "*We have condoms and birth control so now women can sex with fewer and much less serious consequences as well. It makes sense that they would have more sex with the top guys when there's relatively little to risk.*"). False positives occur with non-gendered attacks on individual women targets (e.g., "*This woman is just stupid.*").

Grey Areas with Lacking Context Although a decision was reached by the authors on each manually-inspected instance, 9 of the 24 mislabelled instances were considered a possibly grey area – instances that could fall under a different label depending on context or slight modifications to the guideline interpretation. Individual posts and comments taken in isolation sometimes lack the context to interpret meaning or intent. Annotators may then fill in the context based on their experiences with online sexism or knowledge of the relevant communities. This was particularly evident in criticisms of individual woman, where it is often unclear without context whether the speaker is criticising an individual for their specific actions or using that individual as scapegoat to express general sexist stereotypes and views. This also raises the question of whether something should be flagged if it hints at the speaker's sexist tendency but does not itself convey specific sexist content.

7 Discussion

7.1 Lessons Learnt

Value of Diverse Data Sampling We note that previous SemEval tasks on hate or abuse detection primarily report false positive errors from their participants. For example, in HatEval (Basile et al., 2019), 89.1% of the errors made by the top three systems were false positives. In our analysis, we instead find more false negatives. We believe this flipped distribution of error types is due to the subtle and diverse forms of sexist content in our dataset (such as *Prejudiced Discussion*) which do not necessarily share linguistic form or keywords with more explicit sexist content, making them more challenging to correctly identify.

Challenge of Fine-Grained Predictions Our baselines and participant scores demonstrate that binary sexism detection (Task A) is a substantially easier task than fine-grained sexism detection (Task B and C). The maximum F1 score achieved in Task C is just above 50%. This shows that high scores reported in binary detection tasks of socially-conditioned concepts (like sexism, hate or abuse) obscure fine-grained model failings in distinguishing between forms of content, which may have very different impacts on their targets and society at large. Given the remaining headroom in predicting our categories and vectors of sexism, we are excited to see how future model-centric contributions using our dataset continue to improve performance. Future work could also apply data-centric techniques to increase performance and robustness, such as augmenting the dataset with challenging adversarial examples (Kiela et al., 2021; Vidgen et al., 2021b; Kirk et al., 2022c) or mapping hard-to-learn subspaces with data cartography methods (Swayamdipta et al., 2020).

Efficiency of Continued Pretraining Labelling data for socially-conditioned tasks (like sexism) is challenging, expensive and requires some degree of domain expertise (Kirk et al., 2022b). However, providing unlabelled in-domain data is both cheap, scalable and avoids causing psychological harm on annotators from a greater labelling burden (Kirk et al., 2022a). We are encouraged that our best baselines and top performing participant systems effectively used continued pre-training on the unlabelled in-domain data that we provided (from Reddit and Gab). We encourage future SemEval

tasks to similarly release unlabelled datasets from which they sample smaller labelled datasets.

7.2 Limitations

Class Imbalance Our dataset is more balanced on the binary label (24% sexism) than many previous datasets for hate and abuse detection (Vidgen and Derczynski, 2020). However, there is substantial imbalance in categories and vectors of sexism. Thus, it is hard to confirm whether confusion between categories and vectors is due to inherent features of the data or due to the class imbalance. We encourage future work to examine the effect on performance and cross-vector confusion when balancing the dataset. That said, we explicitly made the decision to not re-balance our dataset because different types of sexism (especially at the fine-grained level) do have very different base rates 'in the wild'. Dealing with class imbalance is then part-and-parcel of the problem we seek to address.

Defining "Explainable" We cast the explainability problem as a classification task on a single text document (without network or user information). In this setting, combining hierarchical classifications is what adds explainability by demonstrating whether a piece of content is *sexist* and for what reason (*category* and *vector*). There are many other promising methods to make automated systems more explainable for socially-sensitive tasks (e.g., sexism, abuse or hate), including classifying spans or masked tokens for rationale prediction (Mathew et al., 2021; Kim et al., 2022); leveraging user and network data (Qian et al., 2018; Wich et al., 2021) or targets data (Kennedy et al., 2020); applying social bias frames for stereotypes (Sap et al., 2020); or encouraging multi-hop and chain-of-thought reasoning (Zhang et al., 2022a; Jin et al., 2022).

Value Specificity and Perspective Our dataset is labelled by UK-residing annotators, who all self-identify as women. We issue prescriptive guidelines, then rely on majority vote and expert adjudication to produce a gold label. Many previous studies reveal that annotator identity is a critical determinant of labelling behaviour (Waseem, 2016; Larimore et al., 2021; Sap et al., 2022) and so majority votes inadequately capture subjective disagreement (Davani et al., 2021). To encourage future work on annotator-specific artefacts, we provide annotator IDs with each labelled entry (Prabhakaran et al., 2021). Nonetheless, our definition of sexism, its subcategories in our taxonomy and how our annotators apply their best judgement are grounded in few cultural viewpoints, primarily from Western-centric and English-speaking communities. Any classifiers built from our dataset will thus inherit our values (as taxonomy and guideline writers) and those of our annotators.[15] Like most toxic content datasets, our data is limited to English (Röttger et al., 2022). The bounds of sexism, however, are culturally contested and people may legitimately disagree in where to draw the line.

8 Conclusion

In this paper, we made three main contributions. First, we provided a new taxonomy for the more explainable classification of sexism in three hierarchical levels – binary sexism detection, category of sexism and fine-grained vector of sexism. This taxonomy is grounded in reviews of prior taxonomies and social science literature, then empirically-validated with in-domain data from two social media platforms. It thus provides a sociotechnical overview of the varied and nuanced landscape of online sexism. Second, we created a high-quality dataset, sampled with an ensemble of techniques to increase the diversity of content and annotated by self-identifying women experts to ensure consistent labels. This labelled dataset is paired with a larger unlabelled dataset to mitigate the constraints that a labelling budget (both in terms of financial cost and psychological cost to annotators) has on the efficacy of trained systems. Finally, we shared baseline models and summarised systems submitted to our SemEval task. This analysis demonstrates the success of techniques that combine state-of-the-art transformer models with continued pre-training and multi-task learning. However, there is still substantial headroom for improving performance on detecting fine-grained forms of sexist content. We hope that our research and resources can contribute towards the design of future systems that make online spaces safer for all.

Acknowledgements

This work was carried out by Rewire, and funded by MetaAI. We are grateful to our annotators, and to the Rewire team for providing feedback on this

[15] Bang et al. (2022) seek to design "value-aligned" sexism detection systems, where their model considers both the content and an individual's value system to return a classification. This technique is a promising way to further improve inclusivity and explainability in sexism detection tasks.

article. We also thank all participants who submitted results to our task, and our reviewers who gave valuable and constructive comments on the system papers.

Ethical Risks and Harm Statement

We release a dataset containing online sexism, where we have demonstrated that state-of-the-art models still err on some specific entries and vector types. Malicious actors could use these fine-grained failures as inspiration for sexist online posts which bypass current detection systems, or in principal use the entries to train a generative model (concerns also shared by Vidgen et al. (2021b) and Kirk et al. (2022c)). We believe this risk is manageable given the benefit facilitated by our dataset in understanding and predicting online sexism.

Following Kirk et al. (2022a)'s advice, we describe the risks from our dataset construction and release. First, there is a risk of harm to data subjects (women targeted by sexist entries) and readers of the paper in reproducing or reinforcing harmful representations, stereotypes, prejudiced discussions or slur usage. We include a content warning on the first page and consistently display quoted examples in italic text (Tab. 1, §6.2). Where possible, we replace vowels in slurs with an asterisk. Due to the labelling intensiveness of this work (20k entries), there is a risk to annotators from repeatedly viewing sexist content, especially violent entries (like 'threats of harm'). We carefully follow protocols for supporting annotator well-being and keep a direct line of communication open to them via a group messaging forum. We survey annotators at the end of the labelling process to understand how the task affected their well-being and how we can do better in the future. Overall, annotators found we adequately supported their mental health and ensured the process was as safe as possible. A few annotators suggested we run a workshop presenting findings of our work. We commit to hosting this de-brief session in the coming months as it is important for annotators to see the significance of their work in supporting online safety.

References

Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. 2021. New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, 39(5):943–960. Publisher: SAGE Publications Inc.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2022. Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values. _eprint: 2210.07652v1.

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

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben

Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9). Publisher: National Acad Sciences.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. ArXiv:2110.05719 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. #NotOkay: Understanding Gender-Based Violence in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):52–61. Number: 1.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96. Association for Computing Machinery, New York, NY, USA.

E Fersini, P Rosso, and M Anzovino. 2018a. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 214–228.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018b. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *EVALITA Evaluation of NLP and Speech Tools for Italian*, pages 59–66. Accademia University Press.

Debbie Ging. 2017. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, 1:20.

Debbie Ging, Theodore Lynn, and Pierangelo Rosati. 2020. Neologising misogyny: Urban Dictionary's folksonomies of sexual abuse. *New Media & Society*, 22(5):838–856. Publisher: SAGE Publications.

Barney Glaser and Anselm Strauss. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, New York.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. ArXiv:2111.09543 [cs].

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ArXiv:2006.03654 [cs].

Greta Jasser, Jordan McSwiney, Ed Pertwee, and Savvas Zannettou. 2021. 'Welcome to #GabFam': Far-right virtual community on Gab. *New Media & Society*, page 146144482110245.

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, Vancouver, Canada. Association for Computational Linguistics.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Joshua B Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems*.

Callum Jones, Verity Trott, and Scott Wright. 2020. Sluts and soyboys: MGTOW and the production of misogynistic online harassment. *New media & society*, 22(10):1903–1921. Publisher: SAGE Publications Sage UK: London, England.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joseph Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wijaya, Xin Wang, Yong Zhang, and Morteza Dehghani. 2018. The Gab Hate Corpus: A Collection of 27k Posts Annotated for Hate Speech. *PsyArXiv*. Publisher: PsyArXiv.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022a. 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, Bertie Vidgen, and Scott Hale. 2022b. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022c. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

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.

Mary Lilly. 2016. *The World is Not a Safe Place for Men: The Representational Politics of the Manosphere*. University of Ottawa. Backup Publisher: University of Ottawa.

Lucas Lima, Julio C.S. Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522. ISSN: 2473-991X.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 118–126, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875. Number: 17.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing Sexism and Misogyny through Neural Approaches. *ACM Transactions on the Web*, 15(4):17:1–17:31.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. ArXiv:2110.05699 [cs].

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. 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 4755–4764. Association for Computational Linguistics. Event-place: Hong Kong, China.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Alison Ribeiro and Nádia Silva. 2019. INF-HatEval at SemEval-2019 Task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The Evolution of the Manosphere across the Web. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:196–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. 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. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:573–584.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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. ArXiv:1911.03891 [cs].

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. ArXiv:2111.07997 [cs].

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Jane M Ussher. 2016. Misogyny. *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pages 1–3. Publisher: Wiley Online Library.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data, a Systematic Review: Garbage in, Garbage Out. *PLOS ONE*, 15(12):e0243300. Publisher: Public Library of Science arXiv: 2004.01670.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

Papers), pages 1667–1682, Online. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Maximilian Wich, Edoardo Mosca, Adrian Gorniak, Johannes Hingerl, and Georg Groh. 2021. Explainable Abusive Language Classification Leveraging User and Network Data. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, Lecture Notes in Computer Science, pages 481–496, Cham. Springer International Publishing.

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.

Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1007–1014, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022a. Rethinking offensive text detection as a multi-hop reasoning problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022c. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. ArXiv:2209.07562 [cs].

Donna Zuckerberg. 2018. *Not All Dead White Men: Classics and Misogyny in the Digital Age*. Harvard University Press, Cambridge, Massachusetts.

Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pages 1–4, New York, NY, USA. Association for Computing Machinery.

A Data Statement

We provide a data statement (Bender and Friedman, 2018) to document the generation and provenance of our EDOS dataset.

A.1 Curation Rationale

The purpose of the EDOS dataset is to train and evaluate automated systems for the fine-grained and explainable detection of online sexism. We curate social media comments across diverse forms of sexist content – varying from explicitly threatening behaviour (e.g., "I'll kill any women that talks to me") through to more subtle forms (e.g., "Women are delicate flowers who need to be cherished").

A.2 Language Variety

All entries are in English, dictated by the expertise of researchers and annotators. We encourage future work to expand our vectors to other languages.

A.3 Speaker Demographics

All entries are collected from two social media platforms: Gab and Reddit. Thus, the speaker demographics approximate these platforms' users in general (Amaya et al., 2021). We expect (and intentionally sample) users who are likely male, western and right-leaning, and hold extreme or far-right views about women, gender issues and feminism (Zannettou et al., 2018; Lima et al., 2018).

A.4 Annotator Demographics

We recruited 19 annotators from our existing network of freelance annotators. They worked over two months in Spring 2022. Annotators were paid £16/hour and expert annotators were paid £20/hour. All annotators were screened on a gold standard set of 200 entries and had previous experience working on hate speech annotation tasks. Annotators could contact the research team at any point using a messaging platform. All annotators self-identified as women. We sent annotators an optional survey to collect further information on their demographics and to understand past experiences with social media and online sexism. Of 19 annotators, 12 responded to our survey. By age, 3 were between 18–24 years old, 8 were between 25–30 years old and 1 was between 31–40 years old. The completed education level was high school for 1 annotator, undergraduate degree for 6 annotators and post-graduate degree for 5 annotators. Annotators came from a variety of nationalities, with 7 British, as well as Swedish, Swiss, Italian and Argentinian. Most annotators identified as White with one annotator identifying as mixed South Asian and White, and one annotator identifying as Black British. The majority identified as heterosexual (7), with others identifying as bisexual, pansexual or asexual. 8 annotators were native English speakers and 4 were non-native but fluent. The majority of annotators used social media for one or more hours a day (7). All annotators had seen others targeted by online sexism, and 7 had been personally targeted.

We also collected annotator-specific attitudes to online sexism, for example (i) what types of online sexism they believe social media companies should prioritise in their content moderation policies, and (ii) whether content should be removed or partially hidden. We ask annotators about their content moderation preferences for content ascribing to each of our 11 vectors. In future work, we plan to map these preferences to our dataset entries to investigate whether attitudes influenced labelling behaviour.

A.5 Speech Situation

The modality of entries is short-form written textual comments from social media, which were interpreted in isolation for labelling i.e., not in a comment thread and without user or network information.

A.6 Text Characteristics

The genre of texts is sexist and non-sexist social media comments. The composition of the final dataset is described in Tab. 2.

B Review of Existing Taxonomies

We reviewed recent articles that propose taxonomies of sexist or misogynist content, in particular: Jha and Mamidi (2017); Samory et al. (2021); Farrell et al. (2019); Zeinert et al. (2021); Guest et al. (2021); Rodríguez-Sánchez et al. (2021); and Parikh et al. (2021). We describe here several key differences between prior work and how our taxonomy (in §2.2) compares:

Differences in Construction The taxonomies we reviewed are either theoretically- or empirically-grounded. Empirically-motivated studies mostly produce a first version of the taxonomies based on NLP literature and then iteratively adjust the taxonomy using new data they collect (Zeinert et al., 2021; Guest et al., 2021). Twitter is the most

widely used source (Jha and Mamidi, 2017; Fersini et al., 2018a; Samory et al., 2021; Zeinert et al., 2021; Rodríguez-Sánchez et al., 2021), followed by Reddit (Farrell et al., 2019; Guest et al., 2021). Data collection was mostly based on keywords, such as slurs and sexist hashtags, except for Farrell et al. (2019) who sampled more widely from the "manosphere" community and Parikh et al. (2021) who based their study on "The Everyday Sexism Project" – a catalogue of accounts of people who experienced sexism. Some studies also included the re-annotation of existing datasets (Jha and Mamidi, 2017; Samory et al., 2021; Guest et al., 2021) and addition of adversarial data (Samory et al., 2021). Our taxonomy is theory- and empirics-grounded – we first draw on previous theory-based taxonomies and then iterate with in-domain data.

Differences in Scope Existing taxonomies differ in scope, which is partly reflected in their focus on either sexism (Samory et al., 2021; Rodríguez-Sánchez et al., 2021; Parikh et al., 2021) or misogyny (Jha and Mamidi, 2017; Fersini et al., 2018a; Farrell et al., 2019; Guest et al., 2021; Zeinert et al., 2021). There are clear inconsistencies in how both of these terms are defined. Zeinert et al. (2021), for example, formally define misogyny as a type of group-directed abuse and hate speech. In practice, their account of misogyny is very similar to the fairly broad account of sexism that we have adopted. Our taxonomy clearly positions sexism as a type of abuse, with a broader scope than misogyny. Existing taxonomies often cover overlapping but different topics. They also differ in how they name and define more fine-grained types of sexism and misogyny – which we call vectors – if those are present at all. Jha and Mamidi (2017), for instance, use a specific theory on benevolent sexism from social psychology to motivate three main types of sexism: Paternalism, Gender Differentiation, and Heterosexuality. In contrast, Parikh et al. (2021) construct 23 vectors of sexism, based on the perspectives of those targeted by sexism. They distinguish, among other things, between role and attribute stereotyping, as well as work-, menstruation-, and motherhood-related gender discrimination. Some work uses descriptive names for vectors (e.g., "physical violence" (Farrell et al., 2019; Guest et al., 2021), while other names are based on social science theories or terms (e.g., "gaslighting" (Parikh et al., 2021), "belittling" (Farrell et al., 2019) and "neosexism" (Zeinert et al., 2021)). Our taxonomy consistently uses descriptive names, with a roughly equivalent degree of granularity within each level.

Differences in Structure Finally, existing taxonomies differ in how they organise vectors of sexist and misogynist content. Most commonly, all vectors are direct subtypes of misogyny or sexism (Fersini et al., 2018a; Zeinert et al., 2021; Farrell et al., 2019; Rodríguez-Sánchez et al., 2021; Parikh et al., 2021), without explicit relationships between them. Jha and Mamidi (2017) and Samory et al. (2021) include additional flags for benevolent/civil vs. hostile/uncivil expressions, and Fersini et al. (2018a) have flags for targets. Guest et al. (2021) provide a multi-level hierarchy of types and vectors of misogyny. Our taxonomy also takes a hierarchical approach, differentiating first between categories of sexism and then between fine-grained sexism vectors within each category.

C Additional Detail on Reddit Data

We identified a long list of 81 relevant subreddits by reviewing seven research papers and three established online lists. Some of the subreddits have been banned but data is still available for them, and they are included within our long list. The 81 subreddits were then assigned to one of four categories, based on the work of Lilly (2016). Lilly's work was also used in Ribeiro et al. (2021), who adapted it to create a five-part taxonomy of manosphere subreddits by separating 'Men's Rights Activists' (MRA) into MRA and 'The Red Pill' (TRP). We do not follow this distinction because conceptually the two categories are very similar and separating them could bias our data collection by, in effect, oversampling from these similar communities. To maximise the diversity of our data we include subreddits which are topically relevant to the four categories (listed below), even if they are nominally dominated by groups other than straight men (such as `r/gaycel` or `r/trufemcels`) or are primarily concerned with critiquing and debating the manosphere (such as `r/purplepilldebates` or `r/thebluepill`).

- **Incels (IC)** Men who are opposed to, and demean, insult or attack women, because they cannot, or believe they believe cannot, get sexual interest or companionship. Many such men believe that they are entitled to female attention.

- **Men Going Their Own Way (MGTOW)** Men who believe that women, and feminism, have corrupted society and that men need to reassert themselves. Many such men can, or believe that they can, achieve companionship and sexual interest from women and are less explicitly spiteful and hateful towards women, in contrast to Incels.

- **Men's Rights Activists (MRA)** Men who typically oppose feminism, believing that women are privileged and men are systematically discriminated against. MRA groups range from moderate positions, such as activists who want greater rights for divorced fathers, to more extreme positions, such as activists who want men to be given state-backed privileges.

- **Pick Up Artists (PUA)** Men who actively attempt to win companionship and sexual interest from women, often with duplicitous or underhand techniques. Many such men hold derogatory views about women, and portray them as sexual objects.

D Bespoke Sexism Classifier for Data Sampling

In our sampling ensemble, we trained a bespoke binary classifier on English and women-targeted entries from seven open source datasets (Zampieri et al., 2019; Rodríguez-Sánchez et al., 2021; Vidgen et al., 2021b; Samory et al., 2021; Guest et al., 2021; Fersini et al., 2018a,b). For each of these datasets, we remove duplicates, clean white space, and convert URLs, emoji and usernames to special tokens (e.g., [URL]). Each dataset is labelled according to its own taxonomy for sexism and/or misogyny, and their structure and definitions are not consistent. To unify the labels, we use the binary top-level label in each dataset (either misogynist vs. not misogynist; or sexist vs. not sexist). We split the combined datasets into train and test (90/10), stratifying by dataset source. We train a BERT-base model with default parameters for 3 epochs, using the transformers library (Wolf et al., 2020). The model achieves 80% Macro-F1 score on the held-out test set. We use this model to weakly label the presence of misogynist or sexist content in the Gab and Reddit data.

E Crowdworkers vs. Experts Experiments

We used a gold standard of 200 entries to test two alternatives for annotation. First, we launched the gold standard on a crowdsourced annotation service, with 7 annotations per entry, and 176 crowdworkers. Taking the majority-voted label for Task A, nearly all sexist entries were labelled correctly (96%) but most non-sexist entries were mislabelled (4% correct). Of the 57 Sexist entries, the crowd majority-vote was 28% correct for Task B and 14% for Task C. Second, we recruited and trained 19 self-identifying women to label the same 200 gold standard entries. Taking the majority-voted label for Task A, the trained annotators were correct 96% of the time. Of the 57 Sexist entries, the trained annotators' majority-vote was 84% correct for Task B and 72% for Task C. Given the sensitivity and complexity of the task, working with trained annotators provides higher-quality labels and poses other advantages that annotator welfare can be monitored and guidelines can be updated to address feedback.

F Confusion Between Vectors

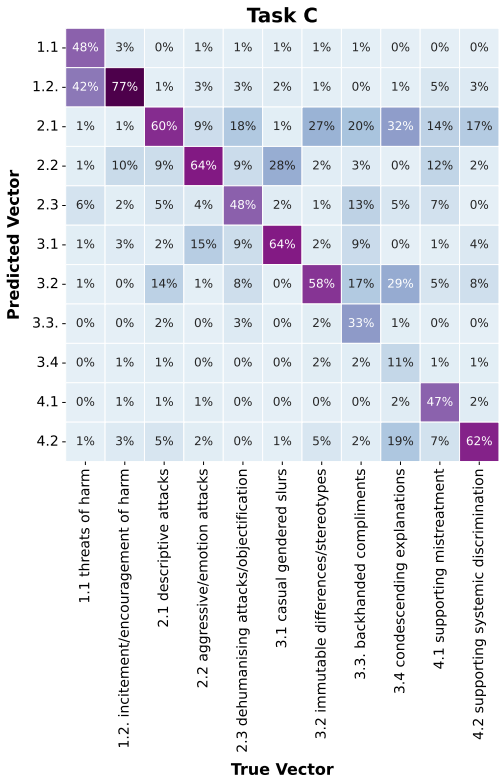

Figure 4: Confusion matrix for vectors of sexism (Task C) across the top 10 performing systems.