USENIX Security '26 Artifact Appendix: Cross-National Information Attacks: A Two-Decade Analysis of Troll Behavior in Korea

Jaehong Kim[1,2,*] Hyeonseung Kim[1,*] Jiseon Kim[1,2]
Alice Oh[1] Thorsten Holz[2] Wonjae Lee[1,†] Meeyoung Cha[2,1,†]
[1]*Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea*
[2]*Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany*

A Artifact Appendix

A.1 Abstract

This artifact contains the dataset, model checkpoint, and source code for the paper. The dataset consists of 112 million South Korean Naver News comments spanning 2006 to 2025, annotated with troll categories. The model checkpoint is a KcELECTRA-based hierarchical classifier distilled from GPT-4.1 annotations, with 18 output heads covering comment-level classification and span-level rationale extraction. The code provides a script for running inference with the trained model on Korean-language comment dataset, and the Jupyter notebook for visualizing the strategy analysis results.

A.2 Description & Requirements

A.2.1 Security, privacy, and ethical concerns

The dataset contains news comment text that may include biased, offensive, or politically sensitive content authored by suspected influence operation accounts. Usernames are partially masked in the released dataset. No destructive steps or elevated privileges are required to run the artifact.

A.2.2 How to access

The artifact is available on Zenodo at `https://doi.org/10.5281/zenodo.20257085`. It contains `data.zip` (the full 112M-comment dataset), `model.pth` (the trained KcELECTRA checkpoint), and the source code.

A.2.3 Hardware dependencies

Inference requires a CUDA-enabled GPU. The reference environment uses an NVIDIA GeForce RTX 3090 24 GB with CUDA 12.5 on Ubuntu 20.04.6 LTS. At least 32 GB of system RAM is recommended for large-scale inference. The included 1,000-comment sample completes in under two minutes on a single GPU. Visualization (A.4.2) requires no GPU and runs on a standard CPU-based machine.

A.2.4 Software dependencies

The reference environment runs Ubuntu 20.04.6 LTS. Miniconda 26.1.1 is used to manage the environment; installation instructions are available at `https://docs.anaconda.com/miniconda/`. Python 3.9 or above is required. All dependencies are listed in `environment.yml`. The primary packages are PyTorch 2.6.0 (CUDA 12.4), Transformers 4.57.1, and Polars 1.26.0. The HuggingFace backbone `beomi/kcelectra-base` is downloaded automatically on first run. `environment.yml` specifies PyTorch with CUDA 12.4. Run `nvidia-smi` to check your CUDA driver version; if your CUDA driver version differs, update the `cu124` entries in `environment.yml` before running `conda env create` (e.g., `cu118` for CUDA 11.8).

A.2.5 Benchmarks

The full dataset (`data.zip`) is part of the artifact. A 1,000-comment username-masked sample (`data/sample.parquet`) is included in the code directory to support immediate functionality testing.

A.3 Set-up

A.3.1 Installation

Download and unzip `code.zip` from Zenodo and navigate to the extracted `code/` directory. `model.pth` and source code are in the root of that directory; `sample.parquet` is in the `data/` subdirectory. Create a Conda environment and install dependencies:

```
conda env create -f environment.yml
conda activate naver
```

A.3.2 Basic Test

Run inference on the included sample:

```
python inference.py
```

The expected output ends with:

```
Using device: cuda:0 Loaded 1000 samples from
data/sample.parquet
Saved 1000 predictions to inference/
predictions.json
```

A.4 Evaluation workflow

A.4.1 Major Claims

We designate the findings in Section 5 (*Strategy Analysis*) as the paper's main claims because they can be reproduced without access to identifying user information. By contrast, reproducing the reported model performance would require the original usernames, which cannot be released because doing so could identify individual users.

(C1): *Condemning Korea* is the dominant high-visibility rhetorical strategy among suspected troll accounts, as supported by three findings reported in Section 5 and verified by experiment (E1):

- *Condemning Korea* achieves higher like ratios than all other strategies (mean = 0.514), being the only strategy to exceed the neutral threshold, while *Condemning Rival* (0.422), *Praising MNS* (0.362), and *Praising Partner* (0.241) all fall below 0.5 (Figure 7a).
- As *Condemning Korea* rhetoric intensity increases, predicted like ratio rises above the neutral threshold at intensity = 0.45 ($\beta = 0.1086$, $p < 0.001$), whereas no other strategy significantly boosts like ratio (Figure 7b).
- Among high-visibility *Condemning Korea* comments (like ratio = 1), 7 of the top-10 most targeted entities are political leaders spanning both ideological camps, suggesting that targeting high-salience political figures amplifies engagement (Figure 8).

A.4.2 Experiments

(E1): *[Strategy Analysis]* [20–30 human-minutes + 10 compute-minutes + 3 GB disk + 32 GB RAM]: The aim of this experiment is to generate the figures present in the paper.

Preparation: Complete the installation steps in Section A.3.1. The required data files (`data/suspected_troll_df.parquet` and `data/suspected_troll_spans.parquet`) are included in `code.zip` and will be present in the `data/` subdirectory after extraction. Open `visualization.ipynb` and set the kernel to `naver`.

Execution: Run all cells in `visualization.ipynb` using the button "Run All Cells".

Results: Resulting figures should match Figures 7(a), 7(b), and 8 present in the paper.

A.5 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/usenixsec2026/.