”Hold on honey, men at work”: A semi-supervised approach to detecting sexism in sitcoms

Smriti Singh*
Manipal Institute of Technology
smritisingh26@yahoo.com

Tanvi Anand*
Manipal Institute of Technology
tanviaanand@gmail.com

Arijit Ghosh Chowdhury
Manipal Institute of Technology
arijit10@gmail.com

Zeerak Waseem
University of Sheffield
z.w.butt@sheffield.ac.uk

Abstract

Television shows play an important role in propagating societal norms. Owing to the popularity of the situational comedy (sitcom) genre, it contributes significantly to the overall development of society. In an effort to analyze the content of television shows belonging to this genre, we present a dataset of dialogue turns from popular sitcoms annotated for the presence of sexist remarks. We train a text classification model to detect sexism using domain adaptive learning. We apply the model to our dataset to analyze the evolution of sexist content over the years. We propose a domain-specific semi-supervised architecture for the aforementioned detection of sexism. Through extensive experiments, we show that our model often yields better classification performance over generic deep learning based sentence classification that does not employ domain-specific training. We find that while sexism decreases over time on average, the proportion of sexist dialogue for the most sexist sitcom actually increases. . A quantitative analysis along with a detailed error analysis presents the case for our proposed methodology.

1 Introduction

Apart from being one of the most popular genres on television [1], sitcoms also attract the adolescent viewership[2] and thus play a vital role in the formation of their thought process (Villani, 2001). Sink and Mastro (2017) argue that documenting the prevalence and quality of television representations of women is a valuable endeavor as television depictions of women is known to influence attitudes and beliefs towards gender. Therefore, these shows would ideally contain a minimal amount of sexist content. However, according to Lee et al. (2019a) and O'Kelly (1974), this may not be the case. For this reason, we present a dataset consisting of dialogue turns labeled as either 'sexist' or 'neutral'. We also build a system that automatically detects instances of sexism present in the dialogue of popular sitcoms. Thus, we attempt to use machine learning to document the gap between activism and social change.

Often, a lack of labeled data can present a considerable challenge for text classification systems. Manual annotation often requires domain knowledge and may be expensive and time-consuming for large datasets. Manual annotation also carries the risk of introducing new annotator biases, privacy-breaches, discrimination, and misunderstanding (Chowdhury et al., 2019). Although dialogue is not the only way that sexism is constructed in TV shows (Brewington, 2019; Mouka and Saridakis, 2015), the more subtle signs of discrimination can be more difficult to detect and analyze. Our work addresses issues of manual annotation by using semi-supervised learning to generate a dataset in a new domain of pseudo-labels from unlabelled data to detect sexism in TV dialogue. This minimizes the need for a manual annotation process while creating large datasets.

We make use of a previously published dataset (Waseem and Hovy, 2016) to create a semi-supervised domain adapted classifier. In general, domain adaptation uses labeled data in one or more source domains to solve new tasks in a target domain. It is a sub-category of transfer learning. Since there is a lack of television show scripts annotated for sexism, we attempt a semi-supervised approach to develop our dataset. Here, our source domain consists of tweets from Waseem and Hovy's (2016)'s 'Hate Speech Twitter Annotations' dataset and our target domain is the dialogue in popular

[1] https://www.statista.com/statistics/1035741/most-in-demand-tv-genres-us-share/

[2] https://www.statista.com/statistics/859722/all-time-tv-shows-millennials/

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 180–185

sitcoms. These two domains are quite different. Tweets are usually short, full of abbreviations, urban slang and grammatical errors. On the other hand, sitcom dialogue turns are descriptive, long, grammatically correct and contextually dependent on the dialogue turns that precede them. These differences warrant the need for a semi-supervised approach in our methodology.

2 Related Work

In the growing body of literature on the automatic detection of sexism in text on social media, Twitter, in particular, has been the object of study and dataset creation.

Waseem and Hovy (2016) created a dataset containing Racist and Sexist tweets. Following this, there have been various efforts towards detecting sexism in English tweets (Sharifirad et al., 2019), (Jha and Mamidi, 2017). (Mishra et al., 2018). Recently, Chiril et al. (2020) developed a dataset for sexism detection in French tweets. While the study of sexism in TV shows has received little attention in natural language processing Lee et al. (2019b), Gala et al. (2020), Xu et al. (2019), it has received significant attention in the field of gender studies (Sink and Mastro, 2017; Glascock, 2003). In gender studies, Sink and Mastro (2017) conducted a quantitative analysis to document portrayals of women and men on prime-time television and Glascock (2003) examines the perception of gender roles on network prime-time television programming. To the best of our knowledge, no previous work has presented a comprehensive dataset for the presence of sexism in TV shows has been created. While efforts have been made to analyse the presence of sexism in TV shows (Nayef, 2016), the question of developing a machine learning based detection system for identifying sexism in scripted TV dialogue remains under-explored. However, Semi-supervised learning has received a lot of attention from the NLP community (Zhai et al., 2019; Xie et al., 2019; Chen et al., 2020). Our method most closely resembles Unsupervised Data Augmentation (Xie et al., 2019), which uses labeled data to annotate unlabeled samples under low resource settings.

3 Dataset

3.1 Collection

The dataset used for this experiment consists of three parts. The first part is the data used for our training dataset. We use a dataset annotated for sexist tweets Waseem and Hovy (2016). To ensure that the classifier can identify non-sexist dialogue correctly, we append $2,000$ tweets that are non-sexist in nature obtained from a web application named 'Tweet Sentiment to CSV'.[3] Before appending these neutral tweets to the dataset, they were manually checked and any tweets that were not in English were removed, along with any ambiguous tweets. To account for our target domain, we collect the dialogues from twenty sitcoms cross-referenced by popularity[4] and script availability[5]. From this set of dialogue scripts, we randomly sample $1,937$ dialogue turns to manually annotate (see subsection 3.2 for annotation guidelines). The final training set consists of $3,011$ tweets labeled as sexist, $2,000$ tweets labeled as neutral, 203 sexist dialogue turns and 926 neutral dialogue turns, henceforth denoted as D_{train}.

For the second part of the dataset, we use the un-annotated dialogue turns from the TV shows to perform semi-supervised learning. We call this dataset $D_{semisupervised}$. Out of these, ten shows aired between 1985 and 1999 (*old shows*) and ten shows aired between 2000 and 2015 (*new shows*).

The third part of our dataset, which is manually annotated and used as a held-out test set, consists of 805 manually annotated dialogues, 411 of that are labeled as neutral and 394 as sexist. This data was annotated by four annotators, achieving a Cohen's Kappa (Cohen, 1960) of 0.87.

3.2 Definition of Sexism

In this section, we describe the guidelines followed during the annotation process. The guidelines of what classifies a tweet as sexist were defined by Waseem and Hovy (2016). We use Glick and Fiske's (1996) definition of sexism to annotate dialogue turns from popular sitcoms. According to this definition, there are three primary dimensions within sexism.

- ***Paternalism:*** Paternalism justifies men being controlling, protective and authoritative over women. E.g. ''*Hold on honey, men at work.*" (Howard Wolowitz, The Big Bang Theory)

- ***Gender Differentiation:*** Gender Differentiation uses biological differences between gen-

[3]https://twitter-sentiment-csv.herokuapp.com/
[4]IMDB: https://www.imdb.com/
[5]https://subslikescript.com/series,
https://transcripts.foreverdreaming.org/

Model	Accuracy	F1	Precision	Recall	AUCROC
NB	0.772 ±0.04	0.776 ±0.02	0.778 ±0.03	0.773 ±0.02	0.765 ±0.03
RF	0.781 ±0.05	0.791 ±0.04	0.784 ±0.03	0.799 ±0.02	0.771 ±0.01
LR	0.777 ±0.02	0.780 ±0.04	0.781 ±0.03	0.779 ±0.02	0.766 ±0.03
SVM	0.783 ±0.04	0.782 ±0.04	0.773 ±0.02	0.793 ±0.03	0.783 ±0.02
BERT	0.773 ±0.04	0.742 ±0.04	0.753 ±0.02	0.713 ±0.03	0.723 ±0.02
Bi-LSTM (Ours)	**0.830 ±0.03**	**0.828 ±0.02**	**0.819 ±0.01**	**0.823 ±0.03**	**0.817±0.04**

Table 1: Results when models are trained on D_{final}. Standard errors are reported after 5 trials.

ders to justify social distinctions. An example of a sexist dialogue turn under this dimension is: *"I think women just have a lower threshold for pain than men.'* (Joey Tribbiani, Friends)

- ***Male Gaze:*** Male Gaze refers to viewing women as sexual objects. An example of a sexist dialogue turn under this dimension is: *"All men want is to see women naked."* (Jerry Seinfeld, Seinfeld)

Apart from this, we have also included dialogue turns that include derogatory terms against women (James (1998)) and dialogue turns that justify stereotypes against women or gender roles (Lauzen et al. (2008)). E.g. "See? Strong women always turn out to be nightmares" (Seinfeld) and "Look I'm sorry but some things are different for men and women." (Chandler Bing, Friends)

We find that within the annotated sexist dialogues in our held-out test set, 27.9% of the dialogues fall under gender differentiation sexism, 33.7% of the dialogues fall under paternalism and 38.4% under male gaze.

3.3 Preprocessing

The following steps were taken as a part of the preprocessing process:

- The names of the characters who said the dialogue were removed from each dialogue turn, to avoid any undue dataset bias pertaining to character names,
- Lines in the transcripts that were not dialogue turns, such as bracketed expressions to convey the settings or scenes, were removed,
- Any numbers that appeared in dialogue turns were removed,
- All words were converted to lowercase, tokenized and lemmatized.

4 Experiment Setup

We begin by training a set of models on D_{train} (section 3.1) to find the best performing model. We make use of a support vector machine (SVM), a logistic regression classifier (LR), a random forest ensemble (RF), a naive Bayes classifier (NB) , fine-tuned BERT, and a bi-directional LSTM (bi-LSTM). We find that the bi-LSTM outperforms the other models by 3.4%, with an accuracy of 76.03% on the held-out test set, D_{test}. Thus, we make use of the bi-LSTM in our proposed semi-supervised approach.

Out of the 20 sitcom show scripts we collect, we use four, namely 'Friends', 'The Big Bang Theory', 'How I Met Your Mother' and 'Seinfeld' for manual annotation (see section 3.1 for more detail). Next, we use the baseline bi-LSTM to make predictions on the other 16 show scripts. Out of these, eight are *new shows* and the other eight are *old shows*. The model classifies 1,639 dialogue turns as sexist. To form $D_{semisupervised}$, we add all dialogue turns identified as sexist by the baseline model and randomly sample 31,944 dialogue turns from the 242,108 dialogue turns identified as neutral. We combine D_{train} and $D_{semisupervised}$ to form D_{final} [6].

Finally, we train a bi-LSTM on D_{final}. We make use of the softmax activation function and the categorical cross entropy loss function while training this bi-LSTM. It consists of an embedding layer, a spatial dropout layer and makes use of the Adam optimizer, with a dropout equal to 0.2. This bi-LSTM attains an accuracy of 83.0% on D_{test}. To offer a fair comparison, we also train other competitive models on D_{final}. Table 1 demonstrates the performance of these models on D_{test} across six evaluation metrics.

To offer some insight on how the amount of sex-

[6] https://github.com/smritisingh26/HHMWdataset

Old Shows	Percentage	New Shows	Percentage
Friends	2.357%	Brooklyn Nine Nine	0.089%
Seinfeld	2.580%	The Big Bang Theory	4.131%
The Simpsons	2.611%	The Office	0.179%
Frasier	1.956%	How I met your Mother	2.343%
Full House	2.299%	Modern Family	2.267%
Everybody Loves Raymond	2.481%	Scrubs	2.168%
Home Improvement	1.956%	Parks and Recreation	1.438%
House	2.556%	New Girl	0.752%
That 70s' Show	2.369%	Two and a Half Men	3.521%
King of Queens	1.478%	Family Guy	1.865%

Table 2: Sexist content in popular sitcoms as classified by the proposed model

ist content in the form of dialogue has developed over the years, we use our proposed model to classify the dialogue turns of all twenty shows.

5 Results

5.1 Model Performance & Content Analysis

In comparing the baseline bi-directional LSTM model trained on D_{train} and the proposed model trained on D_{final}, we observe a gain of 7% in terms of accuracy on D_{test}. Similarly, for all other models, we see an average improvement of 4.67% when they are trained on D_{final}, as compared to their initial performance when they were trained on D_{train}.

The results shown in Table 1 suggest that using an augmented dataset obtained through semi-supervised learning can provide a promising avenue for addressing hate speech in distinct domains that do not have large labeled datasets available.

Furthermore, an analysis of the data labeled by our proposed model (see Table1) reveals that between 1985-1999, the average percentage of sexist dialogue turns in sitcoms is around 2.26%, whereas between 2000-2015, the mean is around 1.87% which shows an overall decrease in the number of sexist dialogue turns by 0.39%. However, it is worth noting that in the shows aired between 1985 and 1999, the show with the greatest percentage of sexist dialogue turns has 2.61% sexist dialogue turns while the proportion of sexist dialogue turns is 4.13% for the worst offender after the turn of the century. This is further complicated by the fact that the shows with the lowest amounts of sexism in the two time periods contain 1.95% and 0.08% for the old and the new shows, respectively.

5.2 Error Analysis

In an analysis of the best-performing model's performance, we identify some confounding variables:

- **Women vs that woman** Aggressively negative statements about a particular woman are marked as sexist. E.g. *"To hell with her! She left me!"* (Friends).
  While such statements may be sexist, our classifier is unable to distinguish the required nuance to make the correct prediction.

- **Sexual content** Some statements that contain extremely sexual terms are marked as sexist. For example: *"And yet you're the one always getting spanked."* (Two and a Half Men)
  This may be because a lot of sentences that contain sexual terms in the underlying datasets are sexist. For instance, dialogue turns in the training dataset like *"Well, most women want to be banged."* (How I met your Mother) and *"Sit with her, hold her, comfort her and if the moment feels right, see if you can cop a feel."* (The Big Bang Theory) are sexist.

- **Marriages** Dialogues that mention women and marriages or weddings are marked as sexist in some cases. For example: *"I know that some lucky girl is going to become Mrs. Barry Finkel."* (Friends)
  This can be attributed to a lack of contextual understanding in the classifier. Perhaps because there aren't that many dialogue turns that mention weddings or marriages.

- **Gendered pronouns for objects** In some cases, the pronoun 'she' is used to refer to objects like vehicles and boats and appear as

sexist to the classifier. For example: *"She really gets going after a while."* where 'she' refers to a car (Family guy).

6 Conclusion

We generate a labeled, real-world dataset and build a classifier using a combination of transfer learning and semi-supervised learning to classify dialogues in sitcoms as sexist or neutral for the purpose of tracking the status of social discrimination. An analysis of the recent content reveals an overall decrease in sexist content over time but an increase in the amount of sexist content in the worst offending TV shows in the recent years.

7 Acknowledgements

Zeerak Waseem has been supported in part by the Canada 150 Research Chair program and the UK-Canada AI Artificial Intelligence Initiative.

References

Morgan Brewington. 2019. Is sexism taking on new forms in movies? an ambivalent sexism perspective.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *CoRR*, abs/2004.12239.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Shah, and Debanjan Mahata. 2019. # youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. *arXiv preprint arXiv:2011.00092*.

Jack Glascock. 2003. Viewer perception of gender roles on network prime-time television. *Communication Research Reports*, 20(2):173–181.

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.

Deborah James. 1998. Gender-linked derogatory terms and their use by women and men. *American Speech*, 73(4):399–420.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data.

Martha M Lauzen, David M Dozier, and Nora Horan. 2008. Constructing gender stereotypes through social roles in prime-time television. *Journal of Broadcasting & Electronic Media*, 52(2):200–214.

Nayeon Lee, Yejin Bang, Jamin Shin, and Pascale Fung. 2019a. Understanding the shades of sexism in popular TV series. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 122–125, Florence, Italy. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Jamin Shin, and Pascale Fung. 2019b. Understanding the shades of sexism in popular TV series. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 122–125, Florence, Italy. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection.

Effie Mouka and Ioannis Saridakis. 2015. *Racism goes to the movies: A corpus-driven study of cross-linguistic racist discourse annotation and translation analysis*, pages 35–69.

Heba Nayef. 2016. Linguistic sexism in tv drama: A linguistic analysis of verbal violence against women in the egyptian sitcom al-kabeer awi. *International Journal of Linguistics*, 4(1):84–103.

Charlotte G. O'Kelly. 1974. Sexism in children's television. *Journalism Quarterly*, 51(4):722–724.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2019. How is your mood when writing sexist tweets? detecting the emotion type and intensity of emotion using natural language processing techniques. *arXiv preprint arXiv:1902.03089*.

Alexander Sink and Dana Mastro. 2017. Depictions of gender on primetime television: A quantitative content analysis. *Mass Communication and Society*, 20(1):3–22.

Susan Villani. 2001. Impact of media on children and adolescents: a 10-year review of the research. *Journal of the American Academy of child & adolescent psychiatry*, 40(4):392–401.

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.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11):e0225385.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. 2019. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485.