A Artifact Appendix

A.1 Abstract

Today's voice personal assistant (VPA) services have been largely expanded by allowing third-party developers to build voice-apps and publish them to marketplaces (*e.g.*, the Amazon Alexa and Google Assistant platforms). In an effort to thwart unscrupulous developers, VPA platform providers have specified a set of policy requirements to be adhered to by third-party developers, *e.g.*, personal data collection is not allowed for kid-directed voice-apps. In this work, we aim to identify policy-violating voice-apps in current VPA platforms through a comprehensive dynamic analysis of voice-apps. To this end, we design and develop SKILLDETECTIVE, an interactive testing tool capable of exploring voice-apps' behaviors and identifying possible policy violations in an automated manner. Distinctive from prior works, SKILLDETECTIVE evaluates voice-apps' conformity to 52 different policy requirements in a broader context from multiple sources including textual, image and audio files. With SKILLDETECTIVE, we tested 54,055 Amazon Alexa skills and 5,583 Google Assistant actions, and collected 518,385 textual outputs, approximately 2,070 unique audio files and 31,100 unique images from voice-app interactions. We identified 6,079 skills and 175 actions potentially violating at least one policy requirement.

A.2 Artifact check-list (meta-information)

- **Algorithm: This work does propose an algorithm for a question-type classifier.**
- **Program: The program components do require many different dependencies which are provided in the repository or listed and easily downloaded.**
- **Data set: All applicable data sets are included in the repository.**
- **Run-time environment:**
- **Hardware:**
- **Metrics: VPA device output data are gathered and after analysis, any potential policy violations should be reported**
- **Output: The outputs consist of VPA device interaction data saved in a data set, collected device output image and audio files. Lastly, after analysis, any potential policy violations should be reported.**
- **Experiments: There are detailed instructions provided for installation and software use. To run the experiment, one would only have to run the software after installation.**
- **How much disk space required (approximately)?: Approximately 300 to 400 GB.**
- **How much time is needed to prepare workflow (approximately)?: It should take no longer than 30 minutes to set up the software.**
- **How much time is needed to prepare workflow (approximately)?: It should take no longer than 30 minutes to set up the software.**
- **Publicly available?: The artifact is available at https://github.com/skilldetective/skilldetective/releases/tag/V0.3**
- **Archived (provide DOI)?: The artifact is archived at https://github.com/skilldetective/**

A.3 Description

A.3.1 How to access

All of the software can be found at https://github.com/skilldetective/

A.3.2 Software dependencies

There is a list of software dependencies provided within the repository. Also, all of the dependency software for the Java components are included in the repository and all of the needed Python dependencies can be easily downloaded and are clearly stated in the instructions documents.

A.4 Installation

Installation of SD requires a java IDE such as NetBeans and a version of Python installed. The chatbot model runs on Java and has a detailed installation guide that walks the user through all the steps necessary such as acquiring a developer account, writing a test app, accessing the testing terminal, installing and running the software and what to expect from the output as well as some troubleshooting. The policy compliance portion of the software package has a detailed installation and user's guide that outlines all needed steps to analyze the outputs of the chatbot.

A.5 Experiment workflow

The chatbot should be installed and run first. We have provided a list of Alexa skill names for testing. Device interactions should be collected automatically for at least 30 minutes to an hour in order to insure an adequate amount of test data get collected. Next, the policy compliance software should be installed and the test data from the chatbot can be used for evaluation.

A.6 Evaluation and expected results

The chatbot runs autonomously once installed and set up. A data set consisting of speech interactions, image files and audio files should be expected as output. These output data can then be analyzed using the policy compliance software. As a final output, the user should expect a list of any suspected policy violations found in the interaction data.

A.7 Experiment customization

The experiment has many different changeable parameters. First, the list of application names can be altered to include any arbitrary set by changing the file Skills.xlsx. The web browser used to collect the interaction data can be changed by altering the selenium code in S5.java. There are a number of software parameters such as how many interactions are allowed per VPA application and possible changes made to the neural network. These are all commented within the source code.

A.8 Notes

Please feel free to contact us at anytime with any concerns or issues. The email address is skilldetectivetroubleshoot@gmail.com. Also, within the repository please make sure to follow the instructions located at:

- https://github.com/skilldetective/skilldetective/tree/master/skilldetective_policy_detector
- https://github.com/skilldetective/skilldetective/blob/master/ChatBot/SkillDetective%20Instructions.pdf