

Classification of Different Types of Sexist Languages on Twitter and the Gender Footprint on Each of the Classes

Sima Sharifirad ¹, Stan Matwin ², and Jack Duffy ³

¹ Faculty of computer science, Dalhousie University, s.sharifirad@dal.ca

² Faculty of computer science, Dalhousie University, stan@cs.dal.ca

³ Rowe School of Business, Dalhousie University, jack.f.duffy@gmail.com

Abstract. Sexism is very prevalent nowadays and have been attracted a lot of media’s attention. However, there is still no complimentary categories of sexism attracting natural language processing techniques. This paper proposes new categories of sexism inspired by social science work and distinguishes itself by looking at more precise categories of sexism, mentioning the challenges of labeling, the affect of gender raters in each category, publishing hashtags for collecting more sexist tweets and finally reports the classification results.

1 Introduction

Language indeed reveals the values of people and their perspectives. There has been a debate about sexist language since the 1960s among female activists [1]. Despite the general concept that sexism is only discrimination about women, there are more concerns about the way women are represented in advertisements, newspapers and social media. There have been many definitions of sexism; for instance, in a more general term in dictionary, it is defined as a way of discrimination against women [2]. However, sexism is not all about discrimination, and what is happening in social media is far from this definition. There is a need to conceptualize sexism on online platforms differently since its manifestations are different.

Mills [1] provides comprehensive types of definitions of sexism. Some of the definitions pertain to presumed activities associated with women and are secondary to men or traditional and stereotypical beliefs about women e.g. fuck off and go back to the kitchen, your husband is hungry. Here it is assumed that there are some duties and jobs specifically for a woman and therefore she is classified less as a person in her own right. Sexism seems to be relatively a complex concept which is not easy to define in lexicon level alone and at the same time in the examples.

Waseem et al., 2016 [3], collects hateful tweets in the categories of sexist, racist and neither, this work was the first touch of natural language processing and sexism. He considered the eighteen conditions for being hate speech but not

bringing specifically definitions for being sexist or racist. In line with the previous research, tweets were categorized into hostile, benevolent and neither by [2]. In their research, hostile sexism tweets are clearly imposing negative emotions and benevolent sexism tweets are indirectly sexist and there is no words of violence in it. Ruth et al.,2017 [4] considers the tweets towards top feminists and collects a range of online abuse and categorize them into classes such as sexual harassment, physical threats, flaming and trolling, stalking, electronic sabotage, impersonation and defamation. Her research inspired us to come up with new categories. Mills [1] divides sexism into overt sexism and indirect sexism, each with their own subcategories.

Online crowdsourcing platforms are considered as an effective tool for the research. Amazon Mechanical Turk (MTurk) was among the first which was perceived as a reliable and cost-effective tool to collect high-quality data for different research purposes [18]. However, in MTurk, there have been participants who complete the questioners more than once and it reduces the credibility of the answers. Concurrently, some other crowdsourcing platforms like Crowd-Flower(CF) and Prolific Academic (ProA) have been introduced in different body of research. Peer et al.,[19] presents a useful and detail discussion about online platforms which have been used for different purposes. Even though the focus of the article is more on behavioural research, it can be generalized well to other types of research.

Inspiring from [4][5], we come up with four major categories of different types of sexism. For each category we come up with the category name, the definitions and the examples. Here you can find different categories.

Indirect harassment (#1):

Definition: These tweets are around stereotypical and traditional believes about women. or showing the inferiority of men over women. or are indirectly sexist and doesn't have any swear words.

Examples:

-A wise woman builds her house but a foolish woman tears it down with her hands. -Yes, we get it. You're pretty. Tone down the self promo and just cook. It's less of #adaywithoutwomen and more of a day without feminists, which to be quite honest, sounds lovely.

Information threat (#2):

Definition: Women faces threats of losing and misusing their information or their information being stolen.

Examples:

-if you are seriously gonna try to change gamin, I will hack your account and put gay porn everywhere. - I have all your names and ledger. 5000\$ a month or I torch you all over the web, forever.

Sexual harassment (#3):

Definition: these tweets contains insulting words, name calling, words of anger and violence or force toward sex.

Examples:

-damn girls, you are fine. -f.k that cunt, I would with my fist.

Physical Harassment (#4):

Definition:

These tweets threatens female biology attributes, beauty of women or consider women as the sex objects or implying sense of ignorance or lack of attractiveness or implying lack of physical or mental ability of women or using humour toward the female body. Tweets can contain insulting words or violence, but the concentration is on the previous 5 points.

Hint:

Female body getting disease, pain, breaks, disgusted. Comments on weight, breast size, hotness, lack of hotness.

Example:

-Just putting it out there, you deserve all those deaths you are getting. -don't pull the gender card. You are not being harassed because you are a woman. It is because you are an ignorant cunt. -the menus look like they were made by a 5 year-old little girl. In this case just the mental age of a 5-year-old girl I guess.

Non-sexist tweets (#5):

Tweets that are not in the above category.

2 Literature review

Introducing sexism as the classification task was first proposed by Waseem, 2016 [4]. He annotated 16 thousand tweets and categorized them as racist, sexist and neither. His dataset is shared in his GitHub with the tweet IDs. However, by the time we downloaded the tweets, most of them were deleted and were no longer available. Out of three thousand tweets collected by Waseem, only two hundred were available at the time of study. Waseem [4] collected tweets around the famous Australian tv show , My Kitchen Rules, and the hashtag #mkr. We found this hashtag useful but after a closer look we saw that the corpus had a bias and did not contain different types of sexism. He tried different methods such as character level grams and word grams and employed logistic regression with 10-fold cross validation. In 2017, Akshita and Radhika [2] introduced a new category of benevolent sexism. Categorizing tweets exhibiting sexism if they have one of the three following features: “protective paternalism”, “complementary gender differentiation” and “heterosexual intimacy”. They used different types of classifiers and reported the result along with other polarity detection methods.

Alessandro et al., [6] tried to automatically detect hate speech, including sexist words along with other types of taboo words by focusing on just words and combinations of nouns, verbs, adjectives and adverbs. Badjatiya et al., [7] took advantage of deep neural networks models and use the Waseem dataset [4]. They used different types of convolutional and recurrent neural networks along with different representation learnings. Long short term memory (LSTM) was combined with random embedding and among them gradient boosted decision trees had the best performance. Park and Fung , 2017 [8] approached the Waseem

dataset and thought of it as a two step classification: First, if the tweet was abusive or not and second if it was sexist or racist. They proposed a hybrid Convolutional Neural Network (CNN) by combining character level and word level CNN. In another study, Clarke and Grieve, 2017 [9] also used the Waseem dataset. In their study, they proposed a multi-dimensional analysis (MDA) including of being interactive, antagonistic and attitudinal, comparing the classes of sexist and racist tweets. Their study investigates a wide range of linguistic features and texts.

Unlike the previous works focusing on the natural language processing, Ruth et al., 2016 [3] focused on female activists who were very controversial and experienced regular trolling and abuse on Twitter. They reported that about eighty-eight percent of them were abused on Twitter and only 60 percent on Facebook. They came up with the top ten categories of abuse as follows: harassment, sexual harassment, threats of physical violence, threats of sexual violence, stalking, flaming and trolling, electronic sabotage, defamation, inciting other to abuse and impersonation. Based on their study, about 40 percent of their audience experienced sexual harassment and the rest sexual violence. In another interesting study, Shaw, 2006 [10], collected information of a campaign “Bye Felipe” which feminist posts the screenshot of harassment and sexual entitlement on online dating websites such as Tinder. They collected information from the reactions that women received after implicitly or explicitly saying no to a man. They argued that women were considered as sex objects. Women mostly experienced rape threats, violent words, facing humorous sexist words and trolling-like participation. In the experiment there were a lot of good examples of different threats and sexist tweets.

Vitis and Gilmour, 2016 [11] brought the definition of Citron [12], mentioning that there are three major components in online harassment including the fact that the victims are women, the harassment is targeted at women and the abuse is in a threatening way. They analyzed different reactions that females had to such harassment. Megarry, 2014 [5] examined the hashtag #mencallmesomething and identified different aspects of sexism on online platforms. They focused on different females from different countries who were harassed on twitter. They argued that there are different types of online harassment. One of them is related to the concept of masculinity which defines the level of being level headed and women were criticized to be emotionally sensitive. The other form is related to violence against the physical features of women body. These violence can manifest as hatred towards the female body or references toward diseases such as “I hope you get the sexually transmitted disease of vagina cancer”. Sexual attacks were another kind of online harassment, such as “f.k that cunt, I would with my fist”. References to perceived unattractiveness is another prevalent category of online abuse. Death threats such as “tape a plastic bag on her head and kill her live on a webcam” is another category mentioned in [5].

After all these studies, one can come to the conclusion that dividing online harassment into merely hostile and benevolent sexism ignores the more fine-grained categories of sexism on online platforms.

3 Data collection, annotation and preprocessing

We collected data using Twitter search API and used a wide range of tweets and hashtags extracted from the previous studies. There were 290 hashtags including: “#mkr”, “#Everydaysexism”, “#instagranniepants”, “#mencallmethings”, “#mcmt-a”, “#gamergate”, “#femfreq”, “#metoo”, “#slutgate”, “#Asian drive”, “#nigger”, “#mkr”, “#sjw”, “#weareequal”, “#womensday” and “#adaywith-outwomen”. The process of collecting data was quite easy using the current hashtags and we also use Latent Dirichlet allocation (LDA) to get the sub hashtags and used them to collect more tweets [13]. List of useful hashtags for collecting sexist tweets shared in author Github¹. We filtered tweets that were not in English, removed duplicate tweets, HTML links, words fewer than three characters, and the spam content. However, we didn’t remove the hashtags because we found them useful for labeling the tweets, they provided context for the tweets.

4 Pilot study, annotation agreement and challenges

We ran the pilot study on 50 tweets by 12 females non-activist and one male non-activist. We calculated the inter-annotator measurement, Fleiss ’kappa score [14] because we had more than two raters to calculate the agreement. The result of Feliss’kappa is 0.70 which is a good score for the agreement. There were some feedbacks from male and female labelers as follows: if they were asked to label the tweets as sexist or if they were asked to label the user as sexist. Some tweets needed context to be categorized properly; raters wanted to know if the tweets were the reply to the previous tweet or was the original tweet. Also, Raters wanted to know the gender of the twitter users. Some tweets were not clear if they were toward a male or a female, so they wanted to know who the audience of the tweet was. There were some abbreviations in the tweets that made understanding the tweets and labeling difficult for them. They found hashtags useful for finding the context of the tweets. They found the hints and false examples useful too. Female labelers show more sensitivity to label the tweets and classifying the tweet into the indirect sexism was easy based on the definition. However, male had difficulty labeling the tweets in the first category and tweets were not sexist for him. There was a big difference between what was in female rater’s mind about the definition of the sexist tweet and the presented definitions and the categories in the instruction, and it made it interesting for them as well. For the male raters, classifying the tweets into category 3 and 4, sexual harassment and physical harassment respectively was difficult. For him, all the tweets were hateful and necessary doesn’t imply the sexist meaning or they were usual speech between male as part of them being masculine and they don’t carry sexist meaning but more hateful.

We addressed some of the challenges in our dataset as follows: for clarifying the gender of the twitter users, we used the lexicon provided by Sap et al., 2014 [15] and we got a number of being positive or negative and it identified if the

¹ <https://github.com/simarad/Sexist-Hashtags.git>

user is female or male. We added some notes and clarified that the task was the multi-class classification of the tweets. When the audience of the tweets were not clear, we asked raters to randomly decide if the tweet was toward women or men.

5 Experiment

We collected about 3243 data and labeled 150 tweets out of them as gold question. The 150 tweets were labeled by the author and another female non-activist. The 150 test questions were useful for testing the performance of the raters and increasing the accuracy of their work. Table 1 shows the amount of data labeled for each of the categories along with their confidence score. Each raters initially answers the gold questions and get an accuracy. Those who get the accuracy of more than 70 percent can label the data. The more their accuracy is, the more they can label. The third column in table 1 shows the confidence score of all raters contribute to each category. As expected, it should be more than 70 percent. About 349 labelers contributed in the study. About 290 were female and 58 were male, females were formed the percentage of 83 percent of the total raters.

Unlike what we expected, most of the data in non-sexist category were in fact indirect sexism which were misclassified. It seems distinguishing the difference between the two classes is a hard task. Crowdfunder presents a dashboard of

Table 1. The Distribution of the Data

Name of the Categories	Data	Confidence score
Indirect harassment(#1)	260	0.75
Information threat(#2)	2	0.77
Sexual harassment(#3)	417	0.75
Physical harassment(#4)	123	0.75
Not sexist(#5)	2440	0.75

contributor satisfaction, table 2 shows the details of information. A number will be assigned for each metric to show the feedback by the labelers.

After running the result, we get a csv file about the distribution of the labelers based on their country, their judgement counts and their submission rate along with their gender. The highest judgement count is 440 from Ukraine with submission rate of 621.

Furthermore, Crowdfunder presents the judgement-count and submission-rate of each raters. It seems there is a relationship between the judgement count and submission rate. Labelers having more judgement-count of 200, coming from countries, USA, IND, VNM, ITA, VEN, RUS, SRB, POL, EGY. Fig1. shows the

Table 2. Details of each task and each of the scales.

Metrics	Scale(/5)
Instruction clear	4
Test question fair	4.1
Ease of job	3.8
pay	4.2
Overall	4

top 22 raters by their judgement-count and submission-rate. Fig 2. shows the submission rate by the country. One interesting point is that, even though raters come from different culture and background, their confidence is an acceptable confidence.

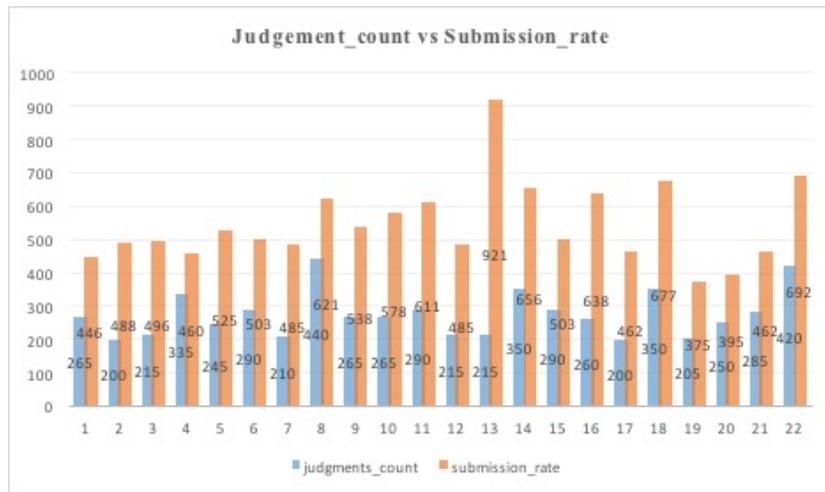


Fig. 1: Judgment-count and submission-rate of the top 22 raters

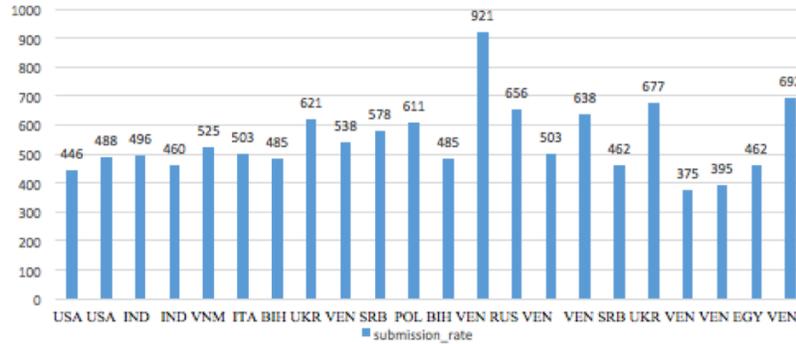


Fig. 2: Submission rate by the country.

In another reports presented by the Crowdfower namely "aggregated" report and "contributors", ones know the confidence of each rater labeling each tweet and the gender of the raters. By simple calculations we have table 3 which shows the contribution of each gender to each category separately along with the total confidence of the tweets in each category.

Table 3. Contribution of each gender classifying the tweets in each category along with the confidence of each category.

Categories	Female contribution	Male contribution	Total confidence of each category
Indirect sexism(#1)	3573(84%)	653(15%)	0.57
Information threat(#2)	316(83%)	63(16%)	0.5
Sexual harassment(#3)	2614(86%)	407(13%)	0.65
Physical harassment(#4)	1739(82%)	373(17%)	0.44
Not sexist(#5)	9373(82%)	2039(17%)	0.80

The highest confidence for classifying the tweet belongs to the fifth category namely not-sexist. Female has the highest contribution in comparison to male. As we argues before, looking through the data in this category, one can understand a wide range of sexist tweets could be classified in indirect classification but they were misclassified into not sexist with the high confidence. It may be because of the fact that distinguishing between the indirect sexism and benevolent sexism is a hard task because it depends more on the culture and maybe because it doesn't have any leading key words such as insulting words. The second highest confidence is related to the third category, one of the reasons is that this type of sexism is a clearly defined types of sexism culturally and it has a lot of key words to help classifying this category. The least confidence is physical harassment and the reason is because distinguishing between the sexual harassment and physical

Table 4. LDA result on each category.

Categories	LDA Results
Indirect sexism(#1)	Women, arrogant, cant, like, think, sassy, dumb
Information threat(#2)	Data, hell, speak, word, name, annoying, proven, babe
Sexual harassment(#3)	Blonde, pigs, bitch, make, sex, hate
Physical harassment(#4)	Pretty, ugly, sane, without, gang, look, attention, smash
Not sexist(#5)	Women, girls, blessed, empowering, lucky, feel

Table 5. Classification results on accuracy

Methods	Algorithm(NB)
bigrams(BOW)	0.73
threegrms(BOW)	0.68
four grams(BOW)	0.66
Two character grams(BOW)	0.79
Three character grams(BOW)	0.80
Four character gram(BOW)	0.79
combination of all grams	0.73
combination of all character-grams	0.81
combination of all character-grams and all the grams	0.75
Word2vec	0.79
Doc2vec	0.77
LSTM	0.91
Character level-CNN	0.93

harassment is a hard task and needs a lot of background and clear key words. As it may seem from the confidence score, not only classifying the tweets is a hard task but also distinguishing between the different types is a hard task and requires more background and maybe training before labeling the tweets. Another interesting step for us was deploying topic modelling on each category separately to report the most frequent content topics using LDA. Table4. Shows the result of LDA on each category separately.

6 Results

When it comes to data preprocessing and classification, we ran our experiment using a wide range of representation learning , neural networks and Naive Bayes as the classifier. The result of these calculations reported in Table 6. We used Long short term memory (LSTM) using Keras , 8 batches of size 30 each, sequence-length of 15, learning rate of 0.001 and size of 128. For character level Convolutional Neural network we used Tensorflow and the code presented in [17].

7 Challenges of study

Collecting the data, labeling and working on the sexism as the working dataset made people look slightly different on you. Their first idea is that they author is a feminist or a girl who witnessed a lot of inequality in her life. Some have reactions of humour, ridicule and irony toward this topic of research. It is also challenging if one can see the sexist sentences as hateful type of speech. They both share some characteristics such as violence and threats. The other challenge is defining the boundaries between the hate speech and sexism tweets. Looking into the tweets, some are just hateful toward feminist.

8 Conclusion

Sexism is a very prevalent in social media. In this study, we present new categories of sexism on social media and address the challenges of coming up with the right categories, right annotation and right instruction. Furthermore, we report our initial classification report on these categories. Our future studies will be considering to collect more data base on the categories, investigating more classification tasks and deploying better representation learnings [19].

References

1. Mills, S.: Language and Sexism Cambridge University Press (2008)
2. Akshita., J., Radhika, M.: When does a compliment become sexist: Analysis and classification of ambivalent sexism using twitter data. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science (2017) 7–16
3. Rowe., L.R.M., Wiper, C.: Online abuse of feminists as an emerging form of violence against women and girls. British Journal of Criminology (2016)
4. Waseem, Z., Hovy, D.: Hateful symbols or hateful people: Predictive features for hate speech detection on twitter. In Proceedings of NAACL-HLT (2016) 88–93
5. Jessica, M.: Online incivility or sexual harassment. conceptualizing women’s experiences in the digital age. Women’s Studies International Forum **47** (2014) 46–55
6. Simonetta., M.A.P.S.V., Pierluigi, V.: Mining offensive language on social media. <http://ceur-ws.org/Vol-2006/paper038.pdf>. (2017)
7. Gupta., P.B.S.G.M., Varma, V.: Deep learning for hate speech detection in tweets. International World Wide Web Conferences Steering Committee (2017) 759–760
8. J.H., P., P, F.: One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206 (2017)
9. Isobelle., C., Jack, G.: Dimensions of abusive language on twitter. Proceedings of the First Workshop on Abusive Language Online (2017) 1–10
10. Shaw, F.: Bitch i said hi: The bye felipe campaign and discursive activism in mobile dating apps. Social Media + Society **2** (2016)
11. L, V., F, G.: Dick pics on blast: A woman’s resistance to online sexual harassment using humour, art and instagram. crime, media, culture. Crime, Media, Culture. Advance online publication (2016)

12. DK, C.: Law's expressive value in combating cyber gender harassment. *Michigan Law Review* **108** (2009) 373–416
13. C., V.E.G.C., D, Z.: Analyzing the political sentiment of tweets in farsi. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)* (2016)
14. Fleiss-kappa <https://en.wikipedia.org/wiki/Fleiss-kappa>
15. CrowdFlower <https://www.crowdflower.com/>
16. Sap, M., Park, G., Eichstaedt, J., Kern, M., Ungar, L., Schwartz, H.A. Developing age and gender predictive lexical over social media In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014) 1146-1151
17. Zhang X. Zhao J. and LeCun Y. Character-level Convolutional Networks for Text Classification *Proceedings of NIPS* (2015)
18. Peer E. Brandimarte L. Samat S. and Acquisti Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online behavioural research *Journal of Experimental Social Psychology* (2017) 153-163
19. Chandler J. Paolacci G. Peer E. Mueller P. Ratliff K. Non-naive participants can reduce effect sizes *Psychological Science* (2015) 131-1139