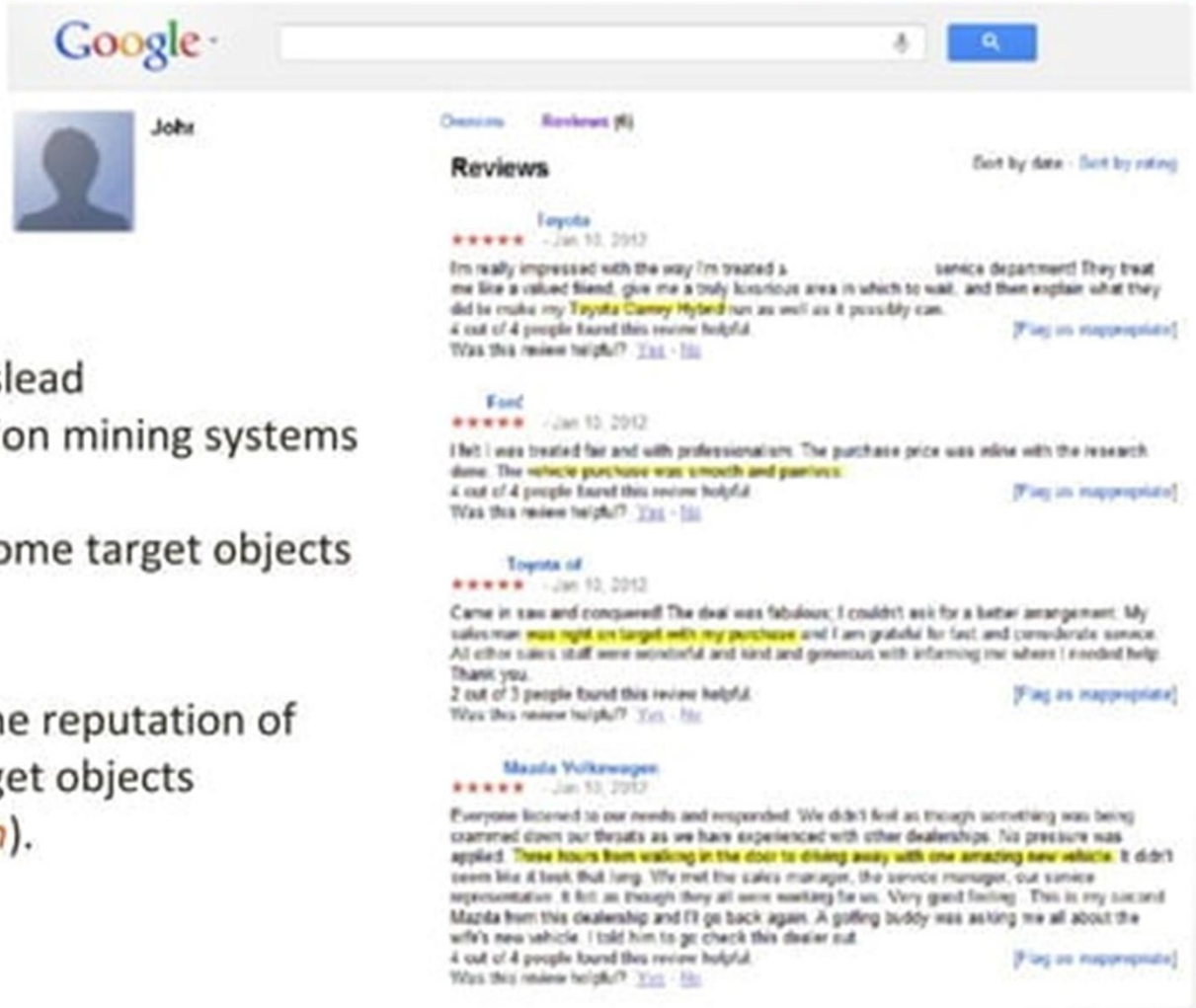
OPINION SPAM AND ANALYSIS

OPINION SPAM DETECTION

three types of spam reviews

- **Type 1 (untruthful opinions)**: also known as *fake reviews* or *bogus reviews*.



deliberately mislead
readers or opinion mining systems

1.To promote some target objects (*hype spam*).

2. To damage the reputation of some other target objects (*defaming spam*).

three types of spam reviews

- **Type 2 (reviews on brands only)**: not comment on the products for the products but *only* the brands, the manufacturers or the sellers.
- **Type 3 (non-reviews)**: two main sub-types:
  - advertisements
  - other irrelevant reviews containing no opinions (questions, answers ..)
- They are not targeted at the specific products and are often biased.

Review Data from amazon.com

- Each amazon.com's review consists of 8 parts
  - *<Product ID> <Reviewer ID> <Rating> <Date> <Review Title> <Revie*
    >

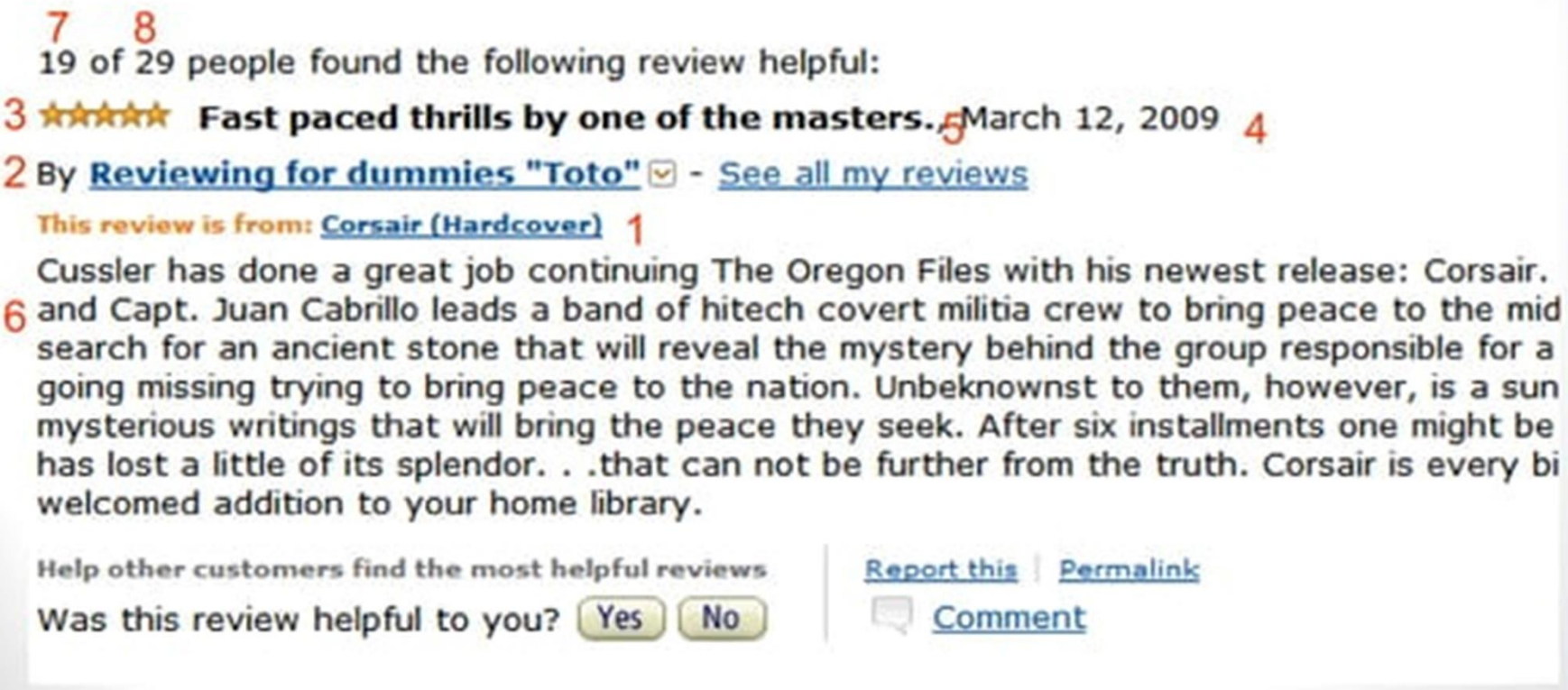


Table 1. Various features of different categories of products

	Number of Reviewed			Total
Category	**Reviews**	**Products**	**Reviewers**	**Products**
All	5838032	1195133	2146048	6272502
Books	2493087	637120	1076746	1185467
Music	1327456	221432	503884	888327
DVD/VHS	633678	60292	250693	157245
mProducts	228422	36692	165608	901913

Reviews, Reviewer, and Products

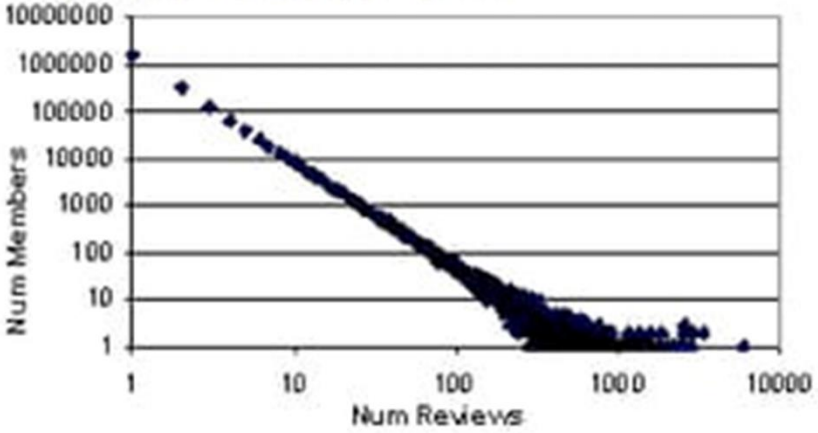


Figure 1. Log-log plot of number of reviews to number of members for amazon.

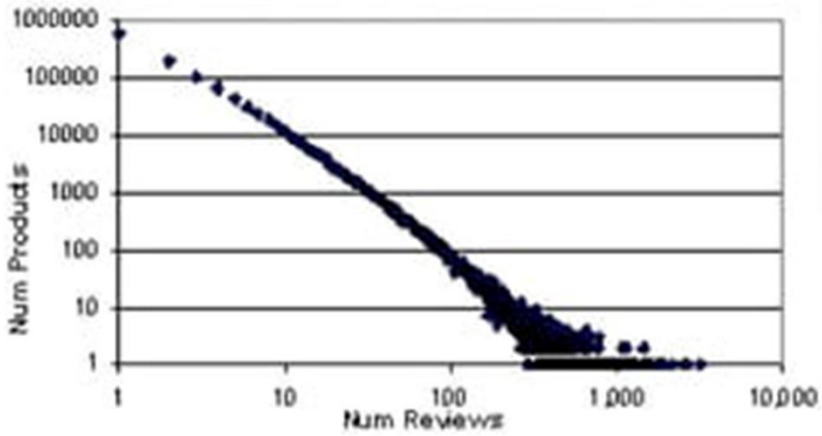


Figure 2. Log-log plot of number of reviews to number of products for amazon.

- There are 2 reviewers with more than 15,000 reviews.
- 68% of reviewers wrote only 1 review.

- 50% of products have only 1 review.
- Only 19% of the products have at least 5 reviews.

Review Ratings and Feedbacks

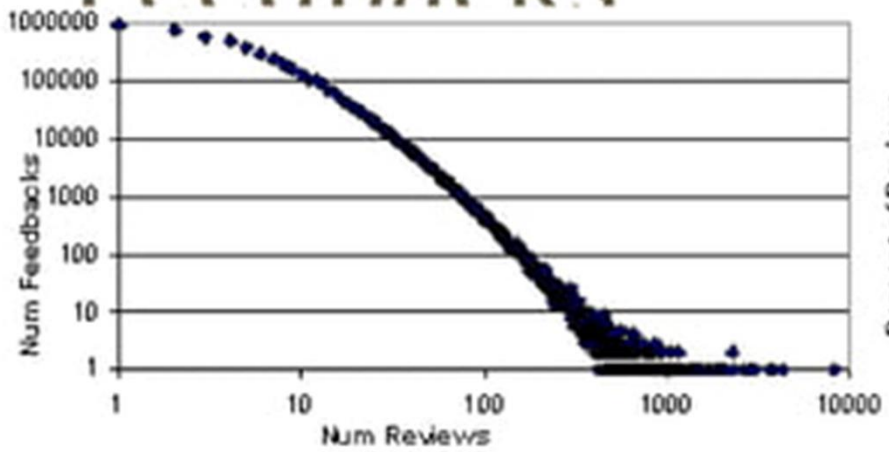


Figure 3. Log-log plot of number of reviews to number of feedbacks for amazon.

Figure 4. Rating vs. percent of reviews

- 60% of the reviews have a rating of 5.0.
- On average, a review gets 7 feedbacks.
- The percentage of positive feedbacks of a review decreases rapidly from the first review of a product to the last.

Spam Detection

- a classification problem with two classes, *spam* and *non-spam*.
- manually label training examples for spam reviews of type 2 and type 3.
- recognizing whether a review is an untruthful opinion spam (type 1) is extremely difficult by manually reading the review.

- found a large number of duplicate and near-duplicate reviews.
  - 1. Duplicates from different userids on the same product.
  - 2. Duplicates from the same userid on different products.
  - 3. Duplicates from different userids on different products.

Detection of Duplicate Reviews

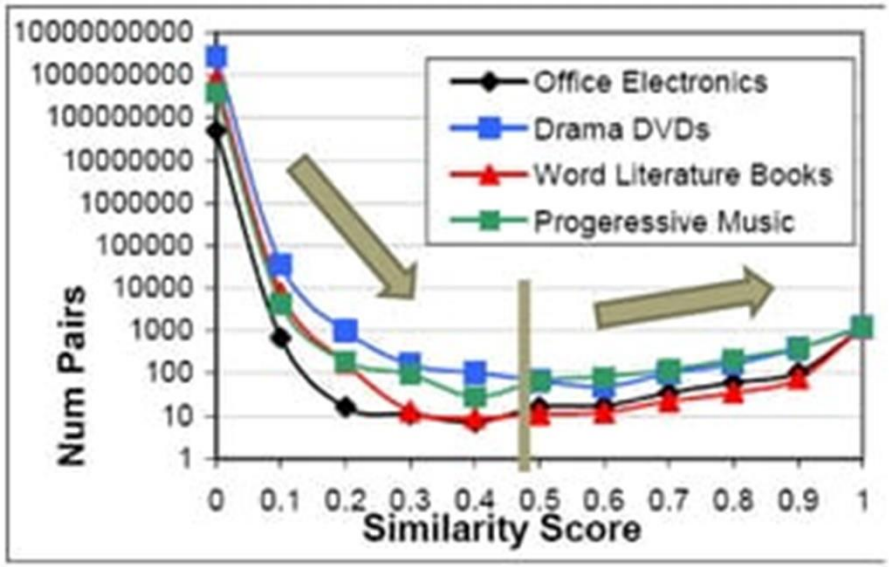


Figure 5. Similarity score and number of pairs of reviews for different sub-categories. Points on X axis are intervals. For example, 0.5 means between interval [0.5, 0.6).

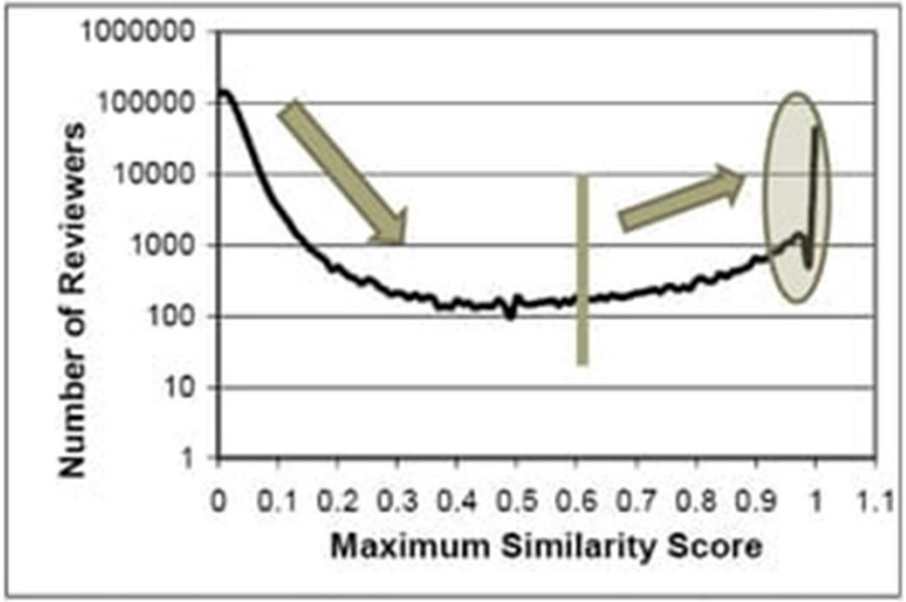


Figure 6. Maximum similarity score and number of members.

- shingle method (2-grams)
- Jaccard distance (Similarity score) > 90% ➔ duplicates.
- *The maximum similarity score* : the maximum of similarity scores between different reviews of a reviewer.

Detecting Type 2 & Type 3

- *Model Building Using Logistic Regression*
  - Manually labeled 40 spam reviews of the two types.
  - It produces a probability estimate of each review being a spam.

Feature Identification and Construction

- There are three main types of information
  - (1) the content of the review,
  - (2) the reviewer who wrote the review,
  - (3) the product being reviewed.
- three types of features:
  - (1) review centric features,
  - (2) reviewer centric features,
  - (3) product centric features.
- three types based on their average ratings
  - *Good* (rating ≥ 4), *bad* (rating ≤ 2.5) and *Average*, otherwise

Review Centric Features

- number of feedbacks(F1)
- number of helpful feedbacks(F2)
- percent of helpful feedbacks(F3)
- length of the review title(F4)
- length of review body(F5)
- Position of the review of a product sorted by date. ascending (F6) and descending (F7)
- the first review (F8)
- the only review (F9)

Review Centric Features -Textual features

- percent of positive bearing words(F10)
- percent of negative bearing words(F11)
- cosine similarity (F12)
- percent of times brand name (F13)
- percent of numerals words(F14)
- percent of capitals words(F15)
- percent of all capital words(F16)

Review Centric Features –Rating related features

- rating of the review (F17)
- the deviation from product rating (F18)
- the review is good, average or bad (F19)
- a bad review was written just after the first good review of the product and vice versa (F20, F21)

Reviewer Centric Features

- Ratio of the number of reviews that the reviewer wrote which were the first reviews (F22)
- ratio of the number of cases in which he/she was the only reviewer (F23)
- average rating given by reviewer (F24)
- standard deviation in rating (F25)
- the reviewer always gave only good, average or bad rating (F26)
- a reviewer gave both good and bad ratings (F27)
- a reviewer gave good rating and average rating (F28)
- a reviewer gave bad rating and average rating (F29)
- a reviewer gave all three ratings (F30)
- percent of times that a reviewer wrote a review with binary features F20 (F31) and F21 (F32).

Product Centric Features

- Price of the product (F33)
- Sales rank of the product (F34)
- Average rating (F35)
- standard deviation in ratings (F36)

Results of Type 2 and Type 3 Spam Detection

- logistic regression on the data using 470 spam reviews for positive class and rest of the reviews for negative class.
- 10-fold cross validation.

Table 3. AUC values for different types of spam

Spam Type	Num reviews	AUC	AUC – text features only	AUC – w/o feedbacks
Types 2 & 3	470	98.7%	90%	98%
Type 2 only	221	98.5%	88%	98%
Type 3 only	249	99.0%	92%	98%

Analysis of Type 1 Spam Reviews

- 1. To promote some target objects (*hype spam*).
- 2. To damage the reputation of some other target objects (*defaming spam*).

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**

Making Use of Duplicates

- the same person writes the same review for different versions of the same product may not be spam.
- We propose to treat all duplicate spam reviews as positive examples, and the rest of the reviews as negative examples.
  - We then use them to learn a model to identify non-duplicate reviews with similar characteristics, which are likely to be spam reviews.

Model Building Using Duplicates

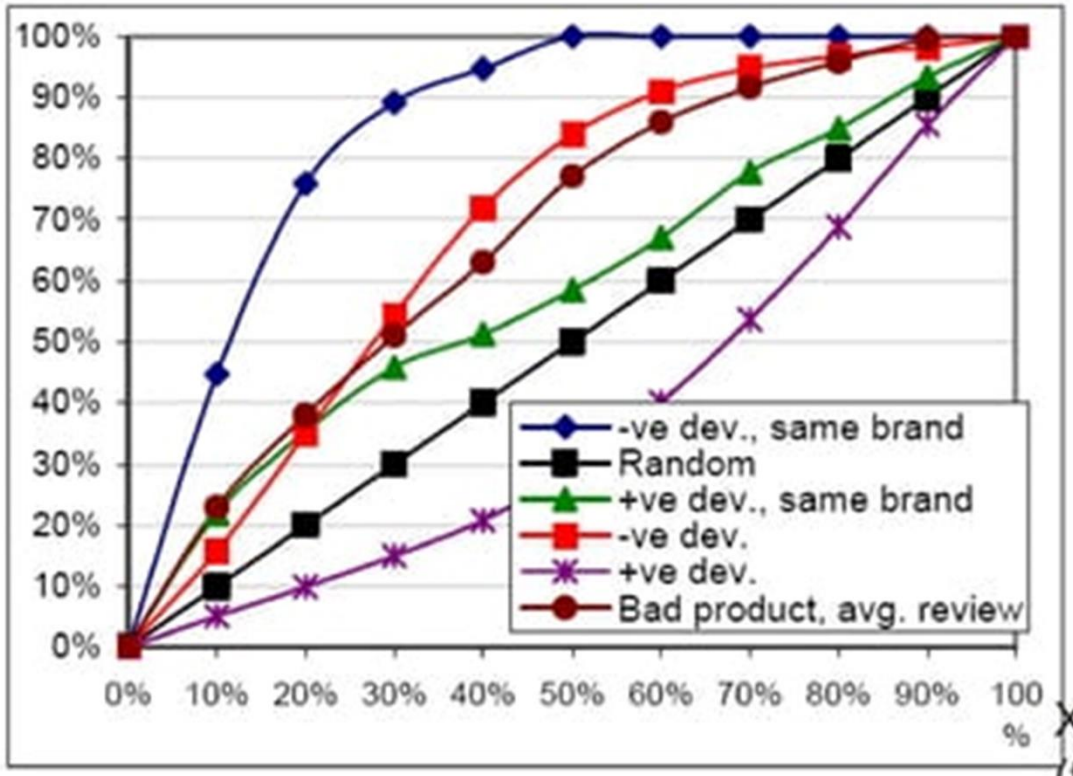
- building the logistic regression model using duplicates and non-duplicates is not for detecting duplicate spam
  - Our real purpose is to use the model to identify type 1 spam reviews that are not duplicated.
  - check whether it can predict outlier reviews

- Outlier reviews : whose ratings deviate from the average product rating a great deal.
- Sentiment classification techniques may be used to automatically assign a rating to a review solely based on its review content.

Predicting Outlier Reviews

- Negative deviation is considered as less than -1 from the mean rating and positive deviation as larger than +1 from the mean rating of the product.
- Spammers may not want their review ratings to deviate too much from the norm to make the reviews too suspicious.

cumulated percentage of reviews of the current bin



X% of reviews (the test data)

Figure 7: Lift curves for reviews with positive and negative deviations. "-ve (+ve) dev., same brand" means those reviews where member wrote multiple reviews on same brand and all reviews have negative (positive) deviation.

Some Other Interesting Reviews - Only Reviews

- We did not use any position features of reviews (F6, F7, F8, F9, F20 and F21) and number of reviews of product (F12, F18, F35, and F36) related features in model building to prevent overfitting.

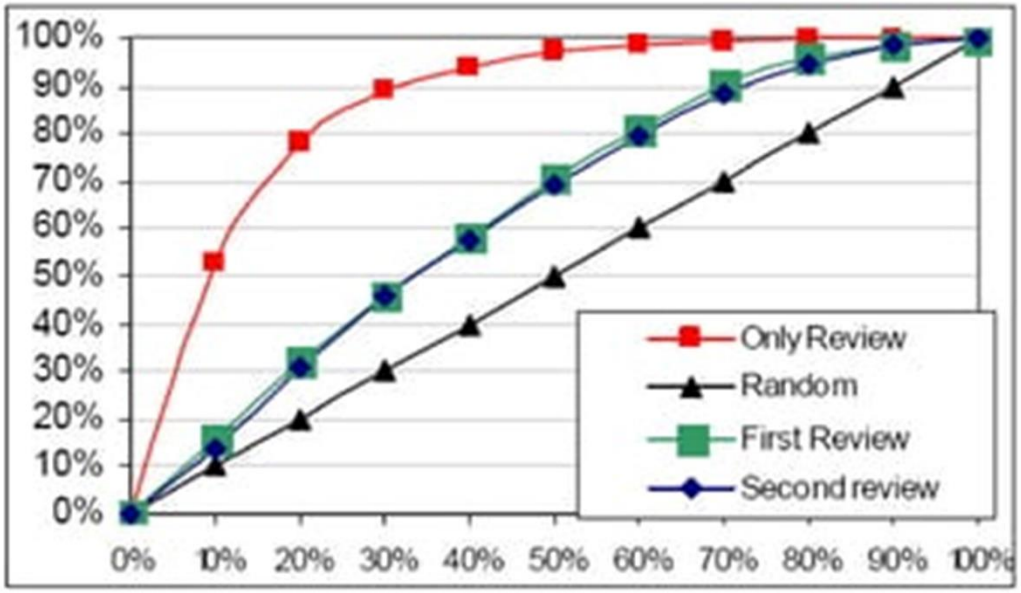


Figure 8: Lift curves for only reviews, first and second reviews of products.

- only reviews are very likely to be candidates of spam

- Reviews from Top-Ranked Reviewers

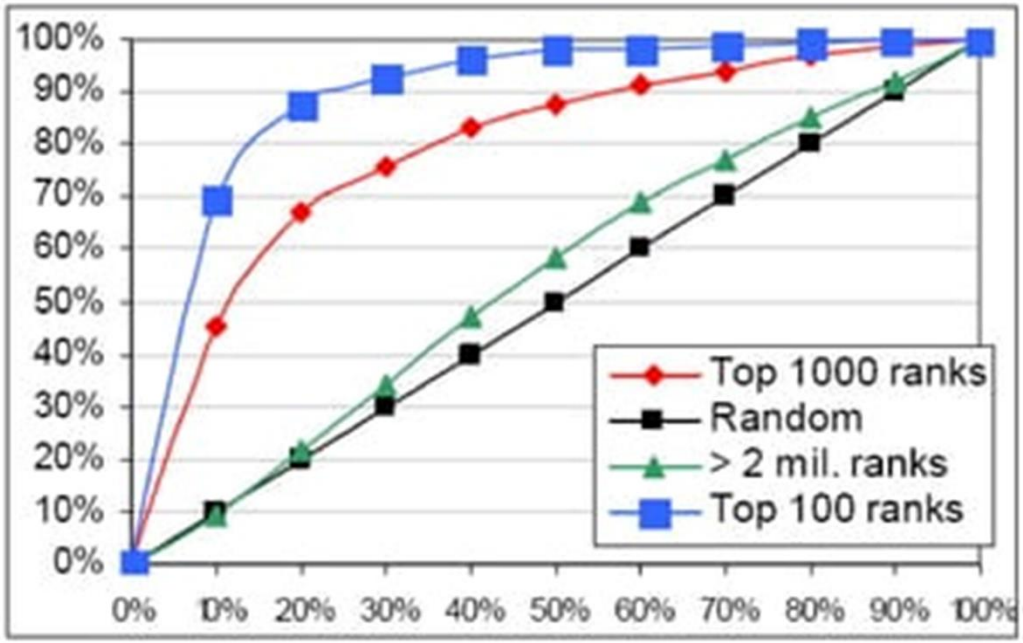


Figure 9: Lift curves for reviewers of different ranks.

- Top-ranked reviewers generally write a large number of reviews, much more than bottom ranked reviewers.
- Top ranked reviewers also score high on some important indicators of spam reviews.

➔ top ranked reviewers are less trustworthy as compared to bottom ranked reviewers.

- Reviews with Different Levels of Feedback

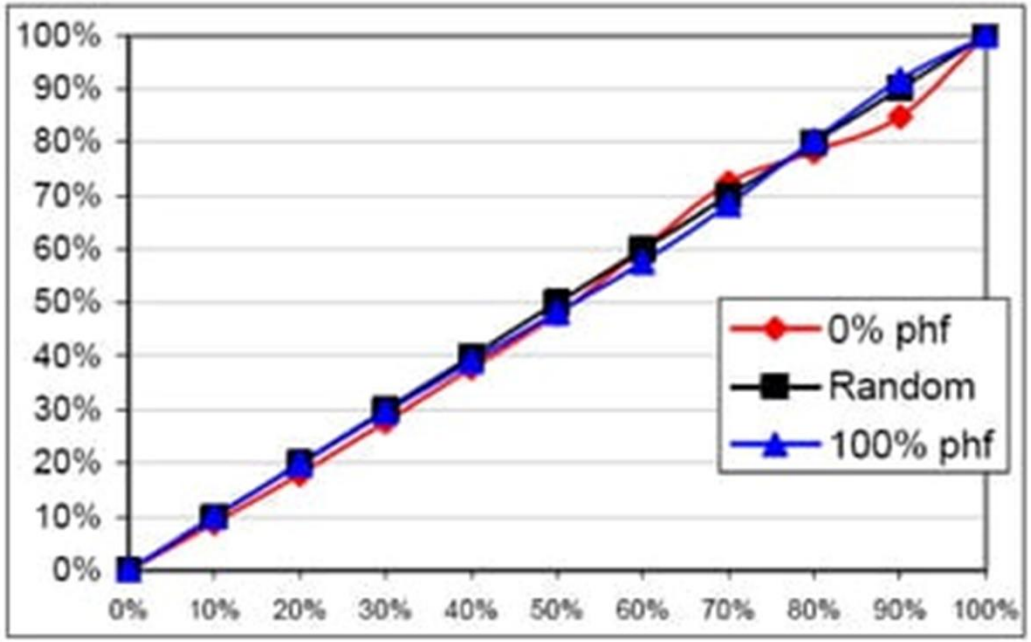


Figure 10: Lift curves for reviews with 0 and 100% positive feedbacks with minimum of 1 feedback.

- If usefulness of a review is defined based on the feedbacks that the review gets, it means that people can be readily fooled by a spam review.

  ➔feedback spam

- Reviews of Products with Varied Sales Ranks

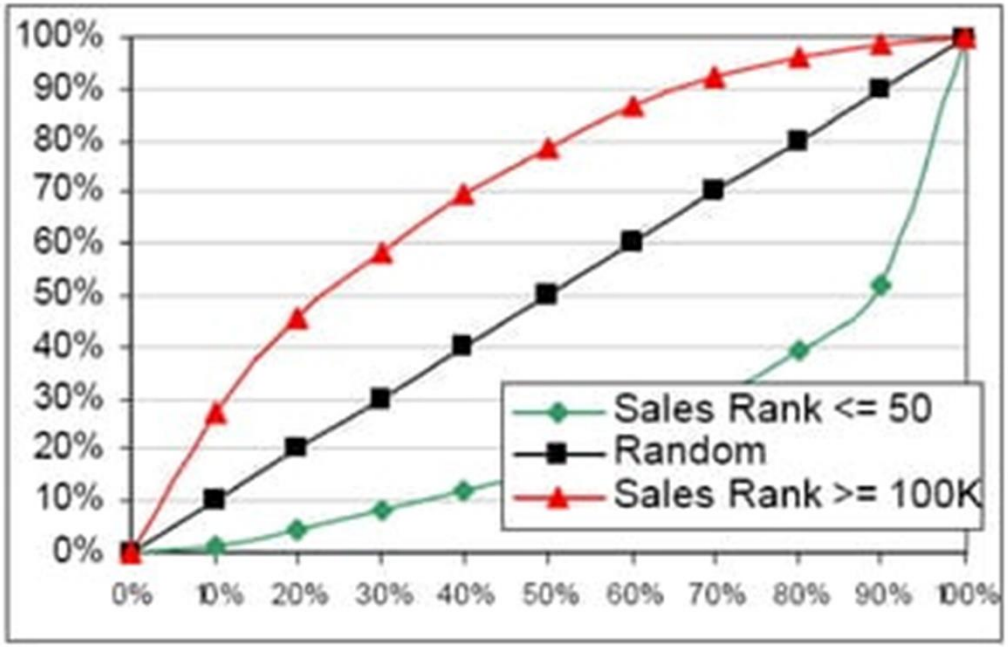


Figure 11: Lift curves for reviews corresponding to products with different sales ranks.

- Spam activities are more limited to low selling products. ➔ difficult to damage reputation of a high selling or popular product by writing a spam review.

Conclusions & Future work

- Results showed that the logistic regression model is highly effective.
- It is very hard to manually label training examples for type 1 spam.
- We will further improve the detection methods, and also look into spam in other kinds of media, e.g., forums and blogs.