Wyatt Kirby
IRK180000
CS 4395.001

A ToxiGen Summary

Social media has made it nearly impossible to ignore the prevalence of derogatory and hurtful language directed at people within various minority groups. It is important for social media to remain a safe space for all that wish to use it and be proactive in the mass radicalization of people present in online communities. Such radicalization limits the user base of a platform as well as massively impacting the world outside of the internet. The paper "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection", by Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar in affiliation with the Massachusetts Institute of Technology, the University of Washington, Microsoft Research, Allen Institute for AI, Carnegie Mellon University, and Microsoft, explores a language model that hopes to improve hate speech detection.

Many models that are in use today flag any text mentioning a minority group instead of those that are attacking said groups. Essentially, instead of trying to understand the meaning of the flagged text, it flags everything and keeps productive conversations from happening and censors the groups that the models are created to help. This also allows subtle hate speech that simply alludes to a group go undetected. To make up for these issues, ToxiGen uses toxic language created by a pretrained model. This model will create a large dataset of both subtle and overt hate speech that is human enough to train ToxiGen well enough to detect a more diverse range of offensive terms. This allows ToxiGen to detect offensive language not containing slurs, specific mentions of minority groups, or generic hate speech and helps detect a wider range of undesirable language. It also allows for a wider range of statements to train on than simply scraping from the internet.

One of the ways that the model was tested was by using a human-validated test set to ensure that the model was correctly classifying text. They also ran a trial where people had to put a sentence on a scale from one to five to determine how harmful the sentence was. This was done to account for the variation in opinion between people. The results were then averaged and compared to those of the ToxiGen model. Both together were a good litmus test for the sensitivity of the model and allowed for later iterations to improve.

Personally, I think that this research is incredibly important. Currently social media allows for echo chambers to be created where individuals push each other's ideal to become more and more radical and isolate themselves farther from everyone else. This allows for hate to grow and multiply, infecting those that would not have otherwise internalized such harmful ideals. By detecting what language is harmful and offensive it could reduce the number of people radicalized greatly. This would allow for platforms to remove such language completely from their platform and make the internet a less mentally taxing and draining experience for all those who are on it. Toxic language like this also has massive real-world impacts, seeping into our politics and policies that get passed here in America. Many of the issues we see surrounding the rise in alt-right ideology could be hindered (though I am doubtful it will ever truly be eradicated) can be attributed to the complicity of social media platforms when offensive language is used. If nothing is done, those with harmful ideologies will continue to fearmonger others into radicalization. By being able to determine if a specific sentiment is harmful and remove it from the platform, it is kept from spreading.

Out of all of the people listed in the authors section, Ece Kamar had the most citations listed in Google Scholars (110). This is most likely because she is a researcher working with

Microsoft Research. Thomas Hartvigsen had 40, Saadia Gabriel had 17, Hamid Palangi had 49, Maarten Sap had 64, and finally Dipankar Ray had 82 citations.