C Artifact Appendix

C.1 Abstract

Throughout the paper, the results obtained in summary statistics, statistical tests, and qualitative coding directly inform our observations on user behavior. To establish the validity of these findings, we provide the data and subsequent analysis to replicate all results reported in this paper. Specifically, we provide our collected data (for review only), our quantitative analysis via an R notebook, and the results of our qualitative analysis via an excel spreadsheet (for review only). Using these, one can recreate all results in tables, figures, statistical tests, and reported code counts throughout the paper.

C.2 Artifact Check-List (Meta-Information)

- **Data set:** Our collected user study data; non-public.
- **Run-time environment:** Tested on macOS 12.0 Monterey and Ubuntu 20.04.
- **Security, privacy, and ethical concerns:** Maintaining the confidentiality of participant data.
- **Metrics:** Perceived trustworthiness of a social media profile; log likelihood of accepting a social connection.
- **Output:** The artifact produces all result-containing tables, figures, and the code counts of the reported user answers.
- **Experiments:** Statistical and qualitative analysis of user responses.
- **How much disk space required (approximately)?:** 5 GBs.
- **How much time is needed to prepare workflow (approximately)?:** 60 minutes.
- **How much time is needed to complete experiments (approximately)?:** 2 minutes.
- **Publicly available (explicitly provide evolving version reference)?:** All scripts and code are made publicly available[16].
- **Code licenses (if publicly available)?:** University of Illinois/NCSA Open Source License.
- **Archived (explicitly provide DOI or stable reference)?:** https://github.com/JaronMink/DeepPhish/releases/tag/USENIX-22-artifact-evaluation

C.3 Description

C.3.1 How to Access

Along with various supplemental material, we make all the scripts and code used to analyze data and perform statistical tests publicly available[16].

As shown in Figure 10, the artifact is comprised of three main parts: (1) the 'data" folder which contains anonymized participant responses; (2) the R notebook "quantitative_analysis.Rmd" which provides the results reported in tables and figures throughout the paper with the expected output "expected_quantitative_analysis.pdf"; and (3) the spreadsheet "qualitative_codings.xlsx" with the script "codebook_kappa.py" which respectively contains the coded responses and calculates the inter-rater agreement of the codes. Additionally, the README.md provides a detailed overview of all files.

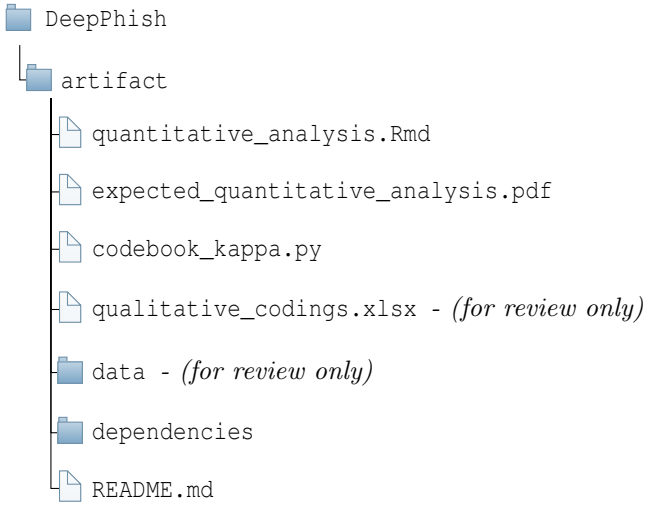

Figure 10: **Artifact File Structure** – We present the file structure of the artifact folder in the paper's supplemental materials[16]. The "qualitative_codings.xlsx" file and "data" folder contain sensitive participant data and thus are only provided for review.

C.3.2 Hardware Dependencies

This analysis requires approximately 5 GBs of disk space.

C.3.3 Software Dependencies

To run the quantitative analysis, we make use of R (4.1.2) RStudio (2021.09.2 Build 382), Pandoc (2.5) and a host of R libraries. We provide download scripts for the specific libraries and dependencies used in macOS and Ubuntu.

To perform the qualitative analysis, we use Python (3.8.10) and Pip (20.0.2) to run the Cohen's-Kappa calculation and Microsoft Excel (16.55) to view the coding spreadsheet (any .xlsx viewer will suffice).

C.3.4 Data Sets

We use the data collected in our user studies. While the data is provided to reviewers, to maintain participant privacy, we do not release this data publicly.

C.3.5 Models

We train our linear mixed-effects model (Section 4.1) and our logistic mixed-effects model (Section 4.2) on the

[16] https://github.com/JaronMink/DeepPhish

gathered user-study data via the R notebook "quantitative_analysis.Rmd".

C.3.6 Security, Privacy, and Ethical Concerns

While there is no inherent risk in our analysis, all participant-provided data should be treated with care. We took steps to anonymize all direct identifiers; however, due to the nature of user and qualitative responses, we cannot exclude the possibility of such data being used to deanonymize participants.

C.4 Installation

C.4.1 Quantitative Analysis

Installation time: ~45 minutes

The quantitative evaluation is performed in an R-notebook and thus requires various software libraries, frameworks, and system dependencies to support it.

Software. R, RStudio, and Pandoc are all publicly available and their instructions for version-specific installation can be found at their respective websites.

System Dependencies. As many R libraries require various system dependencies, we provide scripts to download the required dependencies for Ubuntu ("install_ubuntu_dependencies.sh") and macOS ("install_macos_dependencies.sh").

R Libraries. To install the R libraries used in the analysis, we provide an OS-independent bash script: "install_r_libraries.sh".

C.4.2 Qualitative Analysis

Installation time: ~5 minutes Python

Software. Python (3.8.10) and Pip (20.0.2) are both publicly available and their instructions for version-specific installation can be found at their respective websites.

Python Libraries. To install the utilized Python libraries, we provide a pip3-compatible requirements file: "requirements.txt".

C.5 Evaluation and Expected Results

C.5.1 Quantitative Analysis

Execution Time: ~2 minutes

The results from Section 3, Section 4, Section 5, and Appendix B are produced in the R notebook via the following steps:

1. Open the R Studio application or go to the assigned localhost port with a web-browser (default is 8787).
2. Open the notebook: "quantitative_analysis.Rmd"
3. Produce the results by selecting "Knit → Knit to PDF".
4. Once completed, you may view the produced PDF: "quantitative_analysis.pdf"

The PDF will contain the results for the tables, figures, and information found in the paper which directly inform Observations 1-4.

Section 3.6: Demographic background and time distributions of participants.
Section 4.1: Pairwise correlation of factors, Figure 5 and descriptive statistics, Table 1, ANOVA test and descriptive statistics.
Section 4.2: Figure 6 and descriptive statistics, Table 2.
Section 4.3: Artifact to artifact trust comparison, artifact to artifact acceptance rate comparison.
Section 5.1: Table 3.
Appendix B: Figure 9 and descriptive statistics, Table 4, Sybil trust plot and descriptive stats, Sybil trust modeling.

C.5.2 Qualitative Analysis

Execution Time: N/A

The qualitative results primarily report the counts of the coded qualitative data found in the file "qualitative_codings.xlsx". These code counts along with direct participant quotes inform Observations 5-10.
Section 5: Cohen's-Kappa - To find the interrating reliability of codes, we calculate Cohen's-Kappa for each codebook via the following script:

```
python3 codebook_kappa.py
```

Qualitative Reporting in Sections 5.1-5.3: For each of the following subsections, we note what findings were made, information was reported, and what specific sheet and cells (highlighted in colors) were used to inform these findings.
Section 5.1: Areas of Focus (sheet "factors_by_prompt"; highlighted in red), Artifacts Noticeability (sheet "factors_by_cond"; highlighted in red)
Section 5.2: Perception of Non-Existent Artifacts in Images (sheet "factors_by_cond"; highlighted in blue), Perception of Non-Existent Artifacts in Text (sheet "factors_by_cond"; highlighted in green)
Section 5.3: Noted UIs (sheet "strategies_by_prompt"; highlighted in red), Search for Personal Qualities (sheet "strategies_by_prompt"; highlighted in blue), Search for Inconsistencies (sheet "strategies_by_prompt"; highlighted in green), Reasons for Actions (sheet "strategies_by_prompt"; highlighted in orange).

C.6 Version

Based on the LaTeX template for Artifact Evaluation V20220119.