USENIX Security '25 Artifact Appendix: "Helps me Take the Post With a Grain of Salt:" Soft Moderation Effects on Accuracy Perceptions and Sharing Intentions of Inauthentic Political Content on X

Filipo Sharevski
DePaul University

Verena Distler
Aalto University

Florian Alt
Ludwig-Maximilians-Universität München
University of the Bundeswehr Munich

A Artifact Appendix

A.1 Abstract

The artifact contains two datasets and associated code that analyzes the data for two studies. We performed two studies: **study 1** during the US election 2024 and a replication **study 2** post the elections. The code runs a linear regression model with the soft moderation interventions (warnings/covers in the order of friction) as *independent variable* while the *dependent variable* is: *perceived accuracy* (**RQ1**); *sharing intentions* (**RQ2**). The code also runs a linear regression model with the same variables only controlling for *intention to vote* (**RQ3**) and *demographics*. The data used for the thematic analysis (**RQ5**) is also included in column I titled `sharedqual` and column J titled `accuracyqual` for the sharing intentions and the accuracy perceptions qualitative responses, respectively.

A.2 Description & Requirements

To recreate our analysis for each dataset, RStudio is needed. All the files needed for the functional testing and reproducibility verification is included in the zip archive `manip_media_warnings_replication_files`. The archive contains a `README.md` that describes its structure, consisting of two folders: `study1` and `study2` each including subfolders for `code` and for the *data*, respectively. The `code` subfolders contain an `analyze.qmd` script that runs in RStudio, set up to perform the full quantitative analysis (**RQ1** – **RQ4**) as described in the abstract above (throughout the analysis, we used the customary *p*-value threshold of 0.05, though we included additional thresholds of 0.1 and 0.01 for completeness and transparency). The `data` subfolders contain the `data_clean.Rds` data corresponding to the quantitative values used in the linear regression models, and `data_clean.xlsx` including the qualitative answers used as an input in the thematic analysis (**RQ5**). The `clean` qualifier indicates datasets with removed low quality responses as described in our data collection protocol. Each folder also contains a `tables` subfolder populated with the LaTeX table code export we used for reporting of the results in the paper.

A.2.1 Security, privacy, and ethical concerns

There are no risks for evaluators. The data collection was anonymous and we removed any entries that could potentially lead to identifying any of the participants in both `datasets`. The `code` only implements our linear regression models.

A.2.2 How to access

We are making the dataset as a research artifact available and also provide the scripts we used to analyze the data to allow for both functionality check and reproducibility. The study stimuli are publicly available: https://osf.io/pj895/files/osfstorage?view_only=76a18f0b13bb4281874b0e8d5b0b9bdf. Alternatively, the artifacts can be accessed at: https://doi.org/10.17605/OSF.IO/PJ895

A.2.3 Hardware dependencies

None.

A.2.4 Software dependencies

No specific OS requirements. The software needed for running the linear regression analysis is Rstudio (https://posit.co/download/rstudio-desktop/). Microsoft Excel or any spreadsheet software could be used to open the datasets too, for the purpose of the thematic analysis.

A.2.5 Benchmarks

None.

A.3 Set-up

The analysis for each dataset is executed by running the file `analyze.qmd` in Rstudio. The following package are needed for running the analysis: "qualtRics", "tidyverse", "writexl", "readxl", "ggplot2", "broom", "performance", "see", "patchwork", "lmtest", "car", "estimatr", "sandwich", "stargazer".

You may need to add this line of code in the `analyze.qmd` script if it doesn't load them for some reason:

```
packages <- c("qualtRics", "tidyverse",
"writexl", "readxl", "ggplot2",
"broom", "performance", "see",
"patchwork", "lmtest",
"car", "estimatr",
"sandwich", "stargazer")
install.packages(packages)
```

A.3.1 Installation

Download the replication files from the link above.

A.3.2 Basic Test

The replication files are structured first by study (**study 1** and **study 2**). Within each study, you will find the following folders:

- `code/`: Contains analysis script (`analyze.qmd`).
- `data/`: Contains data from each dataset.
- `tables/`: Contain the LaTeX tables generated by the analysis scripts. Run the code to produce the tables.

Run the analysis script (`analyze.qmd`)

A.4 Evaluation workflow

The replication files are structured first by study (**study 1** and **study 2**). Within each study, you will find the following folders:

- `code/`: Contains analysis script (`analyze.qmd`).
- `data/`: Contains data from each dataset.
- `tables/`: Contain the LaTeX tables generated by the analysis scripts. Run the code to produce the tables.

Run the analysis script (`analyze.qmd`) for (**RQ1–RQ4**). For **RQ5** the qualitative responses are in column I titled `sharedqual` and column J titled `accuracyqual` for the sharing intentions and the accuracy perceptions in the `data_clean.xlsx` for each study. The results for **RQ5** are inherently subjective per the Braun & Clarke's *thematic analysis* methodology.

A.4.1 Major Claims

(C1): RQ1: in both studies, the warning covers (high level friction) and the community notes bundled with warning labels (medium level friction) had a statistically significant negative effect on the perceived accuracy

(C2): RQ2: in both studies, none of the soft moderation interventions had a statistically significant effect on the sharing intentions;

(C3): RQ3: for both studies, we included an interaction term for the soft moderation interventions and the *intention to vote* to check whether the soft moderation's effect varied depending on the voting intentions. The interaction term was not statistically significant

(C4): RQ4: in both studies, frequent content sharing on X was associated with a higher perceived accuracy and of inauthentic content

(C5): RQ4: in study 1, participants who had voted in previous US election cycles showed lower perceived accuracy for the all the stimuli tested compared to participants who had never voted before for a US president

(C6): RQ4: in study 2, younger participants also showed lower perceived accuracy for all the study stimuli tested.

(C7): RQ4: in study 1, participants living in suburbs near a large city reported lower sharing intentions compared to participants living in a large city

(C8): RQ4: in study 2, participants living in rural areas reported lower sharing intentions compared to participants living in a large city.

(C9): RQ4: in both studies, frequent content sharing on X was associated with a higher sharing intentions of inauthentic content

(C10): RQ5: in both studies, the perceptions of accuracy depend on the perceived trustworthiness of the surrounding context. Compared to community notes, labels, and third-party fact-checks were viewed as less credible. A politically incongruent personal disposition toward the subject of the content tended to shift perceived accuracy toward authenticity, even when the content was demonstrably inauthentic.

(C11): RQ5: Sharing inauthentic (soft moderated) content on social media is restrained by the perceived risk of reputation and relationship damage: Sharing might happen as an act of responsibility, though, counting on the warning labels and community notes as a counterbalance against legitimizing misinformation on social media.

A.4.2 Experiments

A short introduction to interpreting linear regression tables (based on Backhaus et al., 2021): In the regression table, the **intercept** represents the expected value of the dependent variable when all independent variables are zero. In multivariate linear analysis, The **coefficients** denote the change in the dependent variable per unit of change in the independent variable, all other independent variables held constant. A positive coefficient indicates a positive relationship and a negative coefficient indicates a negative relationship. Coefficient **standard errors** are in brackets. The standard error of the coefficient is an indicator of the precision (lower standard

errors mean higher precision of estimates). We used the customary statistical **significance level** of 0.05. R-squared values reflect the proportion of variance of the dependent variable explained by the independent variables. In the social sciences, relatively lower R-squared values are often expected because the concepts studied are necessarily influenced by many factors outside of the experimental conditions. The **adjusted R-squared value** accounts for the number of parameters and the sample size, meaning that it penalizes increasing model complexity or overfitting. For more details, see Backhaus et al. (2021), chapter 2.

For each study, once in the `study1` or `study2` folder:

(E1): RQ1 [10 human-seconds + 20 compute-seconds + 0.5MB disk]:
How to: Open the analysis script (`analyze.qmd`) in RStudio
Preparation: `analyze.qmd` is self-contained. No preparation needed.
Execution: `Run all` when the the analysis script (`analyze.qmd`) in RStudio
Results: The results are populated in `model_1_accuracy_warning_table.tex`.
Interpretation: The warning covers and the community notes bundled with warning labels as a medium level friction had a statistically significant negative effect on the perceived accuracy.

(E2): RQ2 [10 human-seconds + 20 compute-seconds + 0.5MB disk]:
How to: Open the analysis script (`analyze.qmd`) in RStudio
Preparation: `analyze.qmd` is self-contained. No preparation needed.
Execution: `Run all` when the the analysis script (`analyze.qmd`) in RStudio
Results: The results are populated in `model_2_share_warning.tex`.
Interpretation: None of the soft moderation interventions had a statistically significant effect on the dependent variable sharing intentions.

(E3): RQ3 [10 human-seconds + 20 compute-seconds + 0.5MB disk]:
How to: Open the analysis script (`analyze.qmd`) in RStudio
Preparation: `analyze.qmd` is self-contained. No preparation needed.
Execution: `Run all` when the the analysis script (`analyze.qmd`) in RStudio
Results: The results are populated in `model_3a_3b_table.tex`.
Interpretation: We included an interaction term for the frictions and the intention to vote to check whether the soft moderation's effect varied depending on the voting intentions. The interaction term was not statistically significant, for any level of soft moderation intervention friction. Warning labels bundled with community notes or warning covers decreased the perceived accuracy of inauthentic content on X.

(E4–E9): RQ4 [10 human-seconds + 20 compute-seconds + 0.5MB disk]:
How to: Open the analysis script (`analyze.qmd`) in RStudio
Preparation: `analyze.qmd` is self-contained. No preparation needed.
Execution: `Run all` when the the analysis script (`analyze.qmd`) in RStudio
Results: The results are populated in `model_4a_4b_table.tex`.
Interpretation: We included an interaction term for the frictions and the intention to vote to check whether the soft moderation intervention varied across the voting intentions. The interaction was not statistically significant. The frictions did not have a statistically significant effect on sharing intention.

(E10): RQ5 [30 human-hours + 5 compute-seconds + 0.5MB disk]:
How to: Open the excel spreadsheet (`data_clean.xlsx`) in Microsoft Excel
Preparation: Locate the column J titled `accuracyqual`
Execution: This is not *execution* per se, but this requires an repetition of the thematic analysis process as described by Braun & Clarke
Results: The results are any resultant themes developed through the thematic analysis process.

(E11): RQ5 [30 human-hours + 5 compute-seconds + 0.5MB disk]:
How to: Open the excel spreadsheet (`data_clean.xlsx`) in Microsoft Excel
Preparation: Locate the column I titled `sharedqual`
Execution: This is not *execution* per se, but this requires an repetition of the thematic analysis process as described by Braun & Clarke
Results: The results are any resultant themes developed through the thematic analysis process.

A.5 Notes on Reusability

The quantitative data in both studies could be reused to run other statistical analyses, given that is rich with variables of many sorts. The data used in the thematic analysis for both studies could be well reused for performing other types of thematic analyses; no adaptation is needed; the data might contain grammatical or syntactical mistakes given it is provided *as is*, at the time of collection.

A.6 Version

Based on the LaTeX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2025/.