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

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

* is for equal contribution

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.

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.

Overview

Fake Review Detection Tool

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  

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score</th>
<th>Feature</th>
<th>Decision</th>
<th>Probability</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>No tool</td>
<td>0.41</td>
<td>Influence</td>
<td>3.69</td>
<td>3.91</td>
<td>1.87</td>
</tr>
<tr>
<td>With tool</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results of Experiments 1-2

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score</th>
<th>Feature</th>
<th>Decision</th>
<th>Probability</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>No tool</td>
<td>0.41</td>
<td>Influence</td>
<td>3.69</td>
<td>3.91</td>
<td>1.87</td>
</tr>
<tr>
<td>With tool</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

• Human capability of fake review detection is unreliable and requires machine assistance.

• Evidence (interpretable attention visualization) is hardly explicable.

• Interpretability is different from explainability.

• Assistive tools need to provide faith-gaining features.