Understanding the Shades of Sexism in Popular TV Series

Nayeon Lee, Yejin Bang, Jamin Shin and Pascale Fung
Human Language Technology Center (HLTC)
Center for Artificial Intelligence Research (CAiRE)
Hong Kong University of Science and Technology
[nyleeaa,yjbang,jay.shin].connect.ust.hk, pascale@ece.ust.hk

Abstract

In the midst of a generation widely exposed to and influenced by media entertainment, the NLP research community has shown relatively little attention on the sexist comments in popular TV series. To understand sexism in TV series, we propose a way of collecting distant supervision dataset using Character Persona information with the psychological theories on sexism. We assume that sexist characters from TV shows are more prone to making sexist comments when talking about women, and show that this hypothesis is valid through experiment. Finally, we conduct an interesting analysis on popular TV show characters and successfully identify different shades of sexism that is often overlooked.

1 Introduction

Popular TV series such as *The Big Bang Theory* and *How I Met Your Mother* constantly influence the viewers through long and continuous exposure to several seasons of episodes. However, some recent works (Sap et al., 2017; Google, 2017) expose the prevalence of *sexism* in such multimedia contents.

In order to analyze the inherent sexism, there has been much attention given in recognizing sexism in Tweets collected with certain hash-tags (Waseem and Hovy, 2016; Park et al., 2018; Jha and Mamidi, 2017) via machine-learning approaches.

However, most sexism detection datasets are biased towards neutral gender-identity terms such as woman, she, and female. In the Twitter dataset proposed by (Waseem and Hovy, 2016), there are more than 50% of sexist samples contain these terms, yet only 10% non-sexist samples do. Thus, machine-learning models tend to predict a sentence to be sexist just by including these neutral terms (Park et al., 2018). Meanwhile, most of the existing models simply view sexist detection as a binary classification problem, i.e., sexist or non-sexist. They mix three dimensions within sexism, which are paternalism, gender differentiation, and heterosexuality as defined in (Glick and Fiske, 1996), into one category.

This leads us to ask *whether we can collect gender-keyword-balanced data for each dimension of sexism*. We hypothesize that sexist characters from TV shows are more prone to making sexist comments when talking about a female, and vice versa for non-sexist characters. Based on this assumption, we collected a distant supervision data that is similar in gender-identity terms for both positive and negative samples. We believe such data would better allow us to analyze and understand the supposedly inherent sexism.

Our contributions are summarized as follows: 1) We collect *gender-keyword-balanced* data about women for distant supervision. The experimental results on benchmark dataset (Waseem and Hovy, 2016) show that our strategy can improve sexism detection. 2) We use the collected data to train classifiers for each dimension of sexism defined by psychologists (Glick and Fiske, 1996), which enable us to analyze the shades of sexism among characters in popular TV shows.

2 Distant Supervision Dataset

Definition of Sexism According to (Glick and Fiske, 1996), there are 3 dimensions within sexism: Paternalism (P) justifies men being authoritative, protective, and controlling over women. Gender Differentiation (G) uses biological differences between gender to justify the social distinctions. Heterosexuality (H) views women as sexual objects.

	Example Tropes	Characters
P	The Patriarch Manipulative Bastard Abusive Parents	Ross Ted Tywin
G	Men Are Better Than Women No Social Skills Lack of Empathy	Sheldon Michael
H	the Casanova Chick Magnet Lovable Sex Maniac	Howard Barney

Table 1: Examples of Character Tropes divided by dimension of sexism. Note that one character can have multiple tropes.

Dataset setting	F1
$\mathcal{D}_{Twitter}$	83.5
+$\mathcal{D}_{random}$	68.8
+$\mathcal{D}_{sexist}$	**85.0**

Table 2: Sexism classification results on the Twitter test set. The overall result improves 1.5% F1 score using $\mathcal{D}_{sexist}$ as additional distant supervision.

Data Collection To construct a gender-keyword-balanced dataset, we collect both "sexist" and "non-sexist" samples from TV scripts [1] that mention female. For cost-effectively obtaining weak-label for these sentences, we draw on the hypothesis: "the comments of sexist characters on a female are more probable for being sexist, and vice versa". Hence, we need to first identify sexist characters. To do this, we utilize the persona information called *character trope*, which is a recognizable element of a story that defines or conveys information about a character. In TVTropes.com [2], thousands of human-annotated tropes are available with descriptions and example characters. The data is collected in three steps.

Firstly, we identify a list of sexist tropes in two different ways. 1) We decide on tropes that fit our definitions of sexism. For example, "The Casanova" trope has a description that aligns with the definition of heterosexuality: "sexual predator a man who relentlessly pursues, lands, loves, and then abandons members of the opposite sex." 2) Based on the list of characters that are known to be sexist, we backtrack and identify tropes associated with them. Such list is obtained from abundantly available websites analyzing sexism in TV series (DeMaria, 2019). Some representative tropes for each dimension are shown in Table 1.

Secondly, we determine characters with sexist tropes based on the character-trope mappings obtained from two sources: TVTropes.com and Character Tropes Dataset (Bamman et al., 2013). We obtain the characters from popular TV series, and we focus on the main characters because they have a higher importance in the series thus more utterances. The selected characters are listed in Table 1.

Lastly, unbiased gender terms such as "she" and "girl" are used to select sentences regarding female.[3] Note that we avoid using biased gender keywords such as "bitch" and "slut".

3 Experimental Setup

Base Model As BERT (Devlin et al., 2018) has shown state-of-the-art performance in many upstream NLP tasks, we decided to use it as our base model to cope with the issue of small data prevalent in the sexism classification task. We use the base version of the pre-trained BERT model to extract feature representation, which is then fed into a linear classifier with 512 hidden dimension.

Sexist Tweets We conduct an experiment using a benchmark corpus ($\mathcal{D}_{Twitter}$) (Waseem and Hovy, 2016) to see if adding our distant supervision data helps the performance of sexism detection. We train our base-model using three different dataset settings:

i) [$\mathcal{D}_{Twitter}$ only]: Using only $\mathcal{D}_{Twitter}$ trainset.

ii) [$\mathcal{D}_{Twitter}$ + $\mathcal{D}_{random}$]: Augmenting $\mathcal{D}_{Twitter}$ with random-sampled utterances regarding women as positive sample.

iii) [$\mathcal{D}_{Twitter}$ + $\mathcal{D}_{sexist}$]: Augmenting $\mathcal{D}_{Twitter}$ with our "sexist" dataset as positive sample.

[1]List of TV series used: ['Friends', 'Gilmore Girls', 'How I met your Mother', 'Seinfeld', 'Game of Thrones', 'The Big Bang Theory', 'The Office']

[2]https://tvtropes.org/

[3]Full list of non-biased gender terms to refer women: 'she','her','woman', 'women', 'girl', 'lady', 'female', 'wife','sister','mom', 'daughter'

Characters	P	G	H	W	Examples of true positives for dominant sexism dimension
Chandler	0.108	0.296	**0.441**	0.228	*[...] before we go snooping around her crotch?*
Goerge	0.952	**0.572**	0.384	0.184	*She knew I didn't have a job, she knew I lived at home. Didn't seem to bother her. I think I could have married this woman.*
Jaqen	**1.000**	0.500	0.278	0.056	*A girl cannot tell a man when exactly he must do a thing.*
Rory	0.182	0.273	0.038	0.186	N/A

Table 3: Scores on three different dimensions of sexism **vs** Score from Sexist Tweets classifier (Waseem)

Character Analysis Based on our newly collected dataset, we try to identify and analyze shades of sexism in unseen test characters via distant supervision. For each dimension of sexism, we chose a test character based on our sexist trope list. For example, Jaqen has paternalism trope and Chandler has heterosexual trope, so we expect each of them to have a high score on the corresponding dimension.

First of all, we train three separate binary classifiers using our base model for each dimensions using the following dataset setting: i) Paternalism = $\{D_P + D_{none}\}$. ii) Gender Differentiation = $\{D_G + D_{none}\}$. iii) Heterosexuality = $\{D_H + D_{none}\}$. We trained a set of binary classifiers instead of one multi-class classifier because it is possible for a character to have multiple dimension of sexism.

Based on these classifiers, we calculate character-level *sexism score* for a set of unseen characters, where *sexism score* is the frequency count of sentences classified as sexist from the pool of sentences mentioning female. For analysis purpose, we also test the classifier trained on a Twitter dataset on the TV scripts.

4 Results

Sexist Tweets From Table 2, using the $\mathcal{D}_{sexist}$ as additional distant supervision helps improve the overall F1 score by 1.5%. The improvements in average performances are not as dramatic as individual components, but still shows improvements. On the other hand, adding $\mathcal{D}_{random}$ data drops the performance by 14.7%, which shows that randomly adding scripts will not necessarily help, rather harm the performance by injecting additional noise. Although the improvement is small, our main goal is to prove that our distant supervision strategy can at least not harm the performance, and at the same time we can gain the ability to analyze different dimensions of sexism.

Character Analysis Table 3 shows the capability of our classifiers to unveil the shades of sexism within characters. By having three separate scores for each dimension of sexism, we can better understand sexist characters. For example, we can analyze that Chandler (who is expected to be heterosexually sexist) is a sexist character, especially in a heterosexual way. We also tested on an arbitrary non-sexist character Rory, we can observe that she has low sexism scores in all dimensions. Few examples of the utterances classified as sexists are listed in Table 3.

If such sub-categorization was absent, the distinction between the level of sexism for each character will not be easily visible. For example, Chandler, George, and Rory all have relatively similar sexism scores using Twitter despite the huge difference that exists in the scores for the different sexism dimensions. In addition, we find that our classifiers tend to predict characters to be paternalism and gender differentiation sexism at the same time, for example, George and Japen have high scores in both dimensions. We infer that it is because these two dimensions are relatively similar compared to heterosexuality.

5 Conclusion

We propose a strategy to collect gender-keyword-balanced data for distant supervision. Our intuition is that sexists TV characters are prone to have sexist comments. In this way, we are able to analyze characters from popular TV shows in three different dimensions of sexism, which are usually overlooked on conventional sexist detection.

References

David Bamman, Brendan OConnor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 352–361.

Meghan DeMaria. 2019. These are the 17 most toxic male characters from your favorite tv shows, May.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Peter Glick and Susan T Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491.

Google. 2017. Using technology to address gender bias in film.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.