Securing Behavior-based Opinion Spam Detection

Shuaijun Ge, Guixiang Ma, Sihong Xie and Philip S. Yu

Dec 13, 2018 @BigData
Fake reviews
Online reviews

Source: https://www.brightlocal.com/learn/local-consumer-review-survey/ based on a pool of representative sample of 1,031 US-based consumers
The challenges

Is it easy to spot if a review is fake?

- Yes, always, 16%
- Yes, sometimes, 54%
- No, 14%
- I don't know, 16%

Existing efforts

1. Feature engineering
2. Detection models

Outcomes + Explanations

Help make decision
Behavior based **Attacking**

- **Spamming**
- **Account Detection**

**Graph**
- **X** - Spammers
- **+** - Normal users
- **g** - Attack gradient
- **X** - False positive

**Axes**
- **Number of 5-star posts per day**
- **Deviation from avg**

**Linear model**
Behavior based **Attacking**

- Accessing knowledge of detector (publications)

*What yelp fake review filter might be doing, ICWSM, 2013*

![Graphs showing various metrics](image)

- (a) MNR: max # of review per day
- (b) PR: Positive ratio
- (c) RL: Review length
- (d) RD: Rating deviation
- (e) MCS: Maximum content similarity

Number of 5-star posts per day vs. Deviation from avg
Behavior based **Attacking**

- Accessing knowledge of detector (Detection websites)
Behavior based **Attacking**

- **Accessing knowledge of detector** *(Released data)*

We use automated software to recommend the reviews we think will be the most helpful to the Yelp community based primarily on quality, reliability and the reviewer’s activity on Yelp. **Advertisers get no special treatment.** The reviews below didn’t make the cut and are therefore not factored into this business’s overall star rating. Watch the video above or check out our FAQ for more details.
Behavior based **Attacking**

To defend: need to generate the attacks.

**Attack parameters:**
- # of 5-star per day = 4
- Dev from avg = 0.5

**Linear model**

**Actionable?**

[Diagram showing a scatter plot with red 'X' markers and blue '+' markers, indicating the deviation from average.]
Behavior based Attacking

To defend: need to generate the attacks. How?

Actionable attack 1
- post 4 ⭐⭐⭐⭐⭐ per day
- post 1 ⭐⭐⭐⭐⭐ per week

Actionable attack 2
- post 3 ⭐⭐⭐⭐⭐ per day

Actionable attack 3
- post 4 ⭐⭐⭐⭐⭐ per day
- post 1 ⭐⭐⭐⭐⭐ per day
- post 1 ⭐⭐⭐⭐⭐ per day

Attack parameters:
- # of 5-star per day = 4
- Dev from avg = 0.5
Behavior based **Attacking**

Spammer objective function = \(\text{(risk of being detected)} - \text{(profit of spamming)}\)

Temporal anomalies

Change in rating

AVG rating

Deviation from predicted avg

Predicted AVG rating
Behavior based Attacking

Spammer objective function = \((\text{risk of being detected}) - (\text{profit of spamming})\)

Rating distribution anomaly

\[
KL(p||\bar{p}) = \sum_{i=1}^{5} p_i \log \frac{p_i}{\bar{p}_i}
\]

\[\begin{array}{cccc}
\text{5 star} & \text{4 star} & \text{3 star} & \text{2 star} & \text{1 star} \\
73\% & 12\% & 5\% & 3\% & 7\%
\end{array}\]

\[\begin{array}{cccc}
\text{5 star} & \text{4 star} & \text{3 star} & \text{2 star} & \text{1 star} \\
83\% & 13\% & 2\% & 1\% & 1\%
\end{array}\]
Behavior based **Attacking**

Spammer objective function = \( \text{risk of being detected} - \text{profit of spamming} \)

Rating distribution anomaly

\[
\begin{align*}
\text{EN}(t) &= - \sum_{i=1}^{5} p_i(t) \log p_i(t) \\
\Delta\text{EN} &= \text{EN}(t + 1) - \text{EN}(t)
\end{align*}
\]

\( \bar{p} \) : Rating dist at time \( t \)

\( p \) : Rating dist at time \( t+1 \)

<table>
<thead>
<tr>
<th>Rating</th>
<th>( \bar{p} )</th>
<th></th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>73%</td>
<td></td>
<td>83%</td>
</tr>
<tr>
<td>4 star</td>
<td>12%</td>
<td></td>
<td>13%</td>
</tr>
<tr>
<td>3 star</td>
<td>5%</td>
<td></td>
<td>2%</td>
</tr>
<tr>
<td>2 star</td>
<td>3%</td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>1 star</td>
<td>7%</td>
<td></td>
<td>1%</td>
</tr>
</tbody>
</table>
Behavior based **Attacking**

Spammer maximizes \( \text{risk of being detected} - \text{profit of spamming} \)
Behavior based **Attacking**

Find amount of promotion

Temporal change in AVG

AVG rating

Manipulated

Predicted

Organic AVG

Deviation from predicted avg

promotion >= 0

Cap of all ratings <= 5

promotion

are set to 80th percentiles of the corresponding changes estimated from the historic data
Behavior based **Attacking**

Find a proper amount of promotion in AVG rating $\delta$

Large temporal change in AVG?

![Diagram showing changes in AVG over time]

are set to 80th percentiles of the corresponding changes estimated from the historic data
Behavior based **Attacking**

find a proper number of spamming ratings $n_\delta$

Large incremental in the number of reviews?

$\leq 80^{th}$ percentile of historic increments

Large absolute number of reviews?

$\leq 80^{th}$ percentile of historic NR
Behavior based **Attacking**

Compute an evasive rating distribution $p$

<table>
<thead>
<tr>
<th>Rating</th>
<th>Background $\bar{p}$</th>
<th>Rating dist at time $t$ $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>73%</td>
<td>83%</td>
</tr>
<tr>
<td>4 star</td>
<td>12%</td>
<td>13%</td>
</tr>
<tr>
<td>3 star</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>2 star</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>1 star</td>
<td>7%</td>
<td>1%</td>
</tr>
</tbody>
</table>

Optimal rating distribution found by the dual problem.
Behavior based **Attacking**

The found evasive rating distribution $\mathbf{p}$

- 5 star: 83%
- 4 star: 13%
- 3 star: 2%
- 2 star: 1%
- 1 star: 1%

$n_\delta = 60$
- x 50
- x 0
- x 0
- x 3
- x 7
Behavior based **Attacking**

Flexible attacks generation.

- Evade time series based detectors
  - E1: NR
  - E2: NR + ∆NR
  - E3: NR + deviation in AVG rating
  - E4: NR + deviation in AVG rating + change in AVG rating

- Evade both time series and rating distribution based detectors
  - E5: KL-DIV
  - E6: KL-DIV + change in entropy
  - E7: ...
  - E8: Max Entropy + change in entropy
  - E9: Max Entropy + change in entropy

- For short history targets
  - E-A: NR + ∆NR + change in AVG rating + Max Entropy
  - E-B: NR + ∆NR + change in AVG rating
Behavior based Attacking

Targets with long review histories
- Products with $\geq 1,000$ reviews
- Reviews span more than 37 months (Yelp) / weeks (Amazon)
- 1,175,088 reviews / 383 products
- 247,117 reviews / 327 restaurants.

Targets with short review histories
- The remaining products / restaurants are used.
- Longitudinal data are too sparse for each target.

Probe parameters
- Long-history data
- last 5 weeks

Attack!

Product 1

Product 383
Behavior based **Attacking**

Average spams posted by each attack

Evasion strategies

- **Amazon**
  - Late negatives
  - Late positives
  - Early negatives
  - Early positives

- **TripAdvisor**
  - Late negatives
  - Late positives
  - Early negatives
  - Early positives
Behavior based **Attacking**

Attacking rate (% of windows can be spammed)

![Bar charts for Amazon and TripAdvisor showing behavior-based attacking rates for different evasion strategies](chart.png)
Behavior based **Attacking**

Promotion in ranking per spam

![Graph for Amazon](image1)

![Graph for TripAdvisor](image2)

Evasion strategies

Late
Early
Behavior based **Attacking**

Secure the detector again

![Graph showing deviation from average vs. number of 5-star posts per day.]

- **Linear model**
- **Re-trained linear model**

- **X** Spammers
- **+** Normal users
- **X** False positive
- **g** Attack gradient
Behavior based **Attacking**

- **Target**
  - First 30 weeks
  - Last 5 weeks

- **Probe parameters** → **Attack simulation** → **Attack in the wild!**

- **Training data**
  - Generated from E1
  - Generated from E2
  - Generated from E9

- **Pooling (DETER)**

- **Model re-training**
  - Model 1
  - Model 2
  - Model 9
  - Ensemble
Behavior based **Attacking**

Full information detection / evasion game: single spammer

<table>
<thead>
<tr>
<th>Game 1</th>
<th>Detector</th>
<th>Game 2</th>
<th>Detector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kl-Div</td>
<td># of reviews</td>
<td>Kl-Div</td>
<td># of reviews</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Equilibrium
Behavior based **Attacking**

![Graph showing AUC of spammed window detection (Yelp)](image)

- **W^m**: Max of signals
- **W^a**: Avg of signals
- **W^r**: Randomly selection
- **EN_A**: Re-train avg
- **EN_M**: Re-train Max
- **DETER**: Re-train Pool
- **Max-min**: Game equilibrium
Behavior based **Attacking**

![Graph showing AUC of spammed window detection (Amazon)]

- $W^m$: Max of signals
- $W^a$: Avg of signals
- $W^r$: Randomly selection
- EN_A: Re-train avg
- EN_M: Re-train Max
- DETER: Re-train Pool
- Max-min: Game equilibrium
Behavior based **Attacking**

- Unsupervised
- Attack agnostic
- Simple and good performance
- Good for long and short review histories
- Can secure the detector!
- Source codes and data available at:
  
  https://bitbucket.org/Doris_Ge/bigdata18_spam_detection
  
  http://www.cse.lehigh.edu/~sxie/codes.html
Thank you