Misinformation detection for e-commerce

Sihong Xie, Assistant Professor
Computer Science and Engineering
Lehigh University
Misinformation are prevalent

Estimated percentage of fake reviews on popular e-commerce websites.

Source: BusinessInsider, ChicagoTribune
Misinformation are hard to spot

Based on a 2017 pool of representative 1,031 US-based consumers

Source: https://www.brightlocal.com/learn/local-consumer-review-survey/
Existing efforts

**How to spot a fake review**

**Likely a fake review:**
- Extreme review
- Few details about the reviewer
- Unestablished reviewer with few, one-sided reviews
- Few "helpful" votes by other consumers
- Unverified purchase

**Pros:**
- Rely on yourselves

**Cons:**
- Not everyone can spot fake reviews

**Likely a genuine review:**
- Balanced review
- Many details about the reviewer
- Established reviewer with many two-sided reviews
- Many "helpful" votes by other consumers
- Verified purchase

**Pros:**
- Convenient and easy

**Cons:**
- Can be gamed

---

**ReviewMeta.com**

1. Feature engineering
2. Detection models

---

**By Damaya on 2 March 2018**

**Verified Purchase**

Very good product and nice this is awesome product and quality is good

One person found this helpful

---

**By Nausad on 3 March 2018**

**Verified Purchase**

Awesome product and quality is good quality and this is nice product and nice this is awesome

2 people found this helpful

---

**By Diviya on 2 March 2018**

**Verified Purchase**

Nice and cool product and quality is good

One person found this helpful

---

**Analysis Details**

**FAIL**

- Suspicious Reviewers

**Take-Back Reviewers**

- 40% of these reviews have 4 stars.

**Single-Day Reviewers**

- 20% of these reviews have 4 stars.

---

**More in-depth automatic analysis**

- Can be gamed
Misinformation detection architecture

**Algorithms**

- Time series motif finding
- Message passing
- Word embedding

**Features**

- Time series
- Graphs
- Texts

**Data**

- Amazon
- Google
- Yelp
- Facebook
- TripAdvisor
Review Spam Detection via Temporal Pattern Discovery

Sihong Xie, Guan Wang, Shuyang Lin, Philip S. Yu
Why detection is so hard

Each account posts just one review.
Can you spot the fake ones?

Number of accounts posting a number of reviews follows a power law distribution.

\[ P(k) \propto k^{-\gamma} \]

---

"My experience with B&H Photo-Video-Pro Audio was excellent. Order arrived sooner than expected and in good shape. I am very satisfied with my new Acer Iconia Tab A200 and accessories, and the Acer tablet was 50$ less than my local Best Buy and paid no tax or shipping to boot."

"Good prices, easy-to-use website, efficient delivery, they make it tough to consider going elsewhere."

"Found everything I was looking for in a single stop. Super fast shipping."
Exploiting an invariance of spamming

**Invariance:**
to manipulate ratings, a **large number of consistent ratings** must be posted in a **short time**.

Temporal features for each window:

- **AVG rating**
- **Review Volume**
- **Singleton Review Volume**

Burst motif detection on all 3 series.
Results

Manual labeling of dishonest businesses

- Hard to evaluate the recall rate.
- Only label the top 53 stores with most reviews.
- Humans background-checked stores on Google and BBB.

<table>
<thead>
<tr>
<th></th>
<th>Evaluator 1</th>
<th>Evaluator 2</th>
<th>Evaluator 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator 1</td>
<td>17</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Evaluator 2</td>
<td>-</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Evaluator 3</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
</tbody>
</table>

Burst detector Performance

- Normal: 29
- Suspicious: 24
- False Positive: 22
- True positives: 14
Case study

A detected 15-day window

AVG rating: 4.4 → 4.79
Volume of reviews: 57 → 154
Ratio of singleton reviews: 61% → 83%

More evidences

Burst of “hurry reviewers”

Burst of positive singleton reviews

Burst of reviews with key phrase “customer service”
Overview of my research

Securing Behavior-based Opinion Spam Detection
Shuaijun Ge, Guixiang Ma, Sihong Xie, and Philip S. Yu
Evading a spam detector

A strategic spammer will be more careful in posting fake reviews.

The strategic spammer will try to avoid the detection while manipulate the rating.

Risk of being detected vs. Profit of spamming

A strategic spammer will be more careful in posting fake reviews.

The strategic spammer will try to avoid the detection while manipulate the rating.
Evading a spam detector

Multiple detection signals need to be evaded:

- Number of reviews
- Change in the number of reviews

- Deviation from baseline average rating
- Change in rating

- Rating distribution
- Change in the rating distribution

\[ n_\delta = 60 \]

\[ \text{Rating optimization} \]

Max: rating
Min: risk

\[ \bar{p} : \text{Background/last distribution} \]

\[ p : \text{Rating dist at time t} \]

<table>
<thead>
<tr>
<th>Star Rating</th>
<th>( \bar{p} )</th>
<th>( 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>
Data augmentation for robust detection

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

First 30 weeks

- Training data generated from Evasion 1
- Training data generated from Evasion 2
- Training data generated from Evasion 9

Pooling and Re-training (DETER)

Model re-training and ensemble

- Linear detector
  - Spammers
  - Normal users
  - Attack direction

This technique is applicable to stores with sparse review data.
Robustness of the re-trained detector

Base detectors using statistics of time windows: number of reviews, positive review ratio, change in rating distribution, ...

\[ W^m: \text{Max of signals} \]
\[ W^a: \text{Avg of signals} \]
\[ W^r: \text{Random selection} \]
\[ EN_M: \text{Re-train Max} \]
\[ EN_A: \text{Re-train Avg} \]
\[ DETER: \text{Re-train Pool} \]
\[ \text{Max-min: Game equilibrium} \]
Overview

Review Graph Based Online Store Review Spammer Detection
Guan Wang, Sihong Xie, Bing Liu, Philip S. Yu
Detecting productive spammers

Can a reviewer with a long history and many reviews be a spammer?

- Diverse review texts
- Diverse ratings
- Spreaded out temporally

A strategic spammer can build credibility over time to hijack ratings.
Dependent trustworthiness on a graph

Algorithm:
iteratively and alternatively calculate Trustworthiness, Honesty, and Reliability, relying on previously computed quantities.

Similar trustworthiness with similar-minded reviews

\[ A(v, \Delta t) = \sum_{i \in S_{v,a}} T(\kappa_i) - \sum_{j \in S_{v,b}} T(\kappa_j) \]

\[ A_n(v, \Delta t) = \frac{2}{1 + e^{-A(v, \Delta t)}} - 1 \]

Reliabilitiy of the business \( s \)

\[ R(s) = \frac{2}{1 + e^{-\xi}} - 1 \]

\[ \xi = \sum_{v \in U, T(\kappa_v) > 0} T(\kappa_v)(\Psi_v - \mu) \]

Honesty of the review \( v \)

\[ H(v) = |R(\Gamma_v)|A_n(v, \Delta t) \]

Trustworthiness of the reviewer \( r \)

\[ T(r) = \frac{2}{1 + e^{-H_r}} - 1 \]

\[ H(r) = \sum_i \text{honest scores of } i\text{-th review by } r \]
Experiments

- Reviews: 408,470
- Reviewers: 343,603
- Stores: 14,561

Dataset size

Distribution of number of reviews per store

\[ \log(P(k)) = -1.43 \times 0.44 \]

\( \text{log} \) (store review number)
Experiments

Convergence of rustworthiness

Quantitative evaluation
- Focused on precision@100

Evaluator inspection outcome and agreement

Evaluator agreement are statistically significant (kappa=60.3%)

Qualitative evaluation

Top reliable stores

<table>
<thead>
<tr>
<th>Store Name</th>
<th>Reseller ratings Rating</th>
<th>BBB Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>TigerDirect</td>
<td>7.44</td>
<td>A</td>
</tr>
<tr>
<td>SuperMediaStore</td>
<td>9.27</td>
<td>A⁺</td>
</tr>
<tr>
<td>OneCall</td>
<td>9.33</td>
<td>A⁺</td>
</tr>
<tr>
<td>Newegg</td>
<td>9.77</td>
<td>A⁺</td>
</tr>
<tr>
<td>Mwave</td>
<td>9.18</td>
<td>B⁺</td>
</tr>
<tr>
<td>LA Police Gear</td>
<td>9.11</td>
<td>A⁻</td>
</tr>
<tr>
<td>iBuyPower</td>
<td>8.33</td>
<td>B⁺</td>
</tr>
<tr>
<td>FrozenCPU</td>
<td>9.44</td>
<td>A⁺</td>
</tr>
<tr>
<td>eWiz</td>
<td>9.08</td>
<td>C</td>
</tr>
<tr>
<td>eForcity</td>
<td>8.55</td>
<td>A⁺</td>
</tr>
</tbody>
</table>

Bottom reliable stores

<table>
<thead>
<tr>
<th>Store Name</th>
<th>Reseller ratings Rating</th>
<th>BBB Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>86th Street Photo</td>
<td>0.30</td>
<td>F</td>
</tr>
<tr>
<td>Best Price Cameras</td>
<td>1.43</td>
<td>F</td>
</tr>
<tr>
<td>Dealer Cost Car Audio</td>
<td>1.23</td>
<td>F</td>
</tr>
<tr>
<td>USA Photo Nation</td>
<td>0.20</td>
<td>F</td>
</tr>
<tr>
<td>Camera Addict</td>
<td>0.59</td>
<td>F</td>
</tr>
<tr>
<td>CCI Camera City</td>
<td>0.44</td>
<td>A⁺</td>
</tr>
<tr>
<td>OC System</td>
<td>3.00</td>
<td>F</td>
</tr>
<tr>
<td>Shop Digital Direct</td>
<td>0.35</td>
<td>F</td>
</tr>
<tr>
<td>Camera Giant</td>
<td>0.21</td>
<td>F</td>
</tr>
<tr>
<td>Infiniti Photo</td>
<td>0.28</td>
<td>F</td>
</tr>
</tbody>
</table>
Reinforcement learning for robust graph-based detection

Robust Spammer Detection by Nash Reinforcement Learning
Yingtong Dou, Guixiang Ma, Philip S. Yu, and Sihong Xie
Reinforcement learning for robust graph-based detection

- **Previous works:**
  - Static dataset
  - Accuracy-based evaluation metric
  - Fixed spamming pattern
  - Single detector

- **Our work:**
  - Dynamic game between spammer and defender
  - Practical evaluation metric
  - Evolving spamming strategies
  - Multiple detectors ensemble
Rating and revenues

In Yelp, product’s rating is correlated to its revenue[1]

Revenue Estimation: \[ f(v; \mathcal{R}) = \beta_0 \times \text{RI}(v; \mathcal{R}) + \beta_1 \times \text{ERI}(v; \mathcal{R}_E(v)) + \alpha \]

Spammer and detector goals

\[ p: \text{Spamming strategy} \]
\[ q: \text{Detector strategy} \]

**Spamming Practical Effect:**
\[ PE(v; R, p, q) = f(v; R(p, q)) - f(v; R) \]

- Revenue after attacks and detection.
- Revenue before attacks

**Spammer’s Goal:**
\[ \max_p \max\{0, PE(v; R, p, q)\} \]

**Defender’s Goal:**
\[ \min_q \mathcal{L}_q = \frac{1}{|R(p, q)|} \sum_{r \text{ is FN}} -C_{\text{FN}}(v, r) \log P(y = 1| r; q) \]

- The cost of false negatives
- The prediction results of detectors

Practical effect and detection recall are not in the same battlefield.
Robust detector: Nash-Detect

Spammer & Attacks
(1) Execute spamming strategies $A(p)$

Defender & Detectors
(2) Remove spams using $D(q)$

(4) Back-propagation
(5) Restart the attack and defense game

(3) Compute Practical Effect

Inference  Learning  Generated spams  Detected Spams
### Experimental settings

#### Base attack algorithms

1. **IncBP**: add reviews using the least suspicious accounts based on MRF.

2. **IncDS**: add reviews using accounts in the least dense block on review graph.

3. **IncPR**: add reviews using the least suspicious accounts based on behavior features.

4. **Random**: randomly select existing accounts to add reviews.

5. **Singleton**: add reviews with new accounts.

#### Base detection algorithms

1. **GANG**: MRF-based detector

2. **SpEagle**: MRF-based detector

3. **fBox**: SVD-based detector for finding subtle changes in a large graph.

4. **Fraudar**: Dense-block detector

5. **Prior**: Behavior-based detector (rating changes, deviations, posting volume, etc.)
• For a fixed detector (Fraudar), the spammer can switch to the spamming strategy with the max practical effect (IncDS/Random)

• The practical effect of detectors configured by Nash-Detect are always less than the worst-case performances.
Transparency

• Model debugging:
  o why my algorithm is not detecting these fake reviews?
  o why the false positive rate is so high?

• Users’ right to know:
  o why these reviews are removed?

• Auditing:
  o privacy
  o fairness
Fairness

• Auditing: company reputation and legal concerns.

• Are businesses treated equally:
  o some businesses may have advantage over others, based on regions, types, size, etc.

• Are customers have equal right to review products/businesses?:
  o It is not right to delete more of the new-comers’ reviews, though they have a high chance to be spammed.