Rajat Sethi

CPSC 8580

**Paper Summary 4 – SkillDetective:**

The Amazon Alexa and Google Assistant are Voice Personal Assistants (VPAs) with several capabilities and features. These devices listen to the owner’s (or user’s) words and subsequently execute the command given. In addition, these assistants allow developers to create custom programs and commands to work with their system, often referred to as “skills.” When creating a skill, the VPA manufacturer provides a series of “policies” that should normally enforce strict regulations and federal guidelines. However, in this paper, the research team discovered that several of these skills violate one or more of these policies. By dynamically testing each available skill with various options, the team organized a comprehensive method to evaluate a VPA and their policy enforcement capabilities.

The key novelty of this paper was their automation of the process. Using Selenium and other input tools, the team was able to script a long list of commands to ask the VPA. Once they received a response, they used NLP and Machine Learning to determine what type of response was given, and what format SkillDetective should reply with. Through this repeated query and answer protocol, the team was able to determine thousands of skills that violated company policy.

There were two limitations that I noticed with this paper. The first issue was the time taken in gathering this data. Scraping data with Selenium and then running a lengthy Machine Learning algorithm is an extremely slow process. The paper even stated that this study took around a year to complete. In future endeavors, the time it takes to research skills (in general, not just for policy-violations), needs to be heavily reduced.

The second limitation was their over-reliance on dynamic, automated testing. As stated in Section 4.3, several policies like trademark violations or misleading information were not included in the study because they would have required human input to prove. This lack of human input may have caused several undetected false negatives in the results. While the study did an excellent job at finding false positives, there were few metrics in the paper to calculate false negatives like specificity. As such, it is important for future endeavors to include more of human input or FN checking.