Learning and Understanding Different Categories of Sexism Using Convolutional Neural Network's Filters

Sima Sharifirad Computer Science Department Halifax, Nova Scotia s.sharifirad@dal.ca

Alon Jacovi Computer Science Department Bar Ilan Univesity, Israel alonjacovi@gmail.com

Stan Matwin Computer Science Department Halifax, Nova Scotia stan@cs.dal.ca

Abstract

Sexism is very common in social media and makes the boundaries of free speech tighter for female users. Automatically flagging and removing sexist content requires niche identification and description of the categories. In this study, inspired by social science work, we propose three categories of sexism toward women as follows: Indirect sexism, Sexual sexism and Physical sexism. We build classifiers such as Convolutional Neural Network (CNN) to automatically detect different types of sexism and address problems of annotation. Even though inherent non-interpretability of CNN is a challenge for users who detect sexism, as the reason classifying a given speech instance with regard to sexism is difficult to glance from a CNN. However, recent research developed interpretable CNN filters for text data. In a CNN, filters followed by different activation patterns along with global max-pooling can help us tease apart the most important ngrams from the rest. In this paper, we interpret a CNN model trained to classify sexism in order to understand different categories of sexism by detecting semantic categories of ngrams and clustering them. Then, these ngrams in each category are used to improve the performance of the classification task. It is a preliminary work using machine learning and natural language techniques to learn the concept of sexism and distinguishes itself by looking at more precise categories of sexism in social media along with an in-depth investigation of CNN's filters.

1 Introduction

In social science, different comprehensive definitions of sexism were provided by Mills (2008). Some of the definitions of online sexism pertain to presumed activities associated with women or stereotypical and traditional beliefs about women and situated women secondary to men. Sexism seems to be a relatively complex concept which is neither easy to define at the lexicon levels alone nor in the examples.

2 Experiment

We ran our first pilot study given participants the instruction for four categories. We asked one male and 12 female non-activists to label 50 tweets and give us feedback about clarity of the instruction, clarity of tweets and also the level of task hardship (Laura Vitis and Fairleigh Gilmour, 2016). We calculated the inter-annotator measurement between raters using Fleiss' kappa score (Amir Ziai, 2017). The score was 0.70 which was a good score of agreement. After the pilot study, we initially used the hashtag #mkr (Waseem et al., 2016) to collect tweets using the Twitter search API. Since hashtags occur together in some tweets, this was a good leading hashtag to point us to other ones to use. Finally, we used a wide range of hashtags to collect more than three thousand tweets. There were 290 hashtags in total, some of the most frequently occurring were #mkr, #Everydaysexism, #instagranniepants, #mencallmethings, #mcmt-a, #gamergate, #femfreq, #metoo, #slutgate, #Asiandrive and #nigger. The list of useful hashtags for collecting sexist tweets was shared in the author's Github. We filtered tweets that were not in English and removed any duplicate tweets, HTML links, words fewer than three characters, and spam content. However, we didn't remove the hashtags because we found them useful for labelling the tweets; they provided context for the tweets using the crowd sourcing platform, Figure Eight. Out of 3240 total

tweets, 260 of them were labeled as indirect harassment, 417 as sexual harassment, 123 as physical harassment and 2440 as not-sexist.

3 Results

Inputs were embedded using pre-trained GloVe Wikipedia 2014 50-dimensional vectors. The convolutional layer used 15 filters of size 3 (fine-tuned from combinations of filters of sizes {2,3,4}). The convolutional layer calculated the inner product between each filter and each n-gram (in our case, 3 words) assigning a score to each ngram. These scores were then fed into a max-pooling layer which selected the top-scoring ngram per filter and a ReLU activation. The score vector was then classified by a linear classifier. For optimization, we used Adam optimizer. The dataset was divided into 85 percent training samples and 15 percent testing. The model achieved an 87 percent accuracy on the test-set. After running the model and fine-tuning, we proceed to interpret the model in accordance with methods introduced by Jacovi et al. (2018). First, we derive the identity class for each filter, based on the respective weights of the filter in the final linear classifier of the model. In essence, the highest identity score shows the class for which the ngrams were chosen by the filter support the classification. The identity number of each class presents the activation number for that class. For each filter, the highest identity number shows the category in which the filter could identify the best supported by ngrams for that category. Next, we investigate informative features in the input: ngrams whose activations pass a certain threshold value (calculated heuristically per previous work) are deemed as *informative ngrams*, while the rest are deemed *uninformative ngrams* which pass the max-pooling layer in the model in spite of their low value in the classifying task. We detail the ngrams that received the strongest activation in the dataset (i.e. the ngrams that serve as the strongest evidence for classification of the filter's identity class). These ngrams are representative of the semantic meaning that the filter detects. After this, we cluster this group of biggest ngrams using Mean Shift Clustering (Yizong Cheng, 1995). Among all the filters, we choose three filters that are strong indications for each class. Later, we add informative ngrams to each class and run the classifier again. Table 1 shows the accuracy.

	Original dataset accuracy	Original dataset + biggest ngrams
SVM	0.76	0.80
Naive Bayes	0.78	0.81
CNN	0.87	0.90

Table 1: Accuracy of classifiers on original and augmented dataset.

4 Conclusion and Future work

We tested convolutional filters on a dataset related to different types of online sexism to test if the filters can help us understand the nature of the dataset where some classes share the same words. The filtering of uninformative ngrams and clustering the informative ngrams helped to understand what the model considers important in the input space of text tweets for the harassment classification task. The biggest ngrams in each category point at different aspects of the tweets and words which are used, for example, filter #zero which presents Indirect harassment category shows two explicit categories of ngrams, those which are related to cooking and those which have an indirect weight of harassing, even though these ngrams in each category separately do not present information but pairing them together can make a large number of indirect harassment tweets. Ngrams in Our second category, sexual harassment are divided into being pejorative and provoking words. These two categories of ngrams together can make a large number of sexual harassment tweets. The third category, physical harassment, has two clusters of ngrams, one of them focused on the physical attribute of the woman while the second class focused on threatening them. As future work, we would like to expand these experiment on available hateful speech dataset and compare the results to the current one to understand the inherent difference between harassment and hateful tweets.

References

Sara Mills. 2008. *Language and sexism.*

Zeerak Waseem, Thomas Davidson, Dana Warmsley and Ingmar Weber, 2017. *Understanding Abuse: A Typology of Abusive Language Detection Sub tasks* Proceedings of the ACL-Workshop on Abusive Language Online. Media, ACL, 2017, vancouver, BC, Canada.,78-84.

Jacovi, Alon and Sar Shalom, Oren and Goldberg, Yoav, 2018. *Understanding Convolutional Neural Networks for Text Classification* Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NL,56-65.

Lewis Ruth, et al., 2016. *Online Abuse of Feminists as an Emerging Form of Violence against Women and Girls.* British Journal of Criminology,1-20.

Yoon Kim., 2014. *Convolutional neural networks for sentence classification.* EMNLP.

Nal Kalchbrenner and Edward Grefenstette and Phil Blunsom, 2014. *Convolutional neural networks for sentence classification.* Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA,, 655-665.

Zhang, Xiang and Zhao, Junbo and LeCun, Yann, 2015. *Character-level convolutional networks for text classification* Advances in neural information processing systems, 649-657.

Laura Vitis and Fairleigh Gilmour, 2016. *Dick pics on blast: A womans resistance to online sexual harassment using humour, art and Instagram.* Crime, Media, Culture. Online. DOI: 10.1177/1741659016652445.

Amir Ziai, 2017. *Inter-rater agreement Kappas.* https://towardsdatascience.com/inter-rater-agreement-kappas-69cd8b91ff75

Yizong Cheng, 1995. *Mean Shift, Mode Seeking, and Clustering* IEEE Trans. Pattern Anal. Mach. Intell., 17(8): 790-799.