Detecting Fake Opinions in Social Media

Bing Liu
Department of Computer Science
University Of Illinois at Chicago
liub@cs.uic.edu

Introduction

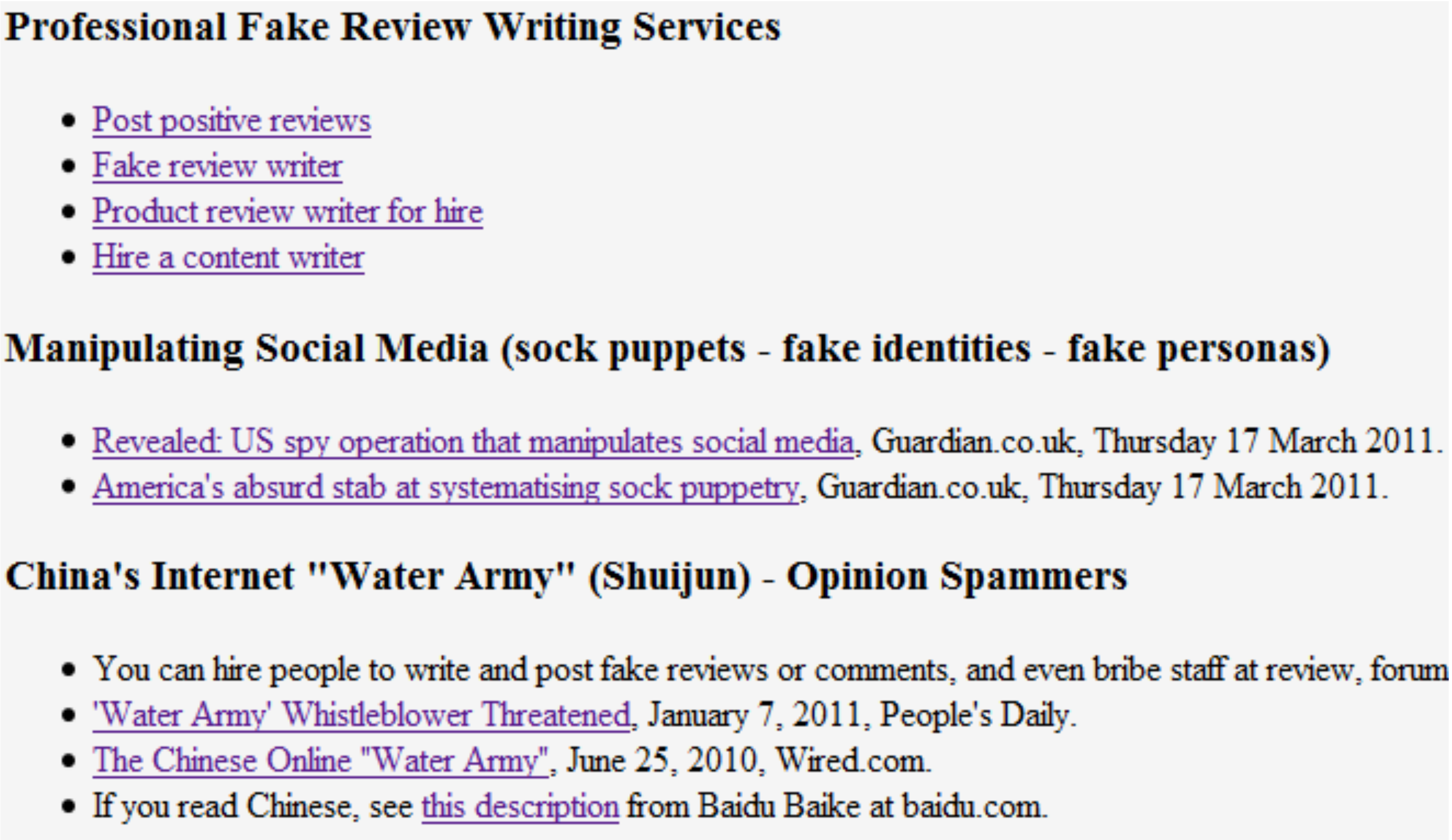
- Opinions from social media are increasingly used by individuals and organizations for
  - making purchase decisions
  - marketing and product design
  - making choices at elections
- Positive opinions often mean profits and fame for businesses and individuals,
  - Unfortunately, this gives strong incentives for people to game the system by posting fake opinions and reviews.

Opinion spam detection

(Jindal and Liu, 2007, 2008)

- Opinion spamming refers to people giving fake or untruthful opinions, e.g.,
  - Write undeserving positive reviews for some target entities in order to promote them.
  - Write unfair or malicious negative reviews for some target entities in order to damage their reputations.
- Opinion spamming has become a business in recent years.
- Increasing number of customers are wary of fake reviews (biased reviews, paid reviews)

Problem is wide-spread



Professional Fake Review Writing Services

- Post positive reviews
- Fake review writer
- Product review writer for hire
- Hire a content writer

Manipulating Social Media (sock puppets - fake identities - fake personas)

- Revealed: US spy operation that manipulates social media, Guardian.co.uk, Thursday 17 March 2011.
- America's absurd stab at systematising sock puppetry, Guardian.co.uk, Thursday 17 March 2011.

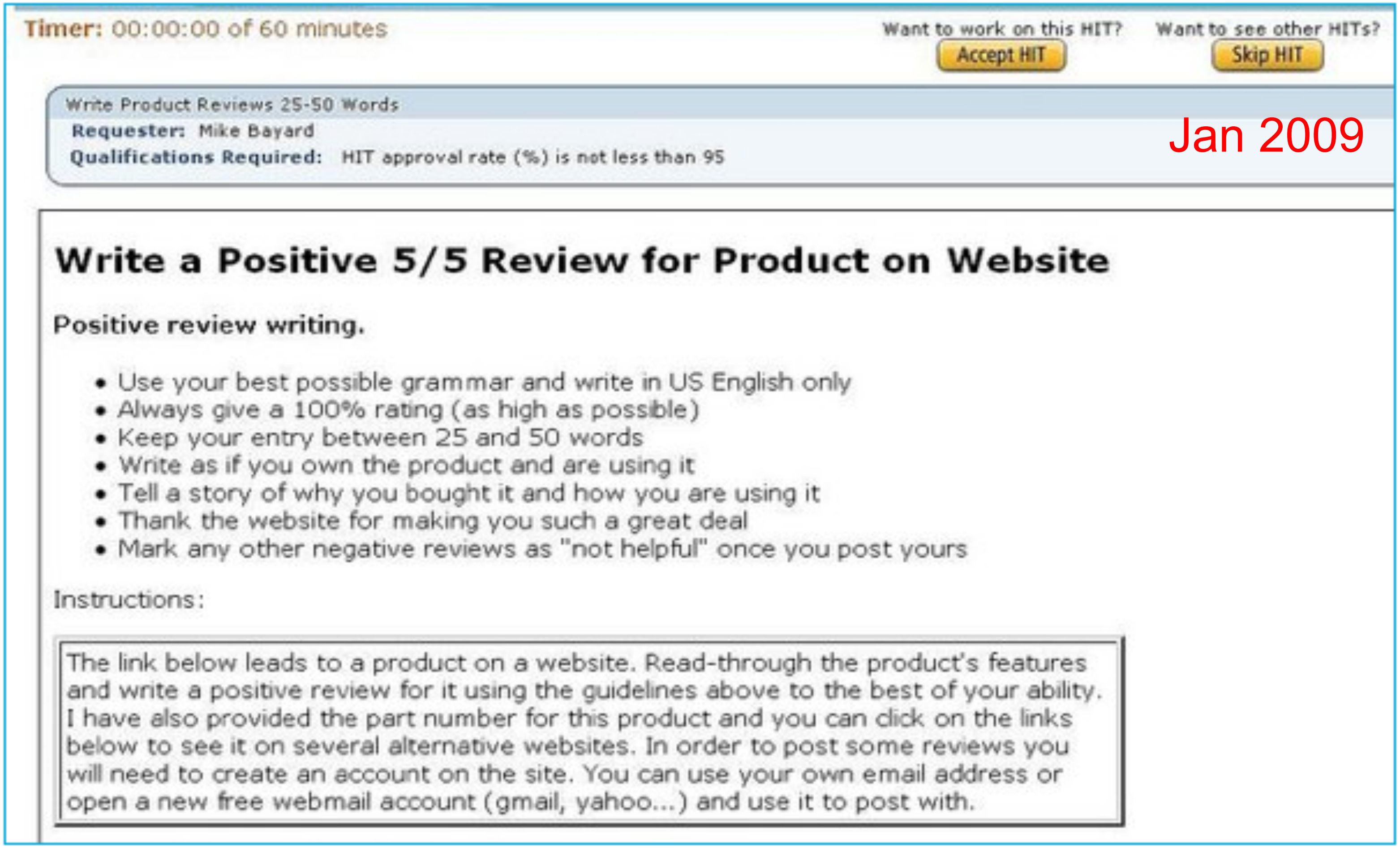
China's Internet "Water Army" (Shuijun) - Opinion Spammers

- You can hire people to write and post fake reviews or comments, and even bribe staff at review, forum
- 'Water Army' Whistleblower Threatened, January 7, 2011, People's Daily.
- The Chinese Online "Water Army", June 25, 2010, Wired.com.
- If you read Chinese, see this description from Baidu Baike at baidu.com.

An example practice of review spam

Belkin International, Inc

- Top networking and peripherals manufacturer | Sales ~ $500 million in 2008
- Posted an ad for writing fake reviews on amazon.com (65 cents per review)



Timer: 00:00:00 of 60 minutes

Want to work on this HIT? Accept HIT

Want to see other HITs? Skip HIT

Write Product Reviews 25-50 Words

Requester: Mike Bayard

Qualifications Required: HIT approval rate (%) is not less than 95

Jan 2009

Write a Positive 5/5 Review for Product on Website

Positive review writing.

- Use your best possible grammar and write in US English only
- Always give a 100% rating (as high as possible)
- Keep your entry between 25 and 50 words
- Write as if you own the product and are using it
- Tell a story of why you bought it and how you are using it
- Thank the website for making you such a great deal
- Mark any other negative reviews as "not helpful" once you post yours

Instructions:

The link below leads to a product on a website. Read-through the product's features and write a positive review for it using the guidelines above to the best of your ability. I have also provided the part number for this product and you can click on the links below to see it on several alternative websites. In order to post some reviews you will need to create an account on the site. You can use your own email address or open a new free webmail account (gmail, yahoo...) and use it to post with.

Is this review fake or not?

I want to make this review in order to comment on the excellent service that my mother and I received on the Serenade of the Seas, a cruise line for Royal Caribbean. There was a lot of things to do in the morning and afternoon portion for the 7 days that we were on the ship. We went to 6 different islands and saw some amazing sites! It was definitely worth the effort of planning beforehand. The dinner service was 5 star for sure. One of our main waiters, Muhammad was one of the nicest people I have ever met. However, I am not one for clubbing, drinking, or gambling, so the nights were pretty slow for me because there was not much else to do. Either than that, I recommend the Serenade to anyone who is looking for excellent service, excellent food, and a week full of amazing day-activities!

What about this?

The restaurant is located inside of a hotel, but do not let that keep you from going! The main chef, Chef Chad, is absolutely amazing! The other waiters and waitresses are very nice and treat their guests very respectfully with their service (i.e. napkins to match the clothing colors you are wearing). We went to Aria twice in one weekend because the food was so fantastic. There are so many wonderful Asian flavors. From the plating of the food, to the unique food options, to the fresh and amazing nan bread and the tandoori oven that you can watch as the food is being cooked, all is spectacular. The atmosphere and the space are great as well. I just wished we lived closer and could dine there more frequently because it is quite expensive.

One more?

Cameraworld is on my list of top photography/video equipment e-tailers. Their reps answer phones from early in the morning through late at night. The service is also first rate and the staff there is knowledgeable on the products they sell. Prices are competitive, although not always the best, but they do price match should you find it cheaper.

I have noticed that some of the products they carry, only a select few that are rare, are not listed on the website even though Cameraworld either stocks or is willing to get for you. This is only a minor inconvenience, and isn't really a bother to me as I normally have other questions that I can get answered when calling.

They also have a "Bonus Bucks" program in which online purchases receive a percentage credit towards a future purchase. I have yet to make a purchase online (always phoned in orders), so no experience with the program.

Detecting fake review is hard

- Different from Web spam or email spam
  - Web spam: link spam and content spam
  - Email spam: mostly commercial ads
- For such spam, when you see it, you know it.
  - Easy to find training data for model building
  - Easy to evaluate the resulting models
- Fake reviews (opinion spam in general)
  - No link or content spam
  - Almost no commercial ads

Detecting fake review is hard (contd)

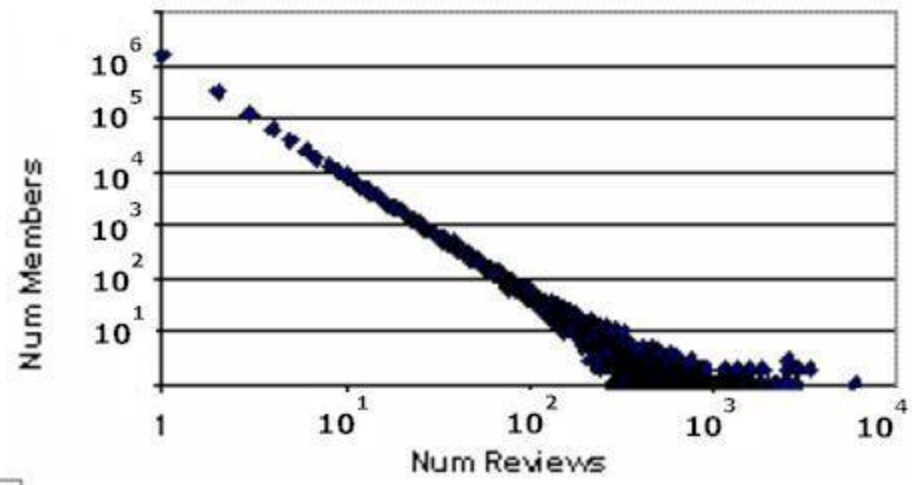
- Fake reviews
  - When you see it, you do not know it.
  - Can only be reliably identified by their authors!
- If one writes carefully, there is almost no way to identify them by their content.
- Logically impossible!
  - I write a truthful 5-star review for a good hotel.
  - But I post the review to another hotel that I want to promote.

A Study of Amazon Reviews

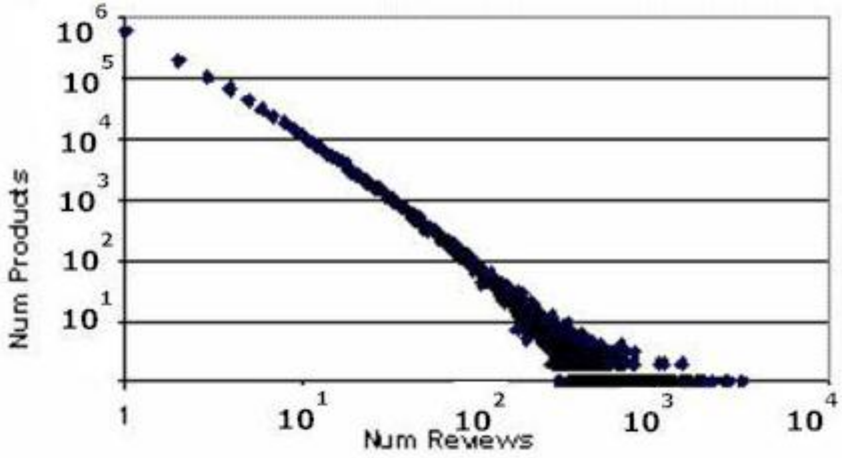
- June 2006
  - 5.8mil reviews, 1.2mil products and 2.1mil reviewers.
- A review has 8 parts
  - *<Product ID>*
  - *<Reviewer ID>*
  - *<Rating>*
  - *<Date>*
  - *<Review Title> <Review Body>*
  - *<Number of Helpful feedbacks> <Number of Feedbacks>*

Log-log plot

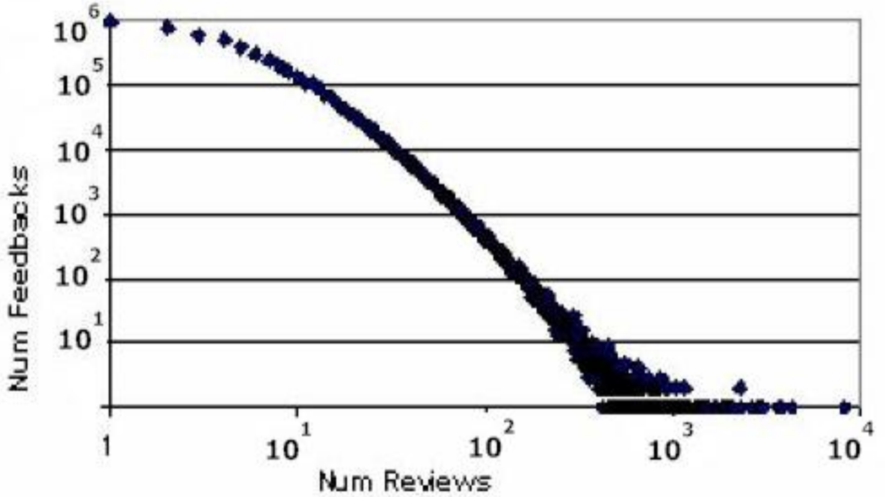
(Jindal and Liu, 2008)



- Fig. 1 reviews and reviewers



- Fig. 2 reviews and products



- Fig. 3 reviews and feedbacks

Star Ratings vs. Percent of Reviews

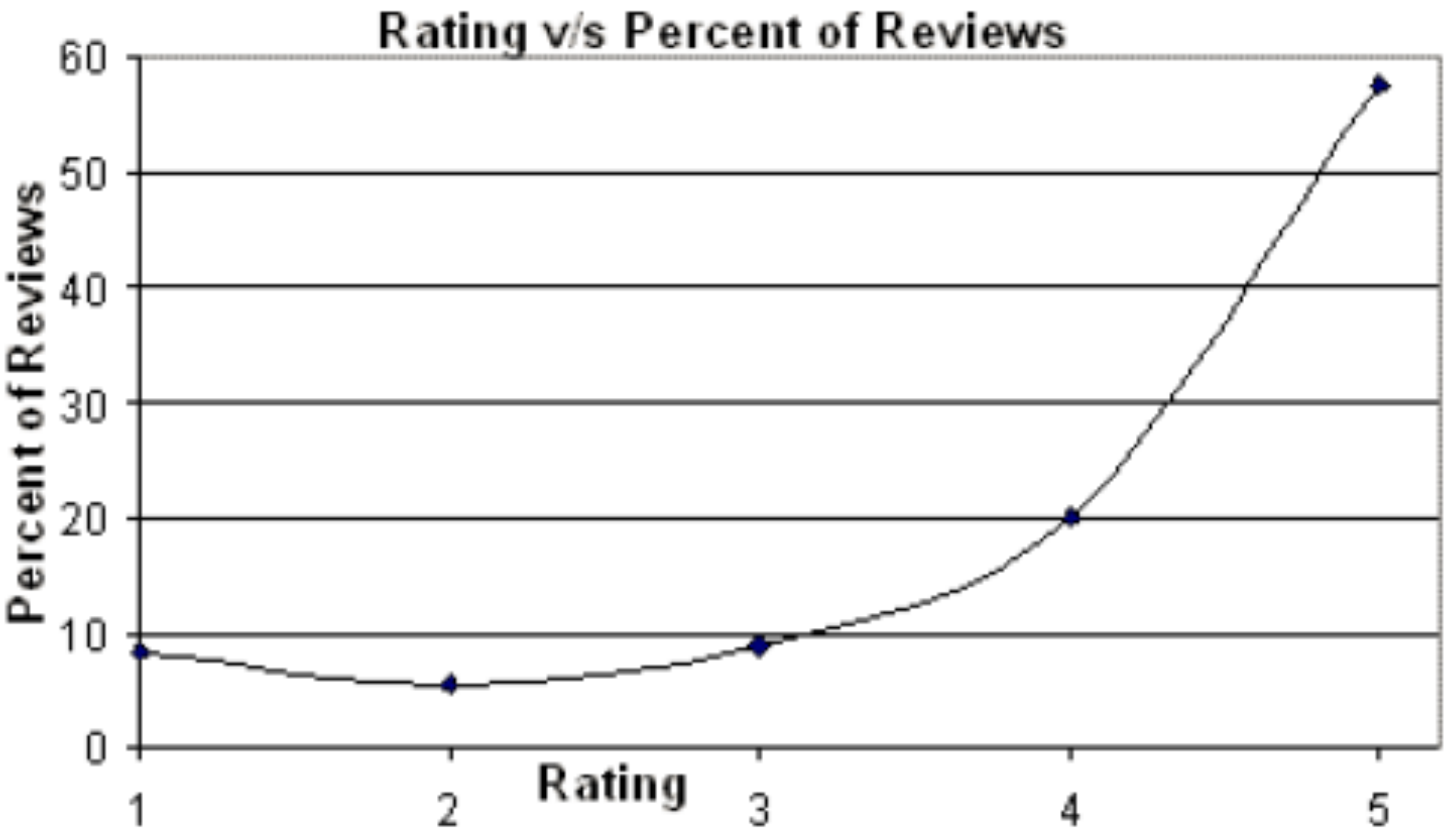


Figure 4. Rating vs. percent of reviews

Categorization of opinion spam

(Jindal and Liu 2008)

- Type 1 (fake reviews)

  Ex:
- Type 2 (Reviews on Brands Only)

  Ex: "*I don't trust HP and never bought anything from them*"
- Type 3 (Non-reviews)
  - Advertisements

    Ex: "*Detailed product specs: 802.11g, IMR compliant, …*"
    "*…buy this product at: compuplus.com*"
  - Other non-reviews

    Ex: "*What port is it for*"
    "*The other review is too funny*"
    "*Go Eagles go*"

Fake reviews vs. product quality

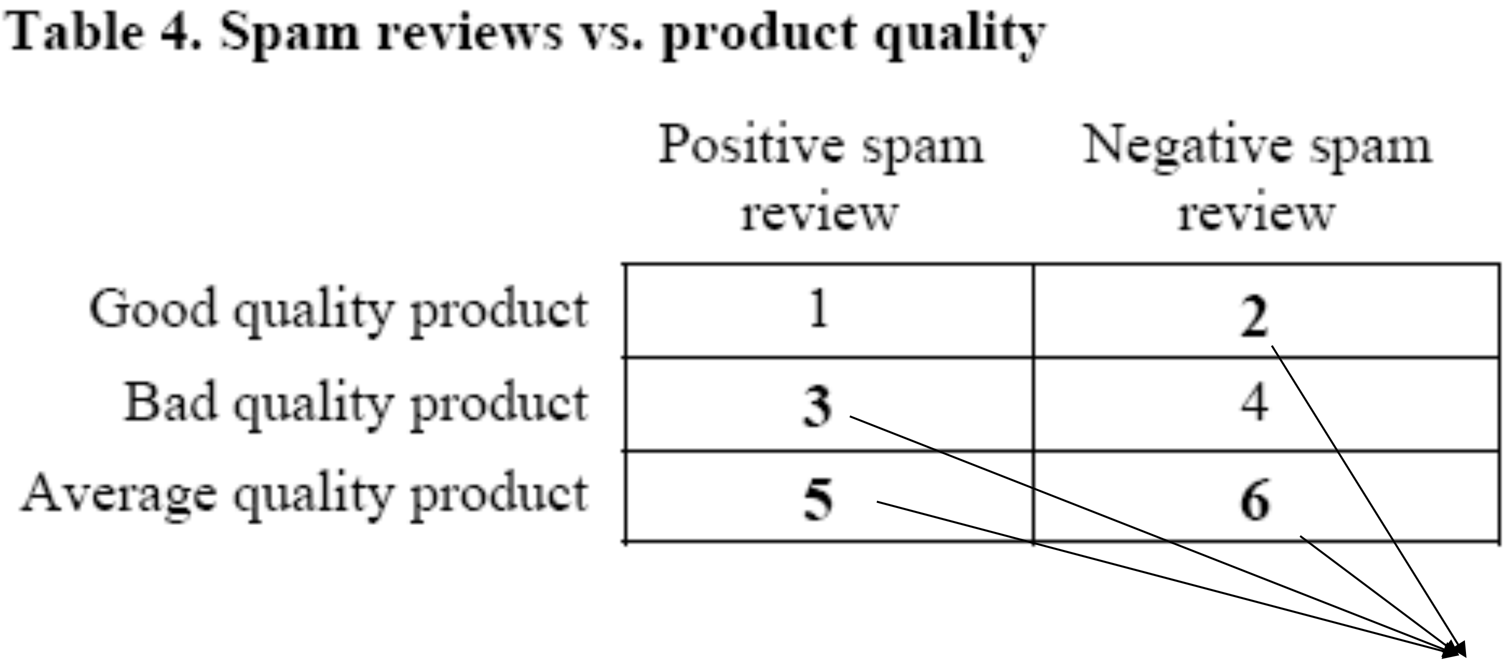


Table 4. Spam reviews vs. product quality

	Positive spam review	Negative spam review
Good quality product	1	**2**
Bad quality product	**3**	4
Average quality product	**5**	**6**

Harmful Regions

Type of spammers

- Individual spammers:
  - The spammer does not work with anyone. He/she just writes fake reviews him/herself using a single user-id, e.g., the author of a book.
- Group spammers
  - A group of spammers (persons) works in collusion
  - A single person registers multiple user-ids (called *sock puppetting*)

Type of data and clues

- Review content:
  - The actual text content of each review, linguistic features and style features
- Meta-data about each reviewer:
  - star rating, user-id,
  - time when a review was posted, and time taken to write/post the review,
  - host IP address and MAC address
  - geo-location of the reviewer
  - sequence of clicks at the review site

Type of data and clues (contd)

- Product information:
  - Information about the entity being reviewed, e.g.,
    - the product description,
    - sales volume
    - sales rank.
- Public data vs. site private (internal) data
  - Site private data, very useful
  - But hard to obtain by outsiders

Spam detection (Jindal and Liu 2008)

- Type 2 and Type 3 spam reviews are relatively easy to detect
  - Supervised learning, e.g., logistic regression
  - It performs quite well, and not discuss it further.
- Type 1 spam (fake) reviews
  - Manual labeling is extremely hard
  - Propose to use duplicate and near-duplicate reviews as positive training data

Four types of duplicates

1. Same userid, same product
2. Different userid, same product
3. Same userid, different products
4. Different userid, different products

- The last three types are very likely to be fake!

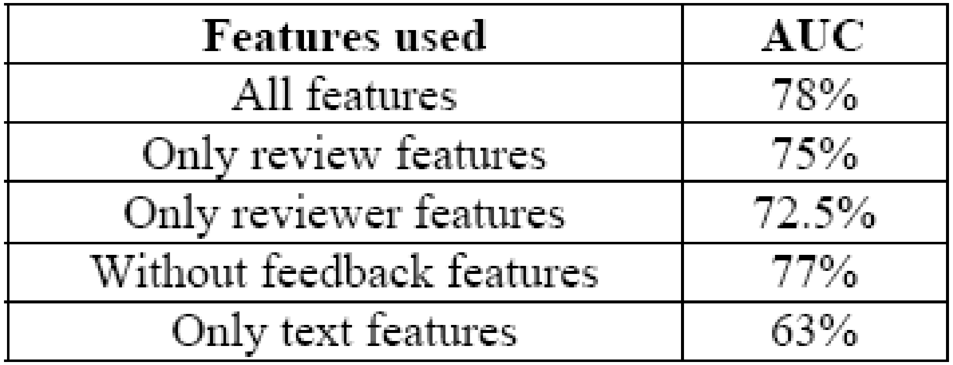
Supervised model building

- Logistic regression
  - Training: duplicates as spam reviews (positive) and the rest as non-spam reviews (negative)
- Use the follow features (clues)
  - Review centric features (content)
    - About reviews (contents (n-grams), ratings, etc)
  - Reviewer centric features
    - About reviewers (different unusual behaviors, etc)
  - Product centric features
    - Features about products reviewed (sale rank, etc)

Predictive power of duplicates

- Representative of all kinds of spam
- Only 3% duplicates accidental
- Duplicates as positive examples, rest of the reviews as negative examples

Table 5. AUC values on duplicate spam reviews.



Features used	**AUC**
All features	78%
Only review features	75%
Only reviewer features	72.5%
Without feedback features	77%
Only text features	63%

  - reasonable predictive power
  - Maybe we can use duplicates as type 1 spam reviews(?)

Tentative classification results

- Negative outlier reviews tend to be heavily spammed
- Those reviews that are the only reviews of products are likely to be spammed
- Top-ranked reviewers are more likely to be spammers
- Spam reviews can get good helpful feedbacks and non-spam reviews can get bad feedbacks
- …

Other Supervised Methods

- Li et al. (2011) built a model similar to that in (Jindal and Liu 2008), but
  - Also use sentiment and some other features
  - Manually labeled data
- Ott et al (2011) also used supervised learning.
  - Use Mechanical Turk to write fake reviews
  - Use n-grams as features
- Yoo and Gretzel (2009) studied deceptive reviews as well.

Finding unexpected reviewer behavior

- Move "behind the scenes"
  - to uncover the "secrets" of reviewers by profiling them based on their posted reviews and behaviors
- Lim et al (2010) and Nitin et al (2010) analyze the behavior of reviewers
  - identifying *unusual review patterns* which may indicate suspicious behaviors of reviewers.
- The problem is formulated as finding unexpected rules and rule groups.

Spam behavior models (Lim et al 2010)

- Several unusual reviewer behavior models were identified.
  - Targeting products
  - Targeting groups
  - General rating deviation
  - Early rating deviation
- Their scores for each reviewer are then combined to produce the final spam score.
- Ranking and user evaluation

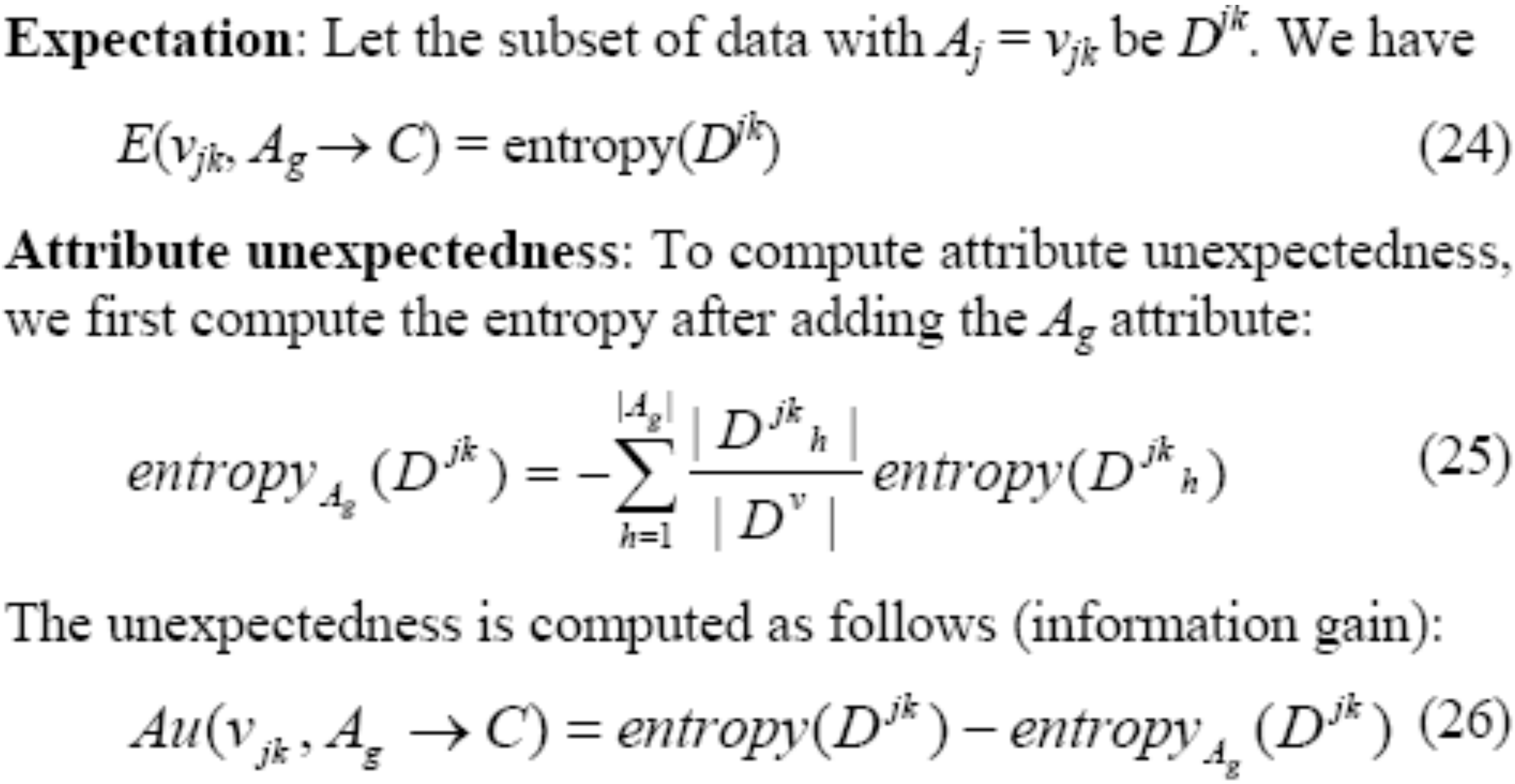
Finding unexpected rules (Jindal, Liu, Lim 2010)

- For example, if a reviewer wrote all positive reviews on products of a brand but all negative reviews on a competing brand …
- Finding unexpected rules,
  - Data: *reviewer-id*, *brand-id*, *product-id*, and a *class*.
  - Mining: class association rule mining
  - Finding unexpected rules and rule groups, i.e., showing atypical behaviors of reviewers.

Rule1: Reviewer-1, brand-1 -> positive (confid=100%)

Rule2: Reviewer-1, brand-2 -> negative (confid=100%)

The example (cont.)



Expectation: Let the subset of data with $A_j = v_{jk}$ be D^{jk}. We have

$$E(v_{jk}, A_g \rightarrow C) = \text{entropy}(D^{jk}) \qquad (24)$$

Attribute unexpectedness: To compute attribute unexpectedness, we first compute the entropy after adding the A_g attribute:

$$entropy_{A_g}(D^{jk}) = -\sum_{h=1}^{|A_g|} \frac{|D^{jk}{}_h|}{|D^v|} entropy(D^{jk}{}_h) \qquad (25)$$

The unexpectedness is computed as follows (information gain):

$$Au(v_{jk}, A_g \rightarrow C) = entropy(D^{jk}) - entropy_{A_g}(D^{jk}) \quad (26)$$

Confidence unexpectedness

Rule: reviewer-1, brand-1 → positive [sup = 0.1, conf = 1]

- If we find that on average reviewers give brand-1 only 20% positive reviews (expectation), then reviewer-1 is quite unexpected.

$$Cu(v_{jk} \rightarrow c_i) = \frac{\Pr(c_i \mid v_{jk}) - E(\Pr(c_i \mid v_{jk}))}{E(\Pr(c_i \mid v_{jk}))}$$

$$E(\Pr(c_i \mid v_{jk}, v_{gh})) = \frac{\Pr(c_i \mid v_{jk})\Pr(c_i \mid v_{gh})}{\Pr(c_i)\sum_{r=1}^{m}\frac{\Pr(c_r \mid v_{jk})\Pr(c_r \mid v_{gh})}{\Pr(c_r)}}$$

Support unexpectedness

Rule: reviewer-1, product-1 -> positive [sup = 5]

- Each reviewer should write only one review on a product and give it a positive or negative rating (expectation).
- This unexpectedness can detect those reviewers who review the same product multiple times, which is unexpected.
  - These reviewers are likely to be spammers.
- Can be defined probabilistically as well.

Detection using review graph
(Wang et al., 2011)

- This study was based on a snapshot of all reviews from resellerratings.com, which were crawled on Oct. 6th, 2010.
  - 343603 reviewers, 408470 reviews, 14561 store
- Form a heterogeneous review graph with three types of nodes,
  - reviewers, reviews and stores,
  - The graph captures their relationships and was used model spamming clues.

The Relationships

- Three concepts were defined and computed,
  - *trustiness* of reviewers,
  - *honesty* of reviews, and
  - *reliability* of stores.
- A reviewer is more trustworthy if he/she has written more honesty reviews
- A store is more reliable if it has more positive reviews from trustworthy reviewers
- A review is more honest if it is supported by many other honest reviews.

Definitions and equations

- Trustiness of a reviewer *r*

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

- Honesty of a review *v*

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

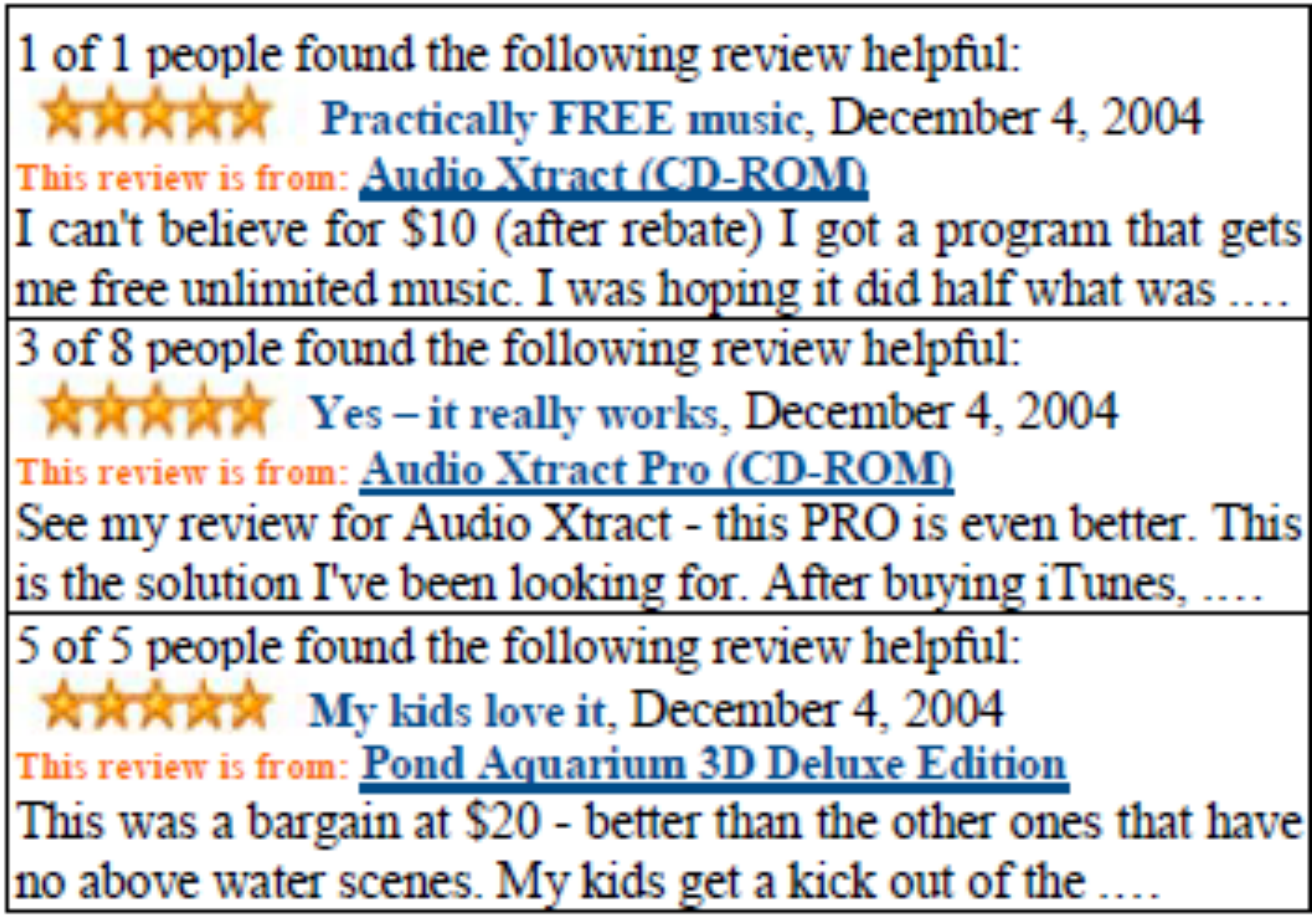
- Reliability of store *s*

$$R(s) = \frac{2}{1 + e^{-\theta}} - 1$$

Detecting group spam (Mukherjee et al WWW-2012)

- A group of people (could be a single person with multiple ids) work together to promote a product or to demote a product.
- Such spam can be very damaging as
  - they can take total control of sentiment on a product
- The algorithm has three steps
  - Frequent pattern mining: find groups of people who reviewed a number of products together.
  - A set of feature indicators are identified
  - Ranking is performed using a relational model

Big John’s Profile



1 of 1 people found the following review helpful:

Practically FREE music, December 4, 2004

This review is from: Audio Xtract (CD-ROM)

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 ….

3 of 8 people found the following review helpful:

Yes – it really works, December 4, 2004

This review is from: Audio Xtract Pro (CD-ROM)

See my review for Audio Xtract - this PRO is even better. This is the solution I've been looking for. After buying iTunes, ….

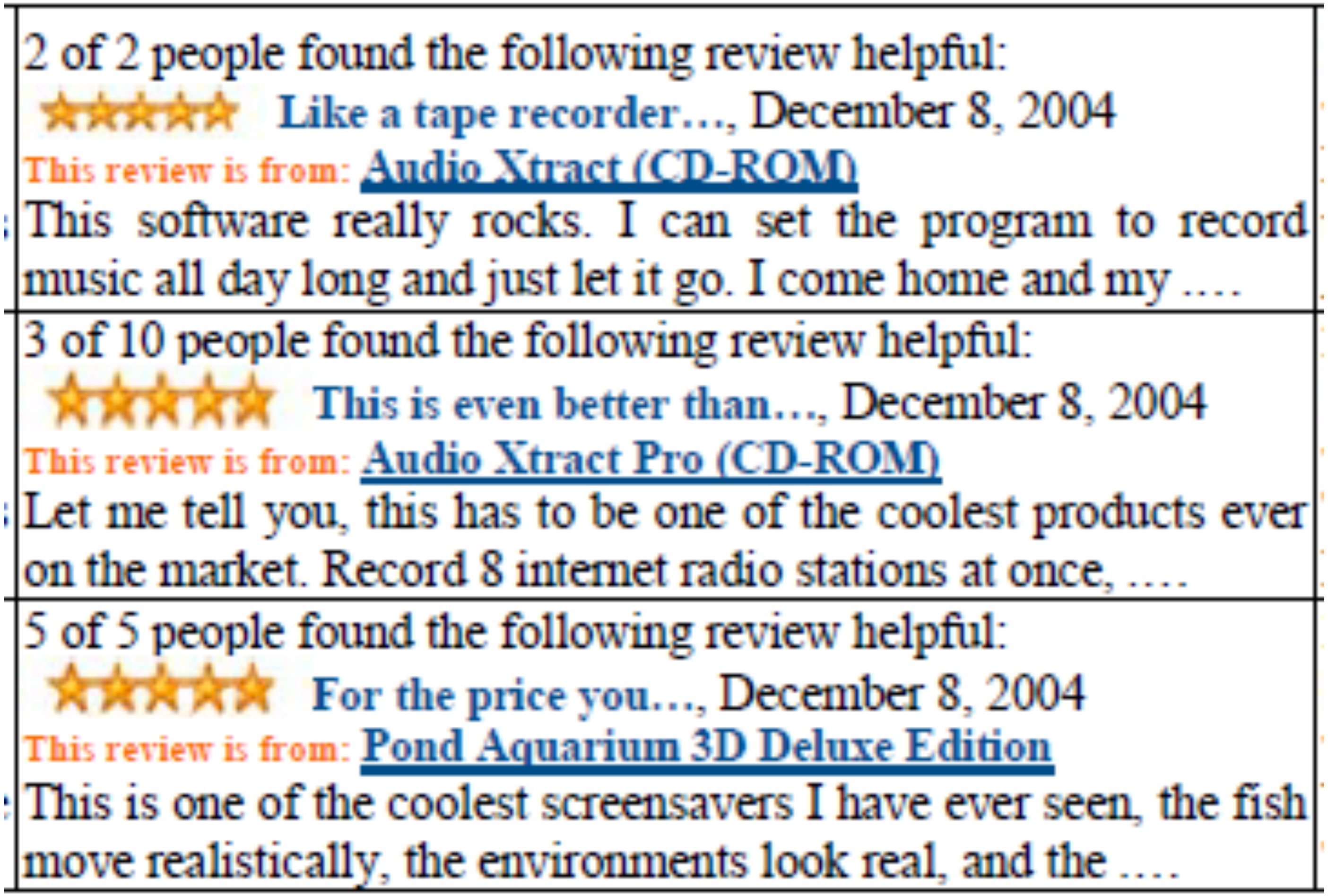
5 of 5 people found the following review helpful:

My kids love it, December 4, 2004

This review is from: Pond Aquarium 3D Deluxe Edition

This was a bargain at $20 - better than the other ones that have no above water scenes. My kids get a kick out of the ….

Cletus' Profile

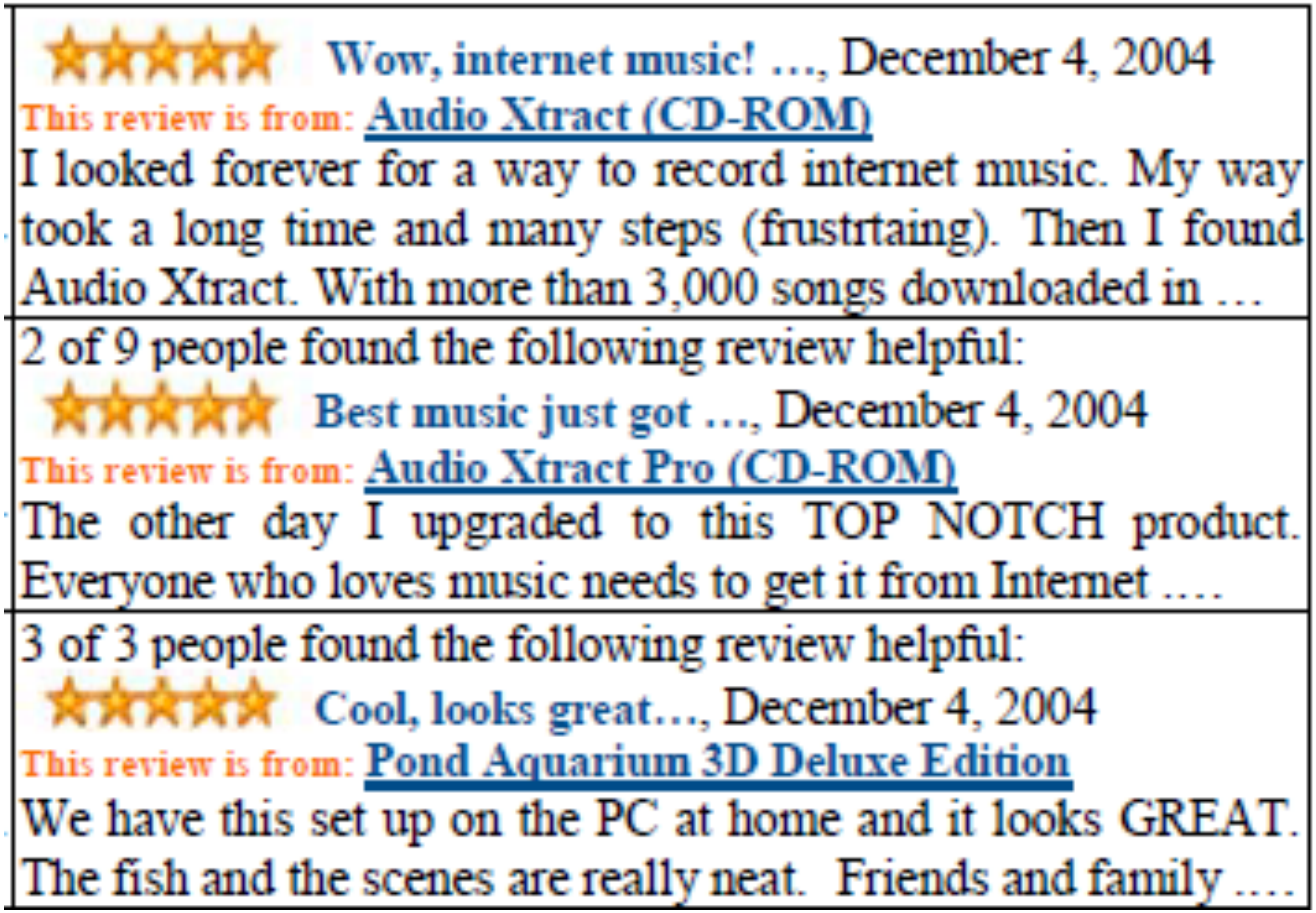


2 of 2 people found the following review helpful:
★★★★★ **Like a tape recorder…**, December 8, 2004
This review is from: **Audio Xtract (CD-ROM)**
This software really rocks. I can set the program to record music all day long and just let it go. I come home and my ….

3 of 10 people found the following review helpful:
★★★★★ **This is even better than…**, December 8, 2004
This review is from: **Audio Xtract Pro (CD-ROM)**
Let me tell you, this has to be one of the coolest products ever on the market. Record 8 internet radio stations at once, ….

5 of 5 people found the following review helpful:
★★★★★ **For the price you…**, December 8, 2004
This review is from: **Pond Aquarium 3D Deluxe Edition**
This is one of the coolest screensavers I have ever seen, the fish move realistically, the environments look real, and the ….

Jake’ s Profile



★★★★★ Wow, internet music! …, December 4, 2004

This review is from: Audio Xtract (CD-ROM)

I looked forever for a way to record internet music. My way took a long time and many steps (frustrtaing). Then I found Audio Xtract. With more than 3,000 songs downloaded in …

2 of 9 people found the following review helpful:

★★★★★ Best music just got …, December 4, 2004

This review is from: Audio Xtract Pro (CD-ROM)

The other day I upgraded to this TOP NOTCH product. Everyone who loves music needs to get it from Internet ….

3 of 3 people found the following review helpful:

★★★★★ Cool, looks great…, December 4, 2004

This review is from: Pond Aquarium 3D Deluxe Edition

We have this set up on the PC at home and it looks GREAT. The fish and the scenes are really neat. Friends and family ….

Finding candidate groups

- Frequent itemset mining
    - Items → Reviewer Ids (rids).
    - Transaction → set of rids for a product
- Frequent itemsets give us
    - “reviewer groups” that have reviewed multiple products together
- Our study was based on Amazon reviews of manufactured products

A set of clues (or features)

- Group Time Window (GTW)
- Group Deviation (GD)
- Group Content Similarity (GCS)
- Group Member Content similarity (GMCS)
- Group Early Time Frame (GETF)
- Group Size Ratio (GSR)
- Group Size (GS)
- Group Support Count (GSUP)

A relational model and algorithm

Algorithm: GSRank

Input: Weight matrices W_{PG}, W_{MP}, and W_{GM}

Output: Ranked list of candidate spam groups

```
1.Initialize $V_G^0 \leftarrow [0.5]_{|G|}$ ; $t \leftarrow 1$;
2.Iterate:
      i.   $V_P \leftarrow W_{PG} V_G^{(t-1)}$ ; $V_M \leftarrow W_{MP} V_P$ ;
      ii.  $V_G \leftarrow W_{GM} V_M$ ; $V_M \leftarrow W_{GM}^T V_G$ ;
      iii. $V_P \leftarrow W_{MP}^T V_M$ ;  $V_G^{(t)} \leftarrow W_{PG}^T V_P$ ;
      iv.  $V_G^{(t)} \leftarrow V_G^{(t)} / \| V_G^{(t)} \|_1$ ;
   until $\| V_G^{(t)} - V_G^{(t-1)} \|_\infty < \delta$
3  Output the ranked list of groups, $V_G^*$
```

Utility or quality of reviews

- Goal: Determining the helpfulness, or utility of each review (not necessarily fake)
  - It is desirable to rank reviews based on utilities or qualities when showing them to users, with the highest quality review first.
- Many review aggregation sites have been practicing this, e.g., amazon.com.
  - "*x of y people found the following review helpful*."
  - Voted by user - "*Was the review helpful to you?*"

Application motivations

- Although review sites use helpfulness feedback to rank their reviews,
    - A review takes a long time to gather enough feedback.
        - New reviews will not be read.
    - Some sites do not provide feedback information.
- It is thus beneficial to score each review once it is submitted to a site.

Regression formulation

(Zhang and Varadarajan, 2006; Kim et al. 2006)

- Formulation: Determining the utility of reviews is usually treated as a regression problem.
  - A set of features is engineered for model building
  - The learned model assigns an utility score to each review, which can be used in review ranking.
- Unlike fake reviews, the ground truth data used for both training and testing are available
  - Usually the user-helpfulness feedback given to each review.

Features for regression learning

- Example features include
  - review length, review rating, counts of some POS tags, opinion words, tf-idf scores, wh-words, product aspect mentions, comparison with product specifications, timeliness, etc (Zhang and Varadarajan, 2006; Kim et al. 2006; Ghose and Ipeirotis 2007; Liu et al 2007)
- Subjectivity classification was applied in (Ghose and Ipeirotis 2007).
- Social context was used in (O' Mahony and Smyth 2009; Lu et al. 2010).

Classification formulation

- Binary classification: Instead of using the original helpfulness feedback as the target or dependent variable,
  - Liu et al (2007) performed manual annotation of two classes based on whether the review evaluates many product aspects or not.
- Binary class classification is also used in (O' Mahony and Smyth 2009)
  - Classes: Helpful and not helpful
  - Features: helpfulness, content, social, and opinion

Personalized review quality prediction

(Moghaddam, Jamali and Ester, 2012)

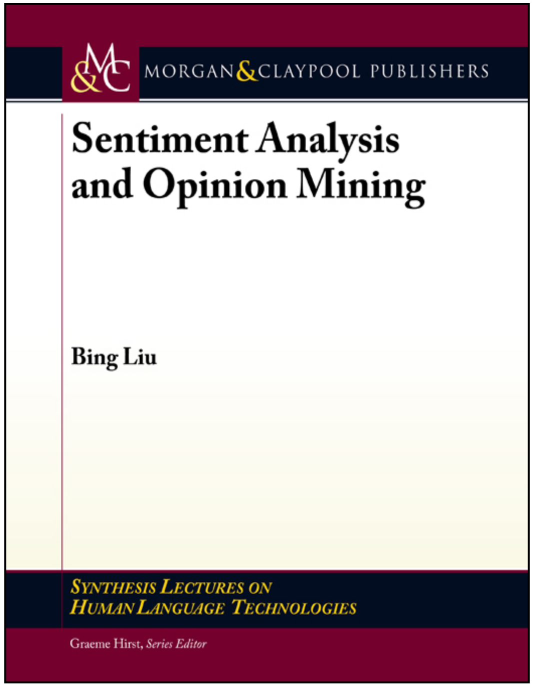
- Personalized review quality prediction for recommendation of helpful reviews. Previous work computes only one helpfulness score for each review, which may not be enough
  - Tensor factorization models were used to find latent features of reviews, reviewers, raters/users, and products.
  - Basically, the authors treated this problem as a personalized recommendation problem.

Summary

- For many businesses, posting fake reviews themselves or employing others to do it has become a cheap way of marketing.
- As social media is increasingly used for critical decision making
  - Detecting fake reviews and opinions is critical.
- Many companies are doing it (internal data)
- Current detection methods are still in their infancy. More research is needed.

More details and references

- **New Book:**
  - B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool publishers. May, 2012.



M&C MORGAN & CLAYPOOL PUBLISHERS

Sentiment Analysis
and Opinion Mining

Bing Liu

SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, *Series Editor*