A Survey on Spam Review Detection

Suna K. Steffen
What Is Spam Review?

- Opinion spams in online review system
  - (1) Untruthful reviews
  - (2) Biased reviews
  - (3) Non-reviews (e.g., a review that is rather asking a question)

- Type (1) is often called as a fake review - usually the source of study

- The truthfulness of a review text is not distinguishable and spam reviewers can easily fake to be innocent reviewers.
Characteristics of the Problem - Review Dataset

- Generally, review data consists of review text, rating, review date, reviewer, product, and feedbacks on reviews.
- All the attributes are supposed to be independent from each other (Jindal 2010).
The distribution of review system follows **power-law** (Jindal 2008)
Characteristics of Spam Review

- **Purpose of Spam Review**
  - In order to *promote* specific product/product group
  - To *hurt reputations* of competitors

An Example of Economic Incentives (Retrieved from Yelp.com’s blog)
Characteristics of Spam Review - Review Contents

- The content of the reviews are more imaginative and fictional (Ott 2011)
- More likely duplicates or similar reviews (Jindal 2008)
- The rating deviates a lot from other reviews for a product, therefore, it has more significant impact on the total rating.
Characteristics of Spam Review

- Reviewers

- Reviewers with too many reviews are more likely spam reviewers (Jindal 2008)

- Reviewers with only one review are more likely spam reviewers (Jindal 2008)

- Spammers can have multiple accounts and write spam reviews as an individual reviewer, which makes it hard to detect (Lim 2010)
Characteristics of Spam Review - Time Attribute

- Spam reviewers tend to leave many reviews in a short time frame for most economic incentive.
- Spam behaviors are more involved in the earlier in the total reviewing time frame, therefore affecting other reviewers. (Luca 2013)
- Or when the business is going down. (Luca 2013)
Characteristics of Spam Review
- Incentives (changes over time)

- **Low ratings increase incentives** for positive review fraud, and high ratings decrease them.

- **Having more review reduces incentives** for positive review fraud (because its impact will be smaller)

- **Restaurants/products with fewer reviews** are more likely to engage in spam review.

- **Chain restaurants** leave fewer spam reviews.

- Businesses signaled by claiming their pages on Yelp.com, will engage in more review fraud.
Challenges

- Hard to get manual labels for training. (Jindal 2008)
- Many duplicate/similar spam reviews - distance based outlier detection wouldn’t work
- Spam reviewers can fake their IDs
- Helpfulness (feedback) of reviews can be manipulated
- There is no gold-standard truth to compare the result
Insights

- Machine can categorize truthful and untruthful texts by natural language processing technique
- Attributes of reviews should be independent
- We can model extreme spamming behaviors without relying on text
- Group spammer behavior looks similar to collective outliers
- Heterogeneous review network can be used
Current Approaches

- Classification Approach
- Ranking Approach
Classification Approach

- A binary classification problem between spam and non-spam reviews.

- Problem: *Labels are expensive*.
  - It is hard to distinguish which is trustful or not by human annotators.
  - It takes a lot of time to read multiple reviews and judge them.

- Solution: *Provide labels! But how?*
(1) Use duplicate/similar reviews as training set (Jindal 2008)

- Insight: spam reviewers would quickly copy and paste their reviews from the past or write similar reviews in a different user name.
- Detect duplicate reviews using the shingle method
  - 2-gram based review content comparison
  - Jaccard distance: measure the similarity score of two reviews by the ratio of intersection of their 2-grams to the union of their 2-grams of the two reviews.
- Review pairs with similarity score of at least 90% were chosen as duplicates
- Treat duplicate/similar reviews as spam review training set
- Drawbacks: The assumption is limited to certain spam behavior
(2) Learn the rules from data’s statistical expectation (Jindal 2010)

- Insight: independence of attributes
- Confidence Unexpectedness
  - Unexpectedness between review rating and reviewers.
  - For example, reviewers (with at least 3 reviews) who wrote only positive reviews and only negative reviews are suspicious or unexpected
- Support Unexpectedness
  - Unexpectedness between the number of reviews and reviewers.
  - For example, if a reviewer wrote 626 reviews and all of them have positive ratings, it is highly unusual.
- Captures unusual phenomenon in review system.
- Drawbacks: It does not fully use many distinctive characteristics of spam review
(3) Create real spam reviews as training set (Ott 2011)

- Create real spam and non-spam reviews from Mechanical Turk
  - large scale gold-standard spam review data set

- Drawbacks
  - This data set is limited to specific domain
  - Costs money to create
  - Static data set - spam reviewers can evolve in their writing strategy and win the game eventually
Ranking Approach

- How to measure spamicity?
- Use heuristic rules based on spam review behavior patterns
- Use distortion as measure
(1) Rank **the reviewers** (Lim 2010)

- **Insight:** *Rank reviewers instead of reviews by themselves*
- Measure spam score functions based on spam review **behavioral patterns**
  - Targeting a specific product/product groups to promote
  - Deviating a lot from other reviews in their reviews for a product. (Lim 2010)
- Combine spam score functions to get one final ranking
(1) Rank the reviewers (Lim 2010)

- 4 Score functions
  - **Single product multiple reviews**: determined by similarity to other reviews for a single product
  - **Single product group multiple ratings**: captures the behavior of a reviewer who has a purpose of promoting its own business and hurting the reputation of competing products in the same category
  - **General Deviation**: Average of all deviations for his reviews (deviation of a rating is its difference from the average rating on the same product)
  - **Early Deviation**: a summation of deviations with the weight that is inversely proportionate to the time of the review

- **Drawback**: Weights in the final ranking function are empirically determined
(2) Rank **the group of reviewers**

*(Mukherjee 2011)*

- **Insight:** Use the pattern of **group behavior** (collective behavior) of spam reviews

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Reviewer's Name</th>
<th>Profile Link</th>
<th>Review Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John</td>
<td><a href="#">Big John's Profile</a></td>
<td>Practically FREE music, December 4, 2004 (Audio Xtract)</td>
</tr>
<tr>
<td>2</td>
<td>Cletus</td>
<td><a href="#">Cletus' Profile</a></td>
<td>Like a tape recorder..., December 8, 2004 (Audio Xtract)</td>
</tr>
<tr>
<td>3</td>
<td>Jake</td>
<td><a href="#">Jake's Profile</a></td>
<td>Wow, internet music! ..., December 4, 2004 (Audio Xtract)</td>
</tr>
</tbody>
</table>

---

**Figure 1:** Big John's Profile

- **Rating:** 5 stars
- **Comment:** Practically FREE music, December 4, 2004
- **Software:** Audio Xtract (CD-ROM)
- **Summary:** I can't believe for $10 (after rebate) I got a program that gets me free unlimited music. I was hoping it did half what was advertised, but it does more than advertised.

**Figure 2:** Cletus' Profile

- **Rating:** 4 stars
- **Comment:** This software really rocks. I can set the program to record music all day long and just let it go. I come home and my computer is filled with music.
- **Software:** Audio Xtract (CD-ROM)
- **Summary:** This review is from Audio Xtract (CD-ROM).

**Figure 3:** Jake's Profile

- **Rating:** 2 stars
- **Comment:** Best music just got ..., December 4, 2004
- **Software:** Audio Xtract (CD-ROM)
- **Summary:** I looked forever for a way to record internet music. My way took a long time and many steps (frustrating). Then I found Audio Xtract. With more than 3,000 songs downloaded in ...

---

**Table Notes:**

- **Figure 1:** John's review details.
- **Figure 2:** Cletus' review details.
- **Figure 3:** Jake's review details.
(2) Rank **the group of reviewers**
(Mukherjee 2011)

- **Step 1:** frequent pattern mining
  - Define an item as a set of transaction data which consist of a product and reviewers.
  - Extract the items from data then mine for frequent items

- **Step 2:** **Rank on Groups**
  - Three heuristic functions
  - Use the results of three ranking functions to SVN ranking to get one final single ranking.

- Drawback: **Heuristic rules can be further tuned.**
(3) Rank on a review network
(Wang et al. 2011-2012)

- Insight: Ranking with three attributes that can mutually enhance each other, giving us an iterative framework.
(3) **Rank on a review network**

(Wang et al. 2011-2012)

- Assign three nodes with their own measures:
  - Reviewer - Truthfulness
  - Review - Honesty
  - Product/Store – Reliability

- **One heuristic rule:** Agreement between reviews (based on time/rating) propagates through the network

- **Drawback**
  - Not taking advantage of the group spam review behavior
  - Not using more information that is available
Use *distortion* as measure (Yoo 2009)

- Insight: distortion is effective in **magnifying the difference** between true spam reviews and non-spam reviews.
- Distortion measure: comparing popularity rankings before and after deletion of singleton reviews from products.
  - Singleton review: reviews from reviewers who have only one review
- Drawback
  - Limited assumption (singleton review could be valid)
  - Not fully using the information from the dataset
Suggestions For the Future Work

- **Better evaluation**: no gold-standard truth to *fairly* evaluate these approaches
- Take advantage of *all the information* we can gain from review network
  - Spam review indicators such as economic incentives that change over time
- **Collective outlier analysis** can be further investigated
- Time attribute has been added to spot group spam review. How about *IP-address*?
Thank you!
References


References

References


References


