Aleena Syed
Norah Khan
CS 4395.001

ACL Paper Summary

"Annotating Online Misogyny"

"Annotating Online Misogyny" is written by Philine Zeinhert, Nanna Inie, and Leon Derczynski under the IT University of Copenhagen, Computer Science.

With the popularity of social media comes the inevitable circulation of potentially abusive language and rhetoric. Current protocols for automatic detection of abusive language depend on many factors: data gathering, data annotation, and bias mitigation. The problem this paper addresses is the difficulty of automatic detection of online misogyny. The research's aim is to develop an iterative annotation process to create a dataset of annotated posts to curb hateful speech towards women online.

Prior work in this field of annotation involves using systems that use labeled training data, directly correlating it with the quality of datasets and their labels. There is no objective framework to define what counts as abuse, let alone more specific kinds of abuse such as misogyny, leading to past research to categorize misogynistic posts in various ways. This can also vary depending on the language of the posts in question. For instance, Chiril et al. (2020) used three different categories to classify misogyny in French (direct sexist content, descriptive sexist content, and reporting sexist content), whereas Anzovino et al. (2018) classify misogyny in five categories (discredit, harassment & threats of violence, derailing, stereotype and objectification, and dominance).

This research focused on the creation of a taxonomy to accurately describe the various ways misogyny can proliferate in online social media posts. This was created as a product of existing research and contextualizing misogynistic posts relative to Danish. To create their dataset, random sampling was used with predefined keywords from relevant posts and comments

on social media. The annotation process involved asking annotators to read and annotate 150 different posts. This would be based on taxonomy with four separate levels: abusive (abusive/not abusive), target (individual, group, others, untargeted), group type (racism/misogyny/untargeted), and misogyny type (harassment/discredit/stereotype and objectification/dominance/neosexism/benevolent). Annotators were given a set of guidelines to annotate their posts with. If annotators disagreed on whether a post was considered abusive or not, they were discussed in weekly meetings and resolved through majority voting. To mitigate biases, labels were selected from past peer-reviewed research and sought diverse annotator profiles. As opposed to some research papers in the past, this team treated neosexism as its own category, which most taxonomies in the past did not separate from general misogyny. Neosexism is a belief that although women have already achieved equality, discrimination of women cannot exist, existing as a more “subtle” form of sexism. This distinction between active and passive misogyny is an important one to make, as seen from the results.

The results of this research culminated in a final dataset of 27,900 comments, 7,500 of which contained abusive language. Neosexism was the most frequently represented class in the tagged posts (1,300 posts out of 7,500), followed by discredit and stereotype & objectification. About half of these posts came from Facebook and Twitter. Topic sampling (searching through specific social media sites) results in more misogynistic content than searching for specific keywords. Neosexism was strongly represented in the data. In the future, research could include taxonomies that distinguish between active vs passive forms of misogyny, as this research shows its prevalence in the labeling of the type of misogynistic posts.

This paper has been cited 14 times on Google scholar. Leon Derczynski has been cited 5494 times, the most of the three authors of this paper. This research is important with the immense popularity of social media and its involvement in dictating and commentating on culture, news, and politics. The ability for anyone to create and post on these platforms raises questions on how we should monitor abuse that can propagate through liking and sharing posts and comments online. It's important to have users follow specified terms of service that prohibit them from sharing content that can be deemed offensive. Having a set of guidelines under a cohesive taxonomy such as what this research suggests can allow for these social media platforms to regulate what is posted on their sites more efficiently.