BeyondGender: A Multifaceted Bilingual Dataset for Practical Sexism Detection

Xuan Luo[1,2], Li Yang[1], Han Zhang[1,3], Geng Tu[1], Qianlong Wang[1], Keyang Ding[1], Chuang Fan[1], Jing Li[2,5], Ruifeng Xu[1,3,4*]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China
[5]Research Centre on Data Science & Artificial Intelligence, Hong Kong, China
gracexluo@hotmail.com, jing-amelia.li@polyu.edu.hk, xuruifeng@hit.edu.cn

Abstract

Sexism affects both women and men, yet research often overlooks misandry and suffers from overly broad annotations that limit AI applications. To address this, we introduce BeyondGender, a dataset meticulously annotated according to the latest definitions of misogyny and misandry. It features innovative multifaceted labels encompassing aspects of sexism, gender, phrasing, misogyny, and misandry. The dataset includes 6.0K English and 1.7K Chinese sexism instances, alongside 13.4K non-sexism examples. Our evaluations of masked language models and large language models reveal that they detect misogyny in English and misandry in Chinese more effectively, with F1-scores of 0.87 and 0.62, respectively. However, they frequently misclassify hostile and mild comments, underscoring the complexity of sexism detection. Parallel corpus experiments suggest promising data augmentation strategies to enhance AI systems for nuanced sexism detection, and our dataset can be leveraged to improve value alignment in large language models.

Introduction

Sexism, prejudice, or discrimination based on one's sex or gender, has exacerbated gender inequality and injustices. Research has been made to assist the detection of sexism at scale (Grosz and Conde-Cespedes 2020; Jiang and Zubiaga 2023; Rizzi et al. 2023; Krenn et al. 2024).[1]

Sexism primarily affects women and **misogyny**[2], a widespread and enduring sexist ideology, has been practiced for thousands of years (Holland 2012). According to Ambivalent Sexism theory (Glick and Fiske 1996), sexism towards women has two sub-components: 1) Hostile Sexism (HS), characterized by overtly negative evaluations and stereotypes, e.g. '*Women belong in the kitchen, not in the engineering department*', and 2) Benevolent Sexism (BS), which may appear subjectively positive, e.g. '*Women are nurturing, caring, or cooperative*'. HS is easily recognizable due to its aggressive expressions, whereas BS often presents itself as positive but ultimately regards women as amiable yet weak. Therefore, BS is a guise of luring women to stay at a lower social status than men (Cowie, Greaves, and Sibley 2019). Although sexism has historically disadvantaged women, men also suffer from the negative consequences of sexism or misogyny, albeit in more subtle ways, including being sexually objectified and facing pressure to conform to masculine norms (Mabrouk 2020; Dafaure 2022). On the other hand, **Misandry**, 'hatred of, contempt for, or prejudice against men or boys', represents women's anger against their oppressors (e.g. '*In fact, a man has less worth than a woman because he has one less place for another man to shove his dick into.*'). It often manifests in portrayals of men as absent, insensitive, or abusive. However, there is a lack of studies addressing the situation of men.

Another issue is the excessively broad annotation found in existing public datasets (e.g., (Jiang et al. 2022; Kirk et al. 2023)), making it impractical to detect harmful discrimination from harmless prejudices. For example, these datasets classify facts or phenomena subject to debates on gender equality[3], as well as profanity words stemming from individual misconduct rather than having directed connection to gender, as instances of sexism. It introduces bias during model training and leads to an excess of false positive predictions, ultimately limiting real-world application.

Objectives and Contributions

We aim to provide a valuable resource and benchmark for the comprehensive detection of sexism. Specifically, it seeks to address the following objectives that have been overlooked in previous research: to raise awareness and facilitate the detection of sneaky misogyny and sexism towards men and foster constructive debates. To achieve these goals, we introduce BeyondGender developed with the following labels: 1) **Sexism**: If the text shows the poster's prejudice or discrimination based on someone's gender. 2) **Gender**:

*Corresponding Author

[1]**Content warning**: This article contains examples of offensive or hateful language to explain concepts and illustrate the annotation guidelines and case study.

[2]As given by Wikipedia, misogyny is 'hatred of, contempt for, or prejudice against women or girls'. It is used to keep women's subordination to men, maintaining the social roles of patriarchy.

[3]Without exhibiting personal bias or endorsing unequal treatment and systemic discrimination.

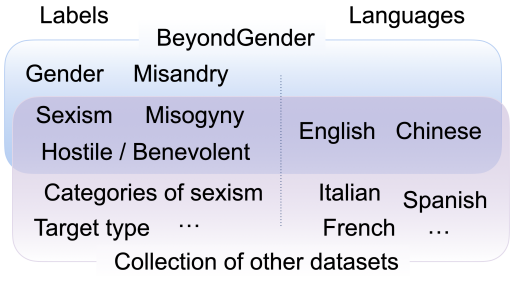


Figure 1: Features of BeyondGender.

The target gender of the text. 3) **Phrasing**: Text's tone, hostile or benevolent. 4) **Misogyny**: If the text expresses hatred of, contempt for, or prejudice against women. 5) **Misandry**: If the text expresses hatred of, contempt for, or prejudice against men.

Four features make BeyondGender [4] a practical dataset for sexism detection: 1) **Novel Facets**: Gender and Misandry labels, which correspond to the state of affairs. 2) **Data Diversity**: data samples from YouTube, Reddit, Gab, and Weibo; a bilingual dataset covering 12.7K English data and 8.4K Chinese data. 3) **Large-scale**: BeyondGender has over 21K data, with a relatively high proportion of sexism data. 4) **Annotation Quality**: clear and detailed annotation guidelines with comprehensive scenarios considered; over 93% inter-annotator agreement.

The data distribution reveals the sexist cultures in English and Chinese. Misogyny remains more pronounced than misandry in both cultures, while they are more likely to be expressed in a hostile manner in English compared to Chinese. Moreover, the high-frequency words are quite different. In addition, We experiment with classic masked language models (MLMs) and large language models (LLMs). It appears that they have a higher performance in identifying misogyny in English and recognizing misandry in Chinese. Nonetheless, they tend to incorrectly categorize hostile comments as misandry/misogyny and benevolent or mild comments as non-misandry/non-misogyny.

The main contributions are as follows: 1) Introduction of a high-quality bilingual dataset for practical sexism detection, 2) The first publicly available dataset for misandry detection in Chinese and English, and 3) Evaluation of baseline models and parallel study, revealing the challenges and a possible solution in detecting target gender, misogyny, and misandry in both languages.

Related Work

In this section, we first compare existing textual datasets (published by journals or proceedings, and most of them are publicly available) from two perspectives. Secondly, we explain why we constructed BeyondGender and the adjustments we made compared to recent annotation codebooks.

Language and source: The majority of sexism detection datasets are available in English, reflecting the extensive research and efforts in this language. Recognizing the global significance of sexism detection in various linguistic contexts and cultures, researchers have expanded their focus to cover multiple languages. (Fersini et al. 2018a) collected a corpus in both English and Italian, while their work in (Fersini et al. 2018b) collected a corpus in English and Spanish. (Bhattacharya et al. 2020) enriched the sexism detection in three languages commonly spoken in India. (Chiril et al. 2020) presented the first French corpus annotated for sexism detection, (El Ansari, Jihad, and Hajar 2020) for Arabic, (Rizwan, Shakeel, and Karim 2020) for Roman Urdu, (Höfels, Çöltekin, and Mădroane 2022) for Romanian, (Zeinert, Inie, and Derczynski 2021) for Danish, (Jiang et al. 2022) for Chinese, and (Krenn et al. 2024) for German. For data collection, Twitter is the most popular platform, followed by Facebook and YouTube. Also, news and document websites are the sources for sexism detection (De Pelle and Moreira 2017; Parikh et al. 2019).

Category and granularity: The mainstream tasks regarding sexism detection are: 1) multi-label Hate speech categorization, where sexism is detected as a sub-category, along with other categories such as racism (Waseem and Hovy 2016; Priyadharshini et al. 2022; Al-Hassan and Al-Dossari 2022); 2) binary sexism identification (Grosz and Conde-Cespedes 2020; Samory et al. 2021; Bertaglia et al. 2023); 3) multi-label sexism type categorization (Jha and Mamidi 2017; Sharifirad and Matwin 2019; Höfels, Çöltekin, and Mădroane 2022); 4) binary misogyny identification (Bhattacharya et al. 2020; Almanea and Poesio 2022), 5) multi-label misogyny type categorization (Fersini et al. 2018a,b; Guest et al. 2021b; Mulki and Ghanem 2021), and 6) other hierarchical classification (Guest et al. 2021a; Jiang et al. 2022; Kirk et al. 2023).

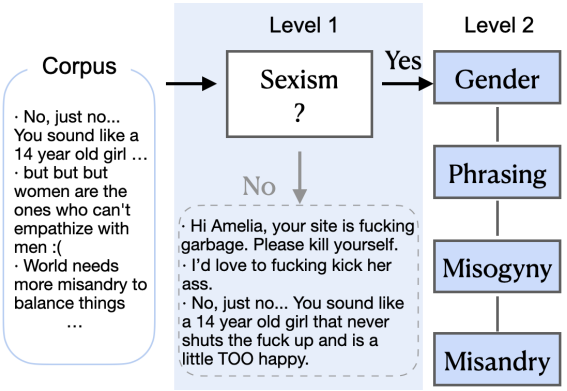
Why BeyondGender? Upon examining the datasets mentioned above, we have identified several areas for improvement in sexism detection: 1) The datasets predominantly focus on sexism towards women, with little data available on misogyny towards men and misandry. 2) Many of these datasets lack clear definitions for their categories, and previous codebooks (Samory et al. 2021; Sultana, Sarker, and Bosu 2021; Sultana 2022) are primarily oriented towards sexism against the opposite gender. 3) A certain proportion of data labeled as sexism are discussions or historical accounts about gender inequality issues rather than personal prejudice or discrimination towards a specific gender. Moreover, aversion caused by misconduct is often roughly categorized as sexism. To address these limitations, we collect and annotate BeyondGender, a dataset designed for sexism detection across both genders. We include misandry and add *women's self-loathing* and *rejection of feminine qualities* to misogyny in our annotation codebook and distinguish those situations and nuances mentioned above when labeling.

Dataset

Collection

BeyondGender is collected from two sources: YouTube comments and previous datasets. In order to acquire comments related to sexism from YouTube, we initiated searches on videos using a predefined set of representative keywords:

[4] BeyondGender will be made available on GitHub.



Examples	Gender	Phrasing	HW	HM
Long term thinking is man's job.	Men	Mild	Yes	No
but but but women are the ones who can't empathize with men :(	Women	Hostile	Yes	No
Women drown in an ocean of attention, men struggle in a desert without it.	Both	Mild	No	No
No, I'm black actually. Go ahead, check my post history to confirm it lol Classic incel tears. Most incels aren't white, white is a *desirable* trait for women.	Women	Mild	Yes	No
Soy boys PMS worse than women cause it's 24/7.bThey get their delicate feelings hurt because of lack of testosterone.. It's a problem!	Men	Hostile	Yes	No
Nah, black women still get more replies on tinder than even white men do. Their female privilege cancels out their black disadvantage.	Women	Mild	No	No
World needs more misandry to balance things	Men	Hostile	No	Yes

Figure 2: The annotation workflow of BeyondGender and several annotated English examples.

1) Sexism culture: *red pill*, *incel*, *manosphere*, *foid*, *misogyny*, *Feminism*[5], etc.,
2) Activities: *marriage*, *parenting*, etc.,
3) Events: *sexual violence*, *sexual harassment*, *#MeToo*, *interview about genders*, etc.,
4) Arts: *Barbie*, *Pride & Prejudice*, etc..
We discard comments which fall below 15 words in length and lack gender-related terminology.

We also leverage recent sexism detection datasets because they represent the contemporary sexism culture and they are collected from different social platforms: 1) English dataset EDOS (Kirk et al. 2023) and 2) Chinese dataset SWSR (Jiang et al. 2022). Given that they broadly categorize critical discussions of gender inequality and aversion stemming from misconduct as instances of sexism, we apply more restricted criteria when determining if it is sexism.

BeyondGender is composed of a total of 21.1K data, comprising 13.2K comments collected from YouTube, 3.1K samples from the Chinese SWSR dataset, and 4.8K samples from the English EDOS dataset.

Annotation

The annotation workflow and examples of annotated data are illustrated in Figure 2. The multifaceted labels are divided into two levels: first, we determine whether the text is sexist or non-sexist. Second, if it is identified as sexism, we annotate the other four labels. The meanings of each label and annotation guidelines are provided as follows:

1. **Sexism: If it conveys prejudice or discrimination based on one's sex or gender.**
   Sexism typically targets a group of people, e.g., '*All women benefit from the actions of violent men*'. Moreover, if the statement is pointing at an individual but can be generalized to that gender group, it is also considered sexism. For example, '*You should do all the heavy lifting since you are a man*' is a **sexism** (label=1), while '*Tell the friend to dump the evil. Let him watch how easily she gets another man to simp for her*' is a **non-sexism** (label=0).
   Previous datasets broadly labeled obscure or controversial conditions as sexism, potentially discouraging discussions about gender issues. Therefore, we make several adjustments and categorize certain situations as non-sexism, including: 1) Hatred directed at an individual due to factors such as race, religion, political views, other than gender. 2) Usage of gender-specific derogatory terms in the context of an event or misconduct not directly related to gender issues.
2. **Gender: The target gender of the text.**
   The values of gender referred to in the text are **men** (label=1), **women** (label=0), and **both** (label=2) (e.g. when the two genders are symmetrically compared). For transgenders, we annotate the gender following the view of the poster.
3. **Phrasing: The manner in which the statement is expressed.**
   It is **hostile** (label=1) if the statement is aggressive, uses derogatory gender terms, or invokes threats. Conversely, it is **benevolent** (label=0) if it is positive or hypocritical. It is **mild** (label=0) if it is neutral or emotionless. Both benevolent and mild instances are labeled as mild.
4. **Misogyny: If it conveys hatred of, contempt for, or prejudice against women.**
   Misogyny is a common sexist ideology in binary gender. The scenarios that reflect misogyny (label=1, otherwise label=0) include, but are not limited to:
   1) Violence against women.
   2) Controlling and punishing women who challenge male dominance, typically differentiating between good women and bad ones.
   3) Rejection of feminine qualities[6], which also extends to the rejection of any aspects of men perceived as feminine or unmanly.
   4) Mistrust of women.
   5) Regarding women as societal scapegoats.
   6) Blaming women for one's own failure in life.
   7) Objectification of women.
   8) Stereotypes suggesting that women weaponize their

[5]Although feminism is not inherently a sexism culture, we can find sexist comments in videos discussing feminism

[6]According to feminists, it holds in contempt institutions, work, hobbies, or habits associated with women.

Categories	# English	# Chinese	# Total
Sexism (Y/N)	6,054 / 6,664	1,691 / 6,710	21,119
Gender (M/W/B)	848 / 5,174 / 31	787 / 878 / 27	7,745
Phrasing (H/B)	5,312 / 742	713 / 978	7,745
Misogyny	4,840	619	5,459
Misandry	954	600	1,554

Table 1: The label distributions in BeyondGender. For gender, *M*, *W*, and *B* represent man, woman, and both genders, respectively. For Phrasing, *H* and *B* represent hostile and benevolent/mild, respectively.

Categories	# English	# Chinese
Sexism? (Yes/No)	163 / 124	70 / 73
Gender (M/W)	298 / 140	74 / 66
Phrasing (H/B)	156 / 209	76 / 65
Misogyny	137	67
Misandry	314	78

Table 2: The average length of comments in BeyondGender.

Language	Label Sexism			Other Labels		
	train	dev	test	train	dev	test
English	10,233	1,000	485	4,733	500	485
Chinese	6,501	700	500	1,099	120	500

Table 3: The size of the train, dev, and test sets.

appearances or that women use seduction to control men.
9) Women's self-loathing, including hating their bodies, disdain for women who are "wives" or "mothers", seeking validation through male approval, etc..

5. **Misandry: If it conveys hatred of, contempt for, or prejudice against men.**
Compared to misogyny, misandry is a minor issue, mainly due to the stress response to misogyny.[7] The scenarios that reflects misandry (label=1, otherwise label=0) include, but not limited to:

1) Violence against men.
2) Women's anger against their oppressors.
3) Opposition to gender-equal laws, such as those related to rape, violence, and divorce.
4) Usage of terms incorporating "man" as a derogatory prefix, such as *mansplaining*, *manspreading*, and *manterrupting*.

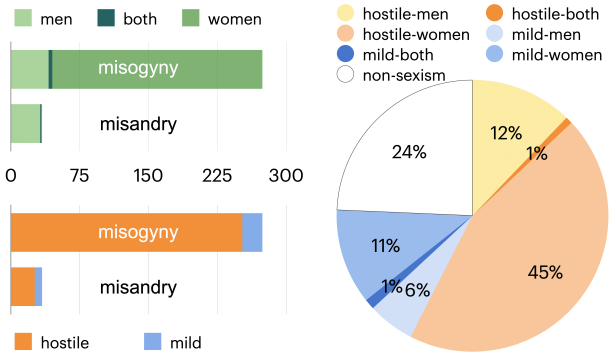
Statistics

The distributions of each label are listed in Table 1. We annotate around 21K data with 7.7K sexism and 13.4K non-sexism. The average length of comments in BeyondGender is listed in Table 2. English data is counted by words, while Chinese data is counted by Chinese characters. Compared to Chinese data, most English sexism data are hostile and

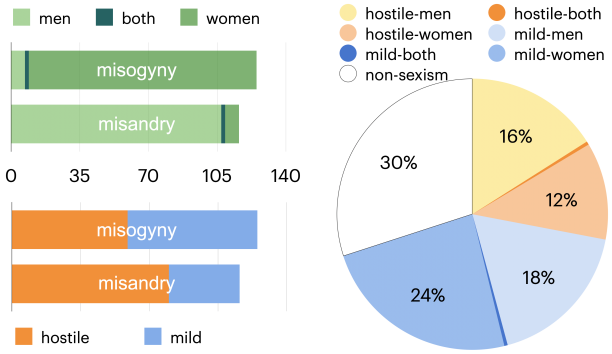
[7]The data could be neither misogyny nor misandry if the stereotype does not blatantly convey hatred of, contempt for, or prejudice against a specific gender.

Given	English		Chinese	
	Miso.	Misa.	Miso.	Misa.
Men	0.47	0.34	0.05	0.63
Women	0.82	0.01	0.65	0.05
Hostile	0.90	0.09	0.42	0.57
Mild	0.25	0.09	0.31	0.17

Table 4: The conditional probabilities in the test set. *Miso.* and *Misa.* represent misogyny and misandry, respectively.



(a) English test set



(b) Chinese test set

Figure 3: Composition of the test set.

misogyny. It can also be revealed by test set composition, demonstrated in Figure 3a and 3b. The test set has 485 English data and 500 Chinese data. The split of the dataset is listed in Table 3.

Table 4 listed the conditional probabilities of misogyny and misandry given specific gender and phrasing manner. Notably, nearly half of the English sexism data directed at men is actually misogyny, while the majority of sexism data targeting women is misogyny. When considering phrasing manner, it becomes evident that English data predominantly exhibit hostile misogyny, whereas Chinese data express a greater degree of hostility towards men.

Cultural Similarities and Differences

After calculating the word frequency in BeyondGender and considering the statistics above, we list representative words in Table 5 and summarize the similarities and differences from three perspectives:

English	Chinese
rape	(*qian*, money)
bitch/whore...	(*hunlv*, marriage donkey)
ugly/pretty	(*gongzuo*, career)
fat	(*shengyu*, childbearing)
stupid	(*lihun*, divorce)
sex	(*jiabao*, domestic violence)
fuck/fucking	(*zhinanai*, male chauvinist)

Table 5: Representative terms in each language. The *Chinese* column displays Chinese characters/words in the format of (transliteration, English translation) pair.

1. Profanity: The English data exhibits a higher proportion of hostile phrasing, primarily attributed to the frequent use of profanities, including terms like *whore*, *fuck*, *bitch*, *shit*, *pussy*, etc., and often combined with sexual references. Although sex-related words are not prevalent in Chinese data, derogatory terms are used, such as describing women as *marriage donkeys* and *marriage object* and animal names, which have the same pronunciation as "men" in Chinese.

2. Gender Bias: In contrast to the widespread practice and extensive history of misogyny, misandry is less pronounced in English data. Conversely, in Chinese data, misandry is more obvious, and misogyny is sneakier than in English data. Moreover, misogyny in English is primarily attributed to men, such as *incel*[8], for reasons 4) - 8) of misogyny in the Annotation section. Conversely, in Chinese data, a notable proportion of misogyny originates from women, particularly in the context of scenario 9). Worth mentioning, misogyny among women is more prevalent in east-Asian cultures than in Western cultures.

3. Topics: English speakers pay more attention to appearance, intelligence, and sex-related actions. People in Chinese culture or even east-Asian culture are more concerned about financial status and marriage-related events, such as childbearing, domestic violence, and divorce.

Quality Evaluation

Annotators are recruited from prestigious university undergraduate and graduate students proficient in both English and Chinese. The team comprises four men and three women to mitigate gender bias. We have provided training to these annotators and initiated the process by annotating a set of 100 data samples. Throughout the pre-annotation phase, we engaged in discussions and made necessary refinements to the annotation rules, resulting in the final version of the guidelines as presented in the Annotation section. As shown in Table 6, we sampled 300 comments from the entire dataset to calculate consistency. Out of these 300 comments, 280 were consistently labeled as sexism. Within the subset of 280 "sexism" comments, the annotation consistency for each label is above 93%. For the entire dataset, the consistency reaches 94, 97, 95, 92, and 95 percentage.

[8]Incel is a portmanteau of "involuntary celibate". The term is associated with a subculture of people who define themselves as unable to get a romantic or sexual partner despite desiring one.

Label	Consisteny	Percentage (%)
Sexism	280 / 300	93
Gender	270 / 280	96
Hostile	263 / 280	94
Misogyny	273 / 280	94
Misandry	262 / 280	94

Table 6: Annotation consistency (#same_label / #sample).

Experiment

Metrics

The metrics we use for classification are **Precision**, **Recall**, **F1-score**, and **Accuracy**. Due to the label settings, the results will have high recall and F1-score if the predictions are all sexism, men, hostile, misogyny, and misandry. Therefore, we also consider the false predictions for better analysis of the shortcomings of models.

Baselines

We evaluate the sexism detection capability of current state-of-the-art and mainstream models. In the monolingual setting, we fine-tune the Masked Language Models (MLMs) using the training set and select the best-performing model based on the dev set. For Large Language Models (LLMs), we adopt in-context learning.

MLMs: 1. BERT (Devlin et al. 2019; Cui et al. 2019), 2. RoBERTa (Liu et al. 2019; Cui et al. 2020), 3. DeBERTa (He, Gao, and Chen 2022).

LLMs: 1. ChatGPT (OpenAI 2022), 2. ChatGLM (Du et al. 2022), 3. Baichuan (Yang et al. 2023), 4. LLama (Touvron et al. 2023), 5. Alpaca (Taori et al. 2023).

Settings

For masked language models, we train five respective classifiers for the five labels. During training, we set the random seed to 42, the learning rate to 1e-5, and the batch size to 16 with Adam optimizer. We try epochs varying from 1, 5, 10, 15, 20, 30, and 40. To simulate the real distribution of comments in social media, we add non-sexism data from previous datasets into training. The randomly sampled train and dev set for labels in level 2 are only those labeled as sexism. The divisions are listed in Table 3. When testing, the classifiers predict all labels for each data. For level-2 labels, only data whose true label is sexism are evaluated.

For large language models, we add several examples in the prompt and combine the data as input. The inputs of LLMs are in the format:

$$Prefix + Data + Suffix \tag{1}$$

where $Prefix$ contains the task description and provides several examples; $Data$ is used to represent each piece of test data; $Suffix$ remains a constant string. Task descriptions declared in the $Prefix$ for other labels will be shared with code and data.

Model	Sexism		Gender		Phrasing		Miso.		Misa.	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	.85	.76	.41	.27	.86	.76	.85	.75	.29	.87
RoBERTa	**.86**	**.78**	.25	**.78**	.87	.78	.86	.77	.32	0.89
DeBERTa	.78	.68	.18	.77	.87	.79	.85	.74	**.33**	**.90**
ChatGLM	**.86**	.75	**.44**	.54	.84	.77	.84	.75	.14	.26
Baichuan	.81	.71	.30	.35	.86	.78	.85	.76	.18	.51
ChatGPT	**.86**	**.78**	.30	.33	**.88**	**.80**	**.87**	**.78**	.23	.47
Llama	.79	.67	.40	.30	.87	.77	.69	.59	.19	.30
Alpaca	.00	.24	.00	.77	.86	.76	.85	.75	.17	.19

Table 7: The test results of the English dataset. Note that ".xx" represents a value of 0.xx.

Model	Sexism		Gender		Phrasing		Miso.		Misa.	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
BERT	.56	.56	**.76**	**.77**	**.75**	**.78**	**.61**	**.72**	**.62**	**.72**
RoBERTa	.28	.41	.66	.50	.66	.70	.59	.63	.51	.43
DeBERTa	.44	.48	.66	.64	.71	.73	.59	.69	.61	.71
ChatGLM	**.82**	**.73**	.53	.51	.57	.65	.49	.49	.50	.59
Baichuan	.75	.68	.53	.49	.64	.66	.51	.40	.49	.54
ChatGPT	.81	.72	.49	.47	.61	.53	.56	.51	.55	.49

Table 8: The test results of the Chinese dataset.

Main Results

The main results of English and Chinese test sets are listed in Table 7 and Table 8, respectively.

For English data, both MLMs and LLMs perform well in detecting sexism, hostility, and misogyny. However, they show only modest performance when it comes to distinguishing gender and detecting misandry. Although model Alpaca achieves similar accuracy as other LLMs, in fact, it predicts all the data as non-sexism (label=0), target women (label=0), hostile (label=1), misogyny (label=1), and misandry (label=1). Therefore, it has 0.0 precision for Sexism and Gender but 1.0 recall for Phrasing, Misogyny, and Misandry labels. However, it cannot be applied to sexism detection. For Chinese data, LLMs significantly outperform MLMs in sexism detection, while MLMs perform better in determining target gender and phrasing and detecting misogyny and misandry than LLMs.

Compared the results of misogyny and misandry detection in two languages, the gaps among MLMs can be explained by the uneven distribution of data. Since the ratio of misogyny to misandry in English is 5 and that in Chinese is 1.3, MLMs trained with these data perform misogyny detection substantially better than misandry detection in English but slightly better in Chinese. On the other hand, LLMs, which are not re-trained with any data, have greater gaps than MLMs in English data. It probably reflects that English misandry data are scarce while Chinese misandry data are sufficient in the corpus for LLM pre-training by then.

False Predictions Analysis

To have a grasp of the improvement direction, we examine the phrasing factor first. We calculate the probability of false predictions of misogyny and misandry given different phrasing manner (ground-truth labels), shown as Table 9 and 10. Comparing the two tables, we have the following findings:

Model	Misa.		Miso.		non-Misa.		non-Miso.	
	H	B	H	B	H	B	H	B
BERT	.07	.06	.85	.85	.77ˆ	.50	.05	.05
RoBERTa	.04	.04	.74	.74	.77ˆ	.63	.05	.09ˆ
DeBERTa	.04	.03	1.0ˆ	.95	.81ˆ	.50	.02	.00
ChatGLM	.83ˆ	.63	.81ˆ	.73	.35ˆ	.38	.08	.18ˆ
Baichuan	.58ˆ	.24	.78	.77	.38	.50ˆ	.05	.18ˆ
ChatGPT	.59ˆ	.53	.70	.76ˆ	.08	.38ˆ	.04	.09ˆ
Llama	.76	.76	.33	.59ˆ	.08	.13ˆ	.37	.41ˆ
Alpaca	1.0	1.0	1.0	1.0	.00	.00	.00	.00

Table 9: False positive (left) and false negative (right) predictions with phrasing factor in English data. *H* and *B* represent hostile and benevolent tones, respectively. Note that ˆ is marked when the difference is equal or larger than 0.03.

Model	Misa.		Miso.		non-Misa.		non-Miso.	
	H	B	H	B	H	B	H	B
BERT	.57ˆ	.16	.26ˆ	.20	.20	.58ˆ	.31	.47ˆ
RoBERTa	.93ˆ	.77	.51ˆ	.38	.06	.14ˆ	.20	.32ˆ
DeBERTa	.50ˆ	.20	.35ˆ	.23	.28	.42ˆ	.29	.47ˆ
ChatGLM	.33	.45ˆ	.69ˆ	.59	.34	.47ˆ	.29	.32ˆ
Baichuan	.55	.53	.86	.86	.33	.31	.10	.14ˆ
ChatGPT	.83ˆ	.69	.75ˆ	.64	.06	.83ˆ	.07	.20ˆ

Table 10: False positive (left) and false negative (right) predictions with phrasing factor in Chinese data.

1. Hostile but not hateful: On the left-hand side are the proportions that are predicted as misandry or misogyny, while the true labels are not misandry or not misogyny. In both languages, data expressed in a hostile manner are more prone to be mistakenly predicted as misandry or misogyny.[9] Because hostility can be conveyed through profanity not related to gender-based hatred, these offensive words introduce noise to the detection.

2. Sneaky misogyny and misandry: On the right-hand side are the proportions that are falsely predicted as not misandry and not misogyny. Mild sexism is more difficult to detect in both languages. In Chinese data, at least $1/7$ remains undetected, with misandry detection suffering more overall. A similar situation exists in English data, except that MLMs are more inclined to predict the misogyny label. These sneaky misogyny and misandry, if left undetected, are more deleterious since they will make acute gender inequality worse than harmless prejudice.

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.

[9]MLMs' false predictions on English data do not follow the pattern, indicating insufficient misandry samples.

Data	S.	G.	P.	HW	HM
Zh: 其實 討厭 另一個 性別 是 由 美國 第二波 女權 發起 的, 約 在 70 年代, 他們 認為 男性 是 “父權” 的 既得利益者 L.S. Qishi taoyan lingyige xingbie shi you Meiguo dierbo nvquan faqi de, yue zai 70 niandai, tamen renwei nanxing shi fuquan de jideliyizhe Trans. In fact, the aversion to the opposite gender was initiated by the second wave of feminism in the United States, around the 1970s. They believed that men were beneficiaries of the 'patriarchy'. **True Labels: Non-sexism.**	1				
Zh: 最 重要 的 是 光 想 不 劳 而 获 不 付出 想 得到 得不到 L.S. zui zhongyao de shi guang xiang bu lao er huo bu fuchu xiang dedao debudao Trans. The most important thing is, just thinking without effort, wanting something without working for it, and ending up with nothing.	0	1	1	0	1
Zh: 妈卖批 什么 好 男人 都 死光 了 什么 累觉不爱 了 什么 一个人 生活 怎么 怎么 了 L.S. mamaipi shenme hao nanren dou siguang le shenme leijuebuai le shenme yigeren shenghuo zenme zenme le Trans. Mother fxxker, saying that all the good men are gone, or that you're too tired to love. And saying that how you live by yourself.	0	1	1	1	1
Zh: 整个 一副 我 有 b 我 有理 的 样子 , 自己 也 不 照照 镜子。 L.S. zhengge yifu wo you b wo youli de yangzi, ziji ye bu zhaozhao jingzi Trans. Acting all self-righteous because of your cunt and not even looking at yourself in the mirror. **True Labels: Sexism, Women, Hostile, and Misogyny.**	1	UNK	1	1	0
	0	UNK	1	1	1
	0	UNK	1	1	1

Figure 4: Model predictions. *Zh*, *L.S.*, and *Trans.* are Chinese, transliteration (Latin script), and Translation. The labels and the model predictions are arranged in the same order as in the main results.

Parallel Study

We conduct two sets of experiments utilizing our bilingual data translation for parallel data. Detailed settings and results analysis are elaborated in the Appendix. The results indicate a promising avenue in data augmentation and support previous findings that balanced and larger amounts of data would improve performance. However, the discrepancy in linguistic features remains challenging in detecting phrasing, misogyny, and misandry.

Case Study

We provide two wrongly predicted Chinese examples in Figure 4. The first example contains sexist words but is a historical narrative, which does not convey personal opinions but is mistakenly predicted as sexism by all models.

The second example only mentions “men”, which “misleads” all MLMs to predict the target gender is men and all LLMs output ”Unknown”. Although they almost correctly predict hostility and misogyny, most of them also predict it as misandry and even non-sexism. Since the models for each label are trained separately, there is a gap in consistent predictions among the whole set of labels, which will be solved in future research.

Implications

The theoretical implications are twofold: first, it delves into the nuances of sexism, particularly in distinguishing between hostile and benevolent sexism. Second, the inclusion of misandry expands the framework, emphasizing the necessity for a comprehensive detection of sexism that goes beyond a single gender. From a practical perspective, by refining annotation criteria and labels, BeyondGender enables more accurate and nuanced detection of harmful discrimination, such as sneaky misogyny, while minimizing false positives. It has significant implications for the development of models and algorithms for automated detection of sexism and large language model alignment, ultimately contributing to the creation of safer and more inclusive environments.

Conclusion

In this paper, we present BeyondGender, a high-quality large-scale bilingual dataset designed for practical sexism detection. We provide comprehensive information on the annotation guidelines and dataset statistics, as well as a comparison of the sexist culture represented in English and Chinese data. In addition, we evaluate the capabilities of masked language models and large language models in detecting sexism, target gender, phrasing manner, misogyny, and misandry. Through a detailed analysis, we shed light on the challenges in identifying misogyny and misandry. Through parallel study, we find data augmentation is a promising solution. For future work, we aim to delve deeper into these challenges and explore potential strategies for enhancing the performance of sexism detection models. Additionally, we plan to expand the scope of our dataset to include more diverse modalities and cultural contexts, thereby enriching the resources available for research.

Ethical Statement

BeyondGender is developed with the aim of improving the distinction between actual sexism and gender-related discussions, as well as between innocuous stereotypes and sneaky sexist ideologies. It is sourced from a combination of previous public datasets and social media, with no personal information collected during this process. The annotation process incorporates perspectives from both male and female annotators to reduce the potential for gender bias. The dataset primarily represents data from recent decades and does not necessarily reflect the historical or future trends in sexism.

BeyondGender is intended solely for academic research purposes and will be made publicly available. We are not responsible for any potential breaches and misuse by others.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of GuangDong 2023A1515012922,

the Shenzhen Foundational Research Funding JCYJ20220818102415032, the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

Xuan Luo and Jing Li were partly supported by the Research Grants Council of the Hong Kong Special Administrative Region (No. PolyU/25200821), the Innovation and Technology Fund (No. PRP/047/22FX), the NSFC Young Scientists Fund (No. 62006203), PolyU Research Centre on Data Science and Artificial Intelligence (No. 1-CE1E), and a gift fund from Huawei Noah's Ark Lab.

References

Al-Hassan, A.; and Al-Dossari, H. 2022. Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 28(6): 1963–1974.

Almanea, D.; and Poesio, M. 2022. ArMIS - The Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements. In *Proceedings of the Language Resources and Evaluation Conference*, 2282–2291. Marseille, France: European Language Resources Association.

Bertaglia, T.; Bartekova, K.; Jongma, R.; et al. 2023. Sexism in Focus: An Annotated Dataset of YouTube Comments for Gender Bias Research. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, 22—28. New York, USA: Association for Computing Machinery. ISBN 9798400702259.

Bhattacharya, S.; Singh, S.; Kumar, R.; et al. 2020. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 158–168. Marseille, France: European Language Resources Association. ISBN 979-10-95546-56-6.

Chiril, P.; Moriceau, V.; Benamara, F.; et al. 2020. An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1397–1403. Marseille, France: European Language Resources Association.

Cowie, L. J.; Greaves, L. M.; and Sibley, C. G. 2019. Sexuality and sexism: Differences in ambivalent sexism across gender and sexual identity. *Personality and Individual Differences*, 148: 85–89.

Cui, Y.; Che, W.; Liu, T.; et al. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Cui, Y.; Che, W.; Liu, T.; et al. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 657–668. Online: Association for Computational Linguistics.

Dafaure, M. 2022. Memes, trolls and the manosphere: mapping the manifold expressions of antifeminism and misogyny online. *European Journal of English Studies*, 26(2): 236–254.

De Pelle, R. P.; and Moreira, V. P. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*.

Devlin, J.; Chang, M.-W.; Lee, K.; et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, Z.; Qian, Y.; Liu, X.; et al. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335. Dublin, Ireland: Association for Computational Linguistics.

El Ansari, O.; Jihad, Z.; and Hajar, M. 2020. A dataset to support sexist content detection in arabic text. In *Image and Signal Processing: 9th International Conference, ICISP 2020*, 130–137. Berlin, Heidelberg: Springer.

Fersini, E.; Nozza, D.; Rosso, P.; et al. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR workshop proceedings*, volume 2263, 1–9. CEUR-WS.

Fersini, E.; Rosso, P.; Anzovino, M.; et al. 2018b. Overview of the task on automatic misogyny identification at IberEval 2018. In *CEUR Workshop Proceedings*, volume 2150, 214–228. CEUR-WS.

Glick, P.; and Fiske, S. T. 1996. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3): 491–512.

Grosz, D.; and Conde-Cespedes, P. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops*, 104–115. Berlin, Heidelberg: Springer.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021a. An Expert Annotated Dataset for the Detection of Online Misogyny. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1336–1350. Online: Association for Computational Linguistics.

Guest, E.; Vidgen, B.; Mittos, A.; et al. 2021b. 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*, 1336–1350. Online: Association for Computational Linguistics.

He, P.; Gao, J.; and Chen, W. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of The Eleventh International Conference on Learning Representations*.

Höfels, D. C.; Çöltekin, Ç.; and Mădroane, I. D. 2022. CoRoSeOf-an annotated corpus of Romanian sexist and offensive tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2269–2281. Marseille, France: European Language Resources Association.

Holland, J. 2012. *A brief history of misogyny: The world's oldest prejudice*. London: Hachette UK.

Jha, A.; and Mamidi, R. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16. Vancouver, Canada: Association for Computational Linguistics.

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

Jiang, A.; and Zubiaga, A. 2023. SexWEs: Domain-Aware Word Embeddings via Cross-Lingual Semantic Specialisation for Chinese Sexism Detection in Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 447–458. Limassol, Cyprus: Association for the Advancement of Artificial Intelligence.

Kirk, H.; Yin, W.; Vidgen, B.; et al. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2193–2210. Toronto, Canada: Association for Computational Linguistics.

Krenn, B.; Petrak, J.; Kubina, M.; and Burger, C. 2024. GERMS-AT: A Sexism/Misogyny Dataset of Forum Comments from an Austrian Online Newspaper. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 7728–7739. Torino, Italia: ELRA and ICCL.

Liu, Y.; Ott, M.; Goyal, N.; et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mabrouk, D. M. M. 2020. The Dilemma of Toxic Masculinity in Eastern and Western Societies; With Reference to the Novel "Men in Prison". *Open Journal of Social Sciences*, 8: 419–437.

Mulki, H.; and Ghanem, B. 2021. Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 154–163.

OpenAI. 2022. ChatGPT3.5. https://openai.com/blog/chatgpt. Accessed: 2023-12-30.

Parikh, P.; Abburi, H.; Badjatiya, P.; et al. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1642–1652. Hong Kong, China: Association for Computational Linguistics.

Priyadharshini, R.; Chakravarthi, B. R.; Cn, S.; Durairaj, T.; Subramanian, M.; Shanmugavadivel, K.; U Hegde, S.; and Kumaresan, P. 2022. Overview of Abusive Comment Detection in Tamil-ACL 2022. In Chakravarthi, B. R.; Priyadharshini, R.; Madasamy, A. K.; Krishnamurthy, P.; Sherly, E.; and Mahesan, S., eds., *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, 292–298. Dublin, Ireland: Association for Computational Linguistics.

Rizwan, H.; Shakeel, M. H.; and Karim, A. 2020. Hate-speech and offensive language detection in roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language processing*, 2512–2522. Online: Association for Computational Linguistics.

Rizzi, G.; Gasparini, F.; Saibene, A.; et al. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5): 103474.

Samory, M.; Sen, I.; Kohne, J.; et al. 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, 573–584. Online: Association for the Advancement of Artificial Intelligence.

Sharifirad, S.; and Matwin, S. 2019. When a Tweet is Actually Sexist. A more Comprehensive Classification of Different Online Harassment Categories and The Challenges in NLP. *ArXiv preprint arXiv:1902.10584*.

Sultana, S. 2022. Identifying Sexism and Misogyny in Pull Request Comments. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–3. New York, NY, USA: Association for Computing Machinery.

Sultana, S.; Sarker, J.; and Bosu, A. 2021. A Rubric to Identify Misogynistic and Sexist Texts from Software Developer Communications. In *Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–6. New York, NY, USA: Association for Computing Machinery.

Taori, R.; Gulrajani, I.; Zhang, T.; et al. 2023. Stanford Alpaca: An Instruction-following LLaMA model.

Touvron, H.; Martin, L.; Stone, K.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics.

Yang, A.; Xiao, B.; Wang, B.; et al. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305*.

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