Biasly: An Expert-Annotated Dataset for Subtle Misogyny Detection and Mitigation

Brooklyn Sheppard[1,*], **Anna Richter**[1,*], **Allison Cohen**[1,†], **Elizabeth Allyn Smith**[2], **Tamara Kneese**[3], **Carolyne Pelletier**[4], **Ioana Baldini**[5], **Yue Dong**[6]

[1] Mila - Quebec AI Institute [2] Université du Québec à Montréal [3] Data & Society Research Institute [4] Reliant AI (work done while at Mantium) [5] IBM Research [6] University of California, Riverside

brooklyn.sheppard1@ucalgary.ca, anna.g.richter@gmail.com
allison.cohen@mila.quebec, smith.eallyn@uqam.ca, tkneese@datasociety.net
carolyne@reliant.ai, ioana@us.ibm.com, yue.dong@ucr.edu [*†]

Abstract

Using novel approaches to dataset development, the Biasly dataset captures the nuance and subtlety of misogyny in ways that are unique within the literature. Built in collaboration with multi-disciplinary experts and annotators themselves, the dataset contains annotations of movie subtitles, capturing colloquial expressions of misogyny in North American film. The open-source dataset can be used for a range of NLP tasks, including binary and multi-label classification, severity score regression, and text generation for rewrites. In this paper, we discuss the methodology used, analyze the annotations obtained, provide baselines for each task using common NLP algorithms, and furnish error analyses to give insight into model behaviour when fine-tuned on the Biasly dataset.

Content Warning: To illustrate examples from our dataset, misogynistic language is used which may be offensive or upsetting.

1 Introduction and Related Work

When using language models (LMs) to perform sensitive, subjective, and socially impactful tasks like misogyny detection, hate speech mitigation, or online content moderation, the quality of the underlying dataset is critical (Bender et al., 2021; Gebru et al., 2021; Hutchinson et al., 2021). Because the model will align to the biases in the dataset, which often reflect the biases, or oversights, of the dataset creators (Sap et al., 2019), it is crucial to include a diverse group of stakeholders in the dataset creation process, including LM domain experts and stakeholders who would be impacted by any deployed model that was trained on the dataset (Dignum, 2020; Abercrombie et al., 2023).

Dataset work in the field of bias, and more specifically sexism or misogyny detection, has mainly focused on the domains of social media, with data stemming from Twitter, Reddit or Gab (Fersini et al., 2018; Guest et al., 2021; Rodrıguez-Sánchez et al., 2022; Kirk et al., 2023) (see Table 1). Additionally, Ratnala (2022) provides a dataset for sexism detection using Reddit, however, sufficient detail to permit a comparison is not provided. While using this type of training data is valuable for detecting the often blatant misogyny appearing in social media forums, we contend that those data sources might not be ideal for detecting subtler forms of misogyny found in everyday spoken language, as they might overshadow the latter during the training process (Reif and Schwartz, 2023). Studies with movie or sitcom subtitles as training data may represent a better balance; in this domain, Singh et al. (2022) focuses on the elimination of all types of bias, and Singh et al. (2021) does not supply sufficient detail to permit comparison.

Even though most datasets provide a more fine-grained classification for different subtypes of misogyny (Fersini et al., 2018; Guest et al., 2021; Samory et al., 2021; Kirk et al., 2023), the detection of misogyny remains, at its core, a classification problem where a (sub)category can be either present or not. We argue that due to its nuanced and subjective nature, a continuous severity score modelled by regression is better suited for the detection of subtle misogyny. Only one misogyny-specific dataset includes a type of of misogyny mitigation (Samory et al., 2021). Their goal was to create adversarial examples that language models would find hard to differentiate from real sexist statements, by applying minimal lexical changes. Our work is methodologically closer to the ParaDetox (Logacheva et al., 2022b) and APPDIA (Atwell et al., 2022) datasets, which released a parallel corpus for detoxicification. To our knowledge, our dataset is the first parallel corpus with the purpose of training language models to rewrite text to mitigate the subtle misogyny contained therein.

* Equal contributing first authors, please cite as Sheppard & Richter et al. (2024)

† Corresponding author: allison.cohen@mila.quebec

Findings of the Association for Computational Linguistics: ACL 2024, pages 427–452

Dataset	Size	Classifi.	Severity	Mitigation	Annotators	StM	# Annot.	Source
EDOS (2023)	20,000	Y	-	-	Trained annotators	Y	19	Reddit, Gab
Guest (2021)	6,567	Y	-	-	Trained annotators	Y	6	Reddit
Ami (2018)	5,000	Y	-	-	Domain experts	Y	6	Twitter
Callme (2021)	13,631	Y	-	Y	Crowdworkers	Y	-	Twitter, Psych. Scales
EXIST (2022)	1,058	Y	-	-	Trained annotators	Y	6	Twitter, Gab
ParaDetox (2022b)	11,939	-	-	Y	Crowdworkers	N		Twitter, Reddit, Jigsaw
APPDIA (2022)	2,000	-	-	Y	Domain experts	N	-	Reddit
Biasly (ours)	10,000	Y	Y	Y	Domain experts	Y	10	Movie subtitles

Table 1: Comparison of misogyny detection and bias mitigation datasets. 'StM' denotes specificity to Misogyny; 'Classifi.' indicates support for classification tasks; '# Annot.' refers to the number of annotators. See Appendix A.3 for baseline results on the majority of these previous datasets.

In this work, we document the creation process of the Biasly dataset, an open-source expert-annotated dataset for the detection and mitigation of subtle forms of misogyny[1]. We first describe our process, including how our interdisciplinary team thoughtfully selected, trained, and engaged with our annotators, ensuring our dataset is both high quality and was created in a socially responsible way. Then, we present a short analysis of the annotated dataset and provide model baseline results for the tasks of binary and multi-label misogyny classification, severity prediction and mitigation. Finally, we provide an error analysis of each of our baseline models to provide insight into model behaviour when fine-tuned on the Biasly dataset.

2 Dataset Creation

Our team consists of experts from the domains most relevant to the development of a misogyny dataset; specifically, specialists from NLP, linguistics and gender studies. We engaged in a collaborative, multi-disciplinary process wherein our decisions were informed by qualitative and quantitative analyses of the data, described briefly in the following sections.

2.1 Dataset Selection and Preprocessing

Contemporary Movie Subtitles: Biasly's data is derived from a movie subtitle corpus available through *English-corpora.org*. The decision to use movie subtitles was motivated by: 1) the presence of both overt and subtle forms of misogyny in good proportion, and 2) its similarity to transcribed conversational speech. Because Twitter, Reddit, and Gab are known to offer an abundance of overt misogyny, it was a concern that these more overt forms would predominate and drown out the effect of subtle examples (Reif and Schwartz, 2023). We sought to complement existing efforts that focus on written language with an analysis of spoken language because differences in communication type lead to differences in misogynistic expression. Though scripts are written and not naturalistic speech, screenwriters try to create fluid verbal interactions, which are then spoken by actors. As such, the subtitles from the films approximate oral communication. Business e-mail corpora were rejected for a lack of misogynistic language in sufficient quantity for analysis. Movie subtitles, however, offered both overt and subtle forms of misogyny in the necessary quantity and proportion. While movie scripts themselves might seem preferable to automatically generated subtitles, this dataset was the only one we found with sufficient quantity for our task. Furthermore, the use of movie dialogue data aligned with existing efforts to address sexism in cinema (the Bechdel test, Geena Davis Institute analyzing time on screen, percentage of lines for female actresses, etc.).

Data Pre-Processing: Given how significantly language evolves over time (Juola, 2003), and how differently some items would be judged in one context versus another (e.g. *lil darlin'*), we restricted our sample to movies released in the last 10 years. We filtered out films that, while contemporary productions, were clearly set in the past (e.g. westerns, period pieces) or otherwise did not reflect contemporary colloquial speech (e.g. documentaries). Similarly, we reduced the sample to films that were American releases, given differences across global varieties of English (Major et al., 2005). We also removed movies for which the subtitles were entirely upper- or lowercase to acknowledge the differences in meaning that this changing case produced (i.e. *Black woman* versus *black woman*, *Karen* versus *karen*, *bitch* versus *BITCH*). Furthermore, we fil-

[1]The full dataset can be accessed through our Hugging Face repository.

tered out explicitly-indicated speaker changes since this variable's inclusion was not constant across subtitles and would have affected the consistency of annotators' assumptions about the speakers and their intentions. Finally, we parsed the data into non-overlapping chunks of three sentences each, subsequently referenced as "datapoints," using the Stanza tokenizer (Qi et al., 2020).

Data Filtration Approach: In order to identify as much misogyny as possible without biasing the dataset with terms that were already potentially misogynistic on their own (e.g. *bitch* or feminine-specific job titles), we further filtered the data as follows: 20% of our datapoints contain the keyword *she*, 20% *her*, 10% *herself*, 10% *women*, and 10% *woman*. The remaining 30% were sampled randomly. This data split roughly respects the bias of "natural" occurrences of these keywords in the dataset (*she* and *her* are used twice as often as the other keywords, reflecting their relative frequency in the overall corpus). Though these keywords may bias the dataset towards instances of misogyny expressed in the third person, second person references such as *you* are not gendered and may not have yielded as many misogynistic datapoints. Instead, in order to capture directly-addressed misogyny as well as other types, we decided that 30% of the dataset would consist of random samples.

2.2 Engaging Expert Annotators

When annotating for misogyny, a nuanced and political task, we wanted to ensure that the interpretation of each datapoint was grounded in expertise. As such, we hired annotators pursuing or having completed their post-secondary degrees in linguistics, gender studies, or both, and compensated them at a rate of $25 CAD per hour. We did not place other demographic limits on recruitment, and our annotators included a range of gender and sexual identities, races, ethnicities, and language backgrounds, though all were located in North America and were fluent in English (see Appendix A.1).

The resulting team of annotators included 5 gender studies and 5 linguistics experts (although expertise between the two groups overlapped), with only one identifying as male. To check for inter-annotator agreement (and ensure quality control), three annotators were assigned to each datapoint. Gender studies and linguistics annotators were intentionally assigned at a 2:1 ratio to each datapoint to ensure a diversity of academic backgrounds were included in each annotation.[2]

The tasks and annotation guidelines described in the next section were conceived by the interdisciplinary team and refined with input from annotators in an iterative, collaborative manner. Annotators stress tested the initial version of the annotation tasks without strict prescriptive direction during workshops and pilot rounds (Röttger et al., 2022). Subsequently, we sought feedback through moderated discussions with the team's gender studies and linguistics experts, who crystallized our approach with prescriptive guidelines led by the annotators' comments. This grounded theory approach informed elements like our misogynistic inference categories, interpretations of severity, and appropriate rewrites (Locke, 2002). When devising the list of misogynistic inference categories, for example, no categories were provided in a pilot round; annotators were asked to consider which misogynistic beliefs were being expressed in the data provided. They brought their observations to a workshop guided by our experts who finalized the category list in light of those group discussions.

Finally, it is worth noting that our team was in close contact with the annotators throughout the process, hosting regular office hours and remaining available over Slack and email. We shared with the annotators the goal for the project and how their labeled data would be used. Annotators connected amongst themselves via Slack, enabling them to discuss strategies for confusing or complex datapoints. We allowed space for real differences of opinion and did not require consensus. This close contact also allowed us to check in with the annotators about the potentially harmful impacts of working with misogynistic texts, which they reported being able to manage well.

2.3 Annotation Tasks/Taxonomy

All annotation tasks discussed in this section are summarized in Figure 1.

Task 1: Annotators were asked to conduct a binary classification (yes/no) of whether the datapoint presented contained misogyny anywhere within it. The annotators referenced the following definition of misogyny: "Hatred of, dislike of, contempt for, or ingrained prejudice against women." Misogyny may be directed at a group or an individual, but it is easier to detect in gener-

[2]Half of the datapoints were labeled by 2 linguists and 1 gender studies expert, while the other half were labeled by 1 linguist and 2 gender studies experts.

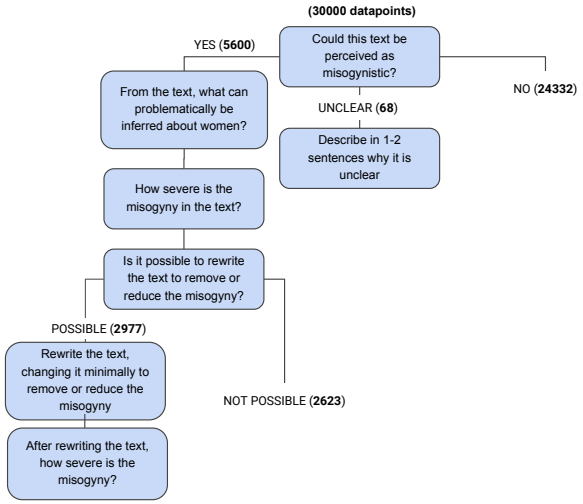


Figure 1: Summary of the annotation tasks for the Biasly dataset. Bracketed numbers in bold represent the number of resulting datapoints for each task.

alizations about groups. For individuals, there is additional verification necessary to be sure that the negative sentiment is, at least in part, associated with the individual by virtue of being a woman. We included reclaimed language, slurs, and potentially humorous utterances as misogynistic so as not to risk creating noise with a seemingly inconsistent dataset. While we recognize that some such speech does indeed have both misogynistic and non-misogynistic uses and that this would be an interesting subject of further study, our goal for this initial step was to identify what *could be* misogynistic and not just what was definitely misogynistic in context, especially since our context was limited to three-sentence windows. Through consultation with annotators, it emerged that when misogynistic slurs were being described rather than used to exhibit misogynistic sentiments (related to the use/mention distinction in linguistics), they did not find the descriptions to be misogynistic. As a result, while there may still be uses of other slurs where even mentioning them does evoke negative and potentially harmful sentiments (cf. Davis and McCready (2020)), for this corpus, we prescribed annotating descriptions as non-misogynistic and uses as misogynistic.

Task 2: Once annotators identified a datapoint as containing misogyny, they were asked to classify the type(s) of misogyny being exhibited in the datapoint from a provided list that they were involved in creating. New categories were devised, ones that uniquely fit our dataset (i.e. gender essentialism and stereotypes). The full list with an explanation of each, is provided in Table 2.

Task 3: In addition to categorizing misogynistic datapoints, annotators were asked to indicate the datapoint's severity on a continuous scale. The continuous scale (rather than ordinal) was intentionally chosen to acknowledge the impossibility of ascribing one definitive number to the severity of a misogynistic statement and to avoid the pitfall of using a discrete metric for a potentially continuous variable (Matejka et al., 2016). The continuous scale allows for a more genuine reflection of human interpretations of misogyny. While annotators only saw the continuous scale with the endpoints of no misogyny/maximum misogyny, the back end was mapped to values between 0 and 1000.

Task 4: Last, annotators were asked, when handling misogynistic datapoints, whether it was possible to remove or reduce the misogynistic inference(s) by rewriting portions of the text while largely retaining the original meaning of the utterance(s). However, the feasibility of this task depended on whether a misogynistic inference was primary (the main point of the utterance) or secondary (e.g. an implicature). When it is primary, the rewrite task is likely impossible in the sense that annotators could not remove the misogyny without losing the core of the original sentence meaning. When the misogynistic inference is secondary, the rewrite was more likely if there was a way to retain the primary intent of the speaker while removing the misogynistic inference.

3 Dataset Analysis

Our resulting dataset consists of 10000 datapoints, each annotated 3 times, for a total set of 30000 annotations. 5600 of the 30000 annotations were labeled as misogynistic according to the first binary classification task (Task 1), with an inter-annotator agreement of 0.4722 according to Fleiss' kappa (Fleiss, 1971). The severity of the original misogynistic datapoints (from 0 to 1000) has a mean of 344.8 with a standard deviation of 209.1, while the severity of all rewritten datapoints has a mean of 53.6 and a standard deviation of 115.8, reflecting a significant reduction in misogyny severity. The most frequent sub-category of misogyny was *Trivialization* with 2227 occurrences, while *Transmisogyny* only appeared 43 times. 1985 misogynistic datapoints were selected to be rewritten to

Name	Description
Anti-feminism	Feminism is a bad idea, feminists are gross and ugly, women shouldn't have equal rights
Dehumanization	Comparing women to animals or objects
Domestic violence and other violence against women	(self-explanatory)
Gender essentialism or stereotypes	Can be both positive, e.g. women are good at childrearing and cooking because they are more nurturing, and negative, e.g. women are untrustworthy and overly emotional because of their hormonal cycles
Gendered slurs	*Chick*, *b*tch*, *c*nt*, etc.
Intersectional, identity-based misogyny	Any other instance of misogyny that is related to race, ethnicity, religion, class, occupation, immigration status, disability, size, etc.
Lacking autonomy or agency	Women are not able to make decisions or must defer to male authorities
Phallocentrism	Focus on penis in organization of social world
Rape and other forms of sexual violence	(self-explanatory)
Sexualization	Outsized focus on appearance, degrading language
Transmisogyny/ Homophobia	Includes mocking individuals or groups for gender nonconformity, e.g. for dressing or acting in a way that does not conform with assumed gender roles; homophobia/transphobia that also contains misogynistic inferences
Trivialization	Infantilizing or paternalistic language, women are not taken seriously

Table 2: Subcategories of misogyny with a short explanation

mitigate the misogyny by one or more annotators, yielding a total of 2977 rewrites.

In order to perform binary classification of misogyny with the annotated dataset, we needed to relate each datapoint to only one label. Since the focus of our dataset is to identify subtle forms of misogyny, we aggregated the binary classifications from all annotators into a single label by deeming the datapoint misogynistic as soon as one of the three annotators labeled it so. This way, we ensure that we are capturing even the subtlest forms of misogyny and prevent overriding minority voices with a 'majority rules' approach. Using this methodology, our dataset contains 3159 misogynistic datapoints, which gives a distribution of 31.59% positive cases and 68.41% negative cases, a more balanced distribution than much previous work in the field has achieved (Guest et al., 2021; Samory et al., 2021; Kirk et al., 2023).

Table 3 provides three examples with differing levels of severity of misogyny. In each case, all three annotators agreed that the datapoint was misogynistic. (1) is an example of trivialization via infantilizing or paternalistic language (referring to a woman as a girl). (2) relies on gender essentialism or stereotypes (that women aren't complete without romantic attachments), and, as with many examples, is intersectional, combining misogyny with aspects of homophobia (or, in other cases, racism, ableism, etc.). Finally, (3) is dehumanizing in its comparison of a woman to an object. Some examples were deemed to be impossible to rewrite, as in (2) where all 3 annotators agreed that mitigation was not possible. Some were possible to rewrite with total mitigation (removal) of the misogynistic inference, as in (1), where the 3 annotators all provided the same rewrite, simply eliminating the problematic item with no significant effect on the dialogue. Lastly, we have examples like (3) where rewriting is possible and can mitigate but not eliminate the misogyny.

4 Automatic Classification, Regression, and Mitigation Experiments

This section describes the experimental setup for machine learning models on our dataset and presents the results. To follow best practice from other work cited below, we provide baseline results for the machine learning tasks of binary and multi-label classification, severity regression, and mitigation. For all models, we used an 80/10/10 train/eval/test split.

#	Sev.	datapoint	Category	Misogyny Mitigation
1	Low	She needs my support. Girl could you give us a second? Really?	Trivialization	She needs my support. Could you give us a second? Really?
2	Mid	I think it's about time that Emanuel had a nice fellow in her life. Why? Were you starting to think that I was a lesbian?	Stereotype	NA
3	High	We passed her mama around like a baton, man. Yeah. You never told me that about your mother.	Dehumanization	We all slept with her mama, man. Yeah. You never told me that about your mother.

Table 3: Example Annotations from the Biasly Dataset.

4.1 Experimental Setup

Binary Classification: For our binary classification experiments, we used four models and report the F1 scores: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DeBERTa v3 (He et al., 2021), and ELECTRA (Clark et al., 2020). In all four cases, we used the base version with a maximum input sequence length of 512, batch size of 32, a learning rate of 2e-5, and 3 epochs for training.

Multi-Label Classification: For multi-label classification, containing 12 classes, we use the union of classes assigned across annotators as the "gold-standard" label for each datapoint. As the binary classification model is already trained to filter out non-misogynistic datapoints and the task of the multi-label model is to predict the type of misogyny present, we only use datapoints that were labeled by at least one annotator as misogynistic. BERT and RoBERTa were configured to train for 15 epochs with batch sizes of 16 for training and 64 for evaluation. For a gradual learning rate increase, warmup steps were set at 50 and, to prevent overfitting, weight decay was applied at 0.01. Evaluations on the validation set were conducted every 50 steps, and the best checkpoint based on validation performance was used for reporting test set performance. We train 15 epochs because, with more classes, it is harder for the model to converge.

Severity: We fine-tuned a BERT regression model to predict the misogyny severity scores (Task 3) in a supervised manner adapting a script from Jiang (2022). Following Samory et al. (2021), we also report the (unsupervised) Perspective API (Lees et al., 2022) toxicity scores for our data. For the regression experiment, as well as to compare the severity to the Perspective API toxicity scores, the original severity values were transformed from a range of [0,1000] to [0,1]. Again, we only used datapoints that were labeled as misogynistic by at least one annotator, this time using the average across the severity scores of all annotators who labeled it as misogynistic as a "gold-standard" label. We fine-tuned a BERT (bert-base-uncased) model for linear regression over three epochs, with a learning rate of 2e-5, a weight decay of 0.1, and a per-device train batch size of 64.

Mitigation: For misogyny mitigation (Task 4), we used each individual rewrite as a datapoint, resulting in a parallel corpus where one original datapoint can have between one and three rewrites mitigating its misogyny. We fine-tuned three baseline models: BART (Lewis et al., 2020), FLAN-T5 (Chung et al., 2022), and Alpaca-LoRA (Wang, 2023). Following the methodology outlined in the ParaDetox paper, all our experiments across various models adhered to specific hyperparameters, including a learning rate of 3e-5, a total of 100 training epochs, and a gradient accumulation step of 1. We employed the base version of each model. We conducted evaluations after each training epoch and selected the checkpoint with the lowest loss on the evaluation set for subsequent prediction tasks.

4.2 Results and Error Analysis

The team's linguistics expert performed an error analysis for each task, assessing true and false positives and negatives to provide insight into model performance when fine-tuned on our dataset.

4.2.1 Binary Classification

Following Fersini et al. (2018); Guest et al. (2021); Samory et al. (2021); Kirk et al. (2023), we evaluate model performance on the binary classification task using macro-F1 score to account for the class imbalance between misogynistic/non-misogynistic datapoints. We provide test set results of the four models BERT, DeBERTa, ELECTRA, and RoBERTa, all fine-tuned on our dataset and averaged across three random seeds in Table 4. Performance of those models compared across four other misogyny/sexism datasets can be found in Appendix A.3. DeBERTa v3 performs best on our dataset with an average F1 score of 0.807. Thus, we used the best

model	Accuracy↑	F1_macro↑	Precision_yes↑	Recall_yes↑	F1_yes↑	Precision_no↑	Recall_no↑	F1_no↑
BERT	0.813	0.781	0.711	0.686	0.698	0.857	0.871	0.864
DeBERTa v3	**0.834**	**0.807**	0.744	**0.725**	**0.734**	**0.874**	0.885	**0.879**
ELECTRA	0.831	0.801	**0.748**	0.700	0.723	0.866	**0.891**	0.878
RoBERTa	0.828	0.799	0.739	0.707	0.722	0.867	0.885	0.876

Table 4: Test results of binary classification models averaged over three runs with different random seeds.

performing run of DeBERTa v3 to analyse model behaviour with an in-depth error analysis.

TARGET / OUTPUT	Misogynist	Non-misogynist	SUM
Misogynist	231 23.10%	80 8.00%	311 74.28% 25.72%
Non-misogynist	85 8.50%	604 60.40%	689 87.66% 12.34%
SUM	316 73.10% 26.90%	684 88.30% 11.70%	835 / 1000 83.50% 16.50%

Figure 2: Confusion matrix for DeBERTa model performance on our test set for binary classification.

As can be seen in Figure 2, of the 165 datapoints incorrectly classified, there were nearly equal numbers of false positives (80) and false negatives (85). This suggests that the model performs well overall but has roughly even difficulty across the classes (rather than disproportionately outputting false negatives due to the larger number of non-misogynistic datapoints in the dataset).

A qualitative analysis of false positives reveals challenges for the model, including: i) failure to distinguish between women and female animals; ii) associating the term *girl* with misogyny (annotators only flagged the term as misogynistic when it was being used to describe an adult), and iii) flagging datapoints with general violence, similar to violence in the true positives but either not oriented to women, or being mentioned in order to criticize it. Labeled examples of these and other datapoints can be found in Figure 5 in Appendix A.4.

There were 85 false negatives, each annotated 3 times, yielding 255 annotations. While the model provided one output in its classification, there was variation among the annotators on each datapoint. There were many (151 of 255) false negatives for which at least one annotator agreed with the model that the datapoint was not misogynistic. Thus, of the 85 datapoints that were incorrectly classified, only one was rated as misogynistic by all three annotators and it was for an instance that required sophisticated reasoning to justify the decision ('giraffe legs' example in Figure 5). There were 14 datapoints for which two annotators disagreed with the model, 1 contained an unclear rating, and the vast majority (69) were datapoints for which only one annotator disagreed with the model. In the case of true positives, we see a number of examples with only one positive rating from the annotators, but in lower proportion. 224/693 (32%) total annotations in that category were negative as compared to 151/255 (59%) for the false negatives. This suggests that we would see even better model performance if we had chosen to use a majority rules system with the annotations, and it also suggests that the model struggles with representing minority opinions, motivating our interest in modelling individual annotators in future work.

4.2.2 Multi-Label Classification

Table 5 presents the ML baseline results on multi-label classification. An error analysis of the multi-label results on RoBERTa shows that the model was unsuccessful in correctly labeling any datapoints containing more than 3 categories. That is, of the 102 datapoints labeled perfectly by the model, 76 contained only 1 category, 24 contained 2 categories, and only 2 datapoints with 3 categories were correctly labeled by the model (see Figure 6 in Appendix A.4). Additionally, 11/102 perfectly labeled datapoints contained the slur *bitch*, while only 2/213 incorrectly labeled datapoints contained this slur. Thus, despite the subtlety of the misogyny contained within our dataset, the multi-label classification model may still rely on the presence of overt slurs to correctly classify the more fine-grained categories of misogynistic inferences.

Model	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
BERT	0.330	0.578	0.386	0.440
RoBERTa	**0.349**	**0.583**	**0.406**	**0.465**

Table 5: Macro-average results of models on multi-label classification with 12 classes.

4.2.3 Severity

We provide severity performance results on the test set for the fine-tuned BERT model and also use regression metrics to compare the Perspective API toxicity scores to the severity scores of our annotators in Table 6.

	mse↓	rmse↓	mae↓	r2↑
perspective_toxicity	0.083	0.288	0.219	-1.215
BERT_test	**0.031**	**0.176**	**0.139**	**0.175**

Table 6: Test results of supervised (averaged across three random seeds) and unsupervised toxic regression models on misogyny regression.

We can see that BERT trained with supervised learning performs better for predicting the level of misogyny as compared to the Perspective AI toxicity score. For BERT, the results were averaged across three runs with different random seeds.

Analyzing severity errors shows that 450/586 are within the confidence interval of +/-0.2, 536/586 within +/-0.3, and 571/586 within +/-0.4. In other words, there are 136 errors for which the model was off by more than 20%, 50 for which it was off by more than 30%, and only 15 errors remaining outside of the 40% range. For each confidence interval, there were fewer overestimations than underestimations, but the distinction was starker with each increase in the confidence interval. These are the respective ratios of overestimations to total errors: 65/136, 14/50, and 2/15.

From the examples in Figure 7 in Appendix A.4 that show the most extreme mispredictions of the model in each direction, it is tempting to think that sentiment analysis is playing a role in biasing the model in cases where we have a datapoint with many positive words but that suggests strong patronizing and controlling of women, or many words suggesting violence, but not towards women. Nevertheless, for severity that is off by 10% as compared to 20%, the differences can be very subtle, and a more in-depth qualitative analysis is left for future work.

4.2.4 Mitigation

We report metrics for the rewrite task in Table 7. The BLEU metric compares the model output to the human generated rewrites via weighted n-gram overlap. We evaluate Content Preservation (SIM) using the cosine similarity between the embeddings of the original text and the output, computed utilizing the model described in Wieting et al. (2019). The Style Accuracy (STA) metric represents the percentage of non-toxic outputs as identified by a style classifier, as detailed in Logacheva et al. (2022a). We use Perspective API to obtain the toxicity scores and compare them to those of the inputs and references. We choose to present the overall toxicity score rather than any of its component categories (insults, threats, sexually-explicit content, etc.) because they all contain gender-based toxicity and none are exclusive to misogyny.

Upon comparing the results across various parallel corpora for toxicity mitigation, all models attain high BLEU and SIM scores on the Biasly dataset, likely due to our annotators' being instructed to alter the text as minimally as possible to make the necessary change. Another notable difference is the small reduction in toxicity scores compared to the inputs. This is a result of the toxicity of the inputs of the Biasly dataset being much smaller than either the ParaDetox or Appdia datasets, reflecting the subtlety of the misogyny present within the Biasly dataset.

Given the challenge in finding effective metrics for evaluating model-generated rewrites, a qualitative analysis comparing the Alpaca-LoRA model's techniques in mitigating misogyny to those of annotators' rewrites was essential. Overall, it looks as though the model's rewrites are promising in many types of cases, producing a rewrite identical to that of an annotator 23% of the time.

Annotators themselves usually chose the same parts of sentences to rewrite as compared to other annotators, but they occasionally used different strategies for the same datapoint, perhaps changing a generalization about women to 'some women' or 'people' depending on whether they thought the context could be expanded to describe those not identifying as women as well. The model performs well in these cases, rewriting most generalizations so as to limit the generalization or to generalize about a larger group (i.e. all humans). The most common strategies for rewriting by both annotators and the model were this type of domain restriction or enlargement and the substitution or the deletion of words.

The examples in Figure 3 give a flavor for the rewrites themselves and the kinds of similarities and differences we see between the model and annotators for one of these categories, that of substitutions. In one case, the model matches the annotator's rewrite exactly, and in another, they

Dataset	Model	BLEU↑	SIM↑	STA↑	Toxicity	Tox-Inp	Tox-Refs
ParaDetox	BART	**56.00**	0.87	0.86	18.57	-55.61	4.22
	FLAN-T5	53.43	0.87	**0.88**	**17.53**	**-56.65**	**3.18**
	Alpaca-LoRA	55.97	**0.89**	0.80	22.08	-52.10	7.74
Appdia	BART	58.80	**0.92**	0.58	41.12	-25.08	18.75
	FLAN-T5	57.21	0.87	0.72	33.69	-32.50	11.33
	Alpaca-LoRA	**60.48**	0.83	**0.73**	**22.08**	**-52.10**	**7.73**
Ours	BART	85.09	**0.97**	0.74	28.51	-3.55	4.69
	FLAN-T5	85.97	**0.97**	0.75	28.33	-3.73	4.51
	Alpaca-LoRA	**86.51**	0.94	**0.80**	**25.23**	**-6.83**	**1.41**

Table 7: Test results of different text generation models for misogyny mitigation. The mean toxicity scores from Perspective AI for the inputs and references of each of the datasets is as follows: ParaDetox Inputs - 74.18, References - 14.35; Appdia Inputs - 66.20, References - 22.37; Ours Inputs - 32.06, References - 23.82.

Mitigation				
Reference Number	**Text**	**Model Output**	**Annotator Rewrite**	**Analysis**
13	Eh, she seemed like a girl who can take care of herself. Look, we can chase this Possum character tomorrow if we get cracking on these obits tonight. Now, Walker says, if we hand in three tonight, we might have a crack at employees of the month.	Eh, she seemed like someone who can take care of herself. Look, we can chase this Possum character tomorrow if we get cracking on these obits tonight. Now, Walker says, if we hand in three tonight, we might have a crack at employees of the month.	Eh, she seemed like someone who can take care of herself. Look, we can chase this Possum character tomorrow if we get cracking on these obits tonight. Now, Walker says, if we hand in three tonight, we might have a crack at employees of the month.	Annotator and model both substitute the same word
14	Thank you. She is even more beautiful than I ever could have imagined. Thank you.	Thank you. She is even more wonderful than I ever could have imagined. Thank you.	Thank you. She is even more amazing than I ever could have imagined. Thank you.	Annotator and model both substitute, but not the same word (though similar)
15	I can't stand her. Well, then get rid of her. If I were in a position to choose, I'd find someone nice.	I can't stand her. Well, then get rid of her. If I were in a position to choose, I'd choose someone nice.	I can't stand her. Well, then don't be with her. If I were in a position to choose, I'd find someone nice.	Annotator substitutes but model produces no change
16	Drinking, drugs, indecent exposure. What I find is a house full of drunkards, prostitutes, and imbeciles... Hey, hey, hey, hey!	Drinking, drugs, indecent exposure. What I find is a house full of drunkards, prostitutes, and idiots... Hey, hey, hey, hey!	Drinking, drugs, indecent exposure. What I find is a house full of drunkards, sex workers, and imbeciles... Hey, hey, hey, hey!	Annotator and model both make substitutions, but not targeting the same thing

Figure 3: Example datapoints from the mitigation task.

both change 'beautiful', but to 'wonderful' and 'amazing', respectively. We would argue that these should both be treated as successful rewrites given that different annotators might have also chosen different near-synonyms that did not reference the appearance of the woman in question. The subsequent two rewrites are less successful. The model sometimes produces a 'rewrite' that is an exact copy of the original datapoint, which is obviously not useful, and it occasionally reduces overall toxicity but misses the misogyny, as in a datapoint where 'imbecile' is changed to 'idiot' but 'prostitute' is left as is, while the annotator changed it to 'sex worker'. The same kinds of trends are found in cases with deletion, and in all cases, there are many more successful than unsuccessful rewrites given our criteria.

Finally, it is worth noting that the model is more successful at things like changing 'girl' to 'woman' or deleting unnecessary diminutives ('thanks sweetie' instead of simply 'thanks') that are either frequent or clearly related to a word that refers to a woman. It is less successful at targeting subtle verb differences, such as 'getting her' to do something versus 'asking her' to do it that can make a big difference to categories such as autonomy.

5 Conclusion

Developed in collaboration among experts in gender studies, linguistics, and NLP, we present Biasly, an open-source dataset for the detection, categorization, severity prediction, and mitigation of subtle misogyny. We provide baseline models for each of these tasks and an error analysis of each. In future work, we aim to employ more advanced modeling techniques, such as those used in Mostafazadeh Davani et al. (2022) to be able to model diverse annotator perspectives, moving beyond our current reliance on a single label. Our hope is that Biasly serves as a model for socially responsible dataset creation for LMs. This process can be readily applied to diverse domains, fostering a broader commitment to responsible AI development.

6 Limitations

Machine learning models trained on our dataset will contain biases associated with: i) the annotators' demographic and academic backgrounds as well as their lived experiences; ii) the subjective, context-dependent and time-bound nature of the task; and, iii) challenges for our annotators in labeling data consistently across time (the task can be performed differently depending on a number of contextual factors that cannot necessarily be controlled for; influencing, among other things, annotators notions of misogyny, severity etc.). Users of this dataset should also be aware that any automatic tools trained on this dataset can never guarantee perfect debiasing. Additionally, the use of movie subtitles as data may be limiting in that this is merely an approximation of natural spoken language. It is unclear, then, how well models trained on this dataset will perform on out-of-domain data such as social media content moderation or misogyny detection and/or mitigation in naturally spoken or written language. While we will do our best to provide insight into the extent to which each of these factors influenced our dataset (through accompanying documentation), we hope that those using the dataset will keep these limitations in mind and not use models fine-tuned on our dataset in ways that fail to acknowledge these limitations.

7 Social Impacts Statement

While we see many beneficial applications of this dataset, namely in building future applications designed to educate the public about misogyny, how it is expressed, and ways it can be removed or minimized, this dataset also presents the risk of being used for nefarious purposes. Specifically, malicious actors could use the dataset to create content that evades traditional toxicity detection models by rendering the misogynistic text more subtle. Furthermore, one could leverage this model to introduce subtle bias into otherwise non-misogynistic statements. This is one of the reasons we're looking to support the development of tools for more robust detection of misogyny, which can identify misogyny in subtle forms as well as overt. In other words, part of our desire to contribute to the domain of subtle misogyny detection is so that this type of misogyny doesn't continue to go unnoticed by traditional toxicity detection tools.

From a development standpoint, the risks centered mostly around our annotators, specifically in terms of their repeat exposure to misogynistic content, particularly datapoints which mentioned violence or suicide. In order to protect our annotators as much as possible, we shared mental health resources accessible through their respective universities, conducted mental health check-ins through surveys, and provided an opportunity to meet with members of our team to discuss the impact any of the work was having on their mental health. Furthermore, our team was easily accessible through platforms that allowed for direct communication. Overall, 70 percent of our annotators said they were fairly comfortable with the task in the context of our project (four people ranked their comfort at 4 out of 5 and three people ranked their comfort at 5 out of 5 in a survey). We made sure to address the feedback we had received in the free text portions of our survey to accommodate the needs expressed (i.e. offering to meet with annotators one-on-one to discuss material they find distressing). We'd like to continue treating annotators as key team members in the project and plan on hosting information sessions to share the impact of their contributions.

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 61st Annual Meeting Of The Association For Computational Linguistics*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. *arXiv preprint arXiv:2209.08207*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Christopher Davis and Elin McCready. 2020. The instability of slurs. *Grazer Philosophische Studien*, 97(1):63–85.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Virginia Dignum. 2020. Ai is multidisciplinary. *AI Matters*, 5(4):18–21.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

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, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Jinhang Jiang. 2022. Linear regression with hugging face. `https://github.com/jinhangjiang/Medium_Demo/tree/main/Transformers_Linear_Regression`.

Patrick Juola. 2003. The time course of language change. *Computers and the Humanities*, 37:77–96.

Hannah Rose 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)*. Association for Computational Linguistics.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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. *CoRR*, abs/1907.11692.

Karen Locke. 2002. The grounded theory approach to qualitative research.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022a. A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 90–101, Dublin, Ireland. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022b. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Roy C Major, Susan M Fitzmaurice, Ferenc Bunta, and Chandrika Balasubramanian. 2005. Testing the effects of regional, ethnic, and international dialects of english on listening comprehension. *Language learning*, 55(1):37–69.

Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5421–5432.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Sowmya Ratnala. 2022. sexismreddit.

Yuval Reif and Roy Schwartz. 2023. Fighting bias with bias: Promoting model robustness by amplifying dataset biases. *arXiv preprint arXiv:2305.18917*.

Francisco Rodrıguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, Marıa Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69:229–240.

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.

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 *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi Sultan, Roshni Ramnani, Anutosh Maitra, et al. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5274–5285.

Smriti Singh, Tanvi Anand, Arijit Chowdhury, and Zeerak Waseem. 2021. "hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms. pages 180–185.

E. J. Wang. 2023. Alpaca-lora. GitHub repository: https://github.com/tloen/alpaca-lora.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Annotator Demographics

Annotators' backgrounds shaped the research process at every step, from the creation of inference categories and the annotations themselves. For this reason, we anonymously collected detailed demographic information from 9 annotators willing to share it, which is summarized below.

Annotators identify with a range of genders and sexual orientations. A majority of annotators identify as queer, bisexual, or gay/homsexual/lesbian (7/9). Several annotators (4/9) identify as genderqueer, trans or nonbinary. One annotator identified as male, and 4 identified as female.

A majority of annotators identified as white (6/9). Two annotators identified as mixed race and one annotator identified as Latino. Half of the annotators (5/9) were from Canada, two were from the United States, one was from Mexico, and one identified as Portuguese/Haitian/Canadian/American.

Annotators' spoken languages may have impacted their interpretations of misogyny. Many annotators were multilingual, and were native or fluent in French and/or Spanish in addition to English (only 1 identified as a non-native English speaker). Annotators indicated that their religious background may have influenced their annotations. Annotators largely self-identified as atheist or agnostic, with cultural backgrounds in Christianity or Catholicism. Two annotators identified as Jewish and one identified as Muslim. Two annotators marked their religion as not applicable or none while three were atheists.

Annotators identified their education levels. Half had already completed a B.A. (5/9), two were pursuing a B.A., one had a Ph.D., and one was pursuing a Ph.D. Annotators hailed from a variety of class backgrounds, including 4/9 identifying as either working or middle class and 5/9 identifying as upper middle class.

Annotators ranged in age from 22 (5/9) to 34 (2/9). One annotator was 28 and one was 30. Finally, other demographic factors may have impacted their annotation decisions, including past experience working in the sex industry, urban living, native language, marital status (unmarried), and neurodivergence.

A.2 Annotation Interface

A screen capture of the annotation interface provided to our annotators is given in Figure 4.

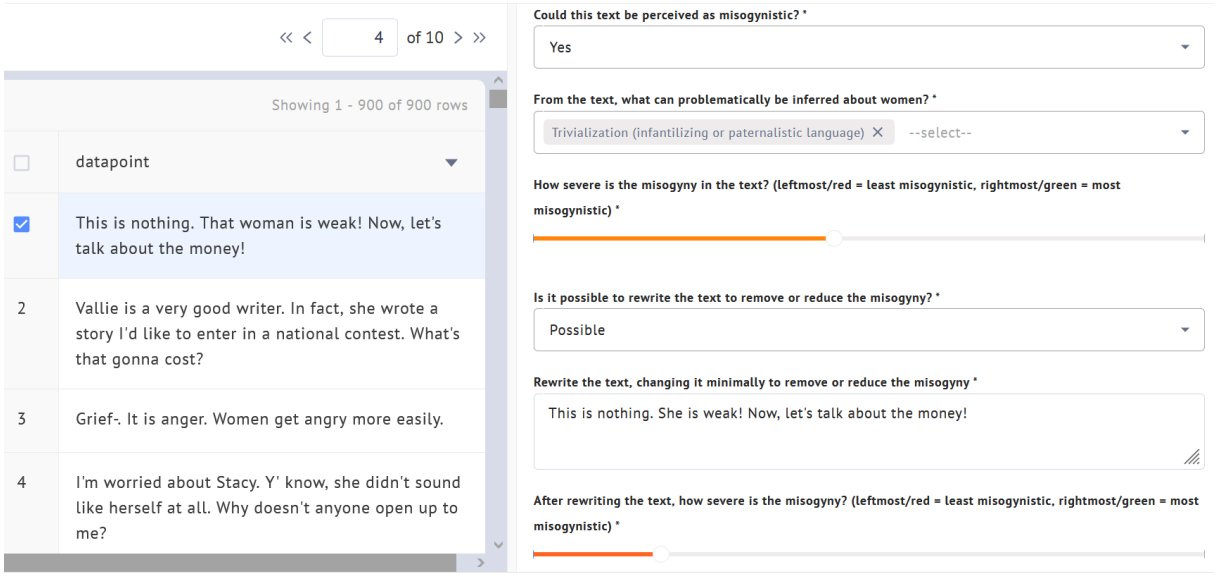


Figure 4: Screen capture of the annotation interface.

A.3 Binary Classification Model Performance Across Datasets

The results of our baseline experiments with other misogyny detection datasets mentioned in this paper are provided in Table 8.

dataset	model	Accuracy↑	F1_macro↑	Precision_yes↑	Recall_yes↑	F1_yes↑	Precision_no↑	Recall_no↑	F1_no↑
EDOS	BERT	0.874	0.826	0.753	0.718	0.735	0.911	0.925	0.918
	DeBERTa v3	0.873	0.824	0.751	0.713	0.732	0.910	0.924	0.917
	ELECTRA	0.877	0.830	**0.763**	0.718	0.740	0.911	**0.929**	**0.920**
	RoBERTa	**0.878**	**0.834**	0.747	**0.749**	**0.748**	**0.920**	0.919	0.919
Guest	BERT	0.937	**0.806**	0.728	**0.581**	**0.647**	**0.955**	0.976	0.965
	DeBERTa v3	**0.938**	0.803	**0.750**	0.558	0.640	0.953	**0.980**	**0.966**
	ELECTRA	0.924	0.765	0.647	0.512	0.571	0.947	0.969	0.958
	RoBERTa	0.929	0.786	0.673	0.558	0.610	0.952	0.970	0.961
Ami	BERT	0.691	0.691	0.631	0.791	0.702	0.773	0.606	0.679
	DeBERTa v3	0.709	0.707	0.635	**0.863**	**0.732**	**0.832**	0.578	0.682
	ELECTRA	0.705	0.704	0.641	0.817	0.718	0.797	0.609	0.690
	RoBERTa	**0.712**	**0.712**	**0.669**	0.739	0.702	0.756	**0.689**	**0.721**
Callme	BERT	0.929	0.865	0.738	0.809	0.772	0.966	0.950	0.958
	DeBERTa v3	0.936	0.877	0.757	0.830	0.792	0.970	0.954	0.962
	ELECTRA	0.930	0.871	0.720	**0.860**	0.784	**0.975**	0.942	0.958
	RoBERTa	**0.937**	**0.880**	**0.763**	0.836	**0.798**	0.971	**0.955**	**0.963**

Table 8: Comparison of binary classification model performance on the test sets of four other misogyny/sexism datasets mentioned in this paper. The setup is identical to the setup we used for the binary classification baselines on our dataset, as specified in section 4.1, averaged over three runs with different random seeds. The best performance for each dataset is highlighted in bold. For reference, Kirk et al. (2023) report in their EDOS paper a F1_macro score of 0.8235 as baseline with a DeBERTa v3 fine-tuned for binary classification; Guest et al. (2021) report Accuracy of 0.89, an F1 score of 0.43, Precision of 0.38, and Recall of 0.5 as baseline for a weighted BERT model fine-tuned for binary classification; Fersini et al. (2018) report in their Ami paper Accuracy of 0.605 as a baseline, based on Support Vector Machines with linear kernels trained on bag-of-words for binary classification; and Samory et al. (2021) report in their Callme paper a Macro-F1 score of approximately 0.83 for a BERT model fine-tuned for binary classification.

A.4 Dataset Examples

Additional dataset examples referenced in Section 4.2 are provided in Figures 5 through 7.

Binary Classification			
Reference Number	**Text**	**Annotator Classification**	**Model Classification**
3	Hey, Hernandez, no plantain for Silly. I'm serious, dude. She shits on the rug, and she feels ashamed of herself.	3 negative annotations	False Positive
4	Girls? I liked girls when I was a kid. I like women now.	3 negative annotations	False Positive
5	I mean, how can you compare me to someone who decapitated women and kept their heads as mementos? I mean, that's the behavior of a truly, truly evil human being. I mean, what comparison is there, ever, possibly in that?	3 negative annotations	False Positive
6	But if we're gonna talk negatives, let's start with the giraffe legs. Okay, let's take a vote on it. Anyone who wants her in, sing a G-sharp.	3 positive annotations	False Negative
7	Sorry, hon. I did it! I did it!	2 positive annotations	False Negative
8	That was a joke. Okay, what if she was to... disappear? I mean, would anybody actually miss her?	1 positive annotation	False Negative

Figure 5: Example datapoints from the binary classification task.

Multi-Label Classification			
Reference Number	**Text**	**Predicted Class Versus Actual**	**Analysis**
9	You are the only woman for me. Wade was telling the truth! You still don't get it, sweetie.	Predicted Label: [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0] Actual Label: [0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]	Example of a typical datapoint with only 1 correctly predicted category. The model correctly predicted the class of trivialization.
10	You must resist that which tears all mortal men apart, women. In a minute, a super-hot sexy bitch is going to come through that door and try to seduce you into telling her the secret word. She will stop at absolutely nothing, to get it out of you.	Predicted Label: [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1] Actual Label: [0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 1. 1.]	Example of a rare 3-category datapoint that was perfectly classified by the model, containing a strong slur word. The model correctly predicted the classes of gendered slurs, gender essentialism, and objectification/sexualization.

Figure 6: Example datapoints from the multi-label classification task.

Severity Regression			
Reference Number	**Text**	**Annotator Severity Ranking**	**Model Severity Ranking**
11	It find it very charming that Sally has convinced herself that she's in control of our relationship. Wait. Horatio Astaire Biggs. Are you, Are you...	0.739	0.08
12	The boy means nothing to me. I can't stand the wailing of women. God, I could watch Tyrion slap the smug off that little fucker's face a hundred times and never get sick of it.	0	0.57

Figure 7: Examples datapoints from the severity task.

Annotation Guidelines

Content Warning:
Examples of misogyny, including sexist language and slurs, violence against women, harassment, transphobia, transmisogyny and sexual assault are included in this document for instructional purposes. This language does not reflect the views of any of its creators.

Tips

- Do not rush. We are paying you by the hour because these tasks are complex so please take your time, especially with complicated or confusing data points.
- If you get tired and/or find the content distressing, please take a break. If necessary, you're also welcome to reach out for support (see Mental Health section).
- We're looking to infuse this project with your expertise. If something strikes you as misogynistic, then you should use your intuition/domain knowledge and annotate it accordingly. Just try to be as clear as possible about the reasoning behind your decision where possible.
- When you're unsure about an annotation, go back to the basic question: Is it misogynistic or could it be considered misogynistic in some circumstances? If you're still unsure, you can of course use the option "Unclear".
- It is possible that data points will include multiple misogynistic elements. In these cases, annotate to include all of the inference categories and rate the severity according to the most severe of the sentences or misogynistic elements. We ask you to flag examples with multiple iterations of misogyny by using Datasaur's "Comments" feature so we can pay special attention to this data.
- In cases where a typo makes the datapoint uninterpretable, please flag these examples using Datasaur's "Comments" feature, and simply annotate "No" to the question "Could this text be perceived as misogynistic?". In cases where there are small grammatical errors, but the text is still understandable, please do not flag it, and instead simply annotate the datapoint as if the grammatical error weren't there.

Welcome

Thank you for being part of the team!

Together, we hope to build an AI application that can systematically identify, flag, and educate individuals about how subtle and explicit forms of misogyny present themselves in written text. As such, these guidelines may be updated over time with the addition of group suggestions, but any such changes will be clearly communicated to you.

Our goal is to:

- Advance research in human bias detection (both blatant and subtle) using AI
- If possible, create a tool that flags misogynistic content, educating people about how they may be perpetuating stereotypes against women with the goal of inspiring them to communicate differently

The goal is to flag (and ultimately remove) misogyny from sentences, as best we can. The goal is not to make every sentence gender neutral. We want to be mindful to not remove/erase discussions of femininity, women, girls, and/or topics/subject matter related to women, girls, and/or the multiplicity of women's and girls' culture(s). However, there may be cases where making a word gender neutral helps to reduce misogyny. Everything is context dependent, which is why we are relying on your expertise!

Definition of Misogyny

Our working definition of misogyny is:

> ***Hatred of, dislike of, contempt for, or ingrained prejudice against women. It is a form of sexism and can be either intentional or unintentional. Misogynistic language is language that reflects or furthers misogyny.***

Misogyny can be directed to an individual or a group, but the key is that the woman or women in question are being shown contempt, hatred, dislike, etc. at least partially **by virtue of being identified as a woman or as women**. If I say that "*I dislike Anne*" because she told everyone something that I had shared with her in confidence, I am technically expressing my dislike of a woman, but the basis for my dislike is her not keeping a secret rather than her being a woman. *On the other hand*, if someone said "*Isn't it just typical; you can't trust a woman like Anne to keep a secret*," we don't necessarily know what is meant by a "*woman like Anne*," but it is more likely that her identity as a woman is involved in her denigration here. And finally, a **generalization** like "*Women can't keep secrets*" would be obviously misogynistic by virtue of the fact that all women are lumped together and associated with a prejudicial characterization. Hillary Clinton is an example of a figure that may be polarizing in part because she is a woman. Sometimes it's hard to know if someone is disliked because of their position/personality or

because they are a woman. But given the pervasiveness of misogynistic discourse about powerful women, err on the side of caution.

Even sentences that appear to be **complimentary** might contain misogynistic elements. For instance, with the set of sentences, "*Let her go. She's a fine woman, highly talented. You may quote me, my dear*", it is unclear if "*fine*" is in reference to a woman's appearance or her character, though the latter might be more likely given other context clues. But even the idea that she is judged in a particular way because she is a woman might be misogynistic (is she being held to different standards because of her identity?). The phrase "*my dear*" might also be considered paternalistic or condescending. With the data point: "*You're kind of pretty. I'd like to draw you, give you a nice pencil twirl. You interested in a nude drawing*?" even though there is no misogynistic key word in this set of sentences, it ***feels* sleazy**, so the inference is still there. The word "*pretty*" to describe a person indicates that the referent is likely a woman. And the "*kind of*" qualifier sounds like something a pickup artist might say. "*Give you a nice pencil twirl*" sounds sexually suggestive, especially when coupled with "*nude drawing*."

As a default, reclaimed language, **slurs**, or potentially humorous utterances should be classified as misogynistic. Look out for new terms which include a more familiar misogynistic pejorative (e.g. "instasluts" contains "sluts"). If these terms are used pejoratively against a person or group of people they should be annotated as misogynistic. In the example, "*I've been great. You finally got tits, bitch! Bitch, the estrogen has been kicking in, the only thing it hasn't broken down was these fucking arms*," the speakers are most likely trans women who are talking to each other, given the reference to estrogen. Even though "*bitch*" and "*tits*" are being used **humorously** here, for our purposes, it makes sense to annotate this as misogynistic because of the use of slurs and slang and to attempt rewriting it to reduce the misogyny.

While slurs or other more obvious forms of misogyny may register more readily, this project also aims to mitigate **subtler misogynistic inferences**. With the set of sentences, *"I wonder if she actually even wanted to. Are you sure she had miscarriages? Because you know there are ways to fake that, right?*", there is an inference that women are untrustworthy and that they are prone to lying about their bodies, including the personal and traumatic medical experience of miscarrying.

Be mindful of the fact that not every **reference to misogyny**, or to sex or violence, is itself misogynistic. In those instances where, for example, misogynistic acts are being described as exhibited by someone else, you do not need to label the data point as misogynistic. This is because we would not want remove every mention of sexual assault out in the world, or every reference to sex acts or violence in general. In fact, describing these incidents will be important to acknowledge and combat misogyny. For example, with a data point like "*So a girl breaks up with you. So you can't take it, so you shame her online*," the reference to someone else shaming a girl online is not necessarily itself misogynistic because it does not reflect the misogynistic intentions of the speaker. However, it's possible that the use of "*girl*" instead of "*woman*" is misogynistic if the referent is an adult. We hope to capture such nuances.

Context and Speaker:

- While the text comes from movie subtitles, our annotations are more concerned with the content of the utterances than the change in speaker or lack thereof. If it is unclear whether the sentences come from one speaker or multiple speakers, feel free to default to a single speaker. When characteristics of the speaker are unclear, assume the language is being expressed *by* a cis man. When the context indicates that the speaker is a woman, annotate accordingly.
- When unclear, assume the language is being expressed *towards* a woman or that the person being described or spoken about is a woman.
- Annotate based on the information that is available in the data. In other words, try not to extrapolate and make too many assumptions about the context other than what comes to you intuitively. Base your response on your reaction to the existing information.
- Our intention is to gently nudge users towards better language choices. So imagine that this language will be used in a workplace email or in another professional context. It is better to err on the side of caution and imagine the worst case scenario rather than giving the speaker the benefit of the doubt.

Inferred Meaning

Here, you will indicate, what from the text can problematically be inferred about women. You will select all of the inference categories that apply from the list below.

- Anti-feminism (feminism is a bad idea, feminists are gross and ugly, women shouldn't have equal rights)
- Dehumanization (comparing women to animals or objects)
- Domestic violence and other violence against women (self-explanatory)
- Gender essentialism or stereotypes (can be both positive, e.g. *women are good at childrearing and cooking because they are more nurturing*, and negative, e.g. *women are untrustworthy and overly emotional because of their hormonal cycles*)
- Gendered slurs (chick, bitch, cunt, etc.)
- Intersectional, identity-based misogyny (any other instance of misogyny that is related to race, ethnicity, religion, class, occupation, immigration status, disability, size, etc.)
- Lacking autonomy or agency (women are not able to make decisions or must defer to male authorities)
- Phallocentrism (focus on penis in organization of social world)
- Rape and other forms of sexual violence (self-explanatory)
- Sexualization (outsized focus on appearance, degrading language)

- Transmisogyny/homophobia (includes mocking individuals or groups for gender nonconformity, e.g. for dressing or acting in a way that does not conform with assumed gender roles. Homophobia/transphobia that also contains misogynistic implication)
- Trivialization (infantilizing or paternalistic language, women are not taken seriously)

Keep in mind, this list is not exhaustive. Rather, it is only a list of *possible* categories. You will be given the option of the "**Other**" category, which will allow you to add a new inference type, in a freeform text box.

You are also given the option of choosing the category ("**Add optional explanation**"). You may select this box in cases where you believe it would be helpful to explain *why* you chose the categories you did if it wasn't immediately intuitive or obvious to you. Once this box is selected, you will be presented with a text box where you can add your explanation.

Severity of Misogyny

Misogyny can exist on a spectrum. We therefore want to know *how* prejudicial the text is against women. To annotate for severity, you will be presented with a sliding scale. The sliding scale increases in misogyny from left to right. That is, the leftmost (red) side of the scale is the least misogynistic, while the rightmost (green) side of the scale is the most misogynistic.[1]

"*Women are good at child-rearing*"

"*Women are only good at child rearing*"

"*Bitches should only have babies*"

Least Severe *Most Severe*

Sentences that use slurs or seem to condone sexual assault or violence against women are more likely to have high severity scores, whereas other more subtle forms of misogyny might have lower scores or be more difficult to assess.

[1] *These are only examples, please use your own judgment when deciding on the severity.*

"*You should see these women, man.*"

"*You should see these chicks, man.*"

"*You should see these bitches, man.*"



Least Severe *Most Severe*

Rewriting the Sentence to Remove Misogyny

Once you have identified a three-sentence block as containing misogyny, we ask you whether or not it is possible to rewrite the sentence to remove or reduce the misogyny. To determine if it is possible to remove or reduce the misogyny, start by asking yourself the following question: "What is the main point of each utterance containing misogyny within the three-sentence block? What work is each doing overall?". If you think the sentences alternate between speakers, there will likely be multiple goals, but if they are a sequence of sentences from a single speaker, it is possible that they are all part of a larger goal.

Whenever the main point is to denigrate women or a woman by virtue of her woman-hood, we indicate that no rewrite is possible. When the main point is not to be misogynistic, we try to rewrite the sentence in such a way that the main point is still conveyed but with less misogyny included.

If you have determined that a rewrite is possible, try to rewrite the sentence in such a way that *all of* the misogyny (in the case of their being multiple misogynistic implications) is removed or at least lessened. This could involve a number of different tactics, including, but not limited to the following (illustrated as modifying the examples above):

Addition of words: **Many** women are good at child-rearing.
Addition of phrases: Women are good at child-rearing, **as are men**.
Substitution of words: **Humans** are good at child-rearing.
Deletion of words: Women are ~~**only**~~ good at child rearing

Other techniques will involve the substitution or deletion of whole phrases, changing the structure of the sentence (such as changing the order of the elements), etc.

Note that in some of these cases, the misogyny has not been eliminated, only reduced. We would still like you to try to remove as much misogyny as possible without drastically altering the principle objective of the speech act represented by the multi-sentence dialogue you have read. In other words, if you think that the main purpose of a sentence is to denigrate women, it is

natural that you won't be able to remove the misogyny. However, as long as there is another meaning to the text beyond just the misogynistic implication, it is worth rewriting the sentence to attempt to lessen (or remove) it.

After you try to rewrite the sentence, you will be presented with another severity scale where you can indicate how misogynistic the rewritten version is. That way, you can put the scale to "Not misogynistic" if you think the misogyny has been completely eliminated or you can indicate how misogynistic the rewritten version is if it was impossible to eliminate the misogynistic content completely.

For example, one could rewrite the sentence "*Bitches should only have babies"* to "*Women should only have babies"*, which would still be highly misogynistic but perhaps less than the original.

Note that many of these rewritten sentences will still have other issues such as racism, ableism, name-calling, etc., but since our objective is to focus on misogyny, the rewrite should not eliminate these other forms of discrimination (unless they are intersectional and contribute to the misogyny itself).

It is important to try to maintain the original function of the sentence as much as possible. What is the spirit of what the sentence is trying to convey? Try to rewrite the sentence in a way that retains the original function. Only rewrite the problematic part of the sentence.

For example, with the sentence, "*You wouldn't know controversy if it pulled up to a middle school, showed you its penis and make you take a blow job*," an effective rewrite would be "*You wouldn't know controversy if it hit you over the head*." This retains the use of an idiom in the sentence in a way that still makes sense, including using a reference to violence, while still reducing the misogynistic inference.

In some cases, changing a word to a gender neutral term can help reduce misogyny. This is the case for inferences where women are assumed to be the primary caregiver for children, for example:

"*Sometimes mommies want daddies to pitch in and help out and do things, and those are called responsibilities*" could be changed to "*Sometimes parents want other parents to pitch in and help out and do things, and those are called responsibilities*."

"*Didn't your mother tell you never to play with knives?*" could become "*Didn't your parents tell you never to play with knives?*"

Severity of Reformulated Language

Once you've changed the original sentence (to remove or reduce the misogyny) you'll be asked about the updated sentence. Specifically, you'll be asked how misogynistic it is. In many cases,

misogyny can't be removed entirely so it might be the case that the severity score is still greater than zero. If you are able to completely remove the misogyny, slide the slider to the leftmost side.

Annotation Task

When you log into [Software Annotation Platform] and click "Start Labeling", you will see a three-sentence block of text. Your job will be to do the following:

1) Read the block of text. It will be roughly 3 sentences long.

2) Indicate whether or not the text could be perceived as misogynistic by selecting "Yes," "No," or "Unclear"

a) If "No, click "Submit answers", and you will be presented with a new block of text to annotate.

b) If "Yes", several more options will appear:

i) "From the text, what can problematically be inferred about women?". For any text that you label as misogynistic, you will be asked to select the most appropriate categories describing why the text is misogynistic. You may select multiple categories, as well as "Other" which, when selected, will open up a text box where you may add any other category as to why the text is misogynistic. Finally, you may optionally provide an explanation as to why you have selected the category/ies by selecting "Add optional explanation".

ii) "How severe is the misogyny?" You will be asked to rate the severity on a sliding scale where the leftmost (red) side is the least misogynistic and the rightmost (green) side is the most misogynistic.

iii) "Is it possible to rewrite the text to remove or reduce the misogyny?" You will be asked to choose either "Not Possible" or "Possible".

(1) If "Not Possible", you will be finished the annotation task and can click "Submit answers"

(2) If "Possible":

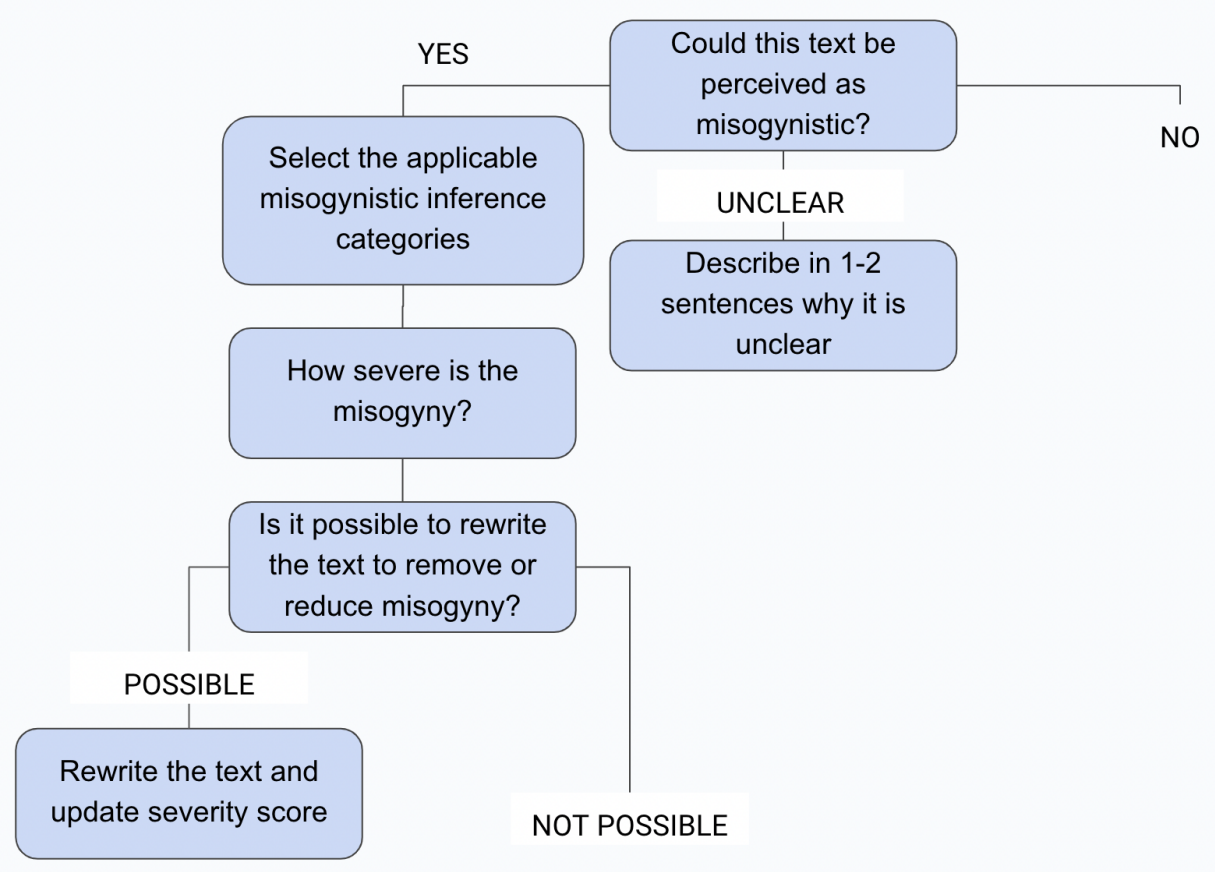
(a) You will be asked to rewrite the text, doing your best to remove or reduce the misogyny. **The text box will be prepopulated with the original text snippet, and your task is to edit this text minimally.**

(b) "After rewriting the text, how severe is the misogyny?" Once again, you will be asked to determine the severity of the rewritten

sentence, with the added option of moving the sliding scale all the way to the left if you were able to entirely remove the misogyny.

c) If you really can't say "Yes" or "No", select the "Unclear" option as a last resort. You will then be prompted to write 1-2 sentences describing why it is unclear.

The annotation task can be summarized by the following flowchart:



You can find a demo of the annotation task and software in the training session recording [removed to preserve anonymity].

Further Important Information

Communication Among Annotators

Throughout the annotation process, you are encouraged to collaborate with your fellow annotators in discussing difficult datapoints. To facilitate this communication, we have created a slack channel where you can post difficult datapoints and engage in a respectful discussion about them. When having these discussions, please keep in mind the following community standards for appropriate communication with one another:

- Be respectful in the way you communicate, both with regards to the content and how you phrase it
- Remember that misogyny is subjective, different opinions will likely occur, this is expected and welcomed
- You will all bring your own personal experiences and expertise, we want you as your "whole self", including your lived experiences and gut feelings, and we value them equally to professional expertise, please bear this in mind during discussions ("your gut feeling or experience x is wrong because textbook y says otherwise" is a no-go)
- In case of different opinions or difficult conversations, stay curious and ask for "why" and "how" without being judgmental
- Assume good faith, always!

Given all the previous points, it is clear that there will not always be consensus at the end of discussions, which is okay, in such cases each annotator can go forward with their own intuition.

Guidance on Outside Sources

With your respective backgrounds, you are all considered experts in your field, and we therefore would like your annotations to be true to your own judgments. In light of that, here are the DOs and DON'Ts of using outside sources to help with your annotations:

DO:

- Ask your fellow annotators for their opinions/thoughts in cases where you are having a hard time making a judgment.
- Use sources like Urban Dictionary or traditional dictionaries like the Oxford English dictionary in cases where you don't understand the language being used.

DON'T:

- Use any kind of generative AI to help guide your decisions. Given that the goal of this project is to build our own AI tool that flags and removes misogyny, it is essential to the integrity of the project for all annotation judgments to come from yourselves.
- Use a grammar checker; this is not a project that requires prescriptive grammar or spelling.

Mental Health Resources

Since annotating offensive language can cause psychological harm, we want to make sure that you are properly supported.

For those who are affiliated with universities, mental health services can be found on your respective university websites, including those of [University X], [University Y] and [University Z]. Additional support from services like [A] or [B] mental health hotline are available at 1-866-585-0445 or you can text WELLNESS to 741741. You are also welcome to reach out to the Project Manager with any questions or concerns.

We will be releasing anonymous mental health forms that can be completed whenever you feel it is necessary and will be mandatory to complete twice during the annotation process. We will also be hosting two office hour sessions (one each month) where you can discuss any concerns with the team.

Practical wellbeing guidance
1. Try to avoid long sessions and instead split the work into smaller, more manageable chunks.
2. Feel free to step away from annotating! We've read that annotators benefit from a short break every 40 minutes.
3. Communicate! Reach out to the rest of your team about challenges you are facing and how to mitigate them.
4. If you begin to feel anxious or uncomfortable during annotation, stop immediately and, if you feel comfortable, reach out for help.