Appendix

Related Datasets

Comparison of reported datasets, focusing or related to sexism detection, are listed in Table 11.

Annotation Examples

In this section are the annotation examples of sexism or not in Table 12, and genders in Table 13.

Baselines' Versions

Table 14 lists the versions of models used in our experiments. Note that DeBERTa Chinese version was released in HuggingFace but the reference is not found. ChatGPT, ChatGLM, and Baichuan are tested on both English and Chinese. LLama and Alpaca can only apply to English.

Hyper-parameters of Best Models

Training epochs for best-performing models are listed in Table 15 for Main Results and Table 16 for Parallel Study.

Prompt Examples

We provide an English prompt example of sexism detection in Table 17 and an English task specification, namely $Prefix$ part, of gender detection in Table 18.

False Predictions Analysis on Gender

Besides phrasing manner, we also examine the gender factor and reach the same conclusion as in the Main Results section that LLMs cannot distinguish misandry well, no matter the target gender is men or women. Moreover, misogyny directed at men is also sneaky to detect. Results are listed in Table 19 and Table 20.

Parallel Study Details

We conduct two sets of experiments on parallel study. We create parallel data of the whole dataset described in the Dataset section by translating it using Baidu API[10]. Note that the size of test set in the parallel study is 500 and 485 for English and Chinese data, respectively.

The first experiment tests the performance of the trained models in the Main Results section of the parallel test set. The results are listed in Table 21 and 22. Compared to the main results (in the Main Results section), we found that LLMs performance is relatively stable and robust compared to MLMs on parallel data, i.e., parallel English data (Table 21) v.s. original Chinese data (Table 8), parallel Chinese data (Table 22) v.s. original English data (Table 7). However, for MLMs, parallel English data significantly improves sexism detection, but there is a drastic drop in the performance of misogyny detection. This phenomenon can be attributed to the differences in phrasing characteristics between English and Chinese. Moreover, MLMs trained on Chinese data poorly predict the parallel Chinese data across all categories except gender. This discrepancy may result from the differences in the linguistic features learned from native speakers in Chinese data compared to those in translated Chinese. Meanwhile, the gender category shows that more balanced training data leads to better performance.

The second experiment is to test a new set of models trained on the parallel corpus, with the best models' hyperparameters listed in Table 16. The results are listed in Table 23 and 24. Even though the translated corpus has shortcomings in translating slang and colloquialism, the performance is similar to the main results, i.e., parallel English data (Table 23) v.s. original Chinese data (Table 8), parallel Chinese data (Table 24) v.s. original English data (Table 7). This indicates a promising avenue in data augmentation. Moreover, the results of the original test set on the new set of models, as shown in Table 25 and 26, further support the promising avenue and previous findings that balanced and larger amounts of data would improve the performance. However, the discrepancy in linguistic features remains challenging in detecting phrasing, misogyny, and misandry.

[10] https://api.fanyi.baidu.com/

Dataset	**Language**	**Categories**	**#Total**	**#Sexism**	**%**	**Source**
(Waseem and Hovy 2016)	en	Racism, Sexism, None	16,914	3,383	20	T
(Jha and Mamidi 2017)	en	Benevolent, Hostile, Others	10,095	2,966	30	T
OFFCOMBR (2017)	pt	Sexism, Racism, ...	1,250	-	-	BW
IberEval-2018 (2018b)	en, es	Misogyny (5) Target classification (2)	8,115	3,915	48	T
Evalita-2018 (2018a)	en, it	Misogyny (5) Target classification (2)	10,000	4,585	46	T
(Sharifirad and Matwin 2019)	en	Sexism (4) or NOT	679	3,119	22	T
(Parikh et al. 2019)	en	Sexism (23) or NOT	13,023	-	-	ESP
(Grosz and Conde-Cespedes 2020)	en	Sexism or NOT	1,142	627	55	T
(Bhattacharya et al. 2020)	en, hi, bn	Misogyny or NOT	12,073	2,092	17	T, F, Y
MeTwo (2020)	es	Sexism, Doubtful, ...	3,600	1,152	32	T
(Chiril et al. 2020)	fr	Sexist content (3) or NOT	11,834	4,047	34	T
RUHSOLD (2020)	ur	Sexism, Religious Hate, ...	10,012	839	8	T
(Guest et al. 2021a)	en	Misogyny(4) or NOT(3) Misogynistic treatment (2) Threatening (3) Disrespectful actions(4)	6,567	696	11	R
LAHM (2021)	en, es, hi,	Sexism, Racism, ...	227,836	-	-	T
CallMeSexist (2021)	en	Sexism or NOT	3,826	-	-	T
(Priyadharshini et al. 2022)	ta, ta-en	Misogyny, Misandry, Homophobia, ...	8,181	1,621	20	-
CoRoSeOf (2022)	ro	Sexist content (3) or NOT	39,245	3,897	10	T
ArMIS (2022)	ar	Misogyny or NOT	964	-	-	T
(Al-Hassan and Al-Dossari 2022)	ar	Sexism, Racism, Hate, ...	11 K	-	6	T
SWSR (2022)	zh	Sexism or NOT Sexism (3) Target (2)	8,969	3,093	34	W
SemEval-2023 EDOS (2023)	en	Sexism or NOT Sexism CAT. (4) Fine-Grained Sexism (11)	20,000	4,854	24	G, R
(Bertaglia et al. 2023)	en	Sexism or NOT	200 K	-	11	Y
GERMS-AT (2024)	de	Sexist/misogynist Level (5)	8,000	-	33	-
MDMD (2024)	ta, ml	Misogyny	2,776	848	31	-
BeyondGender (Ours)	en, zh	Sexism or NOT Gender (2) Phrasing (2) Misogyny or NOT Misandry or NOT	21,119	7,745	37	Y,W G,R

Table 11: Comparison of reported datasets. *#Total* and *#Sexism* is the number of sexism data and the size of the whole dataset. ”-” means unmentioned or unavailable. Source is where the data are collected from: *T*witter, *F*acebook, *Y*ouTube, *G*ab, *R*eddit, *W*eibo, *B*razilian *W*eb, and *E*veryday *S*exism *P*roject.

The social circumstances

No	text	Sexism
1	She has time to hit her man in his mouth if he gets “out of line”. But if a man said he had time to hit his girl in the mouth when she gets out of line, people would be livid. Just more double standards.	0
2	The feminists, family court, #me-too, and divorce rape, are making men in drones withdrawal from women, fatherhood or family. The only way to bring men back to the table is if they change the family laws and divorce laws.	0
3	If anything, it’s the men that are exploited...how much does the average female make in a shoot vs a men? Equal pay for equal work right?	0
4	It’s because in gender equality land, men expect women to be equals and think of them that way. But we’re not equals at all. You’re essentially sexual and social aristocrats and we’re peasants.	0
5	Man kills woman - “That monster!” Woman kills man - “I wonder what he did to her?” Imagine if we responded the same way. “I wonder what she did to him?”	0
6	The deluth model is such to where men visibly beaten with black eyes call the cops and still get arrested while the woman doesn’t have a scratch on her.	0
7	Consequences have never existed for women who falsely accuse men of rape. Women just accuse women of sexual misconduct and go about their day after their lie is exposed..	0
8	there are men in ME and Africa who get killed in wars. If you care about men rights, why don’t you fight for their rights instead of sitting on your first world ass and bitching about how that on woman was too rude to you hence you hate all women	0

Subsets of the gender

No	text	Sexism
1	Not if she wants it. Only **prudish women** hate dirty talk from attractive men from the getgo.	0
2	We should be blaming **the young women entrapping an older man so she can get his money.** These whores are just mad that they did not get the money.	0
3	I didn’t say you had to stay with a crazy bitch ex-wife. Again this was directed towards **men who did not even try to reproduce but are going their own way.**	0

Individual attack but sexism.

No	text	Sexism
1	She’d not even get **one** man who wants to rape her, profile pic checks out, rofl.	1

Table 12: Some examples of the social circumstances, subsets of the gender, and individual attack but sexism cases.

Transgenders.		
No	**text**	**Gender**
1	No **men** in the ladies restroom! Women are the majority in the US, 51%, why should our privacy and safety be compromised for a confused 1.2% of the population.	1
2	That's why, the people with the brains, were so against this BS. It's one story after the other. Heard a male "tranny" raped a girl in a school bathroom! Do not put me in a dressing room, bathroom, etc., with any **man** who claims he is a woman.	1

Equivalent prejudices toward men and women.		
No	**text**	**Gender**
1	The very idea of 'gender violence' is total bullshit, no man on earth beats women just because they're women. There is a reason for violence and gender isn't it.	2
2	It's all projection Women realize they can only offer one thing to a man, so they pretend that men only want that one thing.	2
3	I like how society says that men are not sensitive.. Fuck, men are way more sensitive and considering than those heartless bitches (the vast majority of women)..	2
4	Because women lie and men don't. It be that simple.	2

Table 13: Some examples of gender annotation.

Model	English	Chinese
`bert-large-uncased` (Devlin et al. 2019)	√	
`roberta-large` (Liu et al. 2019)	√	
`deberta-v3-large` (He, Gao, and Chen 2022)	√	
`chinese-bert-wwm-ext` (Cui et al. 2019)		√
`chinese-roberta-wwm-ext-large` (Cui et al. 2020)		√
`Deberta-Chinese-Large`		√
`ChatGPT 3.5` (OpenAI 2022)	√	√
`chatglm2-6b` (Du et al. 2022)	√	√
`Baichuan2-13B-Chat` (Baichuan 2023)	√	√
`Llama-2-13b-chat-hf` (Touvron et al. 2023)	√	
`stanford-alpaca-13b` (Taori et al. 2023)	√	

Table 14: Baseline model for English and Chinese data.

Model	Sexism	Gender	Phrasing	Misogyny	Misandry
`bert-large-uncased`	1	1	5	5	10
`roberta-large`	1	10	5	5	10
`deberta-v3-large`	5	10	5	5	10
`chinese-bert-wwm-ext`	10	10	10	10	10
`chinese-roberta-wwm-ext-large`	40	15	20	30	10
`Deberta-Chinese-Large`	20	15	10	10	10

Table 15: Training epochs for best-performing models in main results.

Model	Sexism	Gender	Phrasing	Misogyny	Misandry
`bert-large-uncased`	5	1	1	1	1
`roberta-large`	15	20	15	20	20
`deberta-v3-large`	20	0	0	0	0
`chinese-bert-wwm-ext`	10	10	2	10	15
`chinese-roberta-wwm-ext-large`	20	20	5	10	20
`Deberta-Chinese-Large`	10	10	2	10	5

Table 16: Training epochs for best-performing models in parallel study.

Var	No	Text
Prefix	1	Sexism is prejudice or discrimination based on one's sex or gender.
	2	Now please complete the sexism detection task. Input is a sentence or paragraph. If the text contains sexism, output "Yes", and if the text does not contain sexism, output "No". For example:
	3	Input: "This kind of heavy lifting should be done by men."
	4	Output: Yes
	5	Input: "Women should cook and do housework."
	6	Output: Yes
	7	Input: "There are many girls putting on makeup on the street."
	8	Output: No
	9	Input: "I am a feminist, but I strongly dislike discrimination against men in law and morality and excessive protection for women."
	10	Output: No
	11	Input: "Girls are weaker, and boys should protect girls."
	12	Output: Yes
Data	13	Input: "Man should perform military service."
Suffix	14	Output:

Table 17: An English prompt example of sexism detection.

Var	No	Text
Prefix	1	Now please determine the target gender in the sentences.
	2	Input is a sentence or paragraph, and the output is "Men" or "Women." For example:
	3	Input: "This kind of heavy lifting should be done by men."
	4	Output: Men
	5	Input: "There are many girls putting on makeup on the street."
	6	Output: Women
	7	Input: "Girls are weaker, and boys should protect girls."
	8	Output: Women

Table 18: Task specification of gender detection.

Model	Misandry			Misogyny			non-Misandry			non-Misogyny		
	Men		Women	Men		Women	Men		Women	Men		Women
BERT	0.05	<	0.08	0.84	<	0.88	0.70	<	1.00	0.02	<	0.05
RoBERTa	0.06	>	0.03	0.71	<	0.75	0.73	<	1.00	0.07	>	0.05
DeBERTa	0.05	>	0.03	0.98	>	0.96	0.73	<	1.00	0.00	<	0.02
ChatGLM	0.76	<	0.78	0.69	<	0.79	0.36	>	0.00	0.09	>	0.08
Baichuan	0.35	<	0.52	0.69	<	0.83	0.42	>	0.00	0.07	>	0.06
ChatGPT	0.59	>	0.56	0.69	<	0.79	0.15	>	0.00	0.02	<	0.04
Llama	0.79	>	0.76	0.41	<	0.65	0.09	>	0.00	0.38	>	0.08
Alpaca	1.00	=	1.00	1.00	=	1.00	0.00	=	0.00	0.00	=	0.00

Table 19: False positive (left) and false negative (right) predictions with gender factor in English dataset.

Model	Misandry			Misogyny			non-Misandry			non-Misogyny		
	Men		Women	Men		Women	Men		Women	Men		Women
BERT	0.23	<	0.28	0.22	<	0.24	0.30	<	0.56	0.78	>	0.36
RoBERTa	0.80	<	0.82	0.43	>	0.40	0.08	<	0.22	0.33	>	0.25
DeBERTa	0.28	>	0.27	0.28	>	0.24	0.32	<	0.33	0.67	>	0.36
ChatGLM	0.53	>	0.38	0.62	<	0.63	0.38	<	0.55	0.44	>	0.31
Baichuan	0.66	>	0.48	0.84	<	0.92	0.32	<	0.44	0.00	<	0.13
ChatGPT	0.69	<	0.74	0.70	>	0.65	0.06	<	0.11	0.33	>	0.12

Table 20: False positive (left) and false negative (right) predictions with gender factor in Chinese dataset.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	**0.77**	**0.64**	0.63	0.47	0.59	0.47	0.34	0.61	0.49	0.38
RoBERTa	**0.77**	**0.66**	0.71	0.64	0.67	0.69	0.16	0.64	0.49	0.40
DeBERTa	**0.81**	**0.69**	0.67	0.54	0.63	0.72	0.18	0.65	0.49	0.35
ChatGLM	**0.82**	0.72	0.65	0.64	0.60	0.67	0.49	0.45	0.47	0.48
Baichuan	0.74	0.66	0.43	0.44	0.63	0.64	0.49	0.43	0.38	0.58
ChatGPT	0.75	0.68	0.64	**0.65**	0.59	0.51	0.54	0.50	0.50	0.57

Table 21: The test results of parallel English data. #test = 500. The results equal to or better than those in Table 8 are in bold.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	0.39	0.40	**0.44**	**0.60**	0.51	0.47	0.13	0.29	0.15	**0.54**
RoBERTa	0.13	0.28	**0.42**	0.26	0.59	0.51	0.70	0.62	0.12	0.14
DeBERTa	0.35	0.37	**0.39**	0.47	0.69	0.60	0.56	0.49	0.14	0.44
ChatGLM	0.84	0.74	0.26	0.27	0.69	0.61	0.73	0.63	0.18	0.52
Baichuan	0.80	0.70	0.40	0.41	0.81	0.73	0.84	0.74	0.16	0.50
ChatGPT	**0.86**	0.76	**0.47**	**0.56**	0.87	0.77	0.86	0.76	0.16	0.19

Table 22: The test results of parallel Chinese data. #test = 485. The results equal to or better than those in Table 7 are in bold.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	0.43	0.48	0.63	0.47	0.57	0.45	0.49	0.42	0.48	0.34
RoBERTa	0.43	0.49	0.72	0.71	0.62	0.75	0.61	0.73	0.52	0.69
DeBERTa	0.42	0.48	0.66	0.49	0.58	0.40	0.53	0.36	0.50	0.33

Table 23: The test results of the second parallel study on parallel English dataset.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	0.80	0.70	0.28	0.77	0.87	0.77	0.85	0.75	0.24	0.83
RoBERTa	0.81	0.73	0.35	0.67	0.86	0.76	0.83	0.73	0.23	0.72
DeBERTa	0.80	0.71	0.27	0.70	0.86	0.76	0.79	0.68	0.22	0.79

Table 24: The test results of the second parallel study on parallel Chinese dataset.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	0.60	0.51	**0.41**	0.27	0.81	0.69	0.76	0.63	0.17	0.11
RoBERTa	0.57	0.47	**0.48**	0.53	0.67	0.59	0.47	0.46	0.19	0.72
DeBERTa	0.39	0.41	**0.42**	0.26	0.86	0.76	0.85	0.75	0.17	0.09

Table 25: The test results of the second parallel study on raw English dataset. The results equal to or better than those in Table 7 are in bold.

Model	Sexism		Gender		Phrasing		Misogyny		Misandry	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	**0.83**	**0.72**	0.68	0.64	0.61	0.48	0.52	0.65	0.52	0.46
RoBERTa	**0.76**	**0.63**	0.69	0.66	0.58	0.40	0.55	0.64	0.52	0.50
DeBERTa	**0.79**	**0.67**	0.56	0.64	0.60	0.45	0.55	0.62	0.53	0.62

Table 26: The test results of the second parallel study on raw Chinese dataset. The results equal to or better than those in Table 8 are in bold.