PHD Defense Reveiw

Social media platforms are some of the most popular websites on the Internet. From making friends to discussing highly controversial news stories, these platforms provide a space for people to connect, discuss ideas, and share their opinions on a variety of different topics. As a result, we can observe online communities that are not just a collection of people discussing the same ideas online; they could also be a collection of managed accounts acting in unison to achieve a malevolent goal. This research primarily focuses on understanding the malicious behavior of online communities.

To this end, the first contribution is the study of inauthentic accounts, also known as troll accounts. Troll accounts on social media interact with one another and appear innocuous to a regular user while covertly being used to spread toxic content and/or disinformation. These accounts are often sponsored by state actors aiming to manipulate public opinion on sensitive political topics. The researcher first study the effect that troll accounts have on online discussions on Reddit and show that state-sponsored troll accounts on Reddit produce threads that attract more toxic comments than other posts on the same subreddit. Next, I build TROLLMAGNIFIER, which is a detection system for troll accounts. TROLLMAGNIFIER learns the typical behavior of known troll accounts and identifies more that behave similarly. He shows that using TROLLMAGNIFIER, one can grow the initial knowledge of potential trolls provided by Reddit by over 300%. Next, the researcher study coordinated attacks, such as cyber-aggression and hate attacks, which are becoming increasingly common on video sharing networks like YouTube. Polarized online communities choose targets on prominent online platforms (e.g., YouTube) and organize their attacks by sending hateful messages to their target. He presents an automated solution to identify attacks on YouTube videos and attribute said attacks to a source community. The attribution is performed through a machine learning classifier based on TF-IDF scores of important keywords. In the future, he plans to extend TROLLMAGNIFIER and understands how various communities spread different ideologies on social media platforms.

In summary, the research focuses on understanding and mitigating the impact of harmful activity by polarized communities, for which the researcher first demonstrate the toxic impact of these communities, then builds a detection system to catch troll campaigns on Reddit and an attribution system to identify aggression attacks on YouTube.