SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection

Aiqi Jiang[1], Xiaohan Yang[2], Yang Liu[1], Arkaitz Zubiaga[1]

[1] *Mile End Road, London E1 4NS,* [2] *Wheatley Campus, Oxford OX33 1HX*

[1]Queen Mary University of London, [2]Oxford Brookes University

Abstract

Online sexism has become an increasing concern in social media platforms as it has affected the healthy development of the Internet and can have negative effects in society. While research in the sexism detection domain is growing, most of this research focuses on English as the language and on Twitter as the platform. Our objective here is to broaden the scope of this research by considering the Chinese language on Sina Weibo. We propose the first Chinese sexism dataset – Sina Weibo Sexism Review (SWSR) dataset –, as well as a large Chinese lexicon SexHateLex made of abusive and gender-related terms. We introduce our data collection and annotation process, and provide an exploratory analysis of the dataset characteristics to validate its quality and to show how sexism is manifested in Chinese. The SWSR dataset provides labels at different levels of granularity including (i) sexism or non-sexism, (ii) sexism category and (iii) target type, which can be exploited, among others, for building computational methods to identify and investigate finer-grained gender-related abusive language. We conduct experiments for the three sexism classification tasks making use of state-of-the-art machine learning models. Our results show competitive performance, providing a benchmark for sexism detection in the Chinese language, as well as an error analysis highlighting open challenges need-

[*] This is to indicate the corresponding author.
Email address: `a.jiang@qmul.ac.uk` ([1]Queen Mary University of London)

ing more research in Chinese NLP. The SWSR dataset and SexHateLex lexicon are publicly available[1].

Keywords: Chinese Sexism Dataset, Sexism Detection, Hate Speech Detection, Abusive Language Detection, Chinese Sexist Lexicon, Natural Language Processing

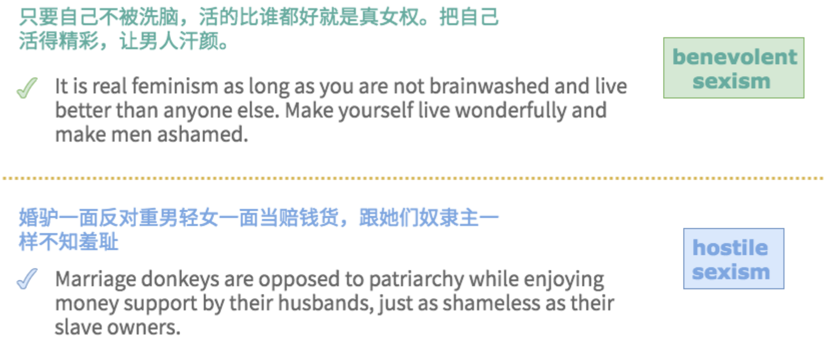
1. Introduction

Along with an unprecedented ability for communication and information sharing, social media platforms provide an anonymous environment which allows users to take aggressive attitudes towards specific groups or individuals by posting abusive language [1]. This leads to increasing occurrences of incidents, hostile behaviours and remarks of harassment, especially for online interactions between people of different genders, nationalities, ethnicities, cultures and physical appearances [2, 3, 4, 5, 6]. Hate speech is one of the most important conceptual categories in anti-oppression politics today [7, 8], referring to using the language to incite violence or to promote hatred against particular groups of people, or to attack, to insult or to disparage members of a group on the basis of specific characteristics [9]. Sexism is a common pattern of hate speech and is currently considered as a deteriorating factor in social networks in China [10, 11, 12]. Sex is a sensitive topic in Asian cultures, hence many women still have a high cognitive and tolerance threshold for hostile gender-biased behaviours [12], which consequently aggravates abusive remarks and violent behaviours online. The task of mitigating hate speech online has attracted the attention of Chinese industries, such as Sina Weibo, to impose strict censorship on the contents of relevant topics [13], but has remain largely understudied in academic research.

In the past few years, due to the increasing amount of user-generated content and the diversity of user behaviour towards women in social media, manual

[1] http://doi.org/10.5281/zenodo.4773875

inspection and moderation of sexist contents becomes unmanageable. The academic community has seen a rapid increase in research tackling the automatic detection of misogynous behaviour and gender-based hatred in both monolingual and multilingual scenarios [14, 15]. The first attempt was by Hewitt et al. [16] who investigated the manual classification of misogynous tweets, and the first survey of automatic misogyny identification in social media was conducted by Anzovino et al. [17]. Nozza et al. [18] attempted to measure and mitigate unintended bias in machine learning models for misogyny detection. An extensive of misogyny detection is then conducted especially in multilingual and cross-domain scenarios [19].



只要自己不被洗脑，活的比谁都好就是真女权。把自己活得精彩，让男人汗颜。

✓ It is real feminism as long as you are not brainwashed and live better than anyone else. Make yourself live wonderfully and make men ashamed.

婚驴一面反对重男轻女一面当赔钱货，跟她们奴隶主一样不知羞耻

✓ Marriage donkeys are opposed to patriarchy while enjoying money support by their husbands, just as shameless as their slave owners.

benevolent sexism
hostile sexism

Figure 1: Examples of hostile and benevolent sexism.

However, misogyny is not always equivalent to sexism, and frequently implies the expression of hostility and hatred against women [19]. As for sexism, Glick and Fiske [20] define the concept of sexism referring to two forms of sexism: hostile sexism and benevolent sexism. Hostile sexism is characterised by an explicitly negative attitude towards women, while benevolent sexism is more subtle with seemingly positive characteristics (see examples in Figure 1). Sexism includes a wide range of behaviours (such as stereotyping, ideological issues, sexual violence, etc.) [21, 17], and may be expressed in different ways: direct, indirect, descriptive or reported [22, 5]. Thus, misogyny is only one case of sexism [21]. Most previous studies concentrate more on detecting hostile and explicit sexism, overlooking subtle or implicit expressions of sexism [10, 17, 11, 19]. Hence, dealing with the detection of sexism in a wide spectrum of sexist attitudes and behaviours is necessary as these are, in fact, the most frequent and dangerous for society [23].

Most relevant studies for identifying online abusive content against women utilise supervised approaches, and recently, deep learning approaches have become more popular, especially transformer-based approaches, which have made state-of-the-art achievements in different languages [24, 15, 5, 25]. Since these approaches for automatic sexism detection are usually established utilising labeled training data, the performance is more dependent on the quality and taxonomy of the available datasets [26]. Some existing studies have made effort to construct sexism-related datasets instead of only collecting explicit misogyny from diverse social platforms in multiple languages (such as English [27], Spanish [15] and French [28]), aiming to improve the task performance of detecting online sexist behaviour in a broad sense. One of the limitations for those approaches is not taking any additional domain knowledge into consideration, like linguistic information from the domain-specific lexicon [29]. Recently, several works have demonstrated the positive influence of infusing external domain knowledge on hate speech detection, a broader research field than sexism detection for online abuse[30], but there is still a lack of more relevant research.

When it comes to sexism-related datasets and resources, however, most efforts have been made for Indo-European languages [10, 3, 31, 32, 33], while the development of Chinese sexism identification is hindered due to the lack of Chinese annotated resources and Chinese sexism-related lexicons. Moreover, the creation of such resources poses several challenges when it comes to data collection and annotation, especially with the diversity of Chinese dialects and the ambiguity brought about by emerging Internet language.

This paper aims to investigate how diverse behaviors, beliefs and attitudes towards women are expressed in social media, and to focus on collecting data resources about sexism in Chinese. Given the modest presence of Chinese content and geographical access restrictions on Twitter, here we focus on the most prevalent microblogging platform in China, Sina Weibo. As a platform integrating the major features of Twitter, Facebook, and Instagram, users of Sina Weibo can share posts (weibos) with texts, photos, and videos, which can trigger replies between users (comments) and endorsement (likes) from others [34, 35]. In this paper, we make the first effort in creating a sexism dataset in Chinese, to the best of our knowledge. By using Sina Weibo to collect sexism-related weibos and comments, we build, annotate and analyse the Sina Weibo Sexism Review (SWSR) dataset. The SWSR dataset consists of two parts: *SexWeibo* and *SexComment*, both of which include the textual content of posts along with anonymised information of users, number of likes and other metadata. The process led to a dataset with 1,527 weibos and 8,969 comments. In addition, with the aim of assisting research in detection and analysis of sexist comments in Chinese, we provide a sexism-related offensive lexicon SexHateLex which aggregates and extends existing lexical resources in Chinese. Furthermore, we present the first experimentation in Chinese sexism detection to provide a benchmark, including the implementation of various machine learning and deep learning methods. Our experiments and methodology for sexism detection aims to further research in this task in Chinese, as well as enables similar research efforts in other types of hate speech detection in Chinese. Our Chinese dataset and lexicon also enable multilingual sexism research which break the restriction of limited language resources. Abundant demographic and Weibo-based features in SWSR empower to exploit relevant studies on online abusive language in different aspects.

1.1. Objectives

To address the problem of the scarcity of Chinese resources in the field of hate speech especially for gender-related content in social media, we focus on the following research objectives:

- *Objective 1*: To define a methodology for the collection and annotation of Chinese online sexism at different levels of granularity, involving both explicitly hostile sexism and implicitly subtle sexism. This research builds on and adapts existing annotation guidelines for other languages, providing the first such effort in Chinese in sexism and hate speech.

- *Objective 2*: To evaluate the effectiveness of existing state-of-the-art models on detecting Chinese sexist content.

- *Objective 3*: To create a Chinese sexist lexicon to assist research in detection and analysis of Chinese sexist content and assess the influence of external lexical knowledge on the model performance.

1.2. Contribution

The main novel contributions of this paper are summarised as follows:

1. We construct and release the first Chinese sexism dataset to our knowledge. The rich features of our SWSR dataset including weibo contents, weibo reviews and basic user information make it possible to detect sexist content with various approaches for better performance and interpretability, as well as enables contextual analysis of sexism.
2. We further provide labels for the sexism category and the type of target of sexist comments, which enables finer-grained investigation of sexist texts.
3. We integrate existing lexical resources and sexism-related terms to build a lexicon including 3,016 sexist and abusive terms, which can support research on Chinese abusive language.
4. We perform an exploratory analysis to validate the quality of the dataset and to understand how sexism is manifested in Chinese.
5. We present preliminary experiments along with an analysis of the results, establishing a benchmark for Chinese sexism detection.

1.3. Paper Structure

This paper is organised as follows. We introduce related work on this task in Section 2, starting with several previous studies for existing sexism-related datasets. We also present recent work and resources focusing on hate-related lexicons. In Section 3 we describe the process of collecting and organising source data from Sina Weibo. Section 4 presents guidelines and evaluation of three annotation tasks for the collected dataset. The procedure of building a sexist lexicon is introduced in Section 5. Then we describe experimental results and analysis for sexism detection in Section 7. Section 9 discusses potential areas of research enabled by our dataset and lexical resources. Section 10 briefly discusses that our work adheres to the 'FAIR' facets. Finally, we present conclusive remarks for our work in Section 11.

2. Related Work

Since no previous datasets in Chinese exist for online abusive language targeting gender groups, we discuss recent sexism-related datasets in non-Chinese language in the literature, and review existing lexical resources in Chinese relevant to sexism and online abuse.

2.1. Existing Non-Chinese Datasets for Sexism Detection

The last few years have witnessed an increase in the interest in and availability of sexism datasets. We provide a summary of the existing sexism datasets in Table 1. The earliest attempt was that by Waseem and Hovy [10], who provided a publicly available dataset of more than 16k tweets for hate speech and annotate it into three categories - racism, sexism and neither. However, it only comprises the expression of hostile sexism towards women, overlooking other kinds of sexism. Chowdhury et al. [36] aggregate experiences of sexual abuse to facilitate a better understanding of social media construction and to bring about social change. These two datasets consist of content in English.

In addition, recent sexism datasets include multilingual content involving Italian, Spanish and Hindi, along with English. The Automatic Misogyny Identification (AMI) competitions in Evalita 2018 [3], Ibereval 2018 [31] and Evalita 2020 [4] provide datasets in English, Spanish and Italian to detect misogynistic content, to classify misogynous behaviour as well as to identify the target of a misogynous text. HatEval@SemEval 2019 [37] is another competition aiming to

Table 1: Existing sexism-related datasets in multiple languages.

Dataset	**Language**	**Offense Type**	**Label**	**Instance**	**Year**	**Ref**
Waseem&Hovy	English	Racism Sexism	racism, sexism neither	16k	2016	[10]
Jha&Mamidi	English	Ambivalent sexism	benevolent hostile, others	22142	2017	[27]

AMI@Evalita	English
	Italian
		Misogyny	misogynous
			not misogynous
				5000(EN)
				5000(IT)
					2018	[3]

AMI@IberEval	English
	Spanish
		Misogyny	misogynous
			not misogynous
				3977(EN)
				4138(ES)
					2018	[31]

Chowdhury et al.	English	Sexual harassment	recollection not recollection	5119	2019	[36]

HatEval	English
	Spanish
		Against immigrants and women	non-hateful
			hateful
				13000(EN)
				6600(ES)
				9091(immigrants)
				10509(women)
					2019	[37]

Parikh et al.	English	Sexism	23 categories of sexism	13k	2019	[24]

TRAC-2	English
	Hindi
	Bangla
		Misogyny
		Aggression
			GEN(Gendered)
			NGEN(Non-gendered)
				25k	2020	[32]

AMI@Evalita	Italian	Misogyny	misogynous not misogynous	7000	2020	[4]
Chiril et al.	French	Sexism	direct, descriptive reporting, non-sexist no decision	12k	2020	[28]
MeTwo	Spanish	Sexism	sexist, not-sexist doubtful	3600	2020	[15]

EXIST@IberLEF	English
	Spanish
		Sexism	sexist
			not sexist
				5644(EN)
				5741(ES)
					2021	-

LeT-Mi	Arabic	Misogyny	misogynistic non-misogynistic	6550	2021	[33]
Guest et al.	English	Misogyny	misogynistic non-misogynistic	6567	2021	[38]
ArMI@HASOC	Arabic	Misogyny	misogynistic non-misogynistic	9833	2021	-
Samory et al.	English	Sexism	benevolent hostile, other callme, scale	16k	2021	[25]
Bajer	Danish	Misogyny	misogyny racism, others	27.9k	2021	[39]

detect hate speech against immigrants and women and further finer-grained features in offensive text, like aggressive attitude and the target harassed in English and Spanish posts from Twitter. Furthermore, Parikh et al. introduce a dataset consisting of accounts of sexism in 23 categories to investigate sexism categorisation as a multi-label classification task [24]. Bhattacharya et al. develope a multilingual annotated corpus of misogyny and aggression in Indian English, Hindi, and Indian Bangla as part of a project studying and automatically identifying misogyny and communalism in social media [32]. The first French dataset [28] and Spanish dataset (MeTwo) [15] have been released for sexism detection, and EXIST@IberLEF 2021[2] proposes the first shared task on sexism identification in social networks (as opposed to misogyny detection), aiming to detect online sexism in English and Spanish. Moreover, Hala and Bilal [33] introduce the first Arabic Levantine dataset for online Misogyny (LeT-Mi) written in the Arabic and Levantine dialect. Then ArMI@HASOC 2021 at FIRE[3] proposes an Arabic Misogyny Identification (ArMI) task with two sub-tasks derived from the Let-Mi dataset[33], which is the first shared task to address the problem of automatic detection of Arabic online misogyny. Guest et al. [38] introduce an expert annotated misogynous dataset collected from Reddit and present a new detailed hierarchical taxonomy for online misogyny, while Zeinert et al. develop the first Danish misogyny dataset, Bajer, under a four-level taxonomy of labels [39]. Besides, Samory et al. provide a sexism dataset using psychological scales and generating adversarial samples to improve construct validity and reliability in sexism detection [25].

Despite the increasing availability of sexism datasets, in an increasing number of Indo-European languages, no dataset exists in the Chinese language [40]. Likewise, we are not aware of previous research in sexism detection in Chinese. To fill this gap, our research here documents our efforts in creating the first such dataset, including Chinese social media posts labelled as sexist or not. For

[2] http://nlp.uned.es/exist2021

[3] https://sites.google.com/view/armi2021/

sexist posts, we annotate the sexism category and target type as well to support deeper investigation on sexism identification.

2.2. Existing Lexical Resources for Online Abuse

Detection of offensive content can be challenging as it not always contains explicit mentions of negative or hateful words [41, 1]. However, there is evidence showing that the use of domain-specific lexical words in classification models can boost model performance [42, 43, 30]. With the expectation that the use of a lexicon can make for a good proxy to improve detection of hate speech, here we develop one in Chinese to support our research in sexism detection. There are many popular lexicons for online abuse, which collect and organise offensive words and phrases. For example, [44] focus on several lists obtained from Wikipedia that are particularly linked to a specific sub-type of hate speech in English, such as ethnic slurs[4] and LGBT slang terms[5]. A popular hate speech lexicon is HateBase,[6] which provides the largest multilingual hate speech lexicon linked to aspects such as religion, gender and ethnicity. It includes 3,635 groups of terms in more than 95 languages [45]. Despite its volume for languages like English, the HateBase lexicon only contains 39 Chinese terms, which is still far from becoming a referential resource. Besides, [46] built a multilingual hate speech lexicon, HurtLex[7], involving over 50 languages. HurtLex provides a larger set of 4,251 terms in Chinese.

However, there is no relevant study for the Chinese sexism scenario, and only a few Chinese lexical resources are designed for offensive language. None of those resources focuses specifically on gender-related contents. Given the scarcity of sexism-specific lexicons as well as the strong relation between those phenomena of offensive language and sexist language [19], we aggregate and expand existing Chinese lexicons to build a large Chinese lexicon consisting of terms that can

[4] https://en.wikipedia.org/wiki/List_of_ethnic_slurs

[5] https://en.wikipedia.org/wiki/List_of_LGBT_slang_terms

[6] https://hatebase.org/

[7] https://github.com/valeriobasile/hurtlex

be generally associated with abusive language as well as gender-specific terms, which can assist by furthering research in Chinese sexism detection.

3. Data Collection

In describing our data collection process, we first describe the key characteristics of the Sina Weibo microblogging platform we use to build our SWSR dataset, discussing the different data harvesting options across the different weibo platforms. Then we delve into the data collection and filtering process.

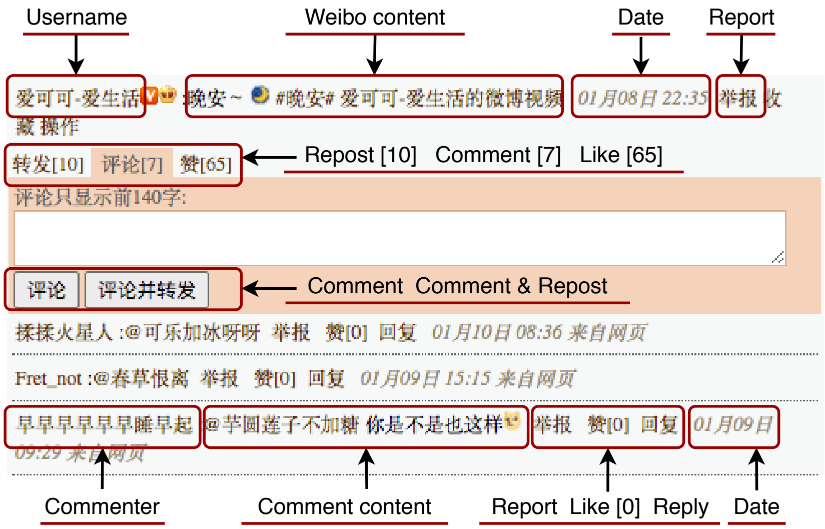
3.1. Sina Weibo

Sina Weibo is the largest microblogging service in China, which has some unique characteristics with respect to Twitter. It is aimed for information sharing, dissemination and information acquisition based on user relationships [34]. Content on Sina Weibo is spread through the 'following-follower' networks established between people [47], for example, allowing users to post comments on someone's Weibo or to reply to other people's comments on someone's Weibo. It allows users to insert images, videos, music, long articles and polls.

Sina Weibo has three main ways of accessing its website, namely weibo.com, weibo.cn and m.weibo.com. We can access Sina Weibo via PC terminal through weibo.com and weibo.cn, and the mobile counterpart is m.weibo.com. The weibo.com is more complex than weibo.cn because its Weibo page presents a richer functionality with more components which weibo.cn doesn't have, such as Top Topic Ranking, Hot Movie Recommendation, advertisements, etc. However, we can see in an example of weibo.cn in Figure 2 that the website structure is simple and straightforward. Both the weibo and its associated comment list can be easily retrieved and parsed for data collection. So we finally decide to use **weibo.cn** as the source website of Sina Weibo.

3.2. Data Collection and Processing

As described above, a Sina Weibo timeline comprises posts (weibos) which receive replies (comments). Initially, we use keyword-driven method to collect



爱可可-爱生活 :晚安~ #晚安# 爱可可-爱生活的微博视频 01月08日 22:35 举报 收藏 操作

转发[10] 评论[7] 赞[65]

评论只显示前140字:

揉揉火星人 :@可乐加冰呀呀 举报 赞[0] 回复 01月10日 08:36 来自网页

Fret_not :@春草恨离 举报 赞[0] 回复 01月09日 15:15 来自网页

早早早早早早睡早起 @芋圆莲子不加糖 你是不是也这样 举报 赞[0] 回复 01月09日 09:29 来自网页

Figure 2: An example of Sina Weibo on weibo.cn

a set of weibos, for which we then collect the associated comments. While the collection of weibos is restricted to those containing the keywords, our focus on the associated comments allows us more flexibility, retrieving content which need not contain the seed keywords. Figure 3 shows an overview of the data collection process, which we introduce further details in the steps below. Our SWSR dataset therefore is made of two tables for weibo and comment data along with some anonymised user information pertaining to the weibos and comments. This user information includes features such as user gender and user location. All personally identifiable information is removed and not disclosed, including user names and mentions.

3.2.1. Step I: Extract Weibo Data

To construct our dataset, we use keyword-driven search to collect gender-related weibos from Sina Weibo platform (weibo.cn). In terms of relevance to the topic and through manual exploration [23, 31], we firstly determine to use seven different keywords related to some hot topics and events of sexism for weibo data collection, namely 婊子(bitch), 女同性恋(lesbian), 女权(feminism),

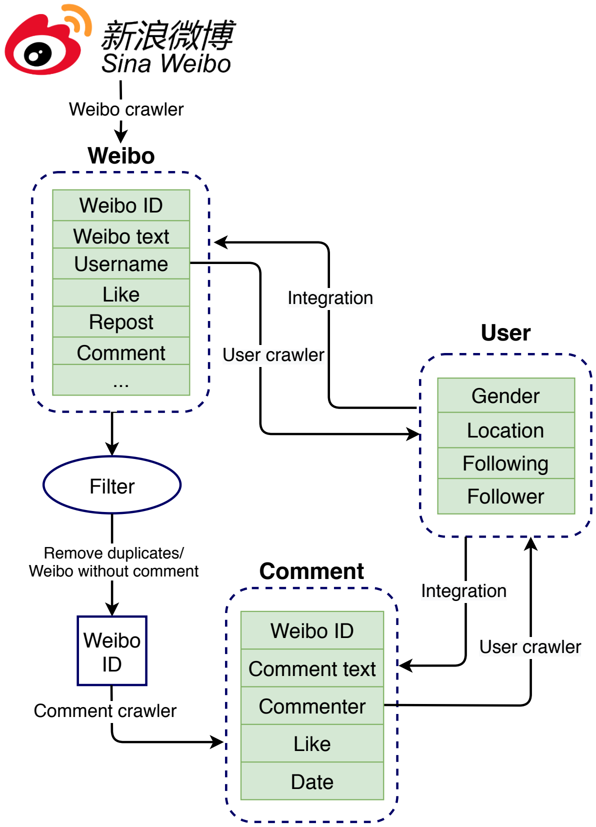


Figure 3: Overview of the data collection process.

厌女(misogyny), metoo运动(metoo movement), 性别歧视(gender discrimination) and 性骚扰(sexual harassment). Then we search and extract weibos containing these keywords. In addition, we retrieve user profiles, which include self-reported values such as gender and location, and other variables such as number of followers. To protect user privacy in the dataset, usernames are anonymised by replacing them with a special token <username>. Then we combine these features into the weibo. The number of weibos collected for each keyword is listed in Table 2, which amounted to a total of 9,087 weibos collected for all keywords. Data collection was limited to posts made between June 2015 to June 2020.

Table 2: Number of weibos collected for each keyword.

Keyword	**Translation**	**Number of Weibos**	**Total**
婊子	bitch	407	**9087**
女同性恋	lesbian	520	
女权	feminism	2255	
厌女	misogyny	1757	
metoo运动	metoo movement	1340	
性别歧视	gender discrimination	1366	
性骚扰	sexual harassment	1442	

3.2.2. Step II: Process Weibo Data

In this step, we process the collected weibos prior to collecting the associated comments in subsequent steps. We remove the weibos that match at least one of the following criteria:

- weibos without any comments. This can be easily done by checking the number of comments for each weibo according to 'weibo_comment' column.
- duplicates which are exact matches of both the 'weibo_id' and 'weibo_text' columns, i.e. weibos collected repeatedly across keywords. We only keep one of these repeated instances.

This led to a final set of 3,856 weibos, along with their associated weibo IDs which we use in the next step to retrieve comments.

3.2.3. Step III: Extract Comment Data

In order to extract comments for the collected weibos, we utilise their weibo ID. This enabled us collection of textual content and metadata of weibos, including user profiles of commenters. This led to the collection of 31,677 comments for the 3,856 weibos.

3.2.4. Step IV: Process Comment Data

For processing the comments collected in the previous step, we remove comments matching at least one of the following criteria:

- Remove duplicate comment texts, keeping only one instance. This is caused by users who copy and paste the same comment repeatedly.
- Remove short comments with commonly identified patterns – fixed tokens on Sina Weibo, e.g. comments solely containing the word '转发' (repost), '回复' (reply) or '举报' (report).
- Remove the remaining short comments (length of less than 5 characters).
- Remove comments without any Chinese character.

Given that users occasionally reply by splitting their texts into multiple comments, we aggregate them. When we find multiple comments from the same user in close temporal proximity, we automatically aggregate them into a single comment.

Finally, we convert all the comments from traditional Chinese to simplified Chinese, which helps ensuring consistency while keeping the same information. We use the Python package chinese_converter[8] to achieve this.

This led to a final set of 8,969 comments linked to 1,527 weibos, whose statistics are shown in Table 5. The final aim of our sexist data collection lies in the retrieval of these comments, which are the ones that we annotate and make up the final dataset. The weibos are solely considered to support the annotation process and, if desired, for context-based analysis of comments.

3.3. Ethics of Data Collection

Due to limitations on the number of weibos that can be crawled at one time and the continuous changes of the Sina Weibo API, we directly obtain the weibo contents via web scraping by using a Python script. Hence, we carefully consider the ethical implications behind the collected data. Posts and comments collected in this dataset are in the public domain and web scraping has been done only for research purposes. Hence, we ensure that no ethics approval is

[8] `https://pypi.org/project/chinese-converter/`

needed for this study [48] and the collected dataset follows acceptable ethical practices by adhering to the following:

- Our dataset does not present any personally identifiable information, as we have anonymised all user names in the dataset, including any user names mentioned in the posts (replaced by the special token <username>)
- Our dataset does not include any private messages between users, and there was no interaction between Weibo users and researchers.
- We rely on publicly available data and carefully collect the data into multiple steps to avoid overloading Sina Weibo servers.
- The Sina Weibo server is publicly accessible.

4. Data Annotation

During the annotation process of our SWSR dataset, we perform three annotation tasks as follows:

1. **Sexism Identification:** whether a text is sexist, as a binary annotation task determining if a comment is sexist (1) or non-sexist (0). Where a comment is deemed sexist, we also perform two additional annotations:
2. **Sexism Category:** We define four categories of sexism, namely stereotype based on appearance (SA), stereotype based on cultural background (SCB), microaggression (MA) and sexual offense (SO).
3. **Target Type:** individual (I) or generic (G).

4.1. Annotation Preparation

In order to reliably identify sexism as well as their corresponding categories and targets, we provide initial annotation guidelines for all three tasks. The annotation guidelines for sexism identification are based on [10, 36], and guidelines for the sexism category and the target type are adapted from [31, 3, 33, 23]. Guidelines were iteratively developed through collective annotation of a small sample of 100 comments by a broader set of five annotators. These annotators met and discussed disagreements between them, which led to revised guidelines.

In most cases, we find that our disagreement for annotation task I was mainly caused by the lack of sufficient context when identifying sexist content. For example, one annotator marks the text 它们的大脑平滑到可以在上面溜冰，真的不是一个物种啊[9] as not sexism because there is no sexist content towards women. But when we check the original Weibo text, we find that 'they' in this text is intended by its author to mean 'some stupid women who insult men for more benefits'. So it should be marked as sexism with consideration of the context. Another common case of disagreement is the misunderstanding of specific words related to sexism. These words commonly appear in sexist text but are not common in general speech. Some annotators did not realise that 婚驴(marriage donkey) is an offensive word specifically towards women. People that use this word have the intention to depict the image of 'women who are as stupid as donkeys in marriage, deprived of a lot of benefits, but still enjoy silly happiness'. Discussions following these agreements led to revisions in the guidelines and improvements in subsequent rounds of annotations. In addition, for the annotation task II determining the sexism category, there were disagreements caused by occasional overlaps in the interpretations of the different labels, which were resolved and led to revision of the guidelines. Annotation III consisting in determining the target type was more straightforward as being easier to label.

In what follows, we reproduce initial guidelines used for the three annotation tasks, which enable annotators to have a better understanding of sexist issues for three annotation tasks and to a large extent improve the final score of inner-annotator agreement.

[9] Translation: Their brains are so smooth that they can skate on them. We are really not the same species

4.2. Annotation Guidelines

Given the difficulty of identifying sexist behaviours, we carefully crafted guidelines for the three annotations tasks based on the insights from the above annotation testing: sexism identification, sexism category and target category, along with examples of annotations by sexism category and target category shown in Table 3.

4.2.1. Annotation I: Sexism Identification

A comment is considered sexist if it belongs to at least one of the following categories:

- explicitly attacks or insults gender groups or individuals using sexist language.
- incites gender-based violence or promote sexist hatred but not directly use a sexual abusive language.
- abuses those who attack or have negative attitudes towards a gender group.
- shows support of problematic incidents or intentions of sexual assault, sexual orientation and sexual harassment.
- negatively stereotypes gender groups by describing physical appeal, oversimplifying image or expressing superiority of men over women.
- expresses underlying gender bias in a sarcastic or tacit way.

The rest of the texts are considered non-sexist. This includes neutral descriptions or testimonies of sex-related events or phenomena.

4.2.2. Annotation II: Sexism Category

Each of the comments marked as sexist in the first task needs to be classified into one of the following, determining the sexism category of the comment:

- *Stereotype based on Appearance (SA):* describes physical appeal, oversimplifies image, or makes comparison with narrow/vulgar standards towards a gender group.

- *Stereotype based on Cultural Background (SCB):* expresses opinions indicating the superiority of men over women and emphasises gender inequality under the concept of a patriarchal society.

- *Microaggression (MA):* intentionally or unintentionally expresses hostile, derogatory or negative attitudes or remarks against gender groups or individuals.

- *Sexual Offense (SO):* incites sexual-related behaviour or attitude against women, such as sexual harassment, sexual assault, rape and violence.

4.2.3. Annotation III: Target Category

Each of the comments marked as sexist in the first task needs to have the type of target identified, which can be one of the following two:

- *Individual (I):* a post with sexist content addressing a specific person.

- *Generic (G):* a post with sexist content addressing a broader group (such as a gender-based group of people).

4.3. Annotator Agreement

All three annotations were performed independently by three annotators, all of them PhD students, including two females and one male. We use the open source text annotation tool *doccano*[10] to facilitate the annotation work and to enable independent annotation effectively by three annotators.

We report inter-annotator agreement rates for the three annotators by using Cohen's kappa as a metric [49]. The inter-annotator agreement of our annotation

[10] https://github.com/doccano/doccano

Table 3: Examples of sexism categories and target types in the dataset.

Example	Translation	Sexism Category	Target
前任的漂亮更清纯甜美一看就是正经人，现在这位一看就很肉的感觉	His ex looks more innocent and beautiful, like a decent person. But the appearance of his current girlfriend makes me a higher libido.	SA	I
还是让女性做些带孩子，鼓励丈夫的工作！	We should let women do more housework, and encourage their husbands' work!	SCB	G
关键是有些女生还没子宫道德，结了婚脑子里自动长了个	The point is that some girls have no uterine morals. There is a dick in their head after they get married.	MA	G
你全家女性送来给我搞一搞，我戴套，保证安全	Send your family's women to me to fuck them, I will wear a condom to ensure safety	SO	I

task I is overall 82.3% (71.8% for the sexist class and 96.1% for non-sexist). For annotation tasks II and III, the inner-annotator agreements reach 76.8% and 85.5% respectively. All these agreement rates can be deemed substantial agreements between the three annotators. Examples of annotations by sexism category and target category are shown in Table 3.

5. Lexicon Collection

We build a large sexism lexicon SexHateLex by aggregating and expanding existing resources, which is a combination of

- profane words and slang,
- sexual abusive words and slang, and
- sexism-related people, websites and events.

SexHateLex is built by integrating four existing lexicons, and augmented by adding typos and synonyms based on integrated sexual-related abusive terms. We aggregate the following lexical resources:

- *Chinese Profanity in Wikipedia*[11]*:* Wikipedia provides a list of Chinese profane words linked to sex, race and sexual orientation. For our purposes we chose the 599 terms for sex and sexual orientation.
- *HateBase*[12]*:* HateBase is the world's largest structured repository of regionalised multilingual hate speech corpora in the field of religion, gender, nationality, ethnicity, etc. We collect 29 Chinese terms from HateBase.
- *TOCP dataset*[13]*:* TOCP is the largest Chinese profanity dataset including 16,450 sentences [50]. All profane words and corresponding locations in each sentence have been labelled in this dataset. A total of 1,014 profane words are extracted.
- *Sexy Lexicon*[14]*:* The repository funNLP provides massive resources to support research in Chinese NLP, one of which is a sexy lexical list in the category of sensitive term datasets. We collect 1,240 terms from this list.

After integrating terms from all the resources above, we get a total of 2,109 terms. Then we combine typo words that users make spell mistakes in the text based on a spell checking method in the 'aion' python package[15], and add the top 5 similar words to each word in the collected lexical list. Word embeddings[16] are leveraged for this step, followed by cleaning all duplicate and incorrect terms. This leads to the final SexHateLex lexicon with 3,016 terms.

6. Data Description

We describe the resulting dataset by first presenting the dataset structure and by then providing descriptive statistics of the dataset.

Table 4: Description of features in the weibo and comment datasets.

Table	Feature
SexWeibo	weibo_id, weibo_text, keyword, user_gender, user_location, user_follower, user_following, weibo_like, weibo_repost, weibo_comment, weibo_date
SexComment	weibo_id, comment_text, gender, location, like, date, label, category, target

6.1. Dataset Structure

The SWSR dataset is organised in two files: *SexWeibo.csv* (SexWeibo) and *SexComment.csv* (SexComment), containing weibos (posts) and comments (replies) respectively. Contents in these two files can be linked through the *weibo_id*. We list all features in *SexWeibo.csv* and *SexComment.csv* files in Table 4 (see more details in Appendix A). Considering the user privacy, all user names in this dataset are anonymous with a special token <username>.

6.2. Dataset Statistics

Table 5: Statistics of the dataset.

	All	Sexist	Non-Sexist
All	8969	3093 (34.5%)	5876 (65.5%)
Average length per comment	71.45	90.34	61.51
Number of comment per weibo	5.87	3.77	4.69

The resulting 8,969 comments are associated with 1,527 weibos. Table 5 shows the statistics of the dataset in terms of the distribution of sexist com-

[11] `https://en.wikipedia.org/wiki/Mandarin_Chinese_profanity#Sex`

[12] https://hatebase.org/

[13] http://nlp.cse.ntou.edu.tw/resources/TOCP/

[14] https://github.com/fighting41love/funNLP/tree/master/data

[15] https://github.com/makcedward/nlp/tree/master/aion

[16] https://fasttext.cc/

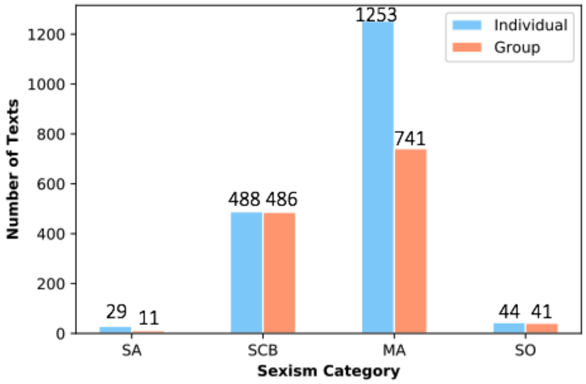


Figure 4: Distribution of sexism categories and target types in the dataset.

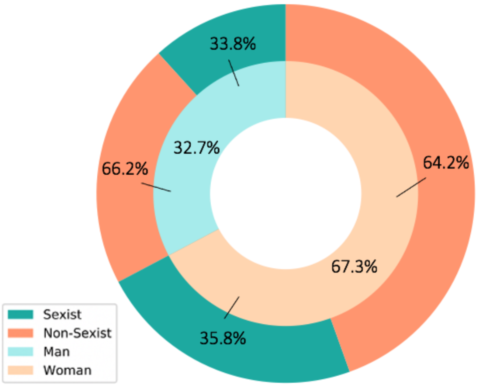


Figure 5: Distribution of user gender across two classes in the dataset.

ments, comment length and number of comments per weibo. We can see that the majority of comments are non-sexist, with nearly twice as many as sexist comments.

Figure 4 depicts the distribution of the sexism category and target type in sexist comments. More than half of the sexist comments are MA, and SCB also takes a large proportion in the sexist class. Besides, the number of comments towards individuals nearly double those towards groups, where sexist texts in the MA category are more frequently abusive towards individuals.

6.2.1. Textual Distribution

We compute the average lengths (in a number of characters) of comments in each category. We see big differences in Table 5 showing that the average length of a sexist comment is 50% bigger than the length of a non-sexist comment. Furthermore, Table 5 presents the averaged number of comments for each weibo. We can see that the number of comments per weibo for both sexist and non-sexist classes are less than that for all data, which is because that one weibo might contain multiple comments in different classes. Hence, the sum of weibo counts for two classes can be larger than the overall number of weibos.

6.2.2. Gender Distribution

According to Figure 5, we find that the gender distribution in sexism is skewed towards women while the probabilities of men and women to send sexist posts are similar. As some weibos are more relevant to topic keywords like misogyny and feminism, female users are more likely to attack or insult males who make malicious remarks or show unequal attitudes towards women.

6.2.3. Word Frequency Distribution

We normalise the data by removing stop words, special markers such as '转发' (Repost), user names, and punctuation marks. Then we select the list of 12 words with the highest frequency in the comments as well as the top 12 words from the SexHateLex lexicon which are most frequent in the comments. We find that the terms frequently occurring in each class differ significantly (see Table 6). The most frequent tokens in the lexicon present negative emotional attitudes while those in the comments are mostly neutral words related to gender topics.

7. Preliminary Experiments: Sexism Detection

To assess the difficulty of computationally detecting sexist comments in SWSR and to provide benchmark experimental results, we conduct both coarse-grained and fine-grained sexism detection experiments, evaluating different features and models. Our experiments are designed in three steps:

Table 6: Description of the 12 most frequent terms in the dataset (DataTerm) and in the lexicon (LexTerm). [尸吊] is a sensitive character which cannot be found in the Latex package. The table presents the character by dividing it into two parts, which can be easily understood in Chinese. PCT denotes the percentage of each term.

DataTerm	**Translation**	**PCT**	**LexTerm**	**Translation**	**PCT**
女权	feminism	29.84%	骂	curse	7.45%
女性	women	25.20%	死	die	2.89%
不是	not	19.04%	搞	flirt	2.75%
男人	man	11.92%	女拳	negative feminism	2.20%
孩子	children	8.78%	歧视	discrimination	2.04%
骂	curse	7.45%	驴	donkey	1.88%
男权	patriarchal	6.14%	[尸吊]	dick	1.78%
极端	extreme	5.90%	逼	pussy	1.45%
结婚	marry	5.26%	强奸	rape	1.44%
姓	surname	5.26%	狗	dog(similar use as pig)	1.23%
权利	right	3.89%	干	fuck	1.08%
平等	equality	3.73%	蛆	maggot	0.89%

1. Sexism identification (Binary): weibo contents are classified as either sexist or non-sexist.
2. Sexism category classification (Multi-class): texts are classified into one of five categories: stereo-type based on appearance (SA), stereotype based on cultural background(SCB), microaggression (MA), sexual offense (SO), or non-sexist.
3. Target classification (Multi-class): texts are classified into either generic, individual, or non-sexist.

7.1. Models

For the three experimental steps, we test various models. As context-based models we utilise different BERT-based models [51] based on transformers. We use three different BERT-based models: (1) BERT, (2) BERT with whole word mask (Bert-wwm), and (3) RoBERTa [52]. Besides, we adopt three different baselines using combinations of unigrams to trigrams as features: (1) a logistic regression (LR), (2) a support vector machine (SVM), and (3) a character-level LR. We also test two content-based models, a CNN and a character-level CNN [53] with FastText word embeddings.

In addition, for the experimental step 1, we test all the models above with and without lexical words from the SexHateLex lexicon, to show its impact on the task. We first count the occurrence of each word, and then convert the count vector from the count frequency to term frequency–inverse document frequency (TF-IDF) [54], indicating how significant a category is to a text in the corpus. Finally, we concatenate the TF-IDF lexical vector with textual embeddings. For BERT-based models, we concatenate lexical embeddings with the output of BERT, and then feed them into a feedforward layer for final classification.

7.2. Experiment Settings

Given that the SWSR dataset is not balanced, especially in the category classification task, we randomly split the comment data into 90% for training and 10% for testing using stratified sampling. Class distribution in the training set includes 34.7% sexist texts and 65.3% non-sexist texts. We perform cross validation experiments on the training data to fine-tune model hyperparameters, choosing the best models for the final experiments. We report global macro F1 and accuracy scores for the three tasks, as well as F1 scores specific to each class for experimental step 1 and weighted F1 scores for steps 2 and 3.

7.3. Experiment Results and Analysis

From the results in Table 7, we see that content-based models (CNN) outperform linguistic ones (LR / SVM) in both word level and character level while context-based models (BERT) perform best. Character-level models (e.g. char-LR / char-CNN) show better performance than word-level models (e.g. LR / CNN), proving them more suitable for a language like Chinese with no space between words. When we incorporate lexical features, most models lead to slight improvements of 0.5-1% in F1 score (with the exception of LR and BERT

Table 7: Sexism detection performance. F1-Sex and F1-Not denote F1 scores respectively for binary labels of sexist or non-sexist. mF1 denotes macro F1 score and Acc denotes accuracy score.

Model	**Original Feature**				**+Lexicon**			
	F1-Sex	**F1-Not**	**mF1**	**Acc**	**F1-Sex**	**F1-Not**	**mF1**	**Acc**
LR + ngram	0.624	0.849	0.737	0.785	0.616	0.846	0.731	0.780
Char-LR + ngram	0.640	0.852	0.746	0.790	0.646	**0.858**	0.752	0.797
SVM + ngram	0.633	0.844	0.739	0.781	0.640	0.842	0.741	0.786
CNN + ft	0.669	0.828	0.749	0.774	0.654	0.844	0.749	0.785
Char-CNN + ft	0.660	0.845	0.753	0.787	0.654	0.850	0.752	0.790
Bert	**0.694**	**0.858**	**0.776**	**0.806**	0.661	0.844	0.752	0.786
Bert-wwm	0.678	0.846	0.762	0.792	0.699	0.851	0.775	0.800
RoBerta	0.685	0.844	0.764	0.792	**0.707**	0.853	**0.780**	**0.804**

models), showing the potential of SexHateLex in improving performance, particularly with the best-performing model RoBERTa. We also observe an overall tendency for achieving 15-23% better prediction on the non-sexist category, highlighting the challenge of detecting sexist comments.

Regarding the category classification task, the results in Table 8 show a different scenario. The best performing model is RoBERTa, with highest weighted and F1 scores, but all three BERT-based models have better performance than others. For the third task, the results in Table 8 show that all the models achieve a competitive performance without a large margin, while RoBERTa performs best across other models. Besides, it can be observed that macro F1 scores for both task 2 and 3 show an averaged lower than weighted F1 scores, which indicates a potential impact of the imbalanced nature of the data among the finer-grained classes. More sampling methods are supposed to be considered before training.

8. Discussion

8.1. Error Analysis

We look at frequent errors across misclassified instances generated from SVM, CNN and BERT, three typical models selected from three types of models

Table 8: Results for the sexism category and target classification tasks. mF1 denotes macro F1 score and wF1 denotes weighted F1 score. Acc denotes accuracy score.

Model	**Category classification**			**Target classification**		
	wF1	**mF1**	**Acc**	**wF1**	**mF1**	**Acc**
LR + ngram	0.628	0.310	0.611	0.663	0.447	0.719
Char-LR + ngram	0.648	0.316	0.646	0.657	0.428	0.721
SVM + ngram	0.647	0.320	0.692	0.661	0.446	0.707
CNN + ft	0.711	0.335	0.716	0.668	0.447	0.711
Char-CNN + ft	0.722	0.347	0.730	0.670	0.448	0.714
Bert	0.732	0.355	**0.736**	0.678	0.457	0.713
Bert-wwm	0.732	0.354	**0.736**	0.682	0.462	0.720
RoBerta	**0.734**	**0.360**	0.732	**0.687**	**0.467**	**0.727**

Table 9: Error analysis for misclassified examples. TL denotes true label and PL denotes predicted label.

Error Type	**Example**	**Translation**	**TL**	**PL**
(1)	如果她自己够优秀就不会在网络上怨天尤人了	If she is excellent enough, she won't blame others on the Internet	1	0
(2)	你这种金针菇明码标价了也只会烂在货架上	Enoki mushrooms like yours will only rot on the shelf even if they are clearly marked	1	0
(3)	田园女权，女拳师，极端女权，是我是我都是我	Pastoral feminist, female boxer, extreme feminist, it's all me	0	1

we used in the experiment step 1 (see Table 9 for examples). Several typical errors appeared in the experiments are summarised below:

(a) Implicit sexism: Errors in those posts lacking explicit sexist expression or context, and most frequent reason for (a) in the misclassified texts is caused by sarcastic expressions. Sarcasm seems to be a suitable way for expressing contempt and subtly offending individuals, which modifies the perception of message, hindering the correct detection of sexism by automatic systems [55]. Example (1) is a sarcastic comment that criticises women who are not successful but insults those people who uphold gender equality. It is difficult to identify sexism when there is no explicit presence of abusive language. Another problem is that the model cannot pick up words with a specific meaning related to gender.

(b) Lack of prior information: It demonstrates that the model cannot identify those contents referring to sexism-related event, people or words/phrases with special meanings as it does not possess prior knowledge. In example (2), 金针菇(enoki mushroom) is a very harmful word specifically towards men associated with some physical characteristics but cannot be directly identified by the model.

(c) Overuse of sexist words: It indicates that sexist words might be overused in one text, leading to the over-dependence of the model on these words, while sexist targets in posts are confounding and hard to be identified. We can see from example (3) that the model can easily identify a text with many sexist words as a sexist text even if there is no specific targeted individual or group attacked by someone.

8.2. External Knowledge Induction

After infusing external domain information to models, most of them present a slight increase in the final performance. We conjecture on a set of factors that may be affecting the performance of using the lexicon:

- **Dataset variety:** The lexical terms found in the randomly split training and test sets might be imbalanced. There may be a certain gap in the quantity of lexical terms extracted in the proportion of the training and test sets, leading to the diverse degree of the influence of lexical terms in the process of the model classification.

- **Term inconsistency between dataset and lexicon:** Terms in the dataset and the lexicon could be inconsistent. The domain-specific lexicon might not be capable of covering all sexism-related terms encountered across datasets.

- **Linguistic characteristics:** Not all posts containing hateful terms are sexist necessarily, due to cases of polysemy or negation.

- **Humour, irony and sarcasm:** Sexist posts with humour, irony and sarcasm are implicit and difficult to be identified, and may contain no explicit hate-related terms.
- **Spelling variation:** Spelling variation is prevalent in social media [56]. Sensitive words sometimes use spelling variations to obfuscate and avoid detection, which do not match those normative words in the lexicon. A certain Chinese character in a word (e.g.绿茶婊→angelic bitch[17]) is often replaced by homophones or pinyin (e.g. 绿茶婊→绿茶表/绿茶biao), or is split into radicals according to the composition rules of Chinese characters (e.g. 绿茶婊→绿茶女表).
- **Quality of lexical features:** TF-IDF frequency features captured from the category of lexical terms might be comparatively sparse and lose information for specific terms. Lexical embeddings derived from pre-trained word embedding models could be beneficial as high-quality word embeddings can be learned efficiently thanks to low space and time complexity [57].
- **Approaches for lexicon induction:** Since the approach for lexicon induction might not fully absorb lexical information by simple concatenation between textual hidden features and lexical features, other forms of fusion can be tested, such as matrix multiplication [58] and cosine similarity [59].

9. Research Applications

The SWSR dataset and the SexHateLex lexicon provide resources for furthering research in a new language in the growing research problem of sexist language. We discuss potential areas of research.

[17] 绿茶婊(angelic bitch) means girls who pretend to be pure and innocent but in fact are manipulative and scheming.

9.1. User-based Sexism Detection

As sexism-related speech belonging to user-generated content online, some investigations are conducted to find out the potential influence of user characteristics like gender and location on sexism detection [10]. User metadata in SWSR, such as gender, location and number of followings, can enable researchers to explore possible correlations between gender-based hateful content and user profiles, furthering user-based studies in the area of sexism detection.

9.2. Explainable Sexism Detection

Providing explanations can make model outputs more convincing and understandable [60, 61]. We provide our dataset with two basic classes to show which text is sexist or not, with fine-grained labels to support furthering detection. Besides, we offer a lexicon composed of abusive words to support detection of offensive content with sexism-specific features.

9.3. Multi-lingual and Cross-lingual Sexism Detection

While most approaches to sexism detection have been proposed for English, other studies have been investigated to deal with this task in other languages such as Spanish, Italian, and Indian, thanks to recent shared tasks [31, 3, 62]. More research is needed in other languages, including Chinese, both in multilingual settings, i.e. proposing models that deal with multiple languages, and cross-lingual settings, i.e. leveraging data in a resource-rich language like English for application in lesser-resourced languages such as Chinese. Our dataset compensates for the lack of sexist speech in Chinese, thereby facilitating the development of sexism identification research in multi-lingual and cross-lingual settings.

9.4. Cross-domain Hate Speech Detection

With the prevalence of identifying hate speech online, some studies concentrate on detecting specific types of hate speech, such as racism or sexism. This differences across types of hate speech make it more challenging to generalise hate speech detection models. Cross-domain detection of hate speech thereby has been a topic of interest to identify common features between distinct hate speech domains, achieving knowledge transfer and model generalization. Our dataset provides gender-related hateful texts with corresponding topic-related keywords, which could enhance research on sexism and facilitate potential research of cross-domain detection in this and other types of hate speech, particularly if additional Chinese hate speech datasets are released.

9.5. Other applications

While most existing research on sexism detection focus on detecting the text to binary classes (sexist or not), our dataset enables investigation of additional, finer-grained perspectives of sexism, thanks to three types of labels provided. Categorising sexism by type as well as identifying the type of targets enable furthering research in sexism detection beyond the widely-studied binary classification task.

10. FAIR

In this section, we show that the SWSR dataset is collected and organised based on 'FAIR' facets: Findable, Accessible, Interoperable and Re-usable. Our dataset is publicly available through Zenodo and can be downloaded completely by using the following citation:

Aiqi Jiang, Xiaohan Yang, Yang Liu, & Arkaitz Zubiaga. (2021). SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection [Dataset]. Zenodo. http://doi.org/10.5281/zenodo.4773875

The dataset files are provided in CSV format for the SWSR dataset and TXT format for the SexHateLex lexicon. A README file is included to explain each file in detail to facilitate the re-use of the dataset.

11. Conclusion

In this paper, we release a comprehensive sexism dataset SWSR along with a large lexicon SexHateLex, to facilitate research on online gender-based speech in Chinese. To the best of our knowledge, this is the first sexism dataset in Chinese. The dataset provides both weibo and comment texts, as well as three types of labels, namely sexist or not, sexism category and target type. The dataset contains two files for *SexComment* and *SexWeibo*, containing sexist comments, original weibos enabling contextual analysis, and anonymised user metadata. We further conduct exploratory analyses of the dataset. Different types of sexism detection approaches are also evaluated on *SexComment*. We experiment with baseline models for sexism detection, which provides a benchmark for further experimentation. We expect our dataset to enable further research in Chinese sexism detection, including a set of possible directions.

12. Acknowledgements

Aiqi Jiang is funded by China Scholarship Council (CSC). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT http://doi.org/10.5281/zenodo.438045. Besides, we are grateful to Peiling Yi, Xia Zeng, Xiaoyu Guo and Wenjie Yin for their assistance discussing the annotation guidelines. We also thank anonymous reviewers provided helpful comments on earlier drafts of the manuscript.

References

[1] P. Fortuna, J. Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?, Information Processing & Management 58 (3) (2021) 102524. doi:https://doi.org/10.1016/j.ipm.2021.102524.
URL https://www.sciencedirect.com/science/article/pii/S0306457321000339

[2] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[3] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami)., in: EVALITA@ CLiC-it, 2018.

[4] E. Fersini, D. Nozza, P. Rosso, Ami@ evalita2020: Automatic misogyny identification, Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org.

[5] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4055–4066. doi:10.18653/v1/2020.acl-main.373.
URL https://www.aclweb.org/anthology/2020.acl-main.373

[6] E. W. Pamungkas, V. Basile, V. Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, Information Processing & Management 58 (4) (2021) 102544. doi:https://doi.org/10.1016/j.ipm.2021.102544.
URL https://www.sciencedirect.com/science/article/pii/S0306457321000510

[7] I. Gagliardone, D. Gal, T. Alves, G. Martinez, Countering online hate speech, Unesco Publishing, 2015.

[8] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, The British Journal of Criminology 60 (1) (2019) 93–117. arXiv:https://academic.oup.com/bjc/article-pdf/60/1/93/31634412/azz049.pdf, doi:10.1093/bjc/azz049.
URL https://doi.org/10.1093/bjc/azz049

[9] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (4) (2018) 85.

[10] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. `doi:10.18653/v1/N16-2013`.
URL `https://www.aclweb.org/anthology/N16-2013`

[11] S. Frenda, B. Ghanem, M. Montes-y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, Journal of Intelligent & Fuzzy Systems 36 (5) (2019) 4743–4752.

[12] X. Shi, Y. Zheng, Perception and tolerance of sexual harassment: An examination of feminist identity, sexism, and gender roles in a sample of chinese working women, Psychology of Women Quarterly 44 (2) (2020) 217–233. `doi:10.1177/0361684320903683`.
URL `https://doi.org/10.1177/0361684320903683`

[13] K. M. DeLuca, E. Brunner, Y. Sun, Weibo, wechat, and the transformative events of environmental activism on china's wild public screens., International Journal of Communication 10.

[14] A. Jha, R. Mamidi, 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, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. `doi:10.18653/v1/W17-2902`.
URL `https://www.aclweb.org/anthology/W17-2902`

[15] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[16] S. Hewitt, T. Tiropanis, C. Bokhove, The problem of identifying misogynist language on twitter (and other online social spaces), in: Proceedings of the 8th ACM Conference on Web Science, 2016, pp. 333–335.

[17] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 57–64.

[18] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 149–155. doi:10.1145/3350546.3352512.
URL https://doi.org/10.1145/3350546.3352512

[19] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (6) (2020) 102360. doi:https://doi.org/10.1016/j.ipm.2020.102360.
URL https://www.sciencedirect.com/science/article/pii/S0306457320308554

[20] P. Glick, S. T. Fiske, Ambivalent sexism, in: Advances in experimental social psychology, Vol. 33, Elsevier, 2001, pp. 115–188.

[21] K. Manne, Down girl: The logic of misogyny, Oxford University Press, 2017.

[22] M. Hellinger, A. Pauwels, 21. language and sexism, in: Handbook of language and communication: Diversity and change, De Gruyter Mouton, 2008, pp. 651–684.

[23] L. Richardson-Self, Woman-hating: On misogyny, sexism, and hate speech, Hypatia 33 (2) (2018) 256–272.

[24] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1642–1652. doi:10.18653/v1/D19-1174.
URL https://www.aclweb.org/anthology/D19-1174

[25] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples, Proceedings of the International AAAI Conference on Web and Social Media 15 (1) (2021) 573–584.
URL https://ojs.aaai.org/index.php/ICWSM/article/view/18085

[26] S. Kiritchenko, I. Nejadgholi, K. C. Fraser, Confronting abusive language online: A survey from the ethical and human rights perspective, arXiv preprint arXiv:2012.12305.

[27] A. Jha, R. Mamidi, 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, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. doi:10.18653/v1/W17-2902.
URL https://www.aclweb.org/anthology/W17-2902

[28] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in French tweets, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1397–1403.
URL https://www.aclweb.org/anthology/2020.lrec-1.175

[29] M. Wiegand, J. Ruppenhofer, A. Schmidt, C. Greenberg, Inducing a lexicon

of abusive words – a feature-based approach, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1046–1056. doi:10.18653/v1/N18-1095.
URL https://www.aclweb.org/anthology/N18-1095

[30] A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti, HurtBERT: Incorporating lexical features with BERT for the detection of abusive language, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 34–43. doi:10.18653/v1/2020.alw-1.5.
URL https://www.aclweb.org/anthology/2020.alw-1.5

[31] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., in: IberEval@ SEPLN, 2018, pp. 214–228.

[32] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 158–168.
URL https://www.aclweb.org/anthology/2020.trac-1.25

[33] H. Mulki, B. Ghanem, Let-mi: An Arabic Levantine Twitter dataset for misogynistic language, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 154–163.
URL https://www.aclweb.org/anthology/2021.wanlp-1.16

[34] Wikipedia, Sina weibo — wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Sina_Weibo, accessed: 2021-01-03.

[35] SinaFinance, Sina weibo monthly active users reach 550 million, revenue exceeds wall street expectations, https://finance.sina.com.cn/stock/usstock/c/2020-05-19/doc-iircuyvi3963989.shtml, accessed: 2021-01.

[36] A. Ghosh Chowdhury, R. Sawhney, R. R. Shah, D. Mahata, #YouToo? detection of personal recollections of sexual harassment on social media, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2527–2537. doi:10.18653/v1/P19-1241.
URL https://www.aclweb.org/anthology/P19-1241

[37] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. doi:10.18653/v1/S19-2007.
URL https://www.aclweb.org/anthology/S19-2007

[38] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, 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, Association for Computational Linguistics, Online, 2021, pp. 1336–1350.
URL https://www.aclweb.org/anthology/2021.eacl-main.114

[39] P. Zeinert, N. Inie, L. Derczynski, Annotating online misogyny, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, in press 2021.

[40] B. Vidgen, L. Derczynski, Directions in abusive language training data,

a systematic review: Garbage in, garbage out, Plos one 15 (12) (2020) e0243300. doi:10.1371/journal.pone.0243300.

[41] X. Han, Y. Tsvetkov, Fortifying toxic speech detectors against disguised toxicity, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7732–7739.

[42] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. doi:10.18653/v1/W17-1101.
URL https://www.aclweb.org/anthology/W17-1101

[43] M. Mladenović, C. Krstev, J. Mitrović, R. Stanković, Using lexical resources for irony and sarcasm classification, in: Proceedings of the 8th Balkan Conference in Informatics, 2017, pp. 1–8.

[44] P. Burnap, M. L. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics, EPJ Data Science 5 (1) (2016) 11.

[45] C. Tuckwood, Hatebase: Online database of hate speech, The Sentinal Project. Available at: https://www. hatebase. org.

[46] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018, Vol. 2253, CEUR-WS, 2018, pp. 1–6.

[47] B. Huberman, D. M. Romero, F. Wu, Social networks that matter: Twitter under the microscope, First Monday.

[48] Q. Xu, Z. Shen, N. Shah, R. Cuomo, M. Cai, M. Brown, J. Li, T. Mackey, Characterizing weibo social media posts from wuhan, china during the early stages of the covid-19 pandemic: Qualitative content analysis, JMIR Public Health and Surveillance 6 (4) (2020) e24125.

[49] J. Cohen, Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit., Psychological bulletin 70 (4) (1968) 213.

[50] H. Yang, C.-J. Lin, TOCP: A dataset for Chinese profanity processing, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 6–12.
URL https://www.aclweb.org/anthology/2020.trac-1.2

[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
URL https://www.aclweb.org/anthology/N19-1423

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692.

[53] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. doi:10.3115/v1/D14-1181.
URL https://www.aclweb.org/anthology/D14-1181

[54] L.-P. Jing, H.-K. Huang, H.-B. Shi, Improved feature selection approach tfidf in text mining, in: Proceedings. International Conference on Machine Learning and Cybernetics, Vol. 2, IEEE, 2002, pp. 944–946.

[55] M. Thomae, A. Pina, Sexist humor and social identity: the role of sexist humor in men's in-group cohesion, sexual harassment, rape proclivity, and victim blame, HUMOR 28 (2) (2015) 187–204. doi:doi:10.1515/

humor-2015-0023.
URL https://doi.org/10.1515/humor-2015-0023

[56] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts, Challenges and frontiers in abusive content detection, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 80–93.

[57] Y. Goldberg, O. Levy, word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, arXiv preprint arXiv:1402.3722.

[58] N. Pappas, J. Henderson, Gile: A generalized input-label embedding for text classification, Transactions of the Association for Computational Linguistics 7 (0) (2019) 139–155.
URL https://transacl.org/ojs/index.php/tacl/article/view/1550

[59] X. Li, J. Song, W. Liu, Label-attentive hierarchical attention network for text classification, in: Proceedings of the 2020 5th International Conference on Big Data and Computing, ICBDC 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 90–96. doi:10.1145/3404687.3404706.
URL https://doi.org/10.1145/3404687.3404706

[60] C. Molnar, Interpretable Machine Learning, 2019, https://christophm.github.io/interpretable-ml-book/.

[61] E. Dai, Y. Sun, S. Wang, Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, 2020, pp. 853–862.

[62] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive

content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17.

Appendix A. SWSR Dataset Format

SWSR dataset consists of two files: ‘SexWeibo.csv’ and ‘SexComment.csv’, containing weibos (posts) and comments (replies) respectively. See more detailed description of features below:

Appendix A.1. SexWeibo.csv

- weibo_id: a string of weibo ID
- weibo_text: a string of weibo content
- keyword: contains sexism-related keyword(s) extracted from the weibo text
- user_gender: the gender of user
- user_location: the location of user
- user_follower: number of users who follow this user’s account
- user_following: number of users whom this user follows
- weibo_like: number of like for the weibo
- weibo_comment: number of comment for the weibo
- weibo_repost: number of repost for the weibo
- weibo_date: the date and time when the weibo is posted

Appendix A.2. SexComment.csv

- weibo_id: the weibo id where the comment is collected
- comment_text: a string of the comment
- gender: the gender of commenter

- location: the location of commenter
- like: number of like for this comment
- date: the date and time when the comment is posted
- label: the comment is sexist(1) or non-sexist(0)
- category: categorise sexism into four classes – Stereotype based on Appearance(SA), Stereotype based on Cultural Background (SCB), MicroAggression (MA) and Sexual Offense (SO)
- target: the type of target who are attacked – Individual (I) or Generic (G)