Can You Distinguish Truthful from Fake Reviews?
User Analysis and Assistance Tool for Fake Review Detection

Jeonghwan Kim*, Junmo Kang*, Suwon Shin® and Sung-Hyun Myaeng
Korea Advanced Institute of Science and Technology, School of Computing

* is for equal contribution

Overview

- Customer review authenticity has become crucial in e-commerce platforms.
- We propose a simple assistance tool, You Only Need Gold (YONG), to detect deceptive reviews and augment user discretion.
- We provide an in-depth user understanding on dealing with fake reviews under the guidance of YONG.

Fake Review Detection Tool

You Only Need Gold (YONG)

- YONG provides the gold indicator, which consists of three intuitive, distinct features:
  I. Model decision (Fake / Gold)
  II. Probability (%)
  III. Evidence (word highlights – the more intense the color highlight, the more important of a role the word plays)

- YONG uses BERT as the backbone model, which is finetuned on the OpSpam dataset, with its attention weights visualized as Evidence.
- We leverage YONG to run user evaluations and on human capabilities and tendencies in detecting deceptive reviews.

Research Questions (RQs) / Experiment

I. Can humans outperform models in fake review detection?
II. Can YONG augment human capability?
III. Which feature in YONG influences human decision the most?

- Separate experiment for each RQ
- 24 participants are required to classify fake reviews with & without YONG.
- The test is single-blind; participants don’t know the ground-truth label

Results of Feature-wise Influence

| Condition | Score | Feature | Decision | Probability | Evidence |
|-----------|-------|---------|----------|-------------|----------|
| No tool   | 0.41  | Influence | 3.69     | 3.91        | 1.87     |
| With tool | 0.54  | Influence | 3.69     | 3.91        | 1.87     |
| Model     | 0.70  | Model    | 3.69     | 3.91        | 1.87     |

Results of Experiments 1-2

I. Humans are not good at detecting fake reviews, underperforming the model by a large margin.
II. With YONG, the accuracy increases substantially. (0.41 → 0.54)
III. Among three features of YONG, probability plays the primary role in convincing users.

Discussion

- Human capability of fake review detection is unreliable and requires machine assistance.
- Evidence (interpretable attention visualization) is hardly explicable.
- Interpretablity is different from explainability.
- Assistive tools need to provide faith-gaining features.